The evolution of mycobacterial pathogenicity: clues from comparative genomics

Roland Brosch, Alexander S. Pym, Stephen V. Gordon and Stewart T. Cole

Comparative genomics, and related technologies, are helping to unravel the molecular basis of the pathogenesis, host range, evolution and phenotypic differences of the slow-growing mycobacteria. In the highly conserved *Mycobacterium tuberculosis* complex, where single-nucleotide polymorphisms are rare, insertion and deletion events (InDels) are the principal source of genome plasticity. InDels result from recombinational or insertion sequence (IS)-mediated events, expansion of repetitive DNA sequences, or replication errors based on repetitive motifs that remove blocks of genes or contract coding sequences. Comparative genomic analyses also suggest that loss of genes is part of the ongoing evolution of the slow-growing mycobacterial pathogens and might also explain how the vaccine strain BCG became attenuated.

Roland Brosch
Alexander S. Pym
Stewart T. Cole*
Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France.
*e-mail: stcole@pasteur.fr

Stephen V. Gordon
Veterinary Laboratories Agency, Woodham Lane, New Haw, Addlestone, Surrey, UK KT15 3NB.

Genomics has the potential to provide a complete understanding of the genetics, biochemistry, physiology and pathogenesis of microorganisms. With the genetic blueprint of an organism in hand, researchers should be able to unravel its biology rapidly, despite the fact that for many fully sequenced genomes no functional information is available for up to 50% of the open reading frames (ORFs). These knowledge gaps are being filled by extensive data-sets generated by functional and comparative genomics, new disciplines employing powerful techniques encompassing proteomics, bioinformatics, structural biology and microarrays. Such approaches are being applied with success to the genus *Mycobacterium*, which includes a wide variety of organisms of medical, veterinary and environmental importance, notably the human pathogens *Mycobacterium tuberculosis* and *Mycobacterium leprae*, which are responsible for tuberculosis (TB) and leprosy, respectively. In this review, we will briefly describe the various genomic approaches that have been used to study the mycobacteria, all of which have been catalyzed by the availability of the genome sequence of the H37Rv strain of *M. tuberculosis*[1], and summarize the findings to date.

Slow-growing mycobacteria

All mycobacteria possess a complex cell envelope that is rich in lipids and glycolipids, and the genus, whose phylogeny is based on analysis of 16S rRNA sequences (Fig. 1), can be subdivided into fast- or slow-growing species[2]. The slow-growing mycobacteria have remarkably long generation times (1–14 days) and include the principal human pathogens, the tubercle and leprosy bacilli. The *M. tuberculosis* complex comprises *M. tuberculosis* together with *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium canettii* and *Mycobacterium microti* and is defined as a single species by DNA–DNA hybridization studies[3], with exceptionally little sequence variation[4,5] (Fig. 1). The subspecies can only be distinguished by a limited number of phenotypic or, more recently, genotypic characteristics, but differ remarkably with respect to their host range and pathogenicity. *M. microti*, for example, is almost exclusively a rodent pathogen and has been used successfully as a live vaccine against TB, whereas *M. bovis* infects a wide variety of mammalian species, including humans. BCG (bacille Calmette–Guérin), a laboratory-attenuated variant of *M. bovis*, has been used extensively since the 1920s as a vaccine against human TB and, more recently, against leprosy. Phylogenetically, the closest relatives of the *M. tuberculosis* complex are the environmental species *Mycobacterium marinum* and *Mycobacterium ulcerans* (Fig. 1). Whereas *M. marinum* is principally an ectothermic pathogen that rarely causes human infections, *M. ulcerans* is responsible for a debilitating cutaneous disease that is reaching epidemic proportions in parts of West Africa. The *Mycobacterium avium* complex, another group of slow-growing mycobacterial subspecies, includes both veterinary and opportunistic human pathogens.

Methods for genome-wide comparisons

Physical and integrated maps

Pulsed-field gel electrophoresis of macro-restriction fragments, the first method enabling whole-genome comparisons, has been used as a tool for molecular epidemiological and population-genetic studies of the *M. tuberculosis* complex[6,7], and for the construction of physical and integrated maps of *M. tuberculosis*[8] and *M. bovis* BCG Pasteur[9]. The physical maps were useful for estimating genome size and structure (e.g. circular chromosome and absence of plasmids), and provided a scaffold for establishing integrated maps of ordered cosmid and bacterial artificial chromosome (BAC) libraries[10,11],

