

Palomino • Leão • Ritacco

TUBERCULOSIS 2007

From Basic Science
to Patient Care



www.TuberculosisTextbook.com

Chapter 4: Genomics and Proteomics

Patricia Del Portillo, Alejandro Reyes, Leiria Salazar, María del Carmen Menéndez and María Jesús García

4.1. Impact of new technologies on *Mycobacterium tuberculosis* genomics

A new wave in the analysis of the physiological secrets of microorganisms started more than a decade ago with the reading of the first complete genome sequence, corresponding to the bacterium *Haemophilus influenzae* (Fleishman 1995). Nowadays, the accessibility to hundreds of bacterial genome sequences has changed our way of studying the bacterial world, including bacterial pathogens such as *M. tuberculosis*.

The overwhelming information displayed by genome sequences started the era of “omics” technologies. These technologies are in accordance to the currently fast times. A quick search in PubMed, limiting results to the last 10 years, showed more than 27,000 papers devoted to “omics” issues: more than three thousand concerning bacteria, and almost three hundred concerning *Mycobacterium tuberculosis*. Up to five different “omics” methodologies have been described so far, all concerning the global study of the target organism, analyzing all its genes, transcriptional products, proteins, etc.

- **Genomics** involves the study of all genes that are present in the genomes
- **Transcriptomics** concerns the analysis of the cellular functions at the messenger ribonucleic acid (mRNA) level
- **Proteomics** refers to the detection and identification of all proteins in a cell
- **Metabolomics** comprises the complete set of all metabolites formed by the cell and its association with its metabolism
- **Fluxomics** compares the cellular networks (Fiehn 2003, Nielsen 2005)

In the tuberculosis (TB) field, only papers concerning genomics, transcriptomics, and proteomics have been published. Integration of data derived from the several “omics” by bioinformatics will probably allow a rational insight into *M. tuberculosis* biology and its interactions with the host, leading to true control of the disease.

Undoubtedly, the biggest step in our knowledge on TB during the last decade was the description of the complete genome sequence of the laboratory reference *M.*

tuberculosis strain H37Rv (Cole 1998a). For example, the identification of genes involved in the bacterial cell wall biosynthesis, the routes for lipid metabolism, the location of insertion sequences and the variability in the PE_PPE genes allowed scientists to merge the fragments of knowledge derived from the pre-genomic era in a more comprehensive way. The sequence of the genome, and its comparison to sequences of other microorganisms reported in several databases, allowed the assignment of precise functions to 40 % of the predicted proteins and the identification of 44 % of orthologues (genes with very similar functions in a different species), leaving 16 % as unique unknown proteins.

The elucidation of complete genome sequences and the development of microarray-based comparative genomics have been powerful tools in the progress of new areas by the application of robotics to basic molecular biology. Comparative genomics and genomic tools have also been used to identify factors associated with the pathogenicity of *M. tuberculosis*, such as virulence factors and genes involved in persistence of the pathogen in host cells. Moreover, these tools allowed a description of the evolutionary scenario of the genus (see Chapter 2).

Structural genomics was the starting point. As more accurate technologies became available, the interest was focused into functional genomics. Thus, information on specific mRNA actively synthesized by bacteria inside macrophages or during *in vitro* starvation, opened ways to the analysis of gene expression. Microarray technology was applied to the detection of global gene activity in *M. tuberculosis* under several environmental conditions. However, bacterial function cannot be understood by looking at the mRNA level alone. A major barrier for genomic studies has been the great number of genes with unknown function that have been identified. Up to 60 % of the open reading frames (ORFs) had unknown functions after the initial annotation of genes (identification of the protein unrevealed by the corresponding ORF's amino acid sequence) (Cole 1998a). The elucidation of protein function was possible with the global analysis of bacterial proteins, giving insights into the functional role of several so far unknown proteins. Thanks to the joint contributions of biochemical techniques and mass spectrometry, up to 1,044 non-redundant proteins were reported in different cellular fractions (Mawuenyega 2005). The upcoming task will be to assign them all a functional role. As more results are obtained from the proteomic analysis, it is expected that the function of more ORFs will be unveiled with the aid of new data on transcriptomics and proteomics.

Genomics and other molecular tools allowed studies on gene expression and regulation, which were unthinkable years ago. *M. tuberculosis* is a restricted human pathogen; therefore it must have developed mechanisms enabling its quick and

efficient adaptation to a variety of “intra-human” environments, which are, in fact, its natural habitat. Understanding how the bacillus regulates its different genes according to environmental changes will probably lead to the comprehension of many interesting aspects of *M. tuberculosis*, including latency and host-adaptation.

This chapter will address the general basics, as well as the state-of-the-art genomics, transcriptomics and proteomics in relation to *M. tuberculosis*. Finally, a general overview will be made on lipids, the most peculiar metabolites of this bacterium.

4.2. *M. tuberculosis* genome

4.2.1. Genomic organization and genes

TB research made huge progress with the availability of the genome sequence of the type strain *M. tuberculosis* H37Rv (Cole 1998a). Expectations were generated on the elucidation of some unique characteristics of the biology of the tubercle bacillus, such as its characteristic slow growth, the nature of its complex cell wall, certain genes related to its virulence and persistence, and the apparent stability of its genome. This first available genome sequence of a pathogenic *M. tuberculosis* strain helped to answer some of these questions and, what is even more stimulating, to open many more. We describe herein the main characteristics of the *M. tuberculosis* genome sequences completed thus far and highlight some of the most interesting questions answered and opened with this advance in TB research.

M. tuberculosis H37Rv (Cole 1998a) was revealed to possess a sequence of 4,411,529 bp, the second largest microbial genome sequenced at that time. The characteristically high guanine plus cytosine (G+C content; 65.5 %) was found to be uniform along most of the genome, confirming the hypothesis that horizontal gene transfer events are virtually absent in modern *M. tuberculosis* (Sreevatsan 1997). Only a few regions showed a skew in this G+C content. A conspicuous group of genes with a very high G+C content (> 80 %) appear to be unique in mycobacteria and belong to the family of PE or PPE proteins. In turn, the few genes with particularly low (< 50 %) G+C content are those coding for transmembrane proteins or polyketide synthases. This deviation to low G+C content is believed to be a consequence of the required hydrophobic amino acids, essential in any transmembrane domain, that are coded by low G+C content codons.

Fifty genes were found to code for functional RNAs. As previously described (Kempell 1992), there was only one ribosomal RNA operon (*rrn*). This operon was found to be located at 1.5 Mbp from the origin of replication (*oriC* locus).

Most eubacteria have more than one *rrn* operon located much closer to the *oriC* locus to exploit the gene-dosage effect during replication (Cole 1994). The possession of a single *rrn* operon in a position relatively distant from *oriC* has been postulated to be a factor contributing to the slow growth phenotype of the tubercle bacillus (Brosch 2000a).

One of the most thoroughly studied characteristic of *M. tuberculosis* is the presence and distribution of insertion sequences (IS). Of particular interest is *IS6110*, a sequence of the IS3 family that has been widely used for strain typing and molecular epidemiology due to its variation in insertion site and copy number (van Embden 1993, see Chapter 9). Sixteen copies of *IS6110* were identified in the genome of *M. tuberculosis* H37Rv; some *IS6110* insertion sites were clustered in sites named insertional hot-spots. The same strain was found to harbor six copies of the more stable *IS1081*, an insertion sequence that yields almost identical profiles in most strains when analyzed by Restriction Fragment Length Polymorphism (RFLP) (Sola 2001, Kanduma 2003). Another 32 different insertion sequences were found, of which seven belonged to the 13E12 family of repetitive sequences; the other insertion sequences had not been described in other organisms (Cole 1998b). Virtually all the ISs found in *M. tuberculosis* so far belong to previously described IS families (Chandler 2002). The only exception is *IS1556*, which does not fit into any known IS family (Cole 1999).

Two prophages were detected in the genome sequence; both are similar in length and also similarly organized. One is the prophage PhiRv1, which in the *M. tuberculosis* H37Rv genome interrupts a repetitive sequence of the family 13E12. This prophage is deleted or rearranged in other *M. tuberculosis* strains (Fleischmann 2002). The genome of *M. tuberculosis* possesses seven potential *att* sites for PhiRv1 insertion, which explains the variability of its position between strains (Cole 1999). The second prophage, PhiRv2 has proven to be much more stable, with less variability among strains (Cole 1999).

