

Comparative genomics of the mycobacteria

Roland Brosch¹, Stephen V. Gordon^{1,2}, Alexander Pym¹, Karin Eiglmeier¹, Thierry Garnier¹, Stewart T. Cole¹

¹ Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

² Present address: Veterinary Laboratories Agency, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, England

Received January 25, 2000 · Revision received January 31, 2000 · Accepted January 31, 2000

Abstract

The genus mycobacteria includes two important human pathogens *Mycobacterium tuberculosis* and *Mycobacterium lepra*. The former is reputed to have the highest annual global mortality of all pathogens. Their slow growth, virulence for humans and particular physiology makes these organisms extremely difficult to work with. However the rapid development of mycobacterial genomics following the completion of the *Mycobacterium tuberculosis* genome sequence provides the basis for a powerful new approach for the understanding of these organisms. Five further genome sequencing projects of closely related mycobacterial species with differing host range, virulence for humans and physiology are underway. A comparative genomic analysis of these species has the potential to define the genetic basis of these phenotypes which will be invaluable for the development of urgently needed new vaccines and drugs. This minireview summarises the different techniques that have been employed to compare these genomes and gives an overview of the wealth of data that has already been generated by mycobacterial comparative genomics.

Key words: comparative genomics – mycobacteria

Preamble

Two of the most important human pathogens are *Mycobacterium tuberculosis*, the etiologic agent of tuberculosis, a chronic infectious disease that every year accounts for three million lives world-wide and *M. leprae*, the causative organism of leprosy. Other mycobacteria, in particular *M. avium-intracellulare* can cause disease in immuno-compromised individuals.

Recent increases in tuberculosis cases in both developing and industrialised countries, together with the emergence of potentially untreatable drug-resistant

strains, have refocused the attention of scientists on the biology of these pathogens. One of the common features of the medically important mycobacteria is the considerable difficulty inherent in their manipulation in the laboratory. These pathogens have to be handled under biohazard-containment facilities, and are difficult to cultivate. *M. tuberculosis* has a generation time of 24 hours in in vitro culture, whilst *M. leprae* cannot be cultivated at all on artificial media and has an extremely long generation time of two weeks or more in experimentally infected animals. In addition mycobacteria possess an exceptionally lipid-rich cell wall that

Corresponding author: Roland Brosch, Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France, Phone: 33-1-45 68 84 49, Fax: 33-1-40 61 35 83, E-mail: rbrosch@pasteur.fr, <http://www.pasteur.fr/recherche/unites/Lgmb/>

confers resistance to desiccation, protects against numerous antibiotics and antimicrobial agents, and has impeded the application of conventional molecular genetic techniques. (Daffe and Draper, 1998). However, during the last decade, the development of efficient vectors and allelic exchange systems has rendered mycobacteria generally more amenable to molecular genetic analysis (Bardarov et al., 1997; Camacho et al., 1999; Pelicic et al., 1997). The availability of the whole genome sequence for *M. tuberculosis* (Cole et al., 1998) allied with the approaches of comparative and functional genomics has the potential to answer fundamental questions concerning the biology of these major human pathogens. Comparative mycobacterial genomics is a field that is developing particularly rapidly, to the extent that it is more advanced than that of most other bacterial genera. In addition to the complete genome sequence of the virulent *M. tuberculosis* reference strain H37Rv there are genome projects for five other slow growing mycobacteria at various stages of completion (Table 1).

The comparison of these genomes with each other and with those of other sequenced organisms will provide an efficient means of identifying the genetic basis of the differences in host range, phenotype and pathogenicity of these inherently difficult to manipulate mycobacteria. This review summarises the techniques currently in use and some of the main advances made to date within the field of comparative mycobacterial genomics.

Genome mapping

Since the development of pulsed-field gel electrophoresis (PFGE) by Schwartz and Cantor (1984) this technique has evolved into a fundamental research tool for a broad spectrum of epidemiological, genomic and post-genomic applications. The main advantage of PFGE compared to standard gel electrophoresis is that it can accurately separate DNA fragments as large as 10 Mb, such as large restriction fragments obtained by cleavage of chromosomal DNA with rare cutting restriction endonucleases. This ability to separate large

molecules of DNA has resulted in physical and genetic maps of a large variety of microorganisms (Cole and Saint Girons, 1994, 1999), including integrated maps for *M. tuberculosis* and *M. bovis* BCG (Philipp et al., 1996a, b).

The basic requirement for the establishment of a physical map is the availability of low-frequency cleavage enzymes that generate a manageable number of fragments. In the case of *M. tuberculosis* the overall GC content is 65.6% and therefore restriction enzymes with AT-rich restriction sites such as DraI (TTTAAA) or *Asn*I (ATTAAT) have been used as rare cutting enzymes, which generated 35 and 47 fragments in *M. tuberculosis* H37Rv, respectively. Sixteen of the identified DraI restriction sites were associated with IS6110 as this insertion element contains a unique DraI site. To establish the order of the restriction fragments, linking clones spanning rare-cutter sites were identified in the libraries and used as probes in hybridisation experiments thus permitting unambiguous linking of most of the fragments. Final confirmation of the established genome map was obtained by two-dimensional PFGE of genomic DNA reciprocally digested with DraI and *Asn*I (Philipp et al., 1996b).

To obtain an integrated genomic map, sets of endsequenced cosmids and/or bacterial artificial chromosomes (BACs) are overlaid on the physical map. Cosmid clones mapped on the integrated map (Philipp et al., 1996b) have played an important role in the *M. tuberculosis* H37Rv genome sequencing project. However, problems such as underrepresentation of certain regions of the chromosome, unstable inserts and the relatively small insert size complicated the production of a comprehensive set of canonical cosmids representing the entire genome.