 0966-842X/01/$ – see front matter PII: S0966-842X(01)02131-X

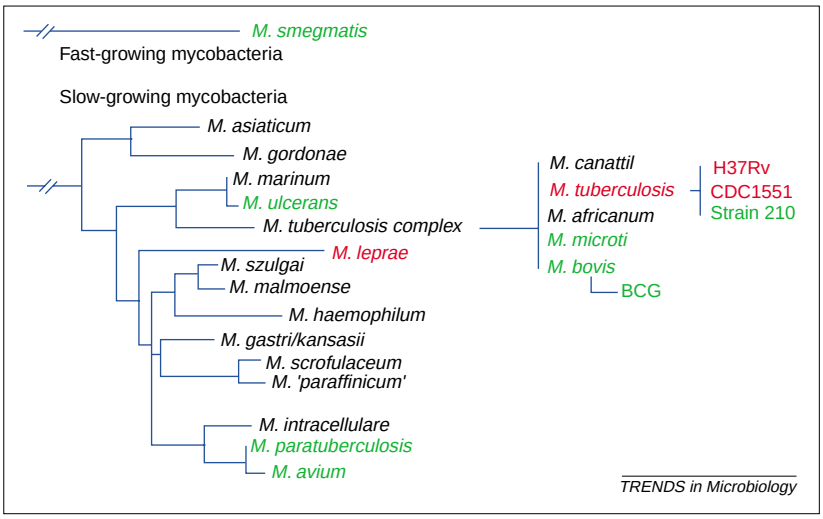

Fig. 1. Phylogenetic tree of selected mycobacteria, based on 16S rRNA sequences[2], including those species whose genome sequences are in progress (green) or completed (red).

which is essential for the rapid completion of genome sequences[1].

A minimal overlapping set of BAC clones spanning the whole genome can also be used to construct comparative genome-hybridization arrays (BAC arrays) or to compare individual BAC clones using libraries from different strains. This combined approach identified several deleted regions (RD1–10) in the genome of *M. bovis* BCG Pasteur compared with *M. tuberculosis* H37Rv, as well as loci (RvD1 and RvD2) that were deleted from the sequenced reference strain *M. tuberculosis* H37Rv relative to other strains of the *M. tuberculosis* complex[10,11]. BAC-to-BAC comparison also allowed the detection of two large tandem duplications (DU1 and DU2) in the genome of *M. bovis* BCG Pasteur[12]. The construction of BAC libraries from *M. tuberculosis*, BCG, *M. bovis*, *M. microti* and *M. ulcerans* has provided a basis for further comparative studies, as well as a resource for functional mycobacterial genomics.

Microarrays

DNA microarrays are also attractive tools for comparing genomes between closely related strains or species. Behr *et al*.[13] were able to identify 14 regions (RD1–14) that were absent from BCG Pasteur relative to *M. tuberculosis* H37Rv, and two deletions (RD15 and 16) specific for particular BCG substrains, by hybridizing whole-genome probes to a microarray of spotted PCR products representing most of the ~4000 *M. tuberculosis* H37Rv ORFs. An oligonucleotide-based Affymetrix GeneChip in conjunction with powerful informatics can accurately identify deletions as small as 300 bp[14,15]. Because of cross-hybridization, the ability of microarrays using spotted PCR products to detect deletions in multi-gene families is somewhat limited. This is particularly important in the *M. tuberculosis* complex as ~10% of the genome contains repetitive DNA[1,16], including the PE and PPE gene families, which often comprise multiple tandem repeats. Genetic rearrangements, insertions, inversions and duplications are also difficult to detect using microarrays.

Subtractive hybridization

Genomic subtraction, another powerful technique for whole-genome comparisons, has been used to compare the genomes of *M. bovis* and *M. bovis* BCG Connaught; three regions (RD1–3) have been identified that were deleted during the attenuation of *M. bovis* BCG (Ref. 17). Unlike microarrays, the strength of this technique is its ability to identify regions that are present in some members of a species but absent from the genome of the sequenced strain.

Bioinformatics

The alignment of complete genome sequences *in silico* is the ultimate DNA-based comparative strategy. With the growing number of completed genome sequences, and advances in bioinformatics, highly refined comparisons of sequence variation between two strains are possible using genome-alignment tools such as MUMmer (Ref. 18) or ACT (http://www.sanger.ac.uk/Software/ACT/). This is by far the most informative approach and its power is increasing as more sequences become available. At the time of writing, complete genome sequences exist for *M. tuberculosis* and *M. leprae*[1,19] and the genome-sequencing projects for *M. bovis*, *M. bovis* BCG, *M. microti*, *M. avium*, *M. paratuberculosis*, *M. smegmatis* and *M. ulcerans* are at various stages of completion (Box 1).

Proteomics

Comparative proteome analysis is based on two-dimensional electrophoresis (2DE) of proteins, mass spectrometry and database comparisons. Jungblut *et al*. used this technique to compare the proteome of *M. tuberculosis* H37Rv with that of BCG (Ref. 20). More recently, 2DE gel comparisons combined with *in silico* analysis of the genome sequences were carried out for *M. tuberculosis* strains H37Rv and CDC1551 (Ref. 21). Comparative proteomics has the potential to detect very subtle differences, for example changes in abundance or pI, as well as post-translational modifications such as methylation, glycosylation or phosphorylation.

