

**REVIEW
ARTICLE****Comparative and functional genomics of the
Mycobacterium tuberculosis complex**

Stewart T. Cole

Tel: +33 1 45688446. Fax: +33 1 40613583. e-mail: stcole@pasteur.fr

Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Keywords: bioinformatics, PE and PPE families, pathogenesis, BCG vaccine, evolution**Background**

Tuberculosis has long been the scourge of humanity, claiming millions of lives. Evidence of its antiquity is available in the form of Egyptian and South American mummies, dating from 3000–5000 years BC, with symptoms typical of Potts disease, a rare tuberculous manifestation affecting the spine (Haas & Haas, 1996; Salo *et al.*, 1994). In Europe, pulmonary tuberculosis was the major cause of death in the 18–19th centuries, and during the industrial revolution its spread was facilitated by poor housing, bad sanitation, overcrowding and malnutrition. As living conditions improved, tuberculosis receded in the Western world but assumed greater prevalence in many developing countries where it had previously been of lesser importance. In part, this was due to demographic factors like those encountered during the industrial revolution, such as displacement of populations and urbanization. More recently, the HIV/AIDS epidemic has greatly exacerbated an already grave situation in the developing world by creating a deadly synergy with tuberculosis that leads to even worse morbidity and mortality (Murray, 1990).

At the present time, the World Health Organization estimates that eight million new cases of tuberculosis occur every year and that 25 million individuals worldwide will lose their lives to the disease in the coming decade (Dye *et al.*, 1999). Although the ultimate solution to the problem of tuberculosis will be socio-economic, many of these deaths could be prevented if better access to treatment were available and if vaccination were more effective. More alarmingly, on the basis of their tuberculin reactivity, a sign of prior infection, it has been calculated that one-third of the world's population has been infected with *Mycobacterium tuberculosis* (Dye *et al.*, 1999), the aetiological agent of the disease. These

individuals are thus at risk of presenting with disease later in life as their immunity wanes due to ageing or as a result of HIV infection (Lillebaek *et al.*, 2002). While immunization with the BCG vaccine prevents tuberculosis, particularly in children in the West, it is of limited efficacy in the developing world where the disease burden is highest (Fine, 1995).

A highly efficient treatment, known as short course chemotherapy, is available to cure the disease. This involves taking a combination of four drugs for a minimum period of 6 months. The lengthy treatment duration is imposed by the exceptionally slow growth of the tubercle bacillus. While high cure rates can be obtained by means of DOTS (Directly Observed Therapy Short-course) (Espinal *et al.*, 1999), this strategy would be even more effective if its duration could be reduced by at least 2 months. Regrettably, despite the efficacy of DOTS, drug resistance is becoming increasingly prevalent for a variety of operational reasons (Dye *et al.*, 2002). Among the challenges facing mycobacteriologists and biomedical researchers are the development of faster-acting drugs that also act on latent disease, and the creation of a vaccine that is universally efficacious. Genomics, the systematic analysis of the complete genetic material found in an organism by means of DNA sequencing and bio-informatics, is opening new avenues for research in these key areas and catalysing discovery.

The *Mycobacterium tuberculosis* complex

In 1882, in a remarkable feat of microbiology, Robert Koch isolated *M. tuberculosis* for the first time, and conclusively demonstrated in the guinea pig that this slow-growing mycobacterium was the agent of a human disease (Koch, 1882). Together with other highly related bacteria, *M. tuberculosis* forms a tightly knit complex, a single species as defined by DNA/DNA hybridization studies (Imaeda, 1985), which is characterized by a singular lack of diversity in the bulk of its genes

This review is based on the 2002 Marjory Stephenson Prize Lecture delivered by the author at the 150th Meeting of the Society for General Microbiology, 9 April 2002.

Table 1. Some properties of tubercle bacilli

TCH, thiophen-2-carboxylic hydrazide; PZA, pyrazinamide; R, resistant; S, sensitive.

Species	Host range	Morphotype	Niacin	Nitrate reduction	TCH	PZA
<i>M. africanum</i>	Humans	Rugose, dysgonic	+/-	+/-	R/S	S
<i>M. bovis</i>	Badgers, cattle, deer, elephants, goats, humans, lions, seals, etc.	Smooth, dysgonic	-	-	S	R
<i>M. bovis</i> BCG	-	Rugose, eugonic	-	-	S	R
<i>M. canettii</i>	Humans	Very smooth	+	+	R	S
<i>M. microti</i>	Voles	Tiny	+	-	S	S
<i>M. tuberculosis</i>	Humans	Rugose, eugonic	+	+	R	S

(Sreevatsan *et al.*, 1997). The *M. tuberculosis* complex comprises six members (Table 1): *M. tuberculosis*, the causative agent in the vast majority of human tuberculosis cases; *Mycobacterium africanum*, an agent of human tuberculosis in sub-Saharan Africa; *Mycobacterium microti*, the agent of tuberculosis in voles; *Mycobacterium bovis*, which infects a very wide variety of mammalian species including humans, and BCG (bacille Calmette–Guérin), an attenuated variant of *M. bovis*; and *Mycobacterium canettii*, a smooth variant that is very rarely encountered but causes human disease. Prior to the introduction of pasteurization of milk, *M. bovis* was responsible for ~6% of total tuberculosis deaths in humans in Europe.

BCG was derived by Calmette and Guérin from a virulent *M. bovis* isolate by 230 serial passages in a broth containing glycerol, potato-extract and bile salts (Calmette, 1927). During the course of these passages the *M. bovis* strain progressively lost its virulence for animals and was first shown to be harmless and protective in a child in 1921. Since that time BCG has been used extensively as a live vaccine against tuberculosis and also protects humans against leprosy (Anon, 1996). Three billion doses have now been administered with negligible side effects and this is strong testimony to the safety of the vaccine (Bloom & Fine, 1994). The attenuation process undergone by BCG probably involved the serial loss of genetic material, rendering reversion to virulence impossible. *M. microti*, the vole bacillus (Wells, 1937), is naturally attenuated for humans and has also been used successfully to protect against tuberculosis (Hart & Sutherland, 1977).