Regarding protein coding genes, it was determined that *M. tuberculosis* H37Rv codes for 3,924 ORFs accounting for 91 % of the coding capacity of the genome (Cole 1998a). The alternative initiation codon GTG is used in 35 % of cases compared to 14 % or 9 % in *Bacillus subtilis* or *Escherichia coli* respectively. This contributes to the high G+C bias in the codon usage of mycobacteria.

A bias in the overall orientation of genes with respect to the direction of replication was also found. On average, bacteria such as *B. subtilis* have 75 % of their genes in the same orientation as that of the replication fork, while *M. tuberculosis* only has 59 %. This finding has led to the hypothesis that such a bias could also be part of

the slow growing phenotype of the tubercle bacillus (Cole 1999). This conjecture, however, does not take into account the fact that *E. coli*, a bacterium that grows much faster than *M. tuberculosis*, has only 55 % of its genes in the same direction as the replication origin (Li 2005).

From the predicted ORFs, all proteins have been classified in 11 broad functional groups (Table 4-1), more precisely classified into COG functional categories (<http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=135>) according to the National Center for Biotechnology Information (NCBI) of the United States (US). The analysis of the codon usage showed a preference for G+C-rich codons. It was also found that the number of genes that arose by duplication is similar to the number seen in *E. coli* or *B. subtilis*, but the degree of conservation of duplicated genes is higher in *M. tuberculosis*. The lack of divergence of duplicated genes is consistent with the hypothesis of a recent evolutionary descent or a recent bottleneck in mycobacterial evolution (Brosch 2002, Sreevatsan 1997, see chapter 2).

From the genome sequence it is clear that *M. tuberculosis* has the potential to switch from one metabolic route to another including aerobic (e.g. oxidative phosphorylation) and anaerobic respiration (e.g. nitrate reduction). This flexibility is useful for survival in the changing environments within the human host that range from high oxygen tension in the lung alveolus to microaerophilic/anaerobic conditions within the tuberculous granuloma. Another characteristic of the *M. tuberculosis* genome is the presence of genes for synthesis and degradation of almost all kinds of lipids from simple fatty acids to complex molecules such as mycolic acids. In total, there are genes encoding for 250 distinct enzymes involved in fatty acid metabolism, compared to only 50 in the genome of *E. coli* (Cole 1999).

Concerning transcriptional regulation, *M. tuberculosis* codifies for 13 putative sigma factors and more than 100 regulatory proteins (see section 4.3 of this chapter).

Among the most interesting protein gene families found in *M. tuberculosis* are the PE and PPE multigene families, which account for almost 10 % of the genome capacity. The names PE and PPE derive from the motifs of Pro-Glu (PE) and Pro-Pro-Glu (PPE) found near the protein N-terminus in most cases. These proteins are believed to play an important role in survival and multiplication of mycobacteria in different environments (Marri 2006). There are about 100 members of the PE family, which is further divided into three sub-families, the most important of which is the polymorphic GC-rich sequences (PGRS) class that contains 61 members. Proteins in this class contain multiple tandem repetitions of the motif Gly-Gly-Ala, hence, their glycine concentration is superior to 50 %. The PE_PGRS proteins

have been found to be exclusive to the *M. tuberculosis* complex (Marri 2006) and resemble the Epstein-Barr virus nuclear antigens (EBNA), which are known to inhibit antigen presentation through the histocompatibility complex (MHC) class I (Cole 1999).

Interestingly, the analysis of the desoxyribonucleic acid (DNA) metabolic system of *M. tuberculosis* indicates a very efficient DNA repair system, in other words, replication machinery of exceptionally high fidelity. The genome of *M. tuberculosis* lacks the MutS-based mismatch repair system. However, this absence is overcome by the presence of nearly 45 genes related to DNA repair mechanisms (Mizrahi 1998), including three copies of the *mutT* gene. This gene encodes the enzyme in charge of removing oxidized guanines whose incorporation during replication causes base-pair mismatching (Mizrahi 1998, Cole 1999).

With the aim of making the information publicly available and the search and analysis of information easier, the Pasteur Institute (<http://www.pasteur.fr/recherche/unites/Lgmb/>) has created a database system incorporating not only all genes and annotation but other search tools such as Blast or FastA, that allow the user to search for homologue sequences of a query sequence inside the *M. tuberculosis* genome. This database is freely available for use on the Internet and is known as the Tuberculist Web Server (<http://genolist.pasteur.fr/TubercuList/>).

As more information was generated, databases grew bigger, more experimental information became available, and better and more accurate algorithms for gene identification and prediction were released. The initial genome annotation in *M. tuberculosis* H37Rv strain soon became out of date. For this reason, a re-annotation of that genome sequence was published in 2002. This re-annotation incorporated 82 additional genes. The gene nomenclature was not altered; the new genes have the name of the preceding gene followed by A, B or D, for example, two new ORFs were described between *Rv3724* and *Rv3725*, hence, they were named *Rv3724A* and *Rv3724B*. The letter C was not included since it usually stands for “complementary”, which means that the gene is located in the complementary strand. As expected, the classes that exhibited the greatest numbers of changes were the unknown category and the conserved hypothetical category (Table 4-1). The re-annotation of the genome sequence allowed the identification of four sequencing errors making the current sequence size change from 4,411,529 to 4,411,532 bp (Camus 2002).

As shown in Table 4-1, the information obtained from a single sequenced genome is enormous. The advances made on the analysis of such information have accelerated TB research.

Table 4-1: Functional classification of *M. tuberculosis* H37Rv and re-annotation*

Class	Function	Number of genes (1998)	Number of genes (2002)
0	Virulence, detoxification, adaptation	91	99
1	Lipid metabolism	225	233
2	Information pathways	207	229
3	Cell-wall and cell processes	516	708
4	Stable RNAs	50	50
5	Insertion sequences and phages	137	149
6	PE and PPE proteins	167	170
7	Intermediary metabolism and respiration	877	894
8	Proteins of unknown function	606	272
9	Regulatory proteins	188	189
10	Conserved hypothetical proteins	910	1,051

* Data taken from Fleischman 2002

4.2.2. Comparative genomics

In recent times, new technologies have been developed at an overwhelming pace, in particular those related to sequencing and tools for genome sequence data management, storage and analysis. As of April 2007, 484 microbial genomes have been finished and projects are underway aimed at the sequencing of other 1,155 microorganisms (<http://www.genomesonline.org/gold.cgi>). Mycobacteria are not an exception in this titanic genome-sequencing race; since 1998, when the first mycobacterial genome sequence was published (Cole 1998a); many genome projects have been initiated. Until April 2007, 34 projects on the genome sequencing of different mycobacterial species are finished or in-process. Of these, 15 are directed towards *M. tuberculosis* strains, and 5 towards other members of the *M. tuberculosis* complex. This information will be invaluable to improve the knowledge about *M. tuberculosis* in the next few years. Currently, there are only two *M. tuberculosis* (H37Rv and CDC1551) and two *M. bovis* (AF2122/97 and BCG Pasteur) genome sequences annotated and published. For this reason, these are the strains that have been used as reference strains for comparative genomics both *in vitro* and *in silico*.

The pioneer of *in vitro* assays of comparative mycobacterial genomics involved comparison of restriction profiles using low frequency restriction enzymes and pulsed-field gel electrophoresis (PFGE). These studies allowed a rough analysis of differences among *M. bovis* bacille Calmette-Guérin (BCG) isolates (Zhang 1995) and most importantly, contributed to the construction of the first physical maps, which were essential for the generation of the first genome sequence (Philipp 1996).