Results of genome-wide comparisons

M. tuberculosis *isolates*

The clinical course and pattern of TB is highly variable, ranging from life-long asymptomatic infection to rapidly progressive pulmonary or disseminated disease. Until recently, this variability was assumed to be the result of differences in an individual's susceptibility, and a variety of acquired and innate host factors have been described that can modify the outcome of *M. tuberculosis* infection[22]. However, the

Box 1. *Mycobacterium* genome-sequencing projects

Mycobacterium smegmatis*, *Mycobacterium avium
http://www.tigr.org/tdb/mdb/mdbinprogress.html

***Mycobacterium ulcerans*, *Mycobacterium microti*, *Mycobacterium bovis* BCG**
http://www.pasteur.fr/recherche/unites/Lgmb/

***Mycobacterium tuberculosis* H37Rv**
http://genolist.pasteur.fr/TubercuList/

***Mycobacterium tuberculosis* CDC1551**
http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmt

***Mycobacterium tuberculosis* strain 210**
http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?

Mycobacterium bovis
http://www.sanger.ac.uk/Projects/M_bovis/

Mycobacterium leprae
http://genolist.pasteur.fr/Leproma/

Mycobacterium paratuberculosis
http://www.cbc.umn.edu/ResearchProjects/AGAC/Mptb/Mptbhome.html

development of robust molecular epidemiological techniques has enabled the identification of clonal populations of clinical isolates, some of which have been associated with microepidemics[23]. This has fueled speculation that some individual strains of *M. tuberculosis* could be 'hypervirulent', and has intensified the search for genetic polymorphisms that could account for such phenotypes.

An *in silico* genome comparison of the virulent laboratory strain H37Rv and the recently isolated epidemic strain CDC1551 revealed a polymorphism rate of approximately 1 in 3000 bp, although 50% of these substitutions were intergenic, giving a lower rate for the coding regions of the genome[21]. This was also found to be the case when the sequences of multiple genes were compared between strains of diverse origin and between members of the *M. tuberculosis* complex[4,5]. Thus, nucleotide substitutions are apparently not a significant source of genetic diversity either between strains of *M. tuberculosis* or between members of the *M. tuberculosis* complex. Are there other genetic polymorphisms in *M. tuberculosis* that could account for the perceived phenotypic differences between strains? The identification of a series of deletions in *M. bovis* relative to *M. tuberculosis* suggested that insertion and deletion events (InDels) could be one such mechanism[11,13]. In a recent microarray study, only one of 16 clinical isolates analyzed had the full complement of H37Rv genes, the other 15 having lost between three and 38 ORFs[14]. Interestingly, in at least five cases, these deletions were associated with IS*6110*, an insertion element present in variable numbers in the majority of *M. tuberculosis* strains[23]. Previously, it has been shown for several mycobacterial species[24–26] that recombination between adjacent insertion elements can lead to the deletion of the intervening genomic segment, suggesting this is an important mechanism of generating diversity. Recombination between an identical 11 bp repeat has also been shown to result in loss of the *katG* gene and resistance to the antimycobacterial drug isoniazid[27].

Further support for the role of InDels comes from an *in silico* full-genome comparison of *M. tuberculosis* H37Rv and CDC1551 (Ref. 21). Forty-three InDels (>3 bp) were reported in the coding region of the genome, of which 21 involved the PE and PPE gene families[1,28], which have not been accurately interrogated by microarrays. These intriguing gene families, totalling >170 proteins, are characterized by a conserved amino-terminal segment with either a proline–glutamic acid (PE) or a proline–proline–glutamic acid (PPE) motif combined with a carboxy-terminal domain comprising varying numbers of short repetitive motifs. Extensive polymorphisms in the PE and PPE genes are also known to be present more generally among strains of *M. tuberculosis*, as DNA probes based on the repetitive motifs of both these families have been successfully exploited as the basis of high-resolution molecular epidemiology[23]. The concentration of InDels in the PE and PPE genes and the observation that these are highly polymorphic suggests that these genes are the principal source of genetic diversity in *M. tuberculosis*. Although their function is currently unknown, various members have been implicated in the pathogenesis of both *M. tuberculosis*[29] and *M. marinum*[30]. It has also been suggested that the repetitive carboxy-terminal region of these proteins is reminiscent of protein families known to generate antigenic diversity in other organisms[1]. As the outcome of infection by *M. tuberculosis* is highly dependent on the efficacy of host immunity, it is tempting to speculate that, if these proteins are indeed antigenic, their polymorphic nature could result in differences in the immune response that would account for the apparent variability of disease among patients infected with *M. tuberculosis*.