Evolution of the *M. tuberculosis* complex

Mycobacteria are abundant in soil and water so the *M. tuberculosis* complex probably arose as the result of an ecological niche change that culminated in pathogenicity for mammals and the apparent disappearance of the last free-living ancestor. It is generally believed that tuberculosis was acquired from cattle following the domestication of livestock at the beginning of the Neolithic period when the hunter–gatherer lifestyle was

replaced by agriculture. Consequently, it is widely accepted that *M. bovis* was the ancestor of *M. tuberculosis* (Haas & Haas, 1996). In seminal work, Musser and his colleagues examined the population genetics of the *M. tuberculosis* complex by multi-locus sequence typing and found remarkably high conservation of gene sequences with little evidence for synonymous substitutions. They concluded that the spread of tuberculosis was young in evolutionary terms and even suggested that *M. tuberculosis* emerged as a human pathogen as recently as 10 000–15 000 years ago, possibly coinciding with the Paleolithic–Neolithic transition (Kapur *et al.*, 1994; Sreevatsan *et al.*, 1997).

Microbiological properties

All members of the *M. tuberculosis* complex have a doubling time close to 24 h and take 3–4 weeks to form colonies on Petri dishes. There are marked differences in colonial morphology as colonies of *M. bovis* and *M. tuberculosis* are flatter and less rugose than those of BCG, which tend to be raised and more compact (Table 1). *M. microti* forms tiny colonies whereas *M. canettii* is smooth due to overproduction of phenolic glycolipid (PGL). In addition to PGL, which is not produced by *M. tuberculosis*, the highly impermeable cell envelope of tubercle bacilli contains a rich variety of lipids, such as the mycolic acids that confer acid-fastness; glycolipids like the inflammatory molecule lipoarabinomannan and its variants; polyketides like phenolphthiocerol, which complexes with mycocerosic acid to form the virulence factor phenolphthiocerol-dimycocerosate, PDIM; and polysaccharides such as arabinogalactan and arabinomannan (Daffé & Draper, 1998). A capsule is also present.

Unlike the other complex members, *M. microti* and *M. bovis* require pyruvate as a growth supplement. There are also differences in the natural resistance to certain antibiotics such as pyrazinamide (PZA), due to a missense mutation in the activating enzyme pyrazinamidase (Scorpio & Zhang, 1996), and thiophen-2-carboxylic hydrazide (TCH), as well as in the production of niacin (Table 1). All virulent members of the complex

are capable of withstanding phagocytosis and replicating within macrophages and monocytes.

Genomics of *M. tuberculosis*

An integrated approach was adopted for the genome project (Fig. 1), which was undertaken with the widely used reference strain *M. tuberculosis* H37Rv (Steenken & Gardner, 1946). Unlike some clinical isolates that often lose virulence after laboratory passaging this strain has retained full virulence in animals since its isolation in 1905. In the early phase of the project, a physical map of the 4.4 Mb chromosome was constructed using PFGE of macro-restriction fragments and this was connected to the gene map by means of hybridization with landmark clones from an ordered cosmid library bearing known sites or genetic markers (Philipp *et al.*, 1996). Subsequently, an ordered library of Bacterial Artificial Chromosome (BAC) clones was constructed containing large inserts of *M. tuberculosis* H37Rv DNA and this enabled near-complete coverage of the *M. tuberculosis* H37Rv genome to be achieved (Brosch *et al.*, 1998). A canonical set of 68 BAC clones carries 98.5% of the genome. Ordered clone libraries, particularly those based on episomal or integrating shuttle vectors (Bange *et al.*, 1999; Jacobs *et al.*, 1991), are invaluable tools for functional genomics of tubercle bacilli. Furthermore,

the importance of having an easily renewable, immortalized source of DNA for a category three pathogen cannot be overstated.

For the genome sequencing project, a combined strategy was employed that involved sequencing selected cosmid and BAC clones, as well as whole-genome shotgun sequencing. The minimally overlapping set of BAC clones containing large inserts of *M. tuberculosis* H37Rv DNA (Brosch *et al.*, 1998) was of critical importance for the timely completion of the *M. tuberculosis* H37Rv genome sequence, as it allowed the extremely G + C-rich areas of the genome, corresponding to the PE-PGRS genes (discussed further below), to be obtained as these were generally under-represented in the small insert shotgun libraries. The complete genome sequence of *M. tuberculosis* H37Rv comprises 4411532 bp and has a mean G + C content of 65.6 mol%. As the findings of the analysis have been described extensively elsewhere (Brosch *et al.*, 2000; Cole, 1999; Cole *et al.*, 1998; Tekaiia *et al.*, 1999), only a brief outline of selected features will be presented here.

The genome contains ~4000 genes distributed fairly evenly between the two strands and accounting for >91% of the potential coding capacity. Genes were classified into 11 broad functional groups and, today, precise or putative functions can be attributed to 52%, with the remaining 48% being conserved hypotheticals or unknown (see Camus *et al.*, 2002). Over 51% of the genes have arisen as a result of gene duplication or domain shuffling events, and 3.4% of the genome is composed of insertion sequences (IS) and prophages (ϕ iRv1, ϕ iRv2). There are 56 copies of IS elements belonging to the well-known IS3, IS5, IS21, IS30, IS110, IS256 and ISL3 families, as well as a new IS family, IS1535, that appears to employ a frameshifting mechanism to produce its transposase (Gordon *et al.*, 1999b). IS6110, a member of the IS3 family, is the most abundant element and has played an important role in genome plasticity.