The next step in comparative genomics was the use of genomic subtractive hybridization or bacteria artificial chromosome hybridization for the identification of regions of difference among the strains under analysis (Mahairas 1996, Gordon 1999). Mahairas *et al.* (Mahairas 1996) used subtractive hybridization to identify regions of difference that account for the avirulent phenotype of the vaccine strain *M. bovis* BCG. As a result of their studies, they identified three regions of difference (RD1-RD3) in the genome of *M. tuberculosis* H37Rv that appeared to be absent from *M. bovis* BCG. Further studies of these regions showed that RD3 corresponded to the prophage PhiRv1, a sequence that has been shown to vary among *M. tuberculosis* clinical isolates and laboratory strains (see section 4.2.1). RD2 was only deleted in isolates of *M. bovis* BCG that were re-cultured after 1925. Finally, RD1 turned out to be the only sequence deleted from all *M. bovis* BCG strains and present in pathogenic strains. However, complementation assays did not reconstitute the full virulent phenotype in *M. bovis* BCG (Mahairas 1996). The RD1 region contains eight ORFs, including members of the Early Secretory Antigenic Target 6 (ESAT-6) gene cluster (Brosch 2000a). The ESAT-6 proteins have been shown to act as potent stimulators of the immune system (Brodin 2002). The genome of H37Rv contains 23 copies of ESAT-6 family proteins distributed in 11 different regions. Except for *esxQ*, all are clustered in pairs belonging to the ESAT-6 and CFP-10 protein families (Stanley 2003, Gey Van Pittius 2001).

Gordon *et al.* (Gordon 1999) used ordered bacteria artificial chromosome arrays to determine genomic differences between *M. tuberculosis* H37Rv and *M. bovis* BCG. As a result, they identified 10 regions of difference, including the three previously described (Mahairas 1996). Interestingly, two of the newly described regions (RD5 and RD8) also contained members of the ESAT-6 family of proteins. In addition, RD5 contained three genes coding for phospholipase C, a gene with a putative role in mycobacterial pathogenesis (Johansen 1996). Several members of the PE and PPE family proteins were also found in the regions of difference. One copy of *IS1532* was identified in RD6 and one copy of *IS6110* in RD5. Furthermore, the study searched for regions present in *M. bovis* BCG but absent from *M. tuberculosis* H37Rv. Two regions with this characteristic were found and were named RvD1

and RvD2 standing for H37Rv Deleted. Almost all ORFs from these regions code for unknown proteins, so the role of these deletions has not been elucidated.

Until 2002, most studies concerning comparative genomics were based on differences among the strain type *M. tuberculosis* H37Rv and other tuberculous bacilli (Behr 1999, Brosch 1999, Brosch 2002). Different approaches using DNA hybridization techniques, such as microarrays, allowed identification of regions of difference with more accuracy and sensitivity than previous methodologies. In total, 16 regions of difference have been found in *M. tuberculosis* H37Rv that were deleted from *M. bovis* BCG. The basic idea behind the identification of regions of difference between the avirulent strain *M. bovis* BCG and the virulent laboratory strain *M. tuberculosis* H37Rv was the identification of specific deletions in all BCG strains that could be responsible for their lack of virulence. However, nine of the regions of difference were also absent in pathogenic isolates of *M. bovis*.

Other studies have been done comparing *M. tuberculosis* H37Rv to its avirulent counterpart *M. tuberculosis* H37Ra (Brosch 1999), in which other Rv-deleted regions were identified. These regions, named RvD3 to RvD5, were found to be products of homologous recombination of adjacent IS6110, as with RvD2. Finally, only RD1 was found to be absent in all *M. bovis* BCG strains and present in other members of the complex.

The regions of difference were used as markers of the molecular evolution of *M. tuberculosis* (Brosch 2002) and are represented in Figure 4-1. The use of deletions as molecular markers has been described in Chapter 2.

Besides the above mentioned deletions, two duplications were identified in the *M. bovis* BCG genome (Brosch 2000b). These duplications, named DU1 and DU2, apparently arose from independent events. DU1 seems to be restricted to the BCG Pasteur strain and comprises the *OriC* locus, indicating that BCG Pasteur is diploid for *OriC* and some other neighboring genes. The DU2 region has been found in all BCG substrains tested and includes the sigma factor *sigH*, which has been related to the heat-shock response (Brosch 2001). Some excellent reviews are available on comparative genomics, made before the publication of the second *M. tuberculosis* genome (Cole 1998a, Brosch 2000a, Brosch 2000c, Brosch 2001, Domenech 2001, Cole 2002a, Cole 2002b).

In 2002, the second *M. tuberculosis* genome sequence was completed, namely the clinical strain CDC1551, which had been previously involved in a TB outbreak. This strain was considered to be highly transmissible and virulent for human beings (Fleischmann 2002). With the sequence of this second strain, a first approach to the bioinformatic analysis of intraspecies variability became possible. In the initial

comparison by sequence alignment, H37Rv presented a total of 37 insertions (greater than 10bp) relative to strain CDC1551; from these, 26 affected ORFs while the remaining 11 were intergenic. On the other hand, CDC1551 presented 49 insertions relative to *M. tuberculosis* H37Rv; 35 affecting ORFs and 14 intergenic. A total of 80 ORFs were inserted in either genome, 25 (31.2 %) of them were hypothetical or conserved hypothetical ORFs, while 36 (45 %) corresponded to the family of PE/PPE proteins, showing the potential role of this family of proteins in antigenic variability and thus in pathogenicity.

Deletion	<i>M. tuberculosis</i> H37Rv	<i>M. africanum</i>	<i>M. microti</i>	<i>M. bovis</i>	<i>M. bovis</i> BCG
RD2					
RD14					
RD1					
RD4					
RD12					
RD13					
RD7					
RD8					
RD10					
RD9					
RvD1					
TbD1					

Figure 4-1: Distribution of deleted regions in *M. tuberculosis* complex members. Dark gray filled cells indicate the presence in all strains tested, light gray indicate the presence in some strains, white is absence from all strains tested. Data taken from (Gordon 1999, Brosch 2002, Brosch 2000b, Marmiesse 2004)

Only one major rearrangement was found, consisting of the PhiRv1 (RD3), which was found in the genome of *M. tuberculosis* H37Rv on coordinates 1,779,312 associated with a protein of the REP13E12 family. On the genome of CDC1551, it was found to be located on the complementary strand at coordinates 3,870,803, also associated with a REP13E12 protein. *M. tuberculosis* CDC1551 was found to have four copies of IS6110 while *M. tuberculosis* H37Rv had 16. Interestingly, four of the 16 IS6110 copies found in *M. tuberculosis* H37Rv lacked the characteristic 3 to 4 base pair direct repeat and were adjacent to regions deleted in *M. tuberculosis* H37Rv relative to *M. tuberculosis* CDC1551, which suggests homologous recombination.

Since 2002, a large number of studies has been based on Large Sequence Polymorphisms (LSPs) and Single Nucleotide Polymorphisms (SNPs), identified by the comparison of the first two *M. tuberculosis* genome sequences (Hughes 2002, Gutacker 2002). These studies have been complemented with data obtained from the genome sequence of a third organism of the *M. tuberculosis* complex. The complete genome of *Mycobacterium bovis* AF2122/97, a fully virulent strain isolated from a diseased cow in 1997 in Great Britain, was published in 2003 (Garnier 2003). This genome was composed of 4,345,492 bp with a G+C content of 65.63 %, 3,952 putative coding genes, one prophage (PhiRv2), and four IS elements. As expected, similarity of more than 99.95 % was found with a complete colinearity, without evidence of extensive rearrangements. With regard to LSP, most of them have been described above as regions of difference. Sequencing confirmed the absence of 11 regions of difference, and the presence of only one insertion in comparison to the sequenced *M. tuberculosis* genomes: the region named *M. tuberculosis* specific deletion 1 (TbD1), is a reflection that deletion events relative to *M. tuberculosis* have shaped the *M. bovis* genome. The comparison of the three genomes reflects the high degree of conservation among the members of the *M. tuberculosis* complex, as well as the divergence of *M. bovis* related to *M. tuberculosis* strains.

For specific proteins or genes that vary between *M. bovis* and *M. tuberculosis*, a detailed list can be found in Garnier *et al.* (Garnier 2003). However, it is important to mention that the greatest degree of variation among these bacilli is found in genes encoding cell wall components and secreted proteins. Extensive variations have been found in genes of the PE/PPE family of proteins as well as in genes from the ESAT-6 family, where six of the more than 20 members are absent or altered in *M. bovis*. Some other changes are registered in genes coding for lipid synthesis and secretion as the *mmpL* and *mmpS* family of genes. Deletions responsible for the *M. bovis* requirement of pyruvate as a carbon source were also identified (Garnier 2003).