Other polymorphic elements have been detected in *M. tuberculosis* that have been exploited for molecular epidemiological purposes, including the direct repeat (DR) region[23], which forms the basis of spoligotyping, and the more recently identified mycobacterial interspersed repetitive units (MIRUs)[31–34]. MIRUs are 40–100 bp elements often found as tandem repeats in intergenic regions of the *M. tuberculosis* complex. At least 12 of these are polymorphic, with variable numbers of the tandem repeats occurring in different strains of

Table 1. Genomic regions that differ between *M. tuberculosis* H37Rv and BCG Pasteur[a,b]

Putative function	Polymorphic genomic region
Putative antigens and virulence factors	
ESAT-6 system	RD1, RD5, RD8
MPT64	RD2
Phospholipases C (MPT40)	RD5, RvD2
PE and PPE proteins	RD1, RD2, RD5, RD6, RD8, RD14, N-RD25, RvD4
Possible invasion operon with six MceP	RD7
Possible lipopolysaccharide synthesis systems	RD4
Mobile elements	
Prophages	RD3, RD11
Insertion elements	RD6, RD11, RD14
Regulatory genes	
Transcriptional regulator	RD2, RD13, RD14, RvD5, DU2
SigI	N-RD18
SigM, SigH	DU1, DU2
Intermediary metabolism	
Precorrin methylase	RD9
Thiosulfate sulfurtransferase, molybdopterin metabolism, cytochrome P450	RD12, RD13, RvD5
Oxidoreductases	RD8, RD9, RD10, RD13
Epoxide hydrolase	RD8
Lipid metabolism	
Enoyl CoA hydratase	RD10
Membrane proteins	
Lipoproteins	RD7, RD8
Unknown putative membrane proteins	RD2, RD9, RD4, N-RD17, RvD1
Conserved hypotheticals	
Conserved hypothetical proteins	RD1, RD2, RD4, RD7, RD10, RD14, N-RD18

[a]Abbreviation: ESAT, early secreted antigenic target.
[b]RD nomenclature as proposed by Refs 17, 11 and 46. More infomation on the exact location of the deleted regions can be obtained via http://www.pasteur.fr/recherche/unites/Lgmb/Deletion.html RD, absent from BCG Pasteur; RvD, absent from M. tuberculosis H37Rv; DU, duplicated in BCG Pasteur; N-RD regions see Ref. 15.

M. tuberculosis[32,33,35]. Their function is unknown but they are similar to short sequence repeats (SSRs), which in other bacterial pathogens can modulate gene expression[36]. However, in a comparative proteomic study of H37Rv and CDC1551 grown *in vitro*, only 15 differences in protein expression were described out of a total of 1750 gel spots identified[21]. Although gene expression *in vivo* might be different and only the soluble protein fractions were compared, this astonishing similarity suggests patterns of gene regulation are highly conserved in *M. tuberculosis*.

BCG

In spite of its extensive use, the reason why BCG is attenuated is still unknown. This remarkably safe live vaccine has never regained virulence, suggesting that irreversible genomic changes, such as deletions, occurred during attenuation. The search for the basis of this attenuation has inspired a variety of comparative studies[10,11,13,15,17]. The combined findings show that there are at least 18 variable (RD) regions ranging from 0.3–12.7 kb, representing 120 genes, that are present in *M. tuberculosis* H37Rv but absent from BCG Pasteur (Table 1). These might account for the phenotypic differences between the vaccine strain and *M. tuberculosis.*

Specific PCR analysis of these RD regions among members of the *M. tuberculosis* complex showed that most of the RD regions absent from BCG were also absent from other strains of *M. bovis*, indicating that some of these variable regions reflect the evolutionary divergence of *M. tuberculosis* and *M. bovis* rather than genomic modifications that were introduced during the attenuation of BCG (Refs 11,13). More insight into the original attenuation process should be gained by constructing recombinant BCG knock-in strains carrying the various RD regions and/or by rational gene knockout in *M. tuberculosis*[37,38]. In addition, this approach could produce more potent TB vaccines.

Comparative genomics has also revealed two large tandem duplications of 29 and 36 kb (DU1 and DU2, respectively) in BCG Pasteur[12] (Fig. 2). These duplications seem to have arisen independently as their presence and/or size varies between the different BCG substrains. Although DU1 appears to be restricted to BCG Pasteur, DU2 has been detected in all BCG substrains tested so far (S. Cole *et al*., unpublished). Interestingly, DU1 comprises the *oriC* locus, indicating that BCG Pasteur is diploid for *oriC*, and several genes involved in replication[12]. For DU2, we know that the tandem duplication resulted in diploidy for 30 genes, including *sigH*, which encodes a sigma factor implicated in the heat-shock response[39]. Gene duplications are a common evolutionary response in bacteria exposed to different selective pressures in the laboratory and presumably in nature[40,41], as they can increase gene dosage, generate novel functions from potential gene-fusion events at duplication endpoints and are a source of redundant DNA for divergence. As such, the duplication events seen in BCG Pasteur might reflect a common adaptation mechanism of mycobacteria to cope with environmental stress, and they could influence the immunogenicity of a particular vaccine strain. It will now be of great interest to extend this analysis to individual *M. tuberculosis* isolates, using appropriate methods to detect duplications.