Genomics and biology

The information gleaned from the genome sequence provided new and valuable insight into the biology of the tubercle bacillus and highlighted the importance of lipid metabolism to its lifestyle as at least 8% of the genome is dedicated to this activity (Cole *et al.*, 1998). While the cell envelope of *M. tuberculosis* was known to contain a remarkable array of lipids, glycolipids, lipoglycans and polyketides (Daffé & Draper, 1998) and the genome sequence revealed many of the genes required for their production, it was a surprise to find numerous genes and proteins that could confer lipolytic functions. Estimates of the concentrations of potential substrates available to a pathogen in host tissues suggest that lipids and sterols are more abundant than carbohydrates (Wheeler & Ratledge, 1994). While *M. tuberculosis* has the prototype β -oxidation cycle required for lipid catabolism, catalysed by the multifunctional FadA/FadB proteins, it also appears to have ~100 enzymes

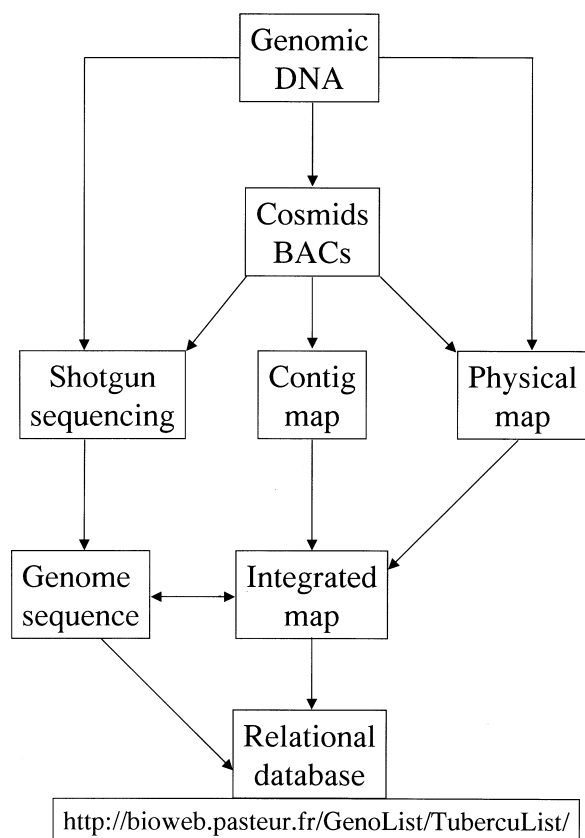


Fig. 1. Strategy used for genome projects involving tubercle bacilli.

potentially involved in alternative lipid oxidation pathways in which exogenous lipids from host cells could be degraded. Such large numbers of lipid-degrading functions have not yet been reported in other bacteria.

Whereas the tubercle bacillus appears to employ lipolysis as its principal catabolic pathway, it has no bias or obvious lesions in its anabolic repertoire. While this is fully consistent with our ability to culture *M. tuberculosis* in defined medium, it is somewhat unusual for an intracellular parasite to have retained such functions as the corresponding metabolites are often scavenged from the host. Although the presence of a complete network of anabolic systems is in agreement with the notion that the tubercle bacillus has only recently emerged as a human pathogen, and thus had insufficient time to adapt to a new host by shedding biosynthetic genes, it may also indicate that the availability of metabolic precursors is limiting within the phagosome. Support for the latter explanation is provided by the finding that genes for anabolic functions have been heavily conserved in the genome of *Mycobacterium leprae*, a related, obligate intracellular pathogen, in the face of massive reductive evolution that may have eliminated as many as 2600 genes (Cole *et al.*, 2001; Eiglmeier *et al.*, 2001).

There are, however, two additional arguments in favour of *M. tuberculosis* recently changing its niche and lifestyle. Firstly, the genome contains numerous genes (> 100) encoding regulatory proteins and signal transduction pathways that control gene expression (Cole *et al.*, 1998). Secondly, there are 20 enzyme systems that are predicted to use cytochrome P450 as a cofactor and these are often involved in the degradation of xenobiotics, or the modification of organic molecules, such as sterols, by means of their mono-oxygenase activity (Aoyama *et al.*, 1998). These enzymes are common in soil organisms where they enable diverse organic matter to be degraded to yield metabolizable sources of carbon and energy (Aoyama *et al.*, 1998; Munro & Lindsay, 1996). Both the regulatory networks and the P450 systems have been subject to massive gene decay in *M. leprae* (Cole *et al.*, 2001; Eiglmeier *et al.*, 2001).

The PE and PPE gene families

One of the major findings of the *M. tuberculosis* genome project was the identification of large gene families which were either unknown previously or poorly understood. Foremost among these were the novel PE and PPE families, comprising 100 and 67 members, respectively (Cole & Barrell, 1998; Cole *et al.*, 1998), which occupy about 8% of the genome. Members of each family share a conserved N-terminal domain of ~110 and 180 amino acid residues, with the characteristic motifs Pro-Glu (PE in single letter code) or Pro-Pro-Glu (PPE) at positions 8–9, or 8–10, respectively. The PE and PPE proteins can be divided into subfamilies on the basis of their C-terminal domains; in some cases these are simple and repetitive in sequence while in others they are of higher complexity. Belonging to the former group are the PE proteins of the PGRS (poly-

morphic GC-rich sequence) class (Poulet & Cole, 1995a) and the PPE proteins of the MPTR (major polymorphic tandem repeat) class. The PGRS encodes the motif AsnGlyGlyAlaGlyGlyAla, or variants thereof, while MPTR encodes Asn-X-Gly-X-Gly-Asn-X-Gly. Multiple tandem repetitions of these motifs are found in the corresponding proteins, which are acidic and exceptionally rich in glycine, and at the gene level variations occur in the repeat copy number and sequence thereby accounting for the genomic polymorphisms observed in hybridization patterns obtained with PGRS or MPTR probes (Hermans *et al.*, 1992; Poulet & Cole, 1995a, b; van Soolingen *et al.*, 1993). Initially, the PGRS and MPTR sequences were thought to correspond to dispersed tandem repeats or microsatellites but the finding that they were part of coding sequences led to reflection about the functions of these proteins.

Variability and possible roles of the PE and PPE multigene families

Whole-genome comparisons and functional genomics have shed new light on the possible roles of the PE and PPE proteins. When the PE genes of *M. tuberculosis* strains H37Rv and CDC1551 were compared *in silico* it was found that the genes encoding a PE domain alone, or a PE domain followed by a unique protein sequence, were identical in both cases (Banu *et al.*, 2002; Betts *et al.*, 2000). By contrast, 39 of the 62 common PE-PGRS proteins displayed variability as a result of in-frame insertion or deletion of different Ala, Gly-rich coding sequences in the PGRS component of the gene, or harboured frameshift mutations. Furthermore, consistent with this finding, size variation was also seen on Western blot analysis of protein samples, prepared from different clinical isolates, using PE-PGRS specific antibodies (Banu *et al.*, 2002). As expected from the conserved repetitive structure, the antibodies cross-reacted with more than one PE-PGRS protein, suggesting that different proteins share common antigenic structures. It is hard to envisage how a protein with enzymic activity could accommodate insertion/deletion of amino acid sequences without losing activity. There is some similarity between structural proteins of insects, such as silk, and the PGRS domain and this suggests that the role of the PE-PGRS proteins may be purely structural.