The analysis of the genome sequence of members of the *M. tuberculosis* complex has led to great advances in the knowledge of the biology and pathogenesis of these bacteria. The sequencing of whole genomes of *Mycobacterium leprae* (Cole 2001), *Mycobacterium avium* subspecies *paratuberculosis* (Li 2005) and of other members of the genus, such as *Mycobacterium smegmatis* and *M. bovis*, has also made huge contributions to the understanding of the lifestyle of mycobacteria. Recently, a report compared the metabolic pathways shared among five of the mycobacterial genomes that have been sequenced (the genome sequence of *M. smegmatis* was not included on this report) (Marri 2006). The characteristics of the sequenced ge-

names of organisms in the genus *Mycobacterium* are presented in Table 4-2. The main differences were found in ISSs, the PE/PPE gene family, genes involved in lipid metabolism and those encoding hypothetical proteins. The members of the *M. tuberculosis* complex had the highest number of IS elements, which might suggest higher intra-species variability in *M. tuberculosis* compared to other species of mycobacteria.

Table 4-2: Features of sequenced genomes of species belonging to the *Mycobacterium* genus*

Feature	<i>M. tuberculosis</i> H37Rv	<i>M. tuberculosis</i> CDC1551	<i>M. bovis</i> AF212/97C	<i>M. leprae</i>	<i>M. avium</i> subsp. <i>paratuberculosis</i>	<i>M. smegmatis</i>
Genome size (bp)	4,411,529	4,403,836	4,345,492	3,268,203	4,829,781	6,988,209
Protein coding genes	3,927	4,186	3,920	1,604	4,350	6,897
G+C (%)	65.6	65.6	65.6	57.79	69.3	67.40
Protein coding (%)	91.3	~ 91	90.8	49.5	91.5	92.42
Gene density (bp/gene)	1,114	1,052	1,099	2,037	1,112	1,013
Average gene length	1,012	952	995	1,011	1,015	936
tRNAs	45	45	45	45	45	47
rRNA operon	1	1	1	1	1	2

*Data taken from Li 2005, Marri 2006

The comparison of the proteins encoded within the five sequenced genomes revealed a core, or a number of shared proteins, of 1,326 proteins, compared to the 219 core genes described by macroarray and bioinformatic analyses (Marmiesse 2004). Unique genes ranged between 966 (*M. avium* subsp. *paratuberculosis*) and 26 (*M. tuberculosis* H37Rv) depending on the genome, and most of these proteins are hypothetical. Regarding the PE/PPE family proteins, it is worth mentioning that *M. tuberculosis* and *M. bovis* contained the highest number of these proteins, while neither *M. leprae* nor *M. avium* subsp. *paratuberculosis* have PE_PGRS proteins. Also, a wide variation has been noted in the *mmpL* gene family, known to partici-

pate in lipid transport and secretion. It has been proposed that these variations could be involved in host specificity (Marsh 2005).

4.2.3. Comparing genomes of clinical strains of *M. tuberculosis*

Genome comparison has shown that gene content can vary between strains of *M. tuberculosis*. The analysis of complete genome sequences identified SNPs, LSPs, and regions of difference (RDs) when clinical isolates of *M. tuberculosis* were compared (Fleischmann 2002, Gutacker 2002, Tsolaki 2004, Filliol 2006).

The microarray approach allows the comparison of a large number of genomes, providing information on the diversity, frequency, and phenotypic effects of polymorphisms in the population (Tsolaki 2004). This kind of genomic analysis is also useful for the investigation of outbreaks. Particularly when applied to genomics, DNA microarrays allow the identification of sequences present in the *M. tuberculosis* reference strain, but absent from different clinical isolates. Unfortunately, the microarray technique cannot detect genes present in a clinical isolate that are absent in the reference strain. These changes can originate from small deletions, deletions in homologous repetitive elements, point mutations, genome rearrangements, frame-shift mutations, and multi-copy genes (Ochman 2001, Schoolnik 2002). Fleischman *et al.* suggested that genetic variation among *M. tuberculosis* strains might denote selective pressure, and therefore might play an important role in bacterial pathogenesis and immunity (Fleischmann 2002). Although associations between host and pathogen populations seems to be highly stable, the evolutionary, epidemiological, and clinical relevance of genomic deletions and genetic variation regions remain ill-defined, as do the molecular bases of virulence and transmissibility (Hirsh 2004).

Up to six *M. tuberculosis* lineages adapted to specific human populations have been described by Gagneux *et al.* using comparative genomics and molecular genotyping tools: the Indo-Oceanic lineage, East-Asian lineage, East-African-Indian lineage, Euro-American lineage, and two West-African lineages (Gagneux 2006, see chapter 2). Specific deletions associated with the hypervirulent Beijing/W strains of *M. tuberculosis* were identified (Tsolaki 2005). Evidently, these differences cannot include sequences present in clinical isolates that are absent from *M. tuberculosis* H37Rv, so they necessarily represent a small part of the total potential genetic variability. Up to 13 complete genome sequences of representative *M. tuberculosis* clinical isolates are currently under progress (<http://www.genomesonline.org/gold.cgi>). That number accounts for near half of all

the mycobacterial strains that are currently undergoing complete genome sequencing.

All major functional categories are represented among deleted genes in clinical isolates of *M. tuberculosis*. Mobile genetic elements (insertion sequences or prophages) are frequently deleted. DNA loss frequently results from the activity of the insertion sequence IS6110 (Brosch 1999, Gordon 1999). The rate of deletion in genes involved in intermediary metabolism and respiration, and in cell wall synthesis is surprisingly high. Some of the genes encoding for potential antigens (*plcA*, *plcD*, *lpqH*, *lppA*, *esx*, or PE/PPE genes) might be deleted under the influence of host selective pressures, which would confer an adaptational advantage during infection or help transmission (Tsolaki 2005). Some of these missing genes (e.g. *esx* genes) encode proteins from the ESAT-6 family (Marmiesse 2004).

The use of microarray-based comparative genomics for the study of the genetic variability of pathogens provides interesting information. Not only the identification of the deleted or absent genes is important, but also the differential hybridization signal between samples is of interest. These differential signals can indicate sequence divergence or a difference in the copy number, which may provide an insight into strain evolution and pathogenesis (Taboada 2005).

4.2.4. Functional genomics

Functional genomics is the analysis of the biological function of the genes and their products within a cell or organism. Unlike genomics and proteomics, functional genomics focus on gene transcription, translation, and protein-protein interactions. Genes operate as long as they are expressed and their expression is regulated at the transcriptional or post-transcriptional level.

Functional genomics uses mRNA expression profiling to provide a picture of the transcriptome in a specific condition or time, in order to identify co-regulated genes that perform common metabolic and biosynthetic functions. A set of co-regulated genes is known as a *regulon*.

Microarrays have additional applications in functional genomics apart from gene expression studies, and other uses have also been reported. Using this new powerful technique, Sassetti *et al.* developed a method to map transposon insertion sites in order to identify essential genes in mycobacteria. The probes were synthesized from a transposon library and then used for hybridization in the array. A number of mutants carrying an insert in each gene were obtained, which were later isolated and identified (Sassetti 2001).

The results of studies on comparative mycobacterial genomics have been validated by functional analysis, involving transcriptomics and proteomics. In fact, gene knock-out followed by transcript analysis and proteome definition seems to be the way to identify essential genes. For example, *M. tuberculosis* genes that encode functions essential for growth are prime choices for further investigation as targets for the development of new drugs or diagnostic methods (Cole 2002b).

Subsequently, research derived from comparative genomic studies was directed towards the study of particular genes. That is the case of the deletion designated as RD750 (corresponding to the *Rv1519* gene) in the genome of the *M. tuberculosis* strain named CH, from a large outbreak that occurred in a community of Indian immigrants in the United Kingdom (Rajakumar 2004) and belonging to the East African-Indian lineage. Complementation and combination of *in vitro* and *in vivo* assay systems indicated the participation of the gene *Rv1519* in the persistence and outbreak potential of this *M. tuberculosis* lineage in human populations (Gagneux 2006, Newton 2006).

