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...
Table 1. Some properties of tubercle bacilli

<table>
<thead>
<tr>
<th>Species</th>
<th>Host range</th>
<th>Morphotype</th>
<th>Niacin</th>
<th>Nitrate reduction</th>
<th>TCH</th>
<th>PZA</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. africana</em></td>
<td>Humans</td>
<td>Rugose, dysgonic</td>
<td>+ / −</td>
<td>+ / −</td>
<td>R/S</td>
<td>S</td>
</tr>
<tr>
<td><em>M. bovis</em></td>
<td>Badgers, cattle, deer, elephants, goats, humans, lions, seals, etc.</td>
<td>Smooth, dysgonic</td>
<td>−</td>
<td>−</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td><em>M. bovis</em> BCG</td>
<td>—</td>
<td>Rugose, eugonic</td>
<td>−</td>
<td>−</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td><em>M. canetti</em></td>
<td>Humans</td>
<td>Very smooth</td>
<td>+</td>
<td>+</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td><em>M. microti</em></td>
<td>Voles</td>
<td>Tiny</td>
<td>+</td>
<td>−</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td><em>M. tuberculosis</em></td>
<td>Humans</td>
<td>Rugose, eugonic</td>
<td>+</td>
<td>+</td>
<td>R</td>
<td>S</td>
</tr>
</tbody>
</table>

(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 africana*, 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 canetti*, 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 10000–15000 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. canetti* is smooth due to overproduction of phenolic glycolipid (PGL). In addition to PGL, which is not produced by *M. tuberculosis* and also protects humans against leprosy, *M. microti* is smooth due to overproduction of phenolic glycolipid (PGL). In addition to PGL, which is not produced by *M. tuberculosis*, the highly permeable 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 mycolic acid to form the virulence factor phenolphthiocerol-dimycocerosate, PDIM; and polysaccharides such as arabino-galactan 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) 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; Tekaia 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 (phiRv1, phiRv2). 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 (Daffe & 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 lysis 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 (polymorphic 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; 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 comparisons (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. canetti 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).

<table>
<thead>
<tr>
<th>Region deleted</th>
<th>Species</th>
<th>ORF, gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD1</td>
<td>All M. bovis BCG</td>
<td>Rv3871–Rv3879c</td>
<td>PE, PPE, ESAT-6, various conserved hypotheticals</td>
</tr>
<tr>
<td>RD3</td>
<td>M. africanum, M. microti, some M. bovis, all M. bovis BCG</td>
<td>Rv1573–Rv1586c</td>
<td>phiRv1 prophage</td>
</tr>
<tr>
<td>RD4</td>
<td>Some M. bovis, all M. bovis BCG</td>
<td>Rv1505c–Rv1516c</td>
<td>Various membrane proteins and enzymes involved in exopolysaccharide synthesis</td>
</tr>
<tr>
<td>RD5</td>
<td>M. microti, M. bovis, M. bovis BCG</td>
<td>Rv2346c–Rv2353c</td>
<td>ESAT-6, PE, PPE family members, phospholipases C</td>
</tr>
<tr>
<td>RD6</td>
<td>Variable in all</td>
<td>Rv3425–Rv3428c</td>
<td>PPE proteins, IS1532</td>
</tr>
<tr>
<td>RD7</td>
<td>All except M. tuberculosis, M. canetti</td>
<td>Rv1964–Rv1977</td>
<td>Various exported and integral membrane proteins, MceP invasins</td>
</tr>
<tr>
<td>RD8</td>
<td>M. microti, M. bovis, M. bovis BCG</td>
<td>ephA–lpqG</td>
<td>Epoxide hydrolase, monoxygenase, lipoprotein, ESAT-6, PE, PPE protein family</td>
</tr>
<tr>
<td>RD9</td>
<td>All except M. tuberculosis, M. canetti</td>
<td>cobL–Rv2075</td>
<td>Precorrin methylase, oxido-reductase, exported protein</td>
</tr>
<tr>
<td>RD10</td>
<td>M. microti, M. bovis, M. bovis BCG</td>
<td>Rv0221–Rv0223c</td>
<td>Enoyl CoA hydratase, aldehyde dehydrogenase</td>
</tr>
<tr>
<td>RD11</td>
<td>M. bovis, M. bovis BCG</td>
<td>Rv2645–Rv2695c</td>
<td>phiRv2 prophage</td>
</tr>
<tr>
<td>RD12</td>
<td>M. bovis, M. bovis BCG</td>
<td>sseC–Rv3121</td>
<td>Thiosulfate sulfurtransferase, molybdo-hydrogenase converting factor, methyltransferase, cytochrome P450</td>
</tr>
<tr>
<td>RD13</td>
<td>M. bovis, M. bovis BCG</td>
<td>Rv1255c–Rv1257c</td>
<td>Transcriptional regulator, cytochrome P450, dehydrogenase</td>
</tr>
<tr>
<td>RD14</td>
<td>Some BCG</td>
<td>Rv1765c–Rv1773c</td>
<td>PE_PGRS, conserved hypotheticals, IclR transcriptional regulator</td>
</tr>
<tr>
<td>RD15</td>
<td>Some BCG</td>
<td>Rv0309–Rv0312</td>
<td>Conserved hypotheticals</td>
</tr>
<tr>
<td>RD16</td>
<td>Some BCG</td>
<td>Rv3400–Rv3405c</td>
<td>Transcriptional regulator, conserved hypotheticals, β-phosphoglucomutase</td>
</tr>
<tr>
<td>TbD1</td>
<td>Modern M. tuberculosis</td>
<td>mmpS6, mmpL6</td>
<td>Membrane proteins</td>
</tr>
</tbody>
</table>

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-
Genomics of the Mycobacterium tuberculosis complex

Tubercle bacilli

M. canetti
M. tuberculosis

RD9

M. africanum
M. bovis
M. microti

mmpL6

M. canetti
M. tuberculosis
(ancestral)

RD4

M. bovis
(caprine, seal)
M. microti

RD1

M. bovis
(clasical)

M. tuberculosis
(modern)

M. bovis

BCG

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.

Protection 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 (Tekaia 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., 1999).

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 desperately 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 Foul-terrace.

References


2926