To overcome these problems, a BAC library containing large (80–100 kb) inserts from *M. tuberculosis* H37Rv-derived DNA was constructed. The BAC cloning system is based on the *E. coli* F-factor, whose replication is strictly controlled and thus ensures stable maintenance of large constructs (Willems and Skurray, 1987; Shizuya et al., 1992). A major advantage of the BAC cloning system (Kim et al., 1996) over cosmid vectors is that the F-plasmid is present in only one or a maximum of two copies per cell, reducing the potential for recombination between DNA fragments and, more importantly, avoiding the lethal overexpression of cloned bacterial genes. For *M. tuberculosis* a canonical set of 68 BAC clones that covers essentially the complete genome (Brosch et al., 1998) provided templates for gap-filling as these clones carried regions of the genome which were underrepresented or missing from cosmid or plasmid libraries. The ability to isolate problem areas, such as those rich in repetitive DNA, in large insert clones has simplified DNA sequence assem-

Table 1. Mycobacterial genome sequencing projects.

1	<i>M. tuberculosis</i> H37Rv	http://www.sanger.ac.uk/Projects/M_tuberculosis/ (Cole et al., 1998)
2	<i>M. bovis</i>	http://www.sanger.ac.uk/Projects/M_bovis/
3	<i>M. bovis</i> BCG	http://www.pasteur.fr/recherche/unites/Lgmb
4	<i>M. lepra</i>	http://www.sanger.ac.uk/Projects/M_leprae/
5	<i>M. tuberculosis</i> CDC1551 (CSU93)	http://www.tigr.org/tdb/mdb/mdb.html#progress
6	<i>M. avium</i>	http://www.tigr.org/tdb/mdb/mdb.html#progress

bly processes considerably. The minimal set of BACs constitutes an immortalised supply of DNA of well-defined genomic sections, and therefore represents a valuable resource for postgenomic studies. Integrated genome maps have also been constructed for *M. bovis* AF2122/97 and *M. bovis* BCG Pasteur (Brosch et al., 1998, Gordon et al., 1999, Philipp et al., 1996a).

Due to the impossibility to isolate DNA from *M. leprae* that was suitable for analysis by PFGE, no physical map could be constructed for this noncultivable organism. Instead, the ordered cosmid library, which was generated from armadillo-derived bacteria, originally isolated from a patient from Tamil Nadu (Eiglmeier et al., 1993), has been used, in conjunction with clones from a whole genome shotgun, to obtain the chromosome sequence.

The *Mycobacterium tuberculosis* complex

The *M. tuberculosis* complex contains several subspecies that cause tuberculosis in mammals. They all are slow growing mycobacteria and are characterised by 99.9% similarity and a singular lack of genetic diversity at the nucleotide level (Sreevatsan et al., 1997). In spite of the conservation of their genomes there are a number of different phenotypic traits, which have been used to distinguish them (Table 2). With the exception of pyrazinamide resistance, the genetic basis of these phenotypes has yet to be elucidated. These subspecies, despite their genomic similarity appear to have adapted to different hosts and possess widely differing degrees of virulence for man. The complex comprises *Mycobacterium tuberculosis*, *Mycobacterium africanum*, *Mycobacterium microti*, *Mycobacterium bovis*, and *Mycobacterium bovis* BCG.

M. tuberculosis

M. tuberculosis is the most common human mycobacterial pathogen and is responsible for the vast majority

of human tuberculosis cases. It only exceptionally causes disease in other mammals.

M. africanum

M. africanum is very closely related to *M. tuberculosis* and has been isolated from patients in Africa. There appear to be at least two variants of *M. africanum* which are of West- and East African origin (Haas et al., 1997; Frothingham et al., 1999). Phenotypically they appear to be intermediate between *M. bovis* and *M. tuberculosis*.

M. bovis

M. bovis has, in contrast to *M. tuberculosis*, a very large host spectrum and is responsible for tuberculous disease in cattle. *M. bovis* can also infect a variety of other domestic and wild-living animals like deer, lions, seals, etc. *M. bovis* is also virulent for laboratory animals. Rabbits, more resistant to *M. tuberculosis* are sensitive to *M. bovis* infections, a fact that can be used to distinguish between these two species. *M. bovis* can also infect humans, but in countries where pasteurisation of milk and control of the tuberculin sensitivity of the live stock are routine practice, human *M. bovis* infections have become very rare.

M. bovis BCG

In an attempt to avoid clumping of *M. bovis* cells, Calmette and Guérin at the Pasteur Institute, Lille, added ox bile to their potato medium, which they were using to grow a bovine isolate of *M. bovis*. After several passages they detected that the culture had become attenuated for guinea pigs and after 230 passages they considered that strain to be attenuated for man (Calmette, 1927; Bloom and Fine, 1994). As such, in 1921 the "Bacille de Calmette et Guerin" (BCG) was obtained and was then used as an attenuated live vaccine against tuberculosis. With approximately 3 billion delivered

Table 2. Phenotypic differentiation criteria of members of the *M. tuberculosis* complex (Grosset et al., 1990).