M. bovis

The availability of the nearly finished sequence of *M. bovis* AF2122/97 has allowed *in silico* comparison of the genomes of *M. tuberculosis* and *M. bovis*[42]. This preliminary analysis confirmed the deletions identified by RD PCR analysis[11], as well as allowing the identification of a deletion (RD17) that causes the truncation of *treY* (*glgY*), a gene encoding a maltooligosyltrehalose synthase in *M. bovis*[42]. More surprisingly, this initial analysis did not

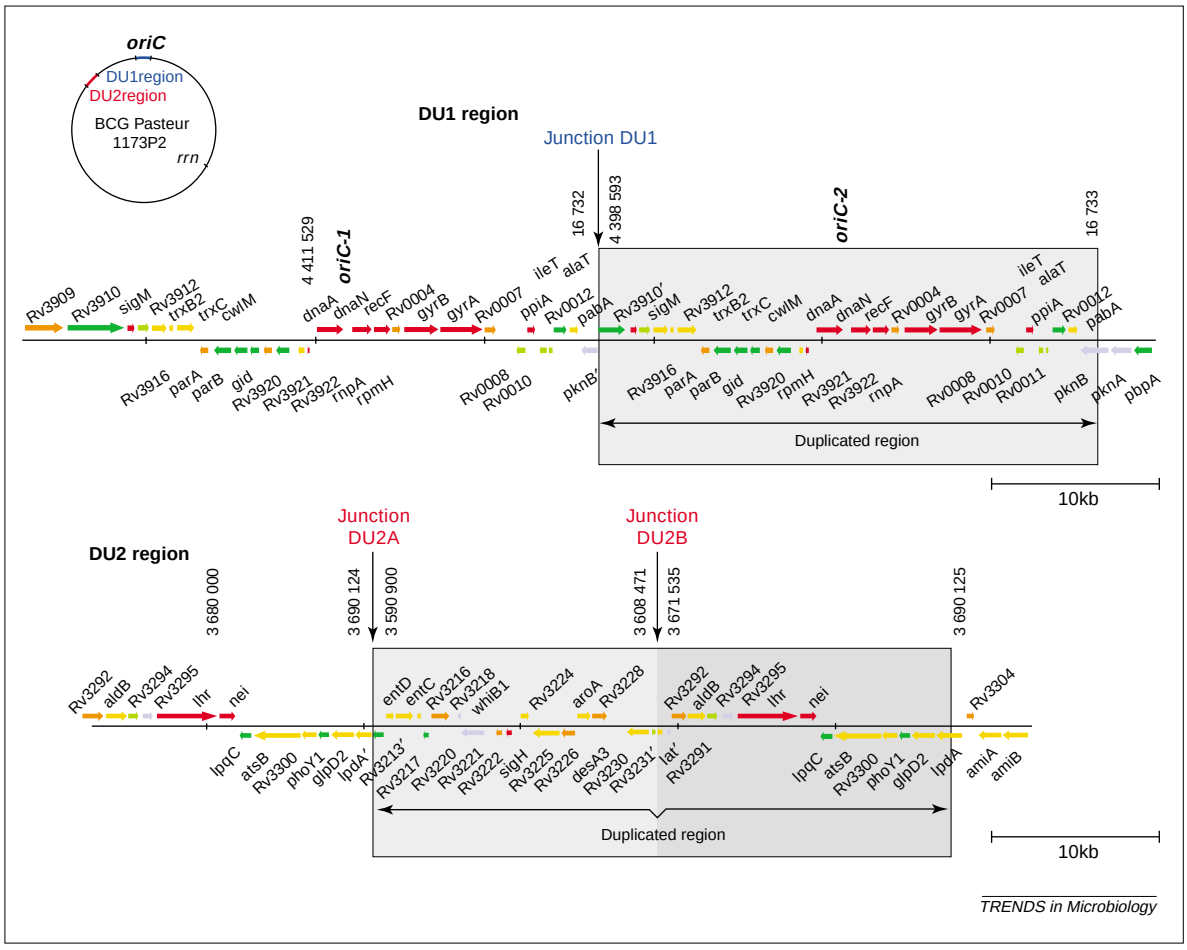

Fig. 2. Gene arrangements in the duplicated genomic regions DU1 and DU2 in *Mycobacterium bovis* BCG Pasteur[12]. Gene nomenclature is based on the highly similar *Mycobacterium tuberculosis* H37Rv genomic nucleotide sequence (Acc. No. AL123456). Further information on putative functions of the duplicated genes can be obtained at http://genolist.pasteur.fr/TubercuList/.

reveal any new gene clusters that were confined specifically to *M. bovis* (i.e. with no counterparts in *M. tuberculosis*), suggesting that the genome of *M. bovis* has less DNA than that of *M. tuberculosis*. It remains to be determined why *M. bovis* – which apparently has the smallest genome of the members of the *M. tuberculosis* complex – has the broadest host range of them all.