There is growing evidence from signature-tagged mutagenesis and micro-array studies that some *M. tuberculosis* PE-PGRS proteins may be involved in pathogenesis (Camacho *et al.*, 1999). In addition, members of the PE-PGRS families have been implicated in the pathogenesis of *Mycobacterium marinum* (Ramakrishnan *et al.*, 2000), where at least two genes were shown to be up-regulated strongly following phagocytosis of the bacterium.

Subcellular fractionation studies and immunogold or fluorescent antibody staining localized some PE-PGRS proteins in the cell wall and cell membrane of *M. tuberculosis* (Banu *et al.*, 2002; Brennan & Delogu,

2002). Disruption of the *M. tuberculosis* gene encoding the PE-PGRS protein Rv1818c resulted in greatly reduced bacterial clumping, suggesting that this protein may mediate cell–cell adhesion, and phagocytosis of the mutant cells by macrophages was also reduced (Brennan *et al.*, 2001). Another PE-PGRS protein, Rv1759c, that varies between strains, binds fibronectin and could thus mediate bacterial attachment to host cells (Espitia *et al.*, 1999; Singh *et al.*, 2001). The PE-PGRS proteins contain no obvious hydrophobic stretch that could act as a trans-membrane anchor and it is difficult to envisage how these proteins cross the cytoplasmic membrane. It has been speculated that a 23-amino-acid sequence that ends the PE domain and precedes the PGRS segment acts in membrane attachment but proof of this is lacking (Brennan *et al.*, 2001).

The immunogenicity of the PE-PGRS protein Rv1818c has been studied extensively in mice (Delogu & Brennan, 2001), where immunization with the PE domain induced Th1-type responses that were not found when the complete PE-PGRS protein was used. Instead, the PGRS part of the protein elicited antibodies and suppressed the Th1 response induced by the PE domain. The PE-PGRS proteins bear some sequence similarity to EBNA, the Epstein–Barr virus nuclear antigens, which block antigen presentation by the MHC class I pathway, through their action as proteasome inhibitors (Cole *et al.*, 1998). It was speculated that PE-PGRS proteins may also have inhibitory activity and it has recently been shown that the PGRS domain, when fused to GFP, confers increased resistance to proteosomal attack (Brennan & Delogu, 2002). If these immunological and adhesive properties are shared among other members of the family, it is conceivable that the extensive variation observed at the gene level could bestow very different phenotypes on the different strains.

The PPE proteins of the MPTR class also show variability (Zhang & Young, 1994), and the largest predicted PPE-MPTR protein detected contains 3300 amino acids. Extensive sequence variation has been reported for PPE proteins between *M. tuberculosis* and *M. bovis* (Gordon *et al.*, 2001a). Little evidence concerning the possible function of the PPE-MPTR proteins exists but one member of the PPE protein family was recently shown to be cell-wall-associated and surface-exposed (Sampson *et al.*, 2001). It seems increasingly likely that both the PPE-MPTR and PE-PGRS proteins may correspond to variable surface antigens (Banu *et al.*, 2002).

Comparative genomics

Several different approaches have been employed to compare the genomes of members of the *M. tuberculosis* complex, extending from various DNA array technologies, which easily identify deletion events but cannot readily uncover insertions (Behr *et al.*, 1999; Gordon *et al.*, 1999a; Kato-Maeda *et al.*, 2001; Salamon *et al.*, 2000), to highly sensitive whole-genome sequence com-

parisons (Brosch *et al.*, 2002; Gordon *et al.*, 2001b), which detect the full range of polymorphisms from single nucleotide polymorphisms (SNPs) to gene rearrangements. Many of these studies have compared virulent and avirulent strains in the hope of uncovering differences linked to changes in pathogenesis. A particularly useful finding from the whole-genome sequence comparison of *M. tuberculosis* and *M. bovis* was the presence of intact *mmpS6* and *mmpL6* genes in *M. bovis*. In most *M. tuberculosis* strains, both of these genes have been truncated and this region, termed TbD1 (Brosch *et al.*, 2002), is a very rare example of *M. tuberculosis* lacking functions that are present in the other members.

SNPs do occur in the genomes of members of the *M. tuberculosis* complex (Table 1) but at a relatively low level for a bacterium of 1 in every 2000–4000 bp (Sreevatsan *et al.*, 1997), depending on the species. Some SNPs, like the point mutation in the *pncA* gene responsible for pyrazinamide resistance (Scorpio & Zhang, 1996), result in phenotypic change but the majority seem to be silent. Consequently, InDels appear to be the most common means of generating diversity. Most of the insertions result from transposition events, generally involving IS6110, or more rarely from gene duplication. No conclusive evidence in favour of recent horizontal gene transfer occurring in the *M. tuberculosis* complex is available and the closest example of this is provided by the prophage genomes, phiRv1 or phiRv2, respectively (Brosch *et al.*, 2000) corresponding to regions of difference (RD) RD3 or RD11.

The deletions fall into two groups, ancient and recent. The ancient deletions occurred at different stages in the speciation process and are widespread whereas the recent deletions have a more restricted distribution. Examples of the latter are the IS6110-mediated deletion of the 7 kb locus RvD2 in *M. tuberculosis* H37Rv, still present in the closely related avirulent derivative H37Ra (Brosch *et al.*, 1999), or loss of the RD2 region encoding the protein antigen MPB64 from some strains of *M. bovis* BCG (Mahairas *et al.*, 1996). The RvD2 region also undergoes great variability in clinical isolates of *M. tuberculosis* and seems to represent a hot-spot for IS6110 transposition events (Ho *et al.*, 2001).

In contrast to these recent deletions, the absence of regions RD7, RD8, RD9 and RD10 from *M. microti*, *M. bovis* and BCG, which are still present in all *M. tuberculosis* strains, seems to be a much older event in evolutionary terms (Table 2). From close inspection of the DNA sequences bordering these RD regions it is apparent that deletions occurred within coding regions. Genes that are present in *M. tuberculosis* in full-length have been disrupted in BCG, *M. bovis* and *M. microti* at exactly the same location, whereas these coding sequences are still intact in *M. tuberculosis* and *M. canettii* strains. This finding rules out the possibility of the DNA in these regions having been acquired by *M. tuberculosis* but, instead, argues strongly in favour of loss of the corresponding genetic material by the other species.