	<i>M. tuberculosis</i>	<i>M. africanum</i>	<i>M. bovis</i>	<i>M. bovis</i> BCG
Lebeck test	aerobe	micro-aerophilic	micro-aerophilic	aerobe
Niacin	+	+/-	-	-
Reduction of nitrates	+	+/-	-	-
Thiophen-carboxylic acid	R	S	S	S
Pyrazinamide	S	S	R	R
Thiosemicarbazone	S	S/R	S	R
Cycloserine	S	S	S	R
Virulence (guinea-pig)	+	+	+	-
Virulence (rabbit)	-	-	+	-

Abbreviations used: + = positive, - = negative, R = resistant to, S = sensitive to.

doses, BCG has become the most widely used vaccine. Although the efficiency of protection against pulmonary tuberculosis induced by BCG is highly variable in various vaccination trials, there is a general agreement that BCG is highly effective against disseminated forms of tuberculosis in children, such as miliary tuberculosis and meningitis (Fine, 1995). In addition, BCG did protect against leprosy in a recent controlled trial (Pon-nighaus et al., 1992). From 1924 on, the Pasteur Institute sent the BCG vaccine to many countries, in order to enable local vaccine production. Due to the fact that efficient methods of conservation of bacteria, like freeze-drying, only became available in the 1960's, BCG was passaged in various laboratories under different conditions resulting in further genetic changes. As such, several variants of BCG carrying subtle genetic differences exist and it has been suggested that these differences may account for the variable protection conferred by BCG found in different clinical trials using different BCG strains. (Behr et al., 1999; Oettinger et al., 1999). Despite almost 50 years of worldwide use the reason for the attenuation of *M. bovis* BCG is still unknown. The elucidation of the mechanism of attenuation of BCG is therefore a primary objective for new vaccine development.

M. microti

M. microti was originally isolated from voles in the UK in the 1930's and causes tuberculosis in rodents (Wells and Oxon, 1937). Several strains have been isolated from voles and these are avirulent for both man and bovines. One strain OV254 (OV stands for Orkney Vole) has been used as a live vaccine by the Medical Research Council (MRC) in the UK in controlled clinical trials involving large human populations (Hart and Sutherland, 1977). Another strain, OV166 has been used as live vaccine in Czechoslovakia in the 1960's. In total, about half a million new-borns were vaccinated with this strain over an 18-year period. No particular health problems due to the vaccination with *M. microti* were reported and immunisation with *M. microti* conferred protection against tuberculosis equivalent to that induced by BCG (Sula and Radkovsky, 1976). These studies clearly show that *M. microti* is not a human pathogen. These data are in conflict with a recent report, where the authors claim the isolation of several *M. microti* strains from severely immuno-compromised humans with tuberculous infections. However, although the spoligotypes of the vole isolates resemble that of human isolates, the IS6110 profiles of the vole isolates and human isolates are completely different (van Soolingen et al., 1998) and as such may not correspond to the same variant of *M. microti*. In order to avoid confusion, in the present minireview we

restrict the name *M. microti* to the avirulent vole isolates, as indicated by its name.

Comparative genomics of the *M. tuberculosis* complex

By elucidating the genomic differences between the virulent and avirulent members of the *M. tuberculosis* complex, comparative genomics has the potential to define mycobacterial virulence determinants specific for humans and other host-range determining factors. Different approaches have been used to perform genome-wide comparisons between members of the *M. tuberculosis* complex.

BAC-arrays

The minimal overlapping set of BAC clones represents a powerful tool for comparative genomics. For example, with each BAC clone containing on average an insert of 70 kb, it should be possible to cover a 1-Mb section of the chromosome with 15 BAC clones. Restriction digests of overlapping clones can then be blotted onto membranes, and probed with radiolabelled total genomic DNA from related strains. By means of BAC-arrays and direct comparison of canonical BACs from ordered libraries several polymorphic genomic regions were uncovered (Brosch et al., 1998; Gordon et al., 1999). To survey the genetic diversity of the tubercle bacilli, oligonucleotides specific for each variable region allowed their presence or absence to be determined in the genomes of *M. tuberculosis* clinical isolates, *M. africanum*, *M. bovis*, *M. bovis* BCG, and *M. microti*.

DNA microarrays

In parallel, Behr and colleagues employed DNA microarrays containing capture probes for 3902 of the coding sequences present in *M. tuberculosis* H37Rv to assess the relatedness of a set of 13 different BCG sub-strains (Behr et al., 1999).

These two studies identified a similar set of genomic deletions in addition to those described in earlier works using different techniques (Mahairas et al., 1996; Philipp et al., 1996a). The regions absent from *M. bovis* BCG relative to the *M. tuberculosis* H37Rv genome were named RD1 to RD16. This review has used the nomenclature proposed by Mahairas and colleagues (1996) and Gordon and colleagues (1999) rather than Behr and colleagues in accordance with publication order. RD1 to RD16 are summarised in Figure 1 and Table 3.