Construction and transcription analysis of the appropriate mutant have revealed the functional role of the *Rv3676* gene, a member of the cyclic adenosine monophosphate (cAMP) receptor protein family of transcription factors. This factor is required for virulence of *M. tuberculosis* in the mouse model. The functional map obtained from the transcriptome revealed information about regulatory pathways. The global transcription profiling experiments, comparing the wild type *M. tuberculosis* H37Rv strain and *Rv3676* mutant grown *in vitro*, identified some of the genes that are co-regulated, directly or indirectly, by *Rv3676* in *M. tuberculosis* (Rickman 2005).

4.3. Gene expression in *M. tuberculosis*

4.3.1. Control of gene expression

The ability of *M. tuberculosis* to survive within host cells requires a complex and tightly controlled gene regulation. The genes that are used under different conditions could be readily inferred from the corresponding mRNAs. Thanks to the development of highly specific and sensitive technologies, such as microarrays and quantitative real-time Polymerase Chain Reaction (qRT-PCR), it is now possible to analyze the global expression from both the bacillus and the infected host. Taken together, all this could help us to understand the adaptative machinery of *M. tuberculosis*.

The deciphering of the complete *M. tuberculosis* genome sequence has unveiled its well-equipped machinery, which accounts for its high degree of adaptability. Thirteen putative sigma (σ) factors and 192 regulatory proteins seem to be involved in the control of *M. tuberculosis* gene expression (Cole 1998a). Interchangeable σ factors regulate the function of RNA polymerase, initiating transcription and conferring promoter specificity to the holoenzyme (Kazmierczak 2005, Mooney 2005). To date, consensus promoter sequences have been proposed for six σ factors, besides the housekeeping σ factor, σ^A (for a review, see Rodriguez 2006). Gene expression levels could be further modified by the action of transcriptional activators and repressors: regulatory proteins (Barnard 2004). These regulatory proteins include 11 two-component systems, five unpaired response regulators, seven *wbl* genes, and more than 130 other putative transcriptional regulators (Cole 1998a).

The differential expression of these regulatory gene products throughout different stages of the lifespan of *M. tuberculosis* must be determinant for the pathogen's successful infection and/or persistence within the human host. In recent years, a number of reports have correlated the response of several of these transcriptional regulators to a variety of environmental stresses (for a summary, see Table 4-3 at [http://www.tuberculosis textbook.com/pdf/Table 4-3.pdf](http://www.tuberculosis textbook.com/pdf/Table_4-3.pdf)), such as cold shock, heat shock, hypoxia, iron or zinc starvation, nitric oxide, surface stress and oxidative stress (Manganelli 1999, Raman 2001, Sherman 2001, Shires 2001, Manganelli 2002, Stewart 2002, Park 2003, Rodriguez 2003, Voskuil 2003, Canneva 2005, Geiman 2006). However, the biological signals that stimulate the expression of the majority of them are still poorly recognized. Likewise, the connections between the different regulatory circuits of the complex network that controls gene expression in *M. tuberculosis* are incompletely established. An example of the intricacy of this network is the genetic regulation of *sigB*, which is induced by σ^E in response to surface stress (Manganelli 2001) or by σ^H under heat shock and oxidative stress (Manganelli 2002). The regulation of *sigB* expression seems to be more complex than the above cited, given that σ^F - and σ^L -dependent promoters were identified in the regulatory promoter region of *sigB*; and σ^L -dependent transcription was originated upstream to *sigB* (Dainese 2006). It has been shown that σ^H is also responsible of the induction of *sigE* after heat shock and exposure to diamine (Raman 2001). DNA microarray experiments with *M. tuberculosis* mutants revealed that some σ factors control the expression of their own structural gene (Manganelli 2002, Geiman 2004, Sun 2004, Raman 2004, Dainese 2006). Autoregulation has also been demonstrated for the six two-component systems studied so far in *M. tuberculosis*, which are *senX3-regX3* (Himpens 2000), *trcRS* (Haydel 2002), *prxAB* (Ewann 2004), *dosRS* (Bagchi 2005), *mprAB* (He 2005), and *phoPR* (Gupta

2006). Two-component signal transduction systems are composed of a histidine kinase sensor and a cytoplasmic response regulator that is activated by the cognate histidine kinase (West 2001). One of these systems, *dosRS*, is induced by hypoxia, exposure to ethanol or the nitric oxide donor S-nitrosoglutathione (Sherman 2001, Voskuil 2003, Kendall 2004). This regulon is responsible for the transcriptional changes during oxygen limitation, which is considered an important stimulus for the entry of *M. tuberculosis* into a dormant state (Wayne 2001). For this reason, the genes included under control of *dosRS* are considered members of the *dormancy regulon*. Recently, the induction of *sigB* and *sigE* has been shown to depend on the two-component system MprA/MprB when the bacilli are subjected to surface stress (He 2006).

The transcriptional regulator WhiB3 seems to positively regulate the expression of the housekeeping σ factor named *sigA*, by interacting with the subregion 4.2 of σ factor (Steyn 2002). WhiB3 is encoded by one of the seven *whiB*-like genes described in the *M. tuberculosis* genome (Cole 1998a) and belongs to the *wbl* family of genes, which encodes putative transcription factors, which are unique to actinomycetes (Molle 2000, Soliveri 2000). A recent analysis has demonstrated that the expression of *M. tuberculosis* *whiB*-like genes is modified in response to antimycobacterial agents and environmental stress conditions (Geiman 2006). Additionally, *whiB1* transcription is regulated by cAMP levels via direct binding of the activated form of the product of *Rv3676* (CRP protein-cAMP) to a consensus site adjacent to the *whiB1* promoter (Agarwal 2006).

A post-translational regulation has also been reported for several σ factors. Antagonist proteins, known as anti- σ factors, can negatively regulate some σ factors by sequestering them and preventing their association with RNA polymerase. Many of these anti- σ factors are located downstream of their cognate σ factor-encoding gene and both genes are usually co-transcribed (Bashyam 2004). The functions of five specific anti- σ factors of *M. tuberculosis* have so far been examined: RseA (Rodrigue 2006); RshA (Song 2003); RslA (Hahn 2005, Dainese 2006); RsbW or UsfX (Beaucher 2002); and RskA (Saïd-Salim 2006). Interestingly, RsbW, the σ^F -specific antagonist, is post-translationally regulated by two anti-anti- σ factors: RsfA and RskB (Beaucher 2002, Parida 2005).

Although the function of many of these mycobacterial transcriptional regulators and signal transduction systems remains poorly defined, recent studies have begun to provide evidence of the biological role of these regulatory circuits throughout each stage of the lifecycle of *M. tuberculosis* inside the human host. The expression of *sigA*, *sigE* and *sigG* (Manganelli 2001, Capelli 2006, Volpe 2006), that of some

two-component systems (Ewann 2002, Haydel 2004, Walters 2006), as well as that of the transcriptional regulator *whiB3* are induced during macrophage infection. The role of these transcriptional regulators in pathogenesis and virulence became even more evident in animal model experiments, where disruption or deletion of these genes was shown to affect *M. tuberculosis* virulence in mice (Parish 2003a, Parish 2003b, Sun 2004, Manganeli 2004, Raman 2004, Hahn 2005, Walters 2006).

Studies on mutagenesis and the expression profile of several regulators during the growth of *M. tuberculosis* in the macrophages and in organs of experimental animal models are currently underway. These regulators can modify bacterial physiology and are able to modulate host-pathogen interactions in response to environmental signals.

4.3.2. *In vitro* gene expression

M. tuberculosis is an obligate mammalian pathogen that is able to infect many different cells, including macrophages, dendritic cells, alveolar-epithelial cells, and neutrophils. It is also able to reside extracellularly in the lung, inside granulomas. As mentioned previously, the tubercle bacillus adapts its transcriptome to the environment in which it replicates. The adaptation of a bacterium to harsh environments involves the transcriptional activation of genes whose final products help the bacterium to reprogram its physiology, thus ensuring survival. Among the genetic determinants that the bacterium must modulate are those involved in intermediary and secondary metabolism, cell wall processes, stress responses and signal transduction pathways.