Table 2. Deleted or truncated genes in the RD regions

The RD nomenclature used in this Table follows publication order and is based on that proposed initially by Mahairas *et al.* (1996) and Gordon *et al.* (1999a) and differs from that used by Behr *et al.* (1999).

Region deleted	Species	ORF, gene	Function
RD1	All <i>M. bovis</i> BCG	Rv3871–Rv3879c	PE, PPE, ESAT-6, various conserved hypotheticals
RD2	Some <i>M. bovis</i> BCG	Rv1978–Rv1988	Methyltransferases, permease, MPT64, ribonucleotide reductase, conserved hypothetical, membrane and secreted proteins, LysR transcriptional regulator
RD3	<i>M. africanum</i> , <i>M. microti</i> , some <i>M. bovis</i> , all <i>M. bovis</i> BCG	Rv1573–Rv1586c	phiRv1 prophage
RD4	Some <i>M. bovis</i> , all <i>M. bovis</i> BCG	Rv1505c–Rv1516c	Various membrane proteins and enzymes involved in exopolysaccharide synthesis
RD5	<i>M. microti</i> , <i>M. bovis</i> , <i>M. bovis</i> BCG	Rv2346c–Rv2353c	ESAT-6, PE, PPE family members, phospholipases C
RD6	Variable in all	Rv3425–Rv3428c	PPE proteins, IS1532
RD7	All except <i>M. tuberculosis</i> , <i>M. canettii</i>	Rv1964–Rv1977	Various exported and integral membrane proteins, MceP invasins
RD8	<i>M. microti</i> , <i>M. bovis</i> , <i>M. bovis</i> BCG	<i>ephA-lpqG</i>	Epoxide hydrolase, monooxygenase, lipoprotein, ESAT-6, PE, PPE protein family
RD9	All except <i>M. tuberculosis</i> , <i>M. canettii</i>	<i>cobL</i> –Rv2075	Precorrin methylase, oxidoreductase, exported protein
RD10	<i>M. microti</i> , <i>M. bovis</i> , <i>M. bovis</i> BCG	Rv0221–Rv0223c	Enoyl CoA hydratase, aldehyde dehydrogenase
RD11	<i>M. bovis</i> , <i>M. bovis</i> BCG	Rv2645–Rv2695c	phiRv2 prophage
RD12	<i>M. bovis</i> , <i>M. bovis</i> BCG	<i>sseC</i> –Rv3121	Thiosulfate sulfurtransferase, molybdopterin converting factor, methyltransferase, cytochrome P450
RD13	<i>M. bovis</i> , <i>M. bovis</i> BCG	Rv1255c–Rv1257c	Transcriptional regulator, cytochrome P450, dehydrogenase
RD14	Some BCG	Rv1765c–Rv1773c	PE_PGRS, conserved hypotheticals, IclR transcriptional regulator
RD15	Some BCG	Rv0309–Rv0312	Conserved hypotheticals
RD16	Some BCG	Rv3400–3405c	Transcriptional regulator, conserved hypotheticals, β -phosphoglucomutase
TbD1	Modern <i>M. tuberculosis</i>	<i>mmpS6</i> , <i>mmpL6</i>	Membrane proteins

Based on the presence or absence of such conserved RD regions, a degree of relatedness to the last common ancestor of the *M. tuberculosis* complex was proposed that shows that the lineages of *M. tuberculosis* and *M. bovis* separated before the *M. tuberculosis* specific deletion TbD1 occurred (Fig. 2). From this analysis it is clear that *M. bovis* cannot have been the ancestor of *M. tuberculosis* but, rather, appears to be descended from *M. tuberculosis* or to have emerged independently (Brosch *et al.*, 2002).

Some of these regions, primarily RD9 and TbD1 but also RD1, RD2, RD4, RD7, RD8, RD10, RD12 and RD13, represent very interesting candidates for the development of powerful diagnostic tools for the rapid and unambiguous identification of members of the *M. tuberculosis* complex (Brosch *et al.*, 2002). Fig. 2 presents a differential scheme for identifying individual species

that relies on the presence of these markers in association with selected SNPs such as the *mmpL6* 551AAC→AAG. This diagnostic strategy offers great promise to the epidemiology and evolutionary biology of the tubercle bacilli.

Functional genomics

One of the objectives of comparative genomics of the *M. tuberculosis* complex was to identify genes or loci that were different or lacking from avirulent or attenuated strains since their characterization would not only help in defining the molecular mechanisms of pathogenicity but might also furnish new leads for vaccine development, particularly in terms of creating new live vaccines. These could be recombinant variants of BCG (Stover *et al.*, 1991, 1992) conferring enhanced pro-