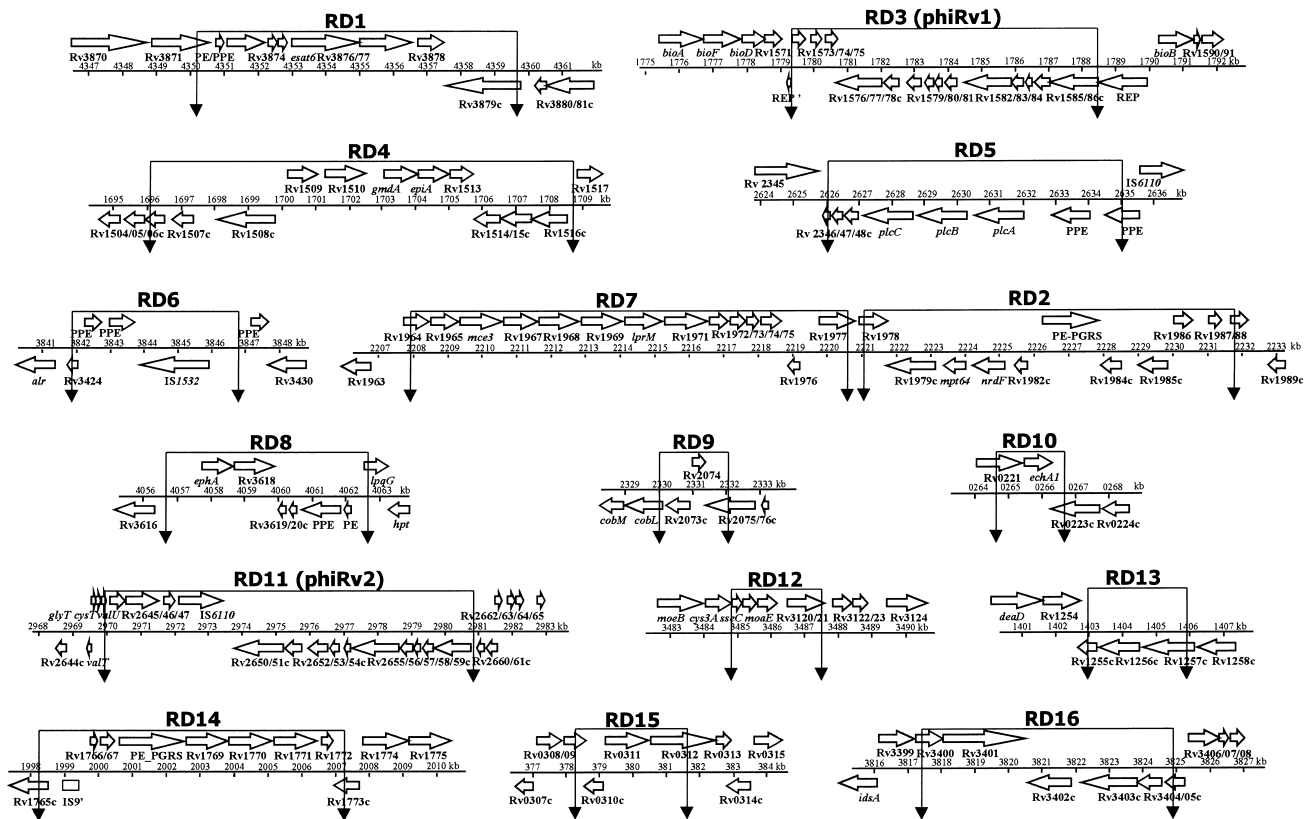


Fig. 1. Overview of the RD1–RD16 deleted regions in *M. bovis* BCG relative to *M. tuberculosis* H37Rv. The regions deleted from the BCG genome are delimited by arrows. ORFs are represented as pointed boxes showing the direction of transcription. For additional information about gene functions consult the website <http://bioweb.pasteur.fr/GenoList/TuberculList>. The RD nomenclature used in this figure follows publication order and is based on that proposed initially by Mahairas et al. (1996) and Gordon et al. (1999) and differs from that used by Behr and coworkers (1999).

According to these studies the RD regions can be clustered into regions that are absent from 1) all BCG substrains, 2) some BCG substrains, 3) *M. bovis* and *M. bovis* BCG strains, 4) *M. bovis*, *M. bovis* BCG, and *M. microti* strains and 5) *M. bovis*, *M. bovis* BCG, *M. microti*, and *M. africanum* strains.

RD1

This deletion, originally described by the group of Ken Stover (Mahairas et al., 1996), is specific for all BCG substrains (Behr et al., 1999). The deleted region contains eight genes, most of which belong to the *esat6* gene cluster (Tekai et al., 1999). ESAT6 has been shown to act as potent stimulator of the immune system and is an antigen recognised during the early stages of infection (Elhay et al., 1998; Horwitz et al., 1995; Rosenkrands et al., 1998). As this 10-kb region is absent from all BCG strains tested so far, but present in virulent *M. bovis* and *M. tuberculosis* strains, the loss of RD1 could be associated with the attenuation of BCG. However, preliminary attempts to comple-

ment BCG with the RD1 region did not restore virulence when the transformants were tested in an animal model (Mahairas et al., 1996).

RD2

The RD2 region, first described by Mahairas and colleagues (Mahairas et al., 1996), is absent from some but not all BCG strains. From the 13 BCG substrains tested by Behr and colleagues, eight have this 10 kb region deleted (Behr et al., 1999). Two of the genes in the RD2 region code for the secreted immunogenic protein Mpb64 and a regulatory protein of the LysR family. Some authors speculate that the loss of RD2 in some BCG substrains may have been responsible for a decrease of protective immunity induced by these BCG substrains (Behr and Small, 1999).

RD3

The RD3 region corresponds to one (phiRv1) of the two prophages (phiRv1 and phiRv2), present in the ge-

Table 3. Distribution of deletions among members of the *M. tuberculosis* complex.

Deletion	<i>M. tuberculosis</i> H37Rv	<i>M. africanum</i>	<i>M. bovis</i>	<i>M. bovis</i> BCG	<i>M. microti</i> OV254
RD1	+	+	+	-	+
RD2	+	+	+	+/-	+
RD3(phiRv1)	+	-	+	-	-
RD4	+	+	-	-	+
RD5	+	+	-	-	-
RD6	+	+	-	-	-
RD7	+	+	-	-	-
RD8	+	+	-	-	-
RD9	+	-	-	-	-
RD10	+	+	-	-	-
RD11(phiRv2)	+	+	+/-	-	+
RD12	+	+	-	-	+
RD13	+	+	-	-	+
RD14	+	+	+	BCG Pasteur -	+
RD15	+	+	+	BCG Frappier BCG Connaught -	+
RD16	+	+	+	BCG Moreau -	+

+ = region is present, - = region is deleted.