One possible explanation is that the presence of certain genes could be detrimental to the pathogenic lifestyle of the bacillus in specific hosts. The deletions from *M. bovis* could therefore represent a fine-tuning of its virulence as it evolved. In the case of *Shigella*, the deletion of *cadA*, encoding lysine decarboxylase, has been shown to be linked to increased virulence[43]. This results from a block in the production of cadaverine, which could protect host cells against the action of *Shigella* enterotoxin. It is also clear that variation in the PE and PPE repertoire is a major source of difference between *M. bovis* and *M. tuberculosis* H37Rv (Ref. 42). An example of variation in the PPE family between strains of *M. tuberculosis* had been described previously[1], and this most probably arose by strand slippage during replication of the underlying repetitive motifs, resulting in a difference in repeat copy number. When the amino acid sequences of the PPE proteins encoded by Rv1753c from *M. tuberculosis* H37Rv and *M. bovis* were compared, the latter was shown to harbour a 52-amino acid deletion and a 50-amino acid insertion. The PPE protein encoded by Rv1917c, which contains major polymorphic tandem repeats (MPTR), shows even greater variation between the bacilli, with the *M. bovis* sequence containing three discrete insertions relative to the *M. tuberculosis* sequence that introduce an extra 481 amino acid residues of highly repetitive sequence.

Genome dynamics and evolutionary clues

Although the genes of the *M. tuberculosis* complex members are well conserved[4], whole-genome comparisons have uncovered several polymorphic regions, some of which reflect recent genetic events (Table 1). Examples are the IS*6110*-mediated deletion of the RvD2 locus in *M. tuberculosis* H37Rv, which is still present in the closely related avirulent derivative H37Ra (Ref. 26), but shows great variability in clinical isolates[44], or the loss of the RD14 locus from BCG Pasteur, which remains in all other BCG substrains[13].

By contrast, the absence of regions RD7–10 from *M. microti*, *M. bovis* and BCG probably reflects a

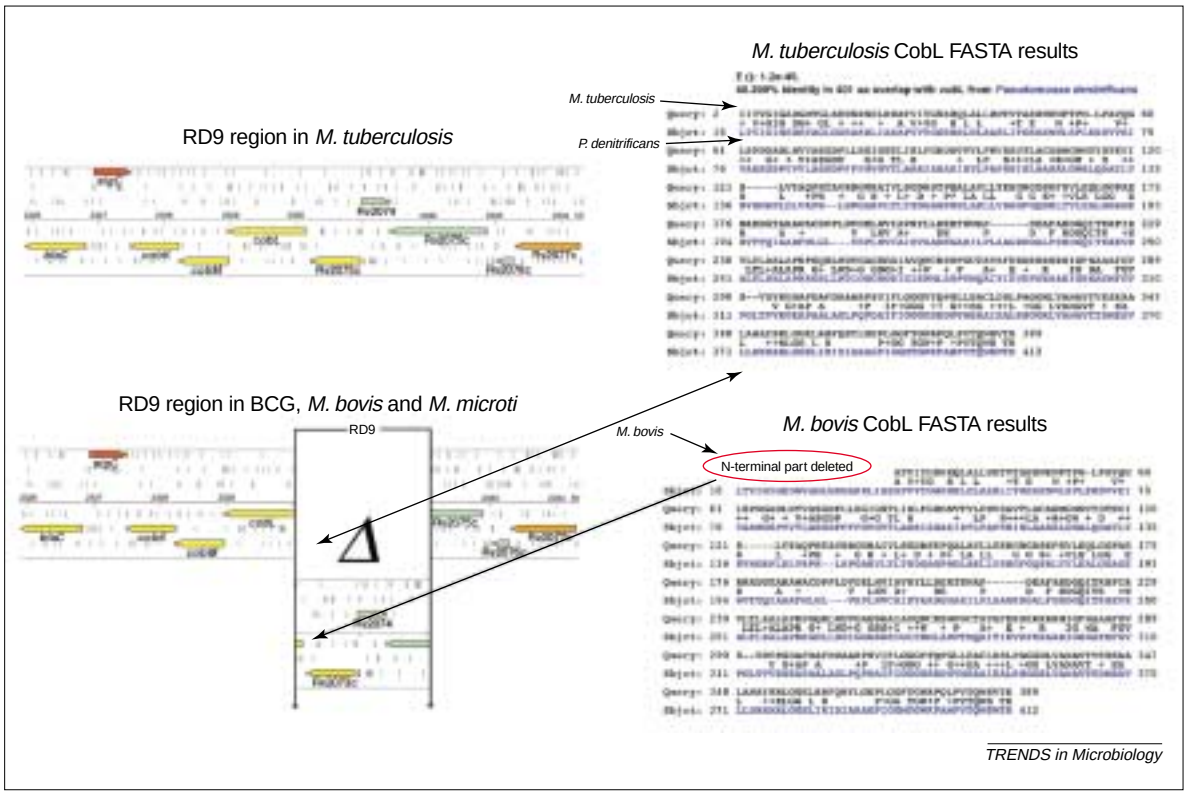

Fig. 3. Map of the RD9 region from *Mycobacterium tuberculosis* and *Mycobacterium bovis* showing the genomic segment absent from BCG, *M. bovis, Mycobacterium microti* and some *Mycobacterium africanum* strains, which is predicted to encode two complete genes Rv2073c and Rv2074 as well as the 5′-end of *cobL*. FASTA analysis of the CobL sequences from *M. tuberculosis* revealed that the amino-terminal region of CobL is highly conserved in a wide variety of bacteria, and an alignment with the sequence from *Pseudomonas denitrificans* is shown. The interruption of CobL indicates that the RD9 polymorphism is caused by the deletion of Rv2073c and Rv2074 from the common ancestor of *M. africanum*, *M. microti*, *M. bovis* and BCG rather than the insertion of these genes into *M. tuberculosis*.