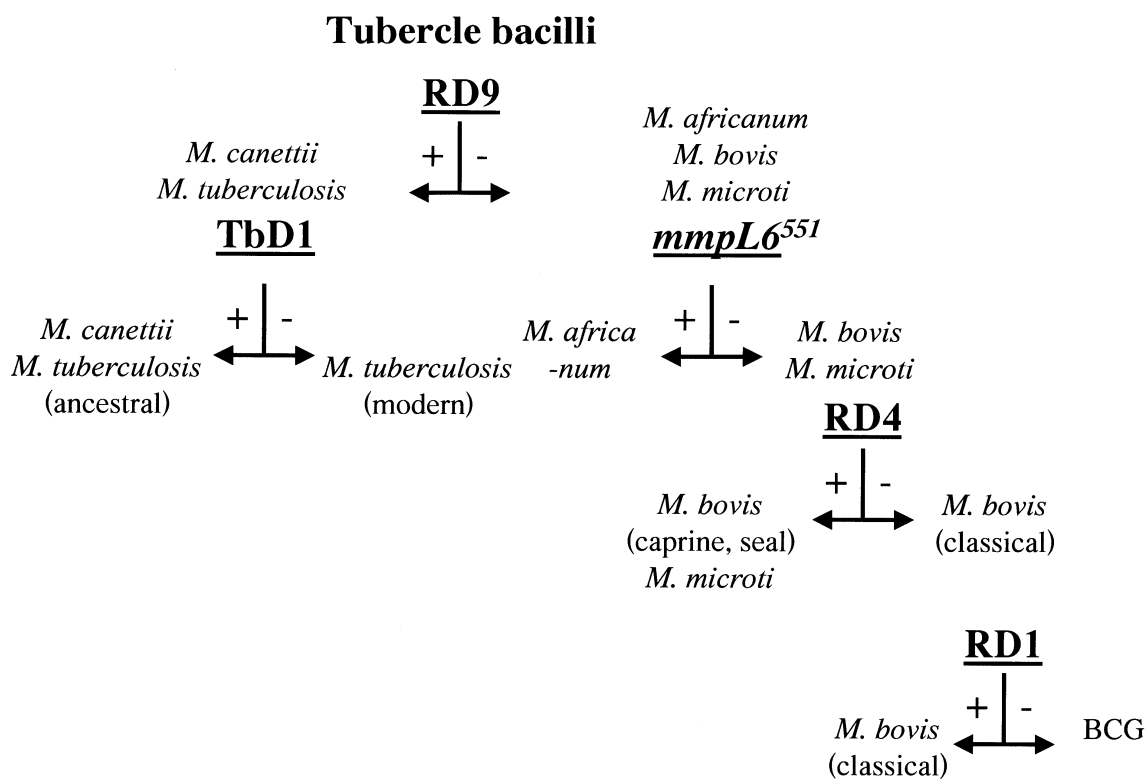


Fig. 2. PCR-based scheme for identifying tubercle bacilli at the species level. See Table 2 for details of the RD loci; +, region present; -, region absent. PCR would first be used with primers RD9-int-F and RD-int-R, and RD9-flank-F and RD9-flank-R (Brosch *et al.*, 2002) to determine whether the RD9 region is present. This splits the mycobacteria into two groups, which can be further subdivided by successive PCRs as shown, thus minimizing the need to perform unnecessary PCR reactions.

tection or even attenuated derivatives of *M. tuberculosis* (Hondalus *et al.*, 2000; Jackson *et al.*, 1999). Another by-product of comparative genomics is better understanding of the basis of host range, for instance why is *M. tuberculosis* confined to humans when *M. bovis* is capable of infecting such a broad range of mammals? Do specific mycobacterial factors determine the outcome? Answers to these questions will be provided by functional genomics and, in recent years, there have been spectacular advances in gene replacement technology (Bardarov *et al.*, 1997; Hinds *et al.*, 1999; Parish & Stoker, 2000; Pelicic *et al.*, 1997). It is now relatively straightforward to construct knockout mutants although this remains a lengthy process owing to the slow growth of tubercle bacilli.

Several of the RD regions described above contain genes that encode potential virulence factors like those characterized in other microbial pathogens (Table 2). These include prophages (RD3, RD11), phospholipases C (RD5), invasins (RD7) and an exopolysaccharide biosynthetic system (RD4). RD1 is the sole region that appears to be missing from the vaccine strains BCG and *M. microti* but is present in all virulent members of the *M. tuberculosis* complex. All *M. microti* strains tested have lost ~14 kb of DNA that has removed or inactivated genes Rv3864–Rv3876 (Brodin *et al.*, 2002)

and this deletion partially overlaps the RD1 locus of *M. bovis* BCG (Rv3871–Rv3879) (Mahairas *et al.*, 1996). However, while the proteins encoded by the corresponding genes belong to prominent mycobacterial protein families (Tekaiia *et al.*, 1999), it has not been possible to predict their functions by bio-informatics. Two of them, ESAT-6 and CFP-10 (Berthet *et al.*, 1998; Harboe *et al.*, 1996; Sorensen *et al.*, 1995), are small proteins, belonging to the ESAT-6 family, which might be secreted by early-exponential-phase cultures. They have attracted considerable immunological interest as a result of potent antigenicity for T cells. Interestingly, two other variable regions (RD5, RD8) also encode ESAT-6 family members, suggesting that there may be strong selective pressure imposed by the immune system for variants from which they have been lost (Gordon *et al.*, 1999a).

To test the biological effect that loss of these regions may have had on the different members of the *M. tuberculosis* complex, two different approaches are being pursued. On the one hand, the corresponding genes can be knocked-out or removed from the genome of *M. tuberculosis* using gene replacement technology or, on the other, they could be knocked-into species such as *M. bovis* BCG from which they are missing. In both cases, the phenotype of the resultant recombinants is

assessed using a combination of *in vitro* and *in vivo* assay systems. These complementary approaches will almost certainly unravel the basis for phenotypic differences among tubercle bacilli and provide insight into their pathogenesis and the attenuation mechanisms at play. Knowledge of the three-dimensional structures of the corresponding proteins and effectors is being generated by structural genomics programmes in which high-throughput technologies are providing datasets at atomic resolution (Cole, 2002). Clearly, all this new information will find rapid application in the development of new diagnostic tests, better drugs and vaccines and, hopefully, help to sway what seems at times a desparately unequal struggle against tuberculosis.

Acknowledgements

I would like to thank my many colleagues who have contributed in different ways to research on mycobacterial genomics, particularly B. G. Barrell, R. Brosch, K. Eiglmeier, T. Garnier, S. V. Gordon, N. Honoré and J. Parkhill. Work described was supported by the Institut Pasteur, the European Community (QLK2-CT1999-01093, QLRT-2000-02018), the Wellcome Trust and the Association Française Raoul Follereau.