By utilizing the microarray technology, quantitative RT-PCR and laboratory generated mutants, studies on *M. tuberculosis* global gene expression have been undertaken using broth cultures, cell cultures, and animal models. None of these models reproduce several key features of TB in the human infection; and, unfortunately, no data is available from human tissues.

Table 4-4 summarizes the most important genes whose expression is modulated by the transcriptional regulators mentioned previously (see section 4.3.1).

Table 4-4: Regulation of cell process genes

Transcriptional regulation	Condition	Cell process genes	
		Up regulated	Down regulated
σ^C	Growth curve	Two-component systems: <i>senX3</i> , <i>mtrA</i> , <i>hspX</i> (α -crystallin), <i>fbpC</i> (antigen 85C)	
σ^D	Exponential growth	<i>rpfC</i> (resuscitation factor), <i>Rv1815</i> , <i>Rv3413c</i>	PE_PGRS family genes
σ^H	Diamine exposure, heat stress ^b	Heat shock proteins: <i>hsp</i> , <i>clp</i> , <i>trxB2C</i> operon, transcriptional regulators: <i>sigE</i> , <i>sigB</i>	
σ^M	Log-phase growth	Esx family genes, PPE1, PPE19	PPE60, <i>kasA-kasB</i> , <i>fas</i> , <i>pks2</i> , <i>pks3</i>
σ^E	Exponential growth, SDS exposure	Icl1 (isocitrate lyase), heat-shock proteins, transcriptional regulators: <i>sigB</i> , <i>mprAB</i>	
σ^L		<i>sigL-rsIA</i> , <i>pks10-pks7</i> , <i>mpt53-Rv2877c</i> , <i>Rv1139c-Rv1138c</i>	
σ^F	Stationary-phase growth	HP and CHP family of proteins, transcriptional regulators: <i>sigC</i> , <i>sigF</i> , and MarA, GntR TetR family, cell envelope: <i>murB</i>	
DosS-DosR	Standing cultures	<i>hspX</i> (α -crystallin), <i>Rv3130c</i> (CHP), <i>Rv1738</i> (CHP), <i>Rv0572c</i> (CHP)	
PhoPR	Exponential growth	Cell envelope components	
MprA	Mid-exponential phase	~95 genes of <i>M. tuberculosis</i> . PE/PPE gene family, HP, CHP	

CHP = conserved hypothetical protein

For example, analysis of a mutant of *M. tuberculosis sigC* showed that this σ factor induces the expression of some virulence-associated genes. On the contrary, the genes *hspX* (encoding the α -crystalline homologue), *senX3* (sensor kinase), *mtrA* (response regulator), and *fbpC* (mycolyl transferase and fibronectine binding protein or antigen 85C) were down-regulated in that mutant strain at different times of the growth curve (Sun 2004). Genes induced by σ^D include the resuscitation promoting factor *rpfC*, several chaperone genes and genes involved in lipid metabolism and cell wall processes (Raman 2004). Thirty-nine genes were shown to be under the control of σ^H . These include genes coding for some heat shock proteins

(*hsp* and *clp*), the *trxB2C* operon and some transcriptional regulators (Manganelli 2002). Quantification of mRNA by primer extension under different stresses demonstrated that the transcription of *trxB2*, *dnaK*, *clp* and *sigE* could be induced from σ^H -dependent promoters located upstream of these genes (Raman 2001). σ^E seems to regulate the expression of proteins involved in fatty acid degradation, such as the isocitrate lyase (coded by *icl1*), two proteins related to fatty acid degradation (*fadE23* and *fadE24*), heat shock proteins, and the transcriptional regulators *sigB* and *mprAB* (Manganelli 2001). Recently, it was reported that σ^M induces the expression of two pairs of secreted proteins of the Esx family and two PPE genes, while it negatively regulates PPE60 expression as well as the expression of several genes involved in surface lipid biosynthesis and transport (Raman 2006).

At least four small operons appear to be directly regulated by σ^L : *sigL-rsIA*, *pks10-pks7*, *mpt53-Rv2877c*, and *Rv1139c-Rv1138c*, which clearly have a σ^L -consensus promoter sequence in their regulatory region (Hahn 2005, Dainese 2006). The *pks* genes are involved in the biosynthesis of phthiocerol dimycocerosate, a component of the cell envelope associated with virulence (Sirakova 2003); and the *mpt* operon contains genes involved in fatty acid transport (Sonden 2005). DNA microarrays of a mutant of *M. tuberculosis* lacking a functional *sigF*, have revealed that this σ factor is able to induce gene expression almost exclusively during the stationary phase of growth, supporting the hypothesis of a major role of σ^F in the adaptation to the stationary phase. Among the σ^F -targeted genes, 50 % coded for hypothetical proteins or proteins of unknown function, some were transcriptional repressor/activators (MarR, GntR and TetR family of DNA binding regulators) and others were found to be involved in the biosynthesis and structure of the cell envelope (Geiman 2004).

A complete genomic microarray analysis has also been performed on *M. tuberculosis* strains mutated in two-component regulatory systems. The dormancy-related two-component system *dosRS* was inactivated in *M. tuberculosis* using different methodologies (Parish 2003b, Park 2003). It was shown that the expression of this two-component system is highly induced under hypoxia (Sherman 2001b, Park 2003). A consensus *dosR*-specific binding motif was reported to be located upstream of hypoxic response genes (Park 2003, Kendall 2004). The microarray expression profiles of mutants in each of the components (*dosR* and *dosS*) showed that DosR is required for the expression of genes usually induced under oxygen limitation, such as *hspX* gene. Several putative operons with unknown function were also strongly regulated by DosRS.

Up to 30 genes were found to be up-regulated, and another 68 genes down-regulated in a mutant of *M. tuberculosis* *senX3-regX3*. However, it has not been clearly determined if the changes found in gene expression were directly or indirectly related to the lack of this two-component regulatory system (Parish 2003a). Recently, the global transcriptional profile of the two-component systems PhoP and MprA has been reported. One of these studies provided evidence that the PhoP/PhoR system is a positive transcriptional regulator of genes involved in the synthesis of the cell envelope of *M. tuberculosis* (Walters 2006). On the other hand, MprA regulates *sigB* and *sigE* and many other genes previously reported to be associated to various stress conditions (He 2006).

In order to analyze the mechanisms involved in bacilli intracellular survival, mycobacterial gene expression was determined in *M. tuberculosis* infected macrophages from different sources. Macrophages play a crucial role in TB infection because they represent both the effector cells for bacterial killing and the primary habitat in which the persisting bacilli reside. Macrophages have been investigated at different time points post-infection for the differential expression of various two-component system regulators (*regX3*, *phoP*, *prnA*, *mprA*, *kdpE*, *tcr*, *devR* and *tcrX*) (Haydel 2004). More recently, the gene expression profile of *M. tuberculosis* grown in human macrophages compared to that of bacteria growing in synthetic culture medium was published (Capelli 2006). In this work, the authors reported that approximately one-third (32 %) of the genes upregulated by *M. tuberculosis* in macrophages correspond to conserved hypothetical proteins, with unknown function; this finding highlights the considerable gap that still remains in the knowledge of how this bacterium survives intracellularly. Genes involved in cell wall processes (19.5 %), regulation and information pathways (16 %), and PE family proteins (3.6 %) were also upregulated. Interestingly, the authors observed high induction of the sigma factor *sigG* and 13 other putative transcriptional regulators. Upregulation of *sigA*, *sigE*, and *sigG* was also reported in a similar study (Volpe 2006). The *whiB3* gene was also induced in *M. tuberculosis* during infection of naïve bone marrow-derived macrophages in comparison to bacteria in broth culture mid-log growth (Banaiee 2006).

4.3.3. *In vivo* gene expression

The use of microarrays for profiling transcriptomes of bacteria inside the host cell has been limited by the paucity of bacterial mRNA in samples containing a preponderance of mammalian RNA. Therefore, while significant work has been per-

formed on the gene expression profile of the host, information on *M. tuberculosis* expression inside infected hosts is still limited.