much older event in evolutionary terms. From close inspection of the flanking sequences it is apparent that deletions occurred – in genes that are still intact in *M. tuberculosis* and *M. canettii* – at exactly the same site in BCG, *M. bovis* and *M. microti* (Fig. 3). This observation argues strongly against the possibility that the RD regions resulted from insertion of genes into *M. tuberculosis*. Interestingly, some *M. africanum* strains were also deleted for region RD9 (Ref. 11). Therefore, these deletions seem to have occurred in a common ancestor of *M. africanum*, *M. microti* and *M. bovis* that resembled *M. tuberculosis* and *M. canettii*. This is an important observation because it contradicts the often proposed hypothesis that human TB evolved from bovine TB by adaptation of an animal pathogen to the human host[45]. According to the distribution of RDs, *M. tuberculosis* strains appear to be more closely related to the common ancestor of the *M. tuberculosis* complex than are *M. bovis* strains. It seems plausible that a separate lineage represented by *M. africanum* (RD9), *M. microti* (RD7, RD8, RD9 and RD10) and *M. bovis* (RD4, RD5, RD7, RD8, RD9, RD10, RD12 and RD13)[46] evolved from the progenitor of today's *M. tuberculosis* isolates and adapted to new hosts. Whether or not at that stage the progenitor of *M. tuberculosis* was already a human pathogen is open to speculation.

To obtain a global evolutionary picture, comparative genomics involving the environmental species *M. avium* and *M. smegmatis* will certainly be of great value for understanding the transition of harmless soil-bacteria into obligate intracellular pathogens. One of the most astonishing findings that can be drawn from a preliminary analysis is that the genome of *M. smegmatis* (http://www.tigr.org/tdb/mdb/mdbinprogress.html) is similar in size (7–8 Mb) to that of the related actinomycete *Streptomyces coelicolor* (http://www.sanger.ac.uk/Projects/S_coelicolor/). This finding, together with results from 16S rRNA sequence data, suggests that the branch of slow-growing mycobacteria represents the most recently evolved part of the genus[47]. It is conceivable that the loss of genes, rather than the acquisition of additional genetic material by horizontal transfer, has been an important factor in slow-growing mycobacteria becoming pathogens. This trend continues, in extreme form, in the genome of *M. leprae*, which has undergone extensive reductive evolution[19], and now comprises only 3.27 Mb and ~1600 genes. The leprosy bacillus could have lost more than half of the genes that were present in the last common ancestor of slow-growing mycobacteria; this downsizing process has most probably defined the minimal gene-set required for a pathogenic mycobacterium.

Acknowledgements
We are grateful to K. Eiglmeier, T. Garnier, T Stinear and G. Hewinson for advice and encouragement. This work was supported by the Association Française Raoul Follereau, the European Union (grants BMH4-CT97-2277, QLK2-CT-1999-01093), the Wellcome Trust and the Institut Pasteur. A.P. is a recipient of a Wellcome Trust Fellowship in Clinical Tropical Medicine and is coaffiliated with the Liverpool School of Tropical Medicine.