References

- Anon (1996).** Randomised controlled trial of single BCG, repeated BCG, or combined BCG and killed *Mycobacterium leprae* vaccine for prevention of leprosy and tuberculosis in Malawi. Karonga Prevention Trial Group. *Lancet* **348**, 17–24.
- Aoyama, Y., Horiuchi, T., Gotoh, O., Noshiro, M. & Yoshida, Y. (1998).** CYP51-like gene of *Mycobacterium tuberculosis* actually encodes a P450 similar to eukaryotic CYP51. *J Biochem* **124**, 694–696.
- Bange, W. R., Collins, F. M. & Jacobs, W. R., Jr (1999).** Survival of mice infected with *Mycobacterium smegmatis* containing large DNA fragments from *Mycobacterium tuberculosis*. *Tuber Lung Dis* **79**, 171–180.
- Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M.-C. & Cole, S. T. (2002).** Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol Microbiol* **44**, 9–19.
- Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R., Bloom, B. R., Hatfull, G. R. & Jacobs, W. R., Jr (1997).** Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* **94**, 10961–10966.
- Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S. & Small, P. M. (1999).** Comparative genomics of BCG vaccines by whole-genome DNA microarrays. *Science* **284**, 1520–1523.
- Berthet, F. X., Rasmussen, P. B., Rosenkrandt, I., Andersen, P. & Gicquel, B. (1998).** A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology* **144**, 3195–3203.
- Betts, J. C., Dodson, P., Quan, S., Lewis, A. P., Thomas, P. J., Duncan, K. & McAdam, R. A. (2000).** Comparison of the proteome of *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC 1551. *Microbiology* **146**, 3205–3216.
- Bloom, B. R. & Fine, P. E. M. (1994).** The BCG experience: implications for future vaccines against tuberculosis. In *Tuberculosis: Pathogenesis, Protection, and Control*, pp. 531–557. Edited by B. R. Bloom. Washington, DC: American Society for Microbiology.
- Brennan, M. J. & Delogu, G. (2002).** The PE multigene family: a molecular mantra for mycobacteria. *Trends Microbiol* **10**, 246–249.
- Brennan, M. J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M. & Jacobs, W. R., Jr (2001).** Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect Immun* **69**, 7326–7333.
- Brodin, P., Eiglmeier, K., Marmiesse, M., Billault, A., Garnier, T., Niemann, S., Cole, S. T. & Brosch, R. (2002).** Bacterial Artificial Chromosome-based comparative genomic analysis identifies *Mycobacterium microti* as a natural ESAT-6 deletion mutant. *Infect Immun* **70** (in press).
- Brosch, R., Gordon, S. V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B. G. & Cole, S. T. (1998).** Use of a *Mycobacterium tuberculosis* H37Rv Bacterial Artificial Chromosome (BAC) library for genome mapping, sequencing and comparative genomics. *Infect Immun* **66**, 2221–2229.
- Brosch, R., Philipp, W., Stavropoulos, E., Colston, M. J., Cole, S. T. & Gordon, S. V. (1999).** Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. *Infect Immun* **67**, 5768–5774.
- Brosch, R., Gordon, S. V., Eiglmeier, K., Garnier, T., Tekaiia, F., Yeramian, E. & Cole, S. T. (2000).** Genomics, biology, and evolution of the *Mycobacterium tuberculosis* complex. In *Molecular Genetics of Mycobacteria*, pp. 19–36. Edited by G. F. Hatfull & W. R. Jacobs, Jr. Washington, DC: American Society for Microbiology.
- Brosch, R., Gordon, S. V., Marmiesse, M. & 12 other authors (2002).** A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA* **99**, 3684–3689.
- Calmette, A. (1927).** *La vaccination préventive contre la tuberculose*. Paris: Masson et cie.
- Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B. & Guilhot, C. (1999).** Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol* **34**, 257–267.
- Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. (2002).** Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148**, 2967–2973.
- Cole, S. T. (1999).** Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* **452**, 7–10.
- Cole, S. T. (2002).** Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J* (in press).
- Cole, S. T. & Barrell, B. G. (1998).** Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. *Novartis Found Symp* **217**, 160–172.
- Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998).** Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.
- Cole, S. T., Eiglmeier, K., Parkhill, J. & 42 other authors (2001).** Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.
- Daffé, M. & Draper, P. (1998).** The envelope layers of mycobacteria with reference to their pathogenicity. *Adv Microb Physiol* **39**, 131–203.
- Delogu, G. & Brennan, M. J. (2001).** Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. *Infect Immun* **69**, 5606–5611.