So far, there is only one publication concerning global mycobacterial transcription expression in the animal model, using microarray as the analytical method (Talaat 2004). Differential expression levels of *M. tuberculosis* during infection in Balb/c or Severe Combined Immunodeficiency (SCID) mice were evaluated and compared to the levels found in mycobacteria grown in broth culture. These authors identified up to 40 genes whose expression significantly changed during Balb/c and SCID mouse infection. These genes include *rubB*, *dinF*, and *fdxA*. The same genes were also found to be induced 24 hours post-infection in murine bone marrow macrophages (Schnappinger 2003). Additionally, several genes were regulated up or down only in Balb/c mice, such as *proZ* (transport system permease protein), *aceAa* (probable isocitrate lyase involved in lipid metabolism), and genes encoding for regulatory proteins, such as *sigK*, *sigE* and *kdpE*. The authors concluded that the expression profile of *M. tuberculosis* in SCID mice resembles the profile found in bacilli grown *in vitro*, while the expression profile in Balb/c mice resembles that reported in multiplication within the macrophages (Schnappinger 2003). Exceptionally, some genes were found to be expressed only in Balb/c mice.

A small number of studies applied quantitative RT-PCR to investigate the expression of mycobacterial genes in the animal model. These studies focused on the analysis of a few particular genes. Examination of lungs of infected C57BL/6 mice showed that the transcriptional regulator genes *whiB3*, *fdxA* (electron transfer), *hspX* (α -crystalline), *acg* (unknown function), *Rv1738* (unknown function), and *Rv2626c* (unknown function) were markedly induced during the course of infection (Banaiee 2006). A gene required for extrapulmonary dissemination (*hbhA*) was also upregulated in the lung but not in the spleen during the early stages of infection (Delogu 2006). While the expression of PE_PGRS16 was up-regulated in the spleens and lungs of infected mice, the expression of PE_PGRS26 was down-regulated (Dheenadhayalan 2006). A study on human lung biopsies revealed a high variability in expression profiles of specific *M. tuberculosis* genes among the specimens analyzed (Timm 2003). The biopsies were obtained from four HIV-negative patients with chronic active TB that was unresponsive to therapy. Although some differences were observed when comparing human and murine lung, the authors admitted that it was difficult to ascertain whether the infection stage in the analyzed human lung specimens could be correlated with the persistent infection in mice.

4.4. *M. tuberculosis* proteome

With the availability of the genomes of *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551, *M. bovis*, and the ongoing sequencing projects, attention in the coming years must be focused on the interpretation of the sequences determining the structure and function of the proteins. Proteomics, the global study of proteins that are translated in a given physiological state is one of the most important and ambitious goals in *M. tuberculosis* research. The proteome of an organism implies not only an inventory of its gene products but also the transduction rate and the post-transcriptional events that occur in the organism (Betts 2002). Classical studies of proteomics involve two dimensional electrophoresis (2-DE), in which proteins are first separated by the isoelectric point and then by the molecular weight (O'Farrel 1975). Every spot of protein is then isolated, hydrolyzed and subjected to techniques of mass spectrometry (MS), tandem MS (MS/MS); matrix-assisted laser desorption/ionization mass spectrometry (MALDI/MS) and, matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF/MS). For a good review on the different techniques used in protein mapping, readers are referred to Patterson *et al* (2000). Techniques different from two dimensional electrophoresis have also been implemented. For instance, the use of one dimension electrophoresis has been shown to be very useful for the separation of hydrophobic proteins (Simpson 2000). Other approaches that do not involve the use of gels, such as two-dimensional liquid chromatography (LC) and the subsequent analysis by MS (2 DLC/MS), have been shown to be very efficient in the identification of hydrophobic and membrane proteins (Isobe 1991). In 1999, the isotope-coded affinity tag (ICAT) technology was reported (Gygi 1999). In this, mixtures of proteins from bacteria in two different conditions are covalently labeled with isotopically labeled heavy or light forms of the reagents. The samples are combined and subjected to proteolysis. After purification of the labeled peptides through affinity tag, which is part of the reagents, they are analyzed by LC-MS/MS. This new technology has proven to be very useful in the quantitation of complex mixtures of proteins.

Before the disclosure of the *M. tuberculosis* genome, antigens and proteins were identified by one- and two-dimension polyacrilamide gel electrophoresis, and the use of cumbersome immunological methods (Nagai 1991, Garbe 1996). With the advance in high-resolution 2-DE and analytical chemistry, *M. tuberculosis* proteome is at present a reality. Pioneering studies in the proteomic field included the mapping of 32 N-terminal sequences by MS and the identification of culture filtrate proteins by 2-DE and immunodetection (Sonnenberg 1997).

Very soon after the publication of its genome, bioinformatic tools were applied to predict the proteomic profile of *M. tuberculosis* (Tekaiia 1999). The in silico analysis showed characteristics of the tubercle bacillus as the duplication of numerous genes, especially those involved in gene regulation and in lipid metabolism, and those coding for the PE/PPE protein family. The study also showed a reduced repertoire of proteins devoted to transport, which might reflect the intracellular lifestyle. The bioinformatics-predicted proteome was compared to 2-DE protein maps obtained from *M. tuberculosis* whole cell lysates, which were separated in a broad pH range between 2.3 and 11.0. This work demonstrated that proteins with a molecular mass below 10 kDa were not predicted from the genome sequence and also that experimentally basic and high molecular mass proteins could not be resolved by 2-DE (Urquhart 1998). Since 1999, the huge amount of data in proteomics has led to the creation of 2-DE databases, where images generated in different laboratories can be stored and analyzed. These databases are accessible on the internet at: <http://web.mpiib-berlin.mpg.de/cgi-bin/pdbs/2d-page/extern/overview.cgi?gel=16> (Mollenkopf 1999) and <http://www.ssi.dk/sw14644.asp> (Rosenkrands 2000b).

4.4.1. Structural proteomics of *M. tuberculosis*

Thanks to recent technological advances, the subcellular protein profile of *M. tuberculosis* can now be drawn. The global analysis of compartmentalized proteins will shed light on host-pathogen interactions, metabolic pathways and cell communication, just to mention some of the mechanisms related to pathogenesis. In addition, many pathogenic bacteria secrete proteins that are involved in virulence (Finlay 1997) and thus culture filtrates of *M. tuberculosis* could be a source for identification of virulence factors. Cell wall proteins play a fundamental role in cell architecture, resistance of the pathogen to chemical injury and dehydration, and many other key functions of this microorganism. Thus, the identification of proteins localized in this subcellular fraction may lead, in the near future, to the development of new diagnostic tests and drugs. Membrane proteins demand special attention, because they are involved in host-pathogen interactions, nutrient transport, quorum sensing mechanisms, etc. Knowledge of these proteins could be the clue to the development of novel vaccines. Finally, the identification of cytosol proteins and the intricate network of their interaction will reveal metabolic pathways that can be targets for the design of rational drugs against TB. Even though we are still far from identifying the almost 4,000 genes predicted by genomics, the number of identified proteins increases each year and shows how genomic and proteomic technologies complement each other.

The biochemical methods developed for the separation of the cell wall, membrane and cytosol fractions have facilitated proteomic studies in *M. tuberculosis* (Hirschfield 1990, Lee 1992). Jungblut *et al.* identified 53 proteins from cell lysates and 54 from culture filtrates using 2-DE and MALDI/MS (Jungblut 1999). These authors performed comparative proteomics in *M. tuberculosis*, which will be discussed later in this chapter.

Reference maps of cellular fractions and culture filtrate proteins were constructed using 2-DE, N-terminal sequencing and antibodies against previously identified antigens (Rosenkrands 2000b). As many as 1,184 spot proteins were visualized after silver staining. Only 10 % of them were identified. In order to map less abundant proteins, different methods were applied for their separation, which allowed the identification of 12 novel proteins, five of them with a known function (Rosenkrands 2000b). The study also showed the identification of a protein that was not predicted by genomics and revealed the presence of alternative start codons.

The implementation of immobilized pH gradient for 2-DE and MALDI/MS allowed the identification of 288 proteins (Rosenkrands 2000a). Six proteins were identified, all of them with molecular masses between 13,200 and 7,200 kDa and with isoelectric point (pI) ranges between 4.5 and 5.9. Five of these proteins were correctly identified in the genome of the clinical strain CDC1551 (Jungblut 2001).