References

1 Cole, S.T. *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544
2 Springer, B. *et al*. (1996) Two-laboratory collaborative study on identification of mycobacteria: molecular versus phenotypic methods. *J. Clin. Microbiol.* 34, 296–303
3 Imaeda, T. (1985) Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium microti* and *Mycobacterium africanum*. *Int. J. Syst. Bacteriol.* 35, 147–150
4 Sreevatsan, S. *et al*. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9869–9874
5 Musser, J.M. *et al*. (2000) Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. *Genetics* 155, 7–16
6 Singh, S.P. *et al*. (1999) Use of pulsed-field gel electrophoresis for molecular epidemiologic and population genetic studies of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 37, 1927–1931
7 Zhang, Y. *et al*. (1995) Genetic differences between BCG substrains. *Tubercul. Lung Dis.* 76, 43–50
8 Philipp, W.J. *et al*. (1996) An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*. *Proc. Natl. Acad. Sci. U. S. A.* 93, 3132–3137
9 Philipp, W.J. *et al*. (1996) Physical mapping of *Mycobacterium bovis* BCG Pasteur reveals differences from the genome map of *Mycobacterium tuberculosis* H37Rv and from *Mycobacterium bovis*. *Microbiology* 142, 3135–3145
10 Brosch, R. *et al*. (1998) Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome (BAC) library for genome mapping, sequencing and comparative genomics. *Infect. Immun.* 66, 2221–2229
11 Gordon, S.V. *et al*. (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol*. 32, 643–656
12 Brosch, R. *et al*. (2000) Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. *Comp. Funct. Genom. (Yeast)* 17, 111–123
13 Behr, M.A. *et al*. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarrays. *Science* 284, 1520–1523
14 Kato-Maeda, M. *et al*. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* 11, 547–554
15 Salamon, H. *et al*. (2000) Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. *Genome Res.* 10, 2044–2054
16 Cole, S.T. (1999) Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett.* 452, 7–10
17 Mahairas, G.G. *et al*. (1996) Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J. Bacteriol.* 178, 1274–1282
18 Delcher, A.L. *et al*. (1999) Alignment of whole genomes. *Nucleic Acids Res.* 27, 2369–2376
19 Cole, S.T. *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011
20 Jungblut, P.R. *et al*. (1999) Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol. Microbiol.* 33, 1103–1117
21 Betts, J.C. *et al*. (2000) Comparison of the proteome of *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC 1551. *Microbiology* 146, 3205–3216
22 Bellamy, R.J. and Hill, A.V. (1998) Host genetic susceptibility to human tuberculosis. *Novartis Found. Symp.* 217, 3–13
23 Kremer, K. *et al*. (1999) 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*. 37, 2607–2618
24 Eckstein, T.M. (2000) A genetic mechanism for deletion of the *ser2* gene cluster and formation of rough morphological variants of *Mycobacterium avium*. *J. Bacteriol.* 182, 6177–6182
25 Fang, Z. *et al*. (1999) IS*6110*-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis*. *J. Bacteriol.* 181, 1014–1020
26 Brosch, R. *et al*. (1999) Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. *Infect. Immun.* 67, 5768–5774
27 Pym, A.S. *et al*. (2001) Regulation of catalase–peroxidase (KatG) expression, isoniazid sensitivity and virulence by *furA* of *Mycobacterium tuberculosis*. *Mol. Microbiol.* 40, 879–889
28 Cole, S.T. and Barrell, B.G. (1998) Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. *Novartis Found. Symp.* 217, 160–172
29 Camacho, L.R. *et al*. (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* 34, 257–267
30 Ramakrishnan, L. *et al*. (2000) Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science* 288, 1436–1439
31 Supply, P. *et al*. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol. Microbiol.* 26, 991–1003
32 Magdalena, J. *et al*. (1998) Specific differentiation between *Mycobacterium bovis* BCG and virulent strains of the *Mycobacterium tuberculosis* complex. *J. Clin. Microbiol.* 36, 2471–2476
33 Magdalena, J. *et al*. (1998) Identification of a new DNA region specific for members of *Mycobacterium tuberculosis* complex. *J. Clin. Microbiol.* 36, 937–943
34 Supply, P. *et al*. (2000) Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol. Microbiol.* 36, 762–771
35 Mazars, E. *et al*. (2001) High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc. Natl. Acad. Sci. U. S. A.*. 13, 1901–1906
36 van Belkum, A. (1999) Short sequence repeats in microbial pathogenesis and evolution. *Cell. Mol. Life Sci.* 30, 729–734
37 Pelicic, V. *et al*. (1997) Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* 94, 10955–10960
38 Parish, T. and Stoker, N.G. (2000) Use of a flexible cassette method to generate a double unmarked *Mycobacterium tuberculosis tlyA plcABC* mutant by gene replacement. *Microbiology* 146, 1969–1975
39 Fernandes, N.D. *et al*. (1999) A mycobacterial extracytoplasmic sigma factor involved in survival following heat shock and oxidative stress. *J. Bacteriol.* 181, 4266–4274
40 Lupski, J.R. *et al*. (1996) Chromosomal duplications in bacteria, fruit flies, and humans. *Am. J. Hum. Genet.* 58, 21–27
41 Riehle, M.M. *et al*. (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 525–530
42 Gordon, S.V. *et al*. (2001) Genomics of *Mycobacterium bovis*. *Tuberculosis* 81, 157–163
43 Maurelli, A.T. *et al*. (1998) 'Black holes' and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3943–3948
44 Ho, T.B.L. *et al*. (2000) Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Comp. Funct. Genom. (Yeast)* 17, 272–282
45 Stead, W.W. *et al*. (1995) When did *Mycobacterium tuberculosis* infection first occur in the new world? *Am. J. Respir. Crit. Care Med*. 151, 1267–1268
46 Brosch, R. *et al*. (2000) Genomics, biology, and evolution of the *Mycobacterium tuberculosis* complex. In *Molecular Genetics of Mycobacteria* (Hatfull, F. and Jacobs, W.R., Jr, eds), pp. 19–36, ASM Press
47 Pitulle, C. *et al*. (1992) Phylogeny of rapidly growing members of the genus *Mycobacterium*. *Int. J. Syst. Bacteriol.* 42, 337–343