- Dye, C., Sheele, S., Dolin, P., Pathania, V. & Raviglione, M. C. (1999). Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. *JAMA (J Am Med Assoc)* **282**, 677–686.
- Dye, C., Espinal, M. A., Watt, C. J., Mbiaga, C. & Williams, B. G. (2002). Worldwide incidence of multidrug-resistant tuberculosis. *J Infect Dis* **185**, 1197–1202.
- Eiglmeier, K., Parkhill, J., Honoré, N. & 12 other authors (2001). The decaying genome of *Mycobacterium leprae*. *Lepr Rev* **72**, 387–398.
- Espinal, M. A., Dye, C., Raviglione, M. & Kochi, A. (1999). Rational 'DOTS plus' for the control of MDR-TB. *Int J Tuberc Lung Dis* **3**, 561–563.
- Espitia, C., Lacleste, J. P., Mondragon-Palomino, M. & 7 other authors (1999). The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? *Microbiology* **145**, 3487–3495.
- Fine, P. E. M. (1995). Variation in protection by BCG: implications of and for heterologous immunity. *Lancet* **346**, 1339–1345.
- Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K. & Cole, S. T. (1999a). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* **32**, 643–656.
- Gordon, S. V., Heym, B., Parkhill, J., Barrell, B. & Cole, S. T. (1999b). New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **145**, 881–892.
- Gordon, S. V., Eiglmeier, K., Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, S. T. & Hewinson, R. G. (2001a). Genomics of *Mycobacterium bovis*. *Tuberculosis (Edinb)* **81**, 157–163.
- Gordon, S. V., Eiglmeier, K., Garnier, T., Parkhill, J., Barrell, B., Cole, S. T. & Hewinson, R. G. (2001b). Genomics of *Mycobacterium bovis*. *Tuber Lung Dis* **6**, 157–163.
- Haas, F. & Haas, S. S. (1996). The origins of *Mycobacterium tuberculosis* and the notion of its contagiousness. In *Tuberculosis*, pp. 4–19. Edited by W. N. Rom & S. Garay. Boston: Little, Brown and Company.
- Harboe, M., Oettinger, T., Wiker, H. G., Rosenkrands, I. & Andersen, P. (1996). Evidence for occurrence of the ESAT-6 protein in *Mycobacterium tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG. *Infect Immun* **64**, 16–22.
- Hart, P. D. & Sutherland, I. (1977). BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Br Med J* **2**, 293–295.
- Hermans, P. W. M., van Soolingen, D. & van Embden, J. D. A. (1992). Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium goodii*. *J Bacteriol* **174**, 4157–4165.
- Hinds, J., Mahenthalingam, E., Kempell, K. E., Duncan, K., Stokes, R. W., Parish, T. & Stoker, N. G. (1999). Enhanced gene replacement in mycobacteria. *Microbiology* **145**, 519–527.
- Ho, T. B. L., Robertson, B. D., Taylor, G. M., Shaw, R. J. & Young, D. B. (2001). Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Yeast* **17**, 272–282.
- Hondalus, M. K., Bardarov, S., Russell, R., Chan, J., Jacobs, W. R., Jr & Bloom, B. R. (2000). Attenuation of and protection induced by a leucine auxotroph of *Mycobacterium tuberculosis*. *Infect Immun* **68**, 2888–2898.
- Imaeda, T. (1985). Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium micoti* and *Mycobacterium africanum*. *Int J Syst Bacteriol* **35**, 147–150.
- Jackson, M., Phalen, S. W., Lagranderie, M., Ensergueix, D., Chavarot, P., Marchal, G., McMurray, D. N., Gicquel, B. & Guilhot, C. (1999). Persistence and protective efficacy of a *Mycobacterium tuberculosis* auxotroph vaccine. *Infect Immun* **67**, 2867–2873.
- Jacobs, W. R., Jr, Kalpana, G. V., Cirillo, J. D., Pascopella, L., Snapper, S. B., Udani, R. A., Jones, W., Barletta, R. G. & Bloom, B. R. (1991). Genetic systems for mycobacteria. *Methods Enzymol* **204**, 537–555.
- Kapur, V., Whittam, T. S. & Musser, J. (1994). Is *Mycobacterium tuberculosis* 15,000 years old? *J Infect Dis* **170**, 1348–1349.
- Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N. & Small, P. M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* **11**, 547–554.
- Koch, R. (1882). Die Aetiologie der Tuberculose. *Berl Klin Wochenschr* **19**, 221–230.
- Lillebaek, T., Dirksen, A., Baess, I., Strunge, B., Thomsen, V. O. & Andersen, A. B. (2002). Molecular evidence of endogenous reactivation of *Mycobacterium tuberculosis* after 33 years of latent infection. *J Infect Dis* **185**, 401–404.
- Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C. & Stover, C. K. (1996). Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J Bacteriol* **178**, 1274–1282.
- Munro, A. W. & Lindsay, J. G. (1996). Bacterial cytochromes P-450. *Mol Microbiol* **20**, 1115–1125.
- Murray, J. R. (1990). Cursed duet: HIV-infection and tuberculosis. *Respiration* **57**, 210–220.
- Parish, T. & 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.
- Pellicic, V., Jackson, M., Reyat, J. M., Jacobs, W. R., Jr, Gicquel, B. & Guilhot, C. (1997). Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* **94**, 10955–10960.
- Philipp, W. J., Poulet, S., Eiglmeier, K. & 7 other authors (1996). An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*. *Proc Natl Acad Sci USA* **93**, 3132–3137.
- Poulet, S. & Cole, S. T. (1995a). Characterisation of the polymorphic GC-rich repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch Microbiol* **163**, 87–95.
- Poulet, S. & Cole, S. T. (1995b). Repeated DNA sequences in mycobacteria. *Arch Microbiol* **163**, 79–86.
- Ramakrishnan, L., Federspiel, N. A. & Falkow, S. (2000). Granuloma-specific expression of mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science* **288**, 1436–1439.
- Salamon, H., Kato-Maeda, M., Small, P. M., Drenkow, J. & Gingeras, T. R. (2000). Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. *Genome Res* **10**, 2044–2054.
- Salo, W. L., Aufderheide, A. C., Buikstra, J. & Holcomb, T. A. (1994). Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. *Proc Natl Acad Sci USA* **91**, 2091–2094.
- Sampson, S. L., Lukey, P., Warren, R. M., van Helden, P. D., Richardson, M. & Everett, M. J. (2001). Expression, charac-

terization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis (Edinb)* **81**, 305–317.

Scorpio, A. & Zhang, Y. (1996). Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat Med* **2**, 662–667.

Singh, K. K., Zhang, X., Patibandla, A. S., Chien, P., Jr & Laal, S. (2001). Antigens of *Mycobacterium tuberculosis* expressed during preclinical tuberculosis: serological immunodominance of proteins with repetitive amino acid sequences. *Infect Immun* **69**, 4185–4191.

Sorensen, A. L., Nagai, S., Houen, G., Andersen, P. & Andersen, A. B. (1995). Purification and characterization of a low-molecular-mass T-cell antigen secreted by *Mycobacterium tuberculosis*. *Infect Immun* **63**, 1710–1717.

Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S. & Musser, J. M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* **94**, 9869–9874.

Steenken, W., Jr & Gardner, L. U. (1946). History of H37 strain of tubercle bacillus. *Am Rev Tuberc* **54**, 62–66.

Stover, C. K., De la Cruz, V. F., Fuerst, T. R. & 11 other authors (1991). New use of BCG for recombinant vaccines. *Nature* **351**, 456–460.

Stover, C. K., de la Cruz, V. F., Bansal, G. P., Hanson, M. S., Fuerst, T. R., Jacobs, W. R., Jr & Bloom, B. R. (1992). Use of recombinant BCG as a vaccine delivery vehicle. *Adv Exp Med Biol* **327**, 175–182.

Tekaia, F., Gordon, S. V., Garnier, T., Brosch, R., Barrell, B. G. & Cole, S. T. (1999). Analysis of the proteome of *Mycobacterium tuberculosis in silico*. *Tuber Lung Dis* **79**, 329–342.

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

Wells, A. Q. (1937). Tuberculosis in wild voles. *Lancet* **232**, 1221.

Wheeler, P. R. & Ratledge, C. (1994). Metabolism of *Mycobacterium tuberculosis*. In *Tuberculosis: Pathogenesis, Protection, and Control*, pp. 353–385. Edited by B. R. Bloom. Washington, DC: American Society for Microbiology.

Zhang, Y. & Young, D. B. (1994). Strain variation in the *katG* region of *Mycobacterium tuberculosis*. *Mol Microbiol* **14**, 301–308.