The RD nomenclature used in this figure follows previous publication order and is based on that proposed initially by Mahairas et al. (1996) and Gordon et al. (1999) and differs from that used by Behr and coworkers (1999), shown in brackets. RD1 = (RD01); RD2 = (RD02); RD3 = (RD03); RD4 = (RD06); RD5 = (RD07); RD6 = (RD11); RD7 = (RD15); RD8 = (RD09); RD9 = (RD12); RD10 = (RD04); RD11 = (RD13); RD12 = (RD05); RD13 = (RD10); RD14 = (RD14); RD15 = (RD08); RD16 = (RD16).

nome of *M. tuberculosis*. PhiRv1 and phiRv2 are both ~10 kb in length, have a similar organisation and some of their gene products show marked similarity to those encoded by certain bacteriophages from *Streptomyces* and saprophytic mycobacteria (Cole et al., 1998). It is possible that during the serial attenuation of *M. bovis* that led to the vaccine strain *M. bovis* BCG, the phiRv1 prophage was lost. The presence of this prophage in other members of the tubercle complex is variable (Table 3).

RD4

The RD4 region (Brosch et al., 1998) represents a deletion of 12.7 kb from both *M. bovis* and all *M. bovis* BCG strains tested. Among the 11 open reading frames present in this region in *M. tuberculosis*, several show similarity to membrane proteins and enzymes involved in the biosynthesis of polysaccharides. These putative lipopolysaccharide-like molecules could be involved in interaction of *M. tuberculosis* with certain receptors of the mammalian host cells.

RD5

The RD5 region being absent from *M. bovis*, *M. bovis* BCG and also from *M. microti* (Gordon et al., 1999) encompasses three (*plcA*, *plcB*, *plcC*) of the

four genes encoding a phospholipase C in *M. tuberculosis* H37Rv (Cole et al., 1998). Phospholipase C has been recognised as an important virulence factor in numerous bacteria, including *Clostridium perfringens*, *Listeria monocytogenes* and *Pseudomonas aeruginosa*, where it plays a role in cell-to-cell spread of bacteria, intracellular survival, and cytolysis (Titball, 1993). Phospholipase C activity in *M. tuberculosis*, *M. microti* and *M. bovis*, but not in *M. bovis* BCG, has been reported (Johansen et al., 1996; Wheeler and Ratledge, 1992). The levels of phospholipase C activity detected in *M. bovis* were much lower than those seen in *M. tuberculosis* consistent with the loss of *plcABC*. It is likely, that *plcD* is responsible for the residual phospholipase C activity in strains lacking RD5, such as *M. bovis* and *M. microti*. In *M. tuberculosis* H37Rv, the *plcD* gene has been shown to be partially deleted (Gordon et al., 1999) as a result of being a hotspot for IS6110 integration. As such, the presence of functional phospholipase C-encoding genes is highly variable among the members of the *M. tuberculosis* complex, a fact which might eventually account for some of the differences seen in their virulence.

In addition, RD5 contains genes encoding proteins belonging to the ESAT-6 family (like RD1 and RD8), of which there are 14 copies organised in 11 distinct loci within the genome sequence (Tekaija et al., 1999).

RD6

RD6 is a region, which contains an insertion element (IS1532) and genes encoding proteins of the highly repetitive PPE family. RD6 is absent from *M. bovis*, *M. bovis* BCG, *M. africanum*, *M. microti* but also from various *M. tuberculosis* clinical isolates (Gordon et al., 1999).

RD7

The RD7 region (Gordon et al., 1999) contains the *mce3* operon. The *mce* gene was described by Riley and colleagues and codes for a putative invasin-like protein in *M. tuberculosis*. By cloning the *mce* gene into *Escherichia coli*, they showed that expression of Mce renders *E. coli* invasive for HeLa cells (Arruda et al., 1993). Twenty-three proteins from the Mce family were identified as part of the genome sequencing project, with their genes occupying the same position in four large, well-conserved operons comprising at least eight genes (Cole et al., 1998; Tekaia et al., 1999). It is difficult to speculate about the effects of the loss of *mce3* (RD7) on *M. bovis*, *M. microti* and *M. bovis* BCG as the remaining three *mce* operons would be expected to complement for any lost activity, unless they were expressed differentially. However, it is intriguing that RD7 is absent from some members of the *M. tuberculosis* complex that are not virulent for man, suggesting that RD7 may play some specific role in human disease (Gordon et al., 1999).

RD8

The RD8 region (Gordon et al., 1999) contains one of the six genes within the genome sequence that encode epoxide hydrolases (EphA-F). Epoxide hydrolases are generally regarded as enzymes of detoxification (Arand et al., 1994). The loss of *ephA* (RD8) may be compensated by other enzymes, although the substrate specificity of the *M. tuberculosis* enzymes is unknown. Epoxide hydrolases may be involved in detoxifying lipid peroxidation products generated by oxygen radicals from the activated macrophage. Like RD1 and RD5, RD8 also contains genes encoding proteins belonging to the ESAT-6 family.

RD9

The RD9 region (Gordon et al., 1999) contains genes encoding a precorrin methylase, an oxidoreductase, and an exported protein. RD9 was found to be deleted from *M. africanum*, *M. bovis*, *M. bovis* BCG and *M. microti* but not from *M. tuberculosis*. Hence, in con-

trast to the other RD regions, at this locus *M. africanum* resembles *M. bovis*, reflecting the postulated intermediate status of this strain between *M. tuberculosis* and *M. bovis* (Heifets and Good, 1994).

RD10

The RD10 region (Gordon et al., 1999) is only 1903 bp in size yet affects three ORFs, Rv0221, *echA1* and Rv0223, that encode a hypothetical protein, enoyl CoA hydratase and an aldehyde dehydrogenase, respectively. PCR reactions revealed that RD10 was absent from *M. microti* as well as *M. bovis* and *M. bovis* BCG.

RD11

The RD11 region corresponds to the second prophage (phiRv2) that has a similar organisation to that of phiRv1 (Cole et al., 1998). The attachment site for phiRv2 is found in the 3' end of *valU*, the tRNA-Val gene, which is part of a tRNA gene cluster, an insertion site similar to those of pathogenicity islands (Hacker et al., 1997). Interestingly, most of the *M. tuberculosis* strains harbor phiRv1 and phiRv2, whereas this prophage seems to be very rare in *M. bovis* (our unpublished findings).

RD12 and RD13

These two regions, each about 2.5 kb in size, contain genes coding for a thiosulfate sulfurtransferase, a molybdopterin converting factor, a methyltransferase, a cytochrome P450 (RD12), a transcriptional regulator, a cytochrome P450, and a dehydrogenase (RD13). Both regions were found to be absent from *M. bovis* and *M. bovis* BCG (Behr et al., 1999), but present in *M. microti* (our unpublished findings). Hence, in contrast to other RD regions, at this locus *M. microti* resembles *M. tuberculosis*, suggesting that this member of the *M. tuberculosis* complex should be intermediate between *M. tuberculosis* and *M. bovis*. Similar to the RD4 locus, RD12 and RD13 differentiate *M. microti* from bovine strains.

RD14, RD15 and RD16

RD14, RD15 and RD16 (Behr et al., 1999) are BCG substrain-specific deletions (Table 3). RD14 was only found in BCG Pasteur 1173P2. RD15 represents a specific deletion of BCG substrains Frappier and Connaught, with RD16 being a specific deletion of BCG Moreau (Behr et al., 1999). Similarly to RD2, these regions could have altered the immunogenicity of these BCG substrains.

Evolution of the *M. tuberculosis* complex

Hypotheses on the evolution of the *M. tuberculosis* complex are often based on the idea that *M. tuberculosis* is derived from *M. bovis*. However, the fact that most of the end-points of the 16 variable regions are located inside genes, suggesting that these regions represent deletions in *M. bovis* or BCG rather than insertions in *M. tuberculosis*, implies that the human tubercle bacillus cannot be descended from the bovine form. Thus, tuberculosis in humans did not result from a zoonosis.

PFGE comparisons

In addition to array-based techniques, comparison of PFGE profiles between related strains can reveal genomic differences. *M. tuberculosis* H37Rv and *M. tuberculosis* H37Ra are a virulent and an avirulent sub-strain of strain H37 (Steenken et al., 1934). Comparison of the DraI macrorestriction profiles of H37Rv and H37Ra revealed that a 480kb DraI fragment present in the genome of H37Rv is replaced by fragments of 260 kb and 220 kb in H37Ra. In addition, an 8 kb fragment present in H37Ra is not present in H37Rv. Comparison of an in silico restriction map of the sequenced strain *M. tuberculosis* H37Rv with the PFGE DraI pattern from *M. tuberculosis* H37Ra allowed the precise localisation of these two polymorphic regions on the genome. Further analyses showed, that these polymorphisms were the result of an IS6110 insertion in strain H37Ra and an IS6110-mediated deletion (RvD2) in strain *M. tuberculosis* H37Rv (Brosch et al., 1999).

PFGE has been described as a powerful tool for molecular epidemiologic and population genetic studies of the *M. tuberculosis* complex (Singh et al., 1999; Zhang et al., 1995). The imminent availability of several mycobacterial genome sequences opens a new perspective for the use of PFGE macrorestriction fragments for comparative genomics. With the whole genome sequence of an organism in hand, one can readily construct an in silico restriction map and compare this map with the PFGE banding pattern from the same strain. By comparison of the PFGE pattern of the sequenced strain with patterns from related strains (e.g. clinical isolates of the same species, or strains from closely related species), variable fragments observed in the patterns can be linked back to the sequence and ultimately to individual genes. These comparisons can identify genomic regions that are conserved in a large number of isolates, as well as identifying hypervariable regions. This simple approach could considerably enrich our knowledge about mycobacterial epidemiology and evolution of the mycobacteria

Comparative genomics – *Mycobacterium leprae*

The impossibility to isolate DNA from *M. leprae* that is amenable for PFGE analysis, restricts comparative genomics to in silico comparison of the *M. leprae* sequence with genome sequences of other organisms. At the time of writing, a contiguous sequence of the single circular chromosome of the leprosy bacillus has just become available. The *M. leprae* genome (size 3268 kb), is almost 1.2 Mb smaller than that of *M. tuberculosis* (4411 kb) and its GC content of ~58% also differs extensively. This low GC value previously led to doubts as to whether the leprosy bacillus was a true member of the Mycobacterium genus. However, with the genome sequence available, the low GC content as well as the smaller size can be attributed partly to the low amount of repetitive DNA in *M. leprae*. At first inspection, *M. leprae* has very few IS elements, although its genome does contain at least three dispersed repetitive sequences termed RLEP, REPLEP and LEPREP, and also lacks the many large genes encoding the PGRS and MPTR proteins present in *M. tuberculosis* (Cole et al., 1998). Preliminary comparison of the two genome sequences reveals a mosaic arrangement in which extended homologous areas are flanked by unrelated genomic segments. In contrast to most other bacteria which carry a copy of *rrn* adjacent to the chromosomal origin of replication *oriC*, resulting in increased rRNA production through increased gene dosage, in both *M. leprae* and *M. tuberculosis*, the single *rrn* operon is situated ~1.3 Mb from *oriC*. It has been suggested that the atypical arrangement seen in these mycobacterial pathogens may be related to their slow growth. The *M. leprae* genome contains a large portion of pseudo-genes, i.e. inactivated versions of genes that are still functional in *M. tuberculosis*. It seems probable that these genes were not essential for the intracellular life of *M. leprae*. This large-scale gene inactivation appears to be the result of the reductive evolution *M. leprae* has undergone during its evolution into an obligate intracellular organism, and is most probably responsible for the inability of *M. leprae* to grow on culture media. Complete analysis of the genome sequence should elucidate the genetic background for the exceptionally slow growth of *M. leprae* and further comparisons with *M. tuberculosis* will define the minimal gene set required by a pathogenic mycobacterium.

Concluding remarks

Genomics has provided a wealth of new information about the tubercle and the leprosy bacilli that enormously enriches our knowledge about these important

organisms. Comparative genomics uncovered a large variety of polymorphic regions in the otherwise highly conserved genomes of the tubercle bacilli. The completion of five additional mycobacterial genome sequences will enable in silico comparisons of whole genomes. This approach will certainly reveal further, probable more subtle genomic differences, that have remained undetected by the current approaches that are described in this review. Functional analysis of all the polymorphic regions will certainly help to elucidate what role these genomic differences play in the phenotypic differentiation, the host range and the virulence of the slow growing mycobacteria. The next scientific challenge will then be to use this information for the development of affordable new drugs and vaccines to combat disease.

Acknowledgments. Special thanks to Bart Barrell, Julian Parkhill and the Pathogen Group at the Sanger Centre. Financial support from the Wellcome Trust, Association Française Raoul Follereau and the Institut Pasteur is gratefully acknowledged.

References

- Arand, M., Grant, D. F., Beetham, J. K., Friedberg, T., Oesch, F., Hammock, B. D.: Sequence similarity of mammalian epoxide hydrolases to the bacterial haloalkane dehalogenase and other related proteins. *FEBS Lett.* 338, 251–256 (1994).
- Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T., Riley, L. W.: Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science* 261, 1454–1457 (1993).
- Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R., Bloom, B. R., Hatfull, G. R., Jacobs, J. W. R.: Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* 94, 10961–10966 (1997).
- Behr, M. A., Small, P. M.: A historical and molecular phylogeny of BCG strains. *Vaccine* 17, 915–922 (1999).
- Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., Small, P. M.: Comparative genomics of BCG vaccines by whole-genome DNA microarrays. *Science* 284, 1520–1523 (1999).
- Bloom, B. R., Fine, P. E. M.: The BCG experience: Implications for future vaccines against tuberculosis. In: *Tuberculosis: Pathogenesis, Protection and Control.* (Bloom, B. R. ed.), pp. 531–557. American Society for Microbiology, Washington DC 1994.
- Brosch, R., Gordon, S. V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B. G., Cole, S. T.: Use of a *Mycobacterium tuberculosis* H37Rv Bacterial Artificial Chromosome (BAC) library for genome mapping, sequencing and comparative genomics. *Infect. Immun.* 66, 2221–2229 (1998).
- Brosch, R., Philipp, W., Stavropoulos, E., Colston, M. J., Cole, S. T., Gordon, S. V.: Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. *Infect. Immun.* 67, 5768–5774 (1999).
- Calmette, A.: *La vaccination contre la tuberculose.* 250 p. Masson et Cie, Paris 1927.
- Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B., Guilhot, C.: Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* 34, 257–267 (1999).
- Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, A., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., Barrell, B. G.: Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544 (1998).
- Cole, S. T., Saint-Girons, I.: Bacterial genomics. *FEMS Microbiol. Rev.* 14, 139–160 (1994).
- Cole, S. T., Saint-Girons, I.: Bacterial genomes – all shapes and sizes. In: *Organization of the prokaryotic genome.* Charlebois, R. L. (ed.), pp. 35–62. American Society for Microbiology, Washington DC 1999.
- Daffe, M., Draper, P.: The envelope layers of mycobacteria with reference to their pathogenicity. *Adv. Microb. Physiol.* 39, 131–203 (1998).
- Eiglmeier, K., Honoré, N., Woods, S. A., Caudron, B., Cole, S. T.: Use of an ordered cosmid library to deduce the genomic organisation of *Mycobacterium leprae*. *Mol. Microbiol.* 7, 197–206 (1993).
- Elhay, M. J., Oettinger, T., Andersen, P.: Delayed-type hypersensitivity responses to ESAT-6 and MPT64 from *Mycobacterium tuberculosis* in the guinea pig. *Infect. Immun.* 66, 3454–3456 (1998).
- Fine P. E. M.: Variation in protection by BCG: implications of and for heterologous immunity, *Lancet* 346, 1339–1345 (1995).
- Frothingham, R., Strickland, P. L., Bretzel, G., Ramaswamy, S., Musser, J. M., Williams, D. L.: Phenotypic and genotypic characterization of *Mycobacterium africanum* isolates from West Africa. *J. Clin. Microbiol.* 37, 1921–1926 (1999).
- Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., Cole, S. T.: Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.* 32, 643–656 (1999).
- Grosset, J., Boisvert, H., Truffot-Pernot, C.: *Mycobactéries.* In: *Le Minor, L., Véron, M. (Eds.) Bactériologie Médicale* 2nd ed, Médecine-Sciences, Flammarion, pp. 965–1011, 1990.
- Haas, W. H., Bretzel, G., Amthor, B., Schilke, K., Krommes, G., Rusch-Gerdes, S., Sticht-Groh, V., Bremer, H. J.: Comparison of DNA fingerprint patterns of isolates of *Mycobacterium africanum* from east and west Africa. *J. Clin. Microbiol.* 35, 663–666 (1997).

- Hacker, J., Blum, O. G., Muhldorfer, I., Tschape, H.: Pathogenicity islands of virulent bacteria: structure, function and impact. *Mol. Microbiol.* 23, 1089–1097 (1997).
- Hart, P. D., Sutherland, I.: BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Br. Med. J.* 2(6082), 293–295 (1977).
- Heifets, L. B., Good, R. C.: Current laboratory methods for the diagnosis of tuberculosis. In: *Tuberculosis: Pathogenesis, Protection and Control*. Bloom, B. R. (ed.), pp. 85–110. American Society for Microbiology, Washington, DC 1994.
- Horwitz, M. A., Lee, B. W., Dillon, B. J., Harth, G.: Protective immunity against tuberculosis induced by vaccination with major extracellular proteins of *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* 92, 1530–1534 (1995).
- Johansen, K. A., Gill, R. E., Vasin, M. L.: Biochemical and molecular analysis of phospholipase C and phospholipase D activity in mycobacteria. *Infect. Immun.* 64, 3259–3266 (1996).
- Kim, U. J., Birren, B. W., Slepak, T., Mancino, V., Boysen, C., Kang, H. L., Simon, M. I., Shizuya, H.: Construction and characterization of a human bacterial artificial chromosome library. *Genomics* 34, 213–218 (1996).
- Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C., Stover, C. K.: Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J. Bacteriol.* 178, 1274–1282 (1996).
- Oettinger, T., Jorgensen, M., Ladefoged, A., Haslov, K., Andersen, P.: Development of the *Mycobacterium bovis* BCG vaccine: review of the historical and biochemical evidence for a genealogical tree. *Tubercle Lung Disease* 79, 243–250 (1999).
- Pelacic, V., Jackson, M., Reyrat, J. M., Jacobs, W. R., Jr., Gicquel, B., Guilhot, C.: Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* 94, 10955–10960 (1997).
- Philipp, W. J., Nair, S., Guglielmi, G., Lagranderie, M., Gicquel, B., Cole, S. T.: 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 (1996a).
- Philipp, W. J., Poulet, S., Eiglmeier, K., Pascopella, L., Subramanian, B., Heym, B., Bergh, S., Bloom, B. R., Jacobs, W. R., Jr., Cole, S. T.: 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 (1996b).
- Ponnighaus, J. M., Fine, P. E., Sterne, J. A., Wilson, R. J., Msosa, E., Gruer, P. J., Jenkins, P. A., Lucas, S. B., Liomba, N. G., Bliss, L.: Efficacy of BCG vaccine against leprosy and tuberculosis in northern Malawi. *Lancet* 339, 636–639 (1992).
- Rosenkrands, I., Rasmussen, P. B., Carnio, M., Jacobsen, S., Theisen, M., Andersen, P.: Identification and characterization of a 29-kilodalton protein from *Mycobacterium tuberculosis* culture filtrate recognized by mouse memory effector cells. *Infect. Immun.* 66, 2728–2735 (1998).
- Schwartz, D. C., Cantor, C. R.: Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* 37, 67–75 (1984).
- Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M.: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* 89, 8794–8797 (1992).
- Singh, S. P., Salamon, H., Lahti, C. J., Farid-Moyer, M., Small, P. M.: Use of pulsed-field gel electrophoresis for molecular epidemiologic and population genetic studies of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 37, 1927–1931 (1999).
- Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S., Musser, J. M.: Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc. Natl. Acad. Sci. USA* 94, 9869–9874 (1997).
- Steenken, W., Oatway, W. H., Petroff, S. A.: Biological studies of the tubercle bacillus. III. Dissociation and pathogenicity of the R and S variants of the human tubercle bacillus (H37). *J. Exp. Med.* 60, 515–540 (1934).
- Sula, L., Radkovsky, I.: Protective effects of *M. microti* vaccine against tuberculosis. *J. Hyg. Epidemiol. Microbiol. Immunol.* 20, 1–6 (1976).
- Tekaia, F., Gordon, S. V., Garnier, T., Brosch, R., Barrell, B. G., Cole, S. T.: Analysis of the proteome of *Mycobacterium tuberculosis* in vitro. *Tubercle Lung Disease* 79, 329–342 (1999).
- Titball, R. W.: Bacterial phospholipases C. *Microbiol. Rev.* 57, 347–366 (1993).
- Van Soolingen, D., van der Zanden, A. G., de Haas, P. E., Noordhoek, G. T., Kiers, A., Foudraire, N. A., Portaels, F., Kolk, A. H., Kremer, K., van Embden, J. D.: Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. *J. Clin. Microbiol.* 36, 1840–1845 (1998).
- Wells, A. Q., Oxon D. M.: Tuberculosis in wild voles. *Lancet* 1221 (1937).
- Wheeler, P. R., Ratledge, C.: Control and location of acyl-hydrolysing phospholipase activity in pathogenic mycobacteria. *J. Gen. Microbiol.* 138, 825–830 (1992).
- Willems, N., Skurray, R.: Structure and function of the F-factor and mechanism of conjugation. In: *Escherichia coli* and *Salmonella Typhimurium*: Cellular and Molecular Biology Neidhardt, F. C. (ed.), Vol. 2, pp. 1110–1133. Am. Soc. Microbiol., Washington, DC 1987.
- Zhang, Y., Wallace, R. J., Jr., Mazurek, G. H.: Genetic differences between BCG substrains. *Tubercle Lung Disease* 76, 43–50 (1995).