In spite of the enormous usefulness of 2-DE in proteomic studies, there are certain disadvantages inherent to its technique, such as the low resolution of proteins with very high or very low molecular masses, or proteins that are very acidic, very basic or hydrophobic. But in particular, the technique is biased towards the preferential identification of the most abundant proteins. Therefore, less abundant proteins, such as transcriptional regulators, are rarely detected when whole cell lysates are analyzed (Gygi 1999). To overcome these inconveniences, alternative techniques have been applied in proteomic studies. For example ICAT reagent method and LC-MS/MS were used as a complement of 2-DE-MS/MS. Using these approaches, 388 *M. tuberculosis* proteins were quantified and identified. Each one of these techniques has been shown to be adequate for the identification of certain classes of proteins. For example, the ICAT method performed better for the identification of cell membrane and high molecular mass proteins, while 2-DE showed better results in the identification of low molecular mass and cysteine-free proteins (Schmidt 2004). Interestingly, none of these techniques allowed the identification of proteins in the following subclasses: cell division, IS elements, repeated sequences, phages, PE/PPE families, cytochrome P450 enzymes, cyclases, and chelatas.

By 2004, only about 400 proteins had been identified in the proteome of *M. tuberculosis*, probably due to the limitations of 2-DE-based separation methods. The use of automated two-dimensional, capillary high-performance liquid chromatography (HPLC) coupled with MS gave a wider and more accurate proteomic profile of *M. tuberculosis* (Mawuenyega 2005). Proteins of the cell wall, membrane and cytosol subcellular fractions could be identified. The number of identified proteins increased to 1,044 non-redundant proteins, 67 % more than those obtained by conventional 2-DE. This study identified proteins in extreme pI ranges, among them the most acidic proteins (PE_PGRS, *Rv3512*) with a pI of 3.89 and the most basic proteins (*rps2*, a 30S ribosomal protein) with a pI of 12.18. Proteins of high molecular mass, such as the 230,621 Da polyketide synthase *ppsC*, were also identified. A total of 705 proteins were identified in the membrane, 306 were localized in the cell wall, and 356 in the cytosol fraction. Forty-seven were present in all analyzed fractions. The study also included a computational analysis of protein networks, one of the most exciting fields in the coming years. Readers are invited to consult the supplementary table of this work (Mawuenyega 2005).

M. tuberculosis is an intracellular pathogen, the bacillus is engulfed by alveolar macrophages where it can survive and grow by altering the intracellular compartments to preclude the normal maturation to phagolysosomes or to prevent fusion of phagosomes to lysosomes (Clark-Curtiss 2003). The interaction between host and pathogen is thought to be mediated by membrane proteins. Therefore, the characterization of membrane proteins is a topic of intensive research. As mentioned before, most of the studies regarding *M. tuberculosis* proteomics have been carried out by 2-DE. However, the number of membrane and membrane-associated proteins has been underestimated by the 2-DE technology due to the hydrophobic nature of this class of proteins and their low solubility. In order to overcome these problems, fractions of cellular membranes were prepared by differential centrifugation and separated by one-dimensional electrophoresis. The separated bands were then excised and hydrolyzed prior to LC and MS/MS (Gu 2003). This approach allowed the identification of up to 739 membrane and membrane-associated proteins. Very hydrophobic proteins, including those with 15 transmembrane helices, were detected in this study. The use of alternative solubilizing agents, such as Triton X-114, has proven to be a good choice for membrane fractionation. The detergent was shown to be useful in the identification of nine novel proteins that have been already incorporated in the *M. tuberculosis* proteome (Sinha 2005). Interestingly, when analyzing the interferon-gamma (IFN- γ) response of BCG-vaccinated healthy individuals from an endemic area to these newly identified proteins, the strongest response was found to be that against ribosomal proteins. Other mem-

brane associated proteins, such as ESAT-6, did not contribute significantly to the T-cell response in these individuals.

4.4.2. Comparative proteomics

The comparative proteomic analysis using 2-DE and MALDI/MS was applied to compare proteins present in two virulent laboratory *M. tuberculosis* strains (H37Rv and Erdman strains) with those present in two *M. bovis* BCG strains (Chicago and Copenhagen BCG strains) (Jungblut 1999). The results showed that, as expected, the two *M. tuberculosis* strains differed from each other in only a few proteins. Of the 18 variant proteins, 16 were identified. L-alanine dehydrogenase (*Rv2780*) was not detected in the Erdman strain, and the protease IV was absent in this strain. On the other hand, the Soj protein and the hypothetical protein *Rv2641* were absent from the *M. tuberculosis* H37Rv proteome. Some of the 18 proteins were over-expressed in one or the other strain, and some shifted their mobility probably due to the presence of amino acid substitutions. The comparison of *M. tuberculosis* H37Rv with *M. bovis* BCG revealed the presence of 13 protein spots exclusive to the tubercle bacilli, six of which were identified. The differential proteins comprised L-alanine dehydrogenase (40 kDa protein), isopropyl malate synthase nicotinate-nucleotide pyrophosphatase (*Rv1596*), MPT64 (*Rv1980c*), and two hypothetical conserved proteins (*Rv2449c* and *Rv0036c*). On the other hand, *M. tuberculosis* H37Rv lacked eight spots compared to *M. bovis* BCG.

In another study using 2-DE and MS, a comparison of the proteins present in *M. tuberculosis* and *M. bovis* BCG revealed the presence of 56 unique protein spots in *M. tuberculosis* and 40 in the attenuated strain BCG (Mattow, 2001). Of these, 32 were identified as exclusive proteins of *M. tuberculosis*, of which 12 had been previously reported to be deleted in *M. bovis* BCG. The remaining 20 spots were newly identified as absent from *M. bovis* BCG.

A third comparative proteomic study of *M. tuberculosis* and *M. bovis* BCG was performed using 2-DE and ICAT technology (Schmidt 2004). This work demonstrated the presence of only three exclusive proteins in *M. tuberculosis* H37Rv. One is *Rv0223c*, a protein belonging to the aldehyde dehydrogenase family. The second is *Rv0570*, a ribonucleotide reductase class II. The third is a hypothetical protein named *Rv1513*.

The studies on comparative proteomics allowed the identification of isopropyl malate synthase exclusively in the *M. tuberculosis* proteome. Recently, this enzyme was included in a new class of virulence factors known as ‘anchorless adhesins’

(Kinhicard 2006) that were absent from the avirulent BCG, thus proving the usefulness of this methodology.

Of special interest in the coming years will be the proteomic comparison between circulating *M. tuberculosis* strains differing in virulence, transmissibility, tissue tropism, and/or ability to acquire drug resistance. As a matter of fact, the proteomic profile of *M. tuberculosis* H37Rv has already been compared with that of the clinical strain CDC1551 at different time points during *in vitro* growth (Betts 2000). Subscribing the low structural DNA polymorphism observed in *M. tuberculosis* (Sreevatsan 1997), the resulting patterns of the protein-spot were found to be both highly reproducible and highly similar between the two strains during growth. One unique protein was identified in *M. tuberculosis* CDC1551, namely *Rv0927c*, a probable alcohol dehydrogenase. Similarly, a spot corresponding to the HisA protein, which is involved in the histidine biosynthetic pathway, was detected in the *M. tuberculosis* H37Rv proteome but was absent from the *M. tuberculosis* CDC1551 protein profile. Oddly enough, both genes were found to be present in both *M. tuberculosis* strains. Thus, the described proteomic differences between H37Rv and CD1551 might be ascribed to post-translational events or to degradation during the manipulation of the specimens. Another interesting feature in the same study was the mobility variations of the transcriptional regulator MoxR, which the authors attributed either to amino acid changes or to post-translational modifications. A BlastP analysis of both genomes showed a single amino acid substitution of histidine in *M. tuberculosis* H37Rv to asparagine in *M. tuberculosis* CDC1551 that might explain the variation in mobility.

Transcriptional regulation differences between strains might be the key to understanding how virulence factors are involved in a variety of roles, including host-cell invasion, survival within the host cell, and long-term persistence. Therefore, comparative proteomic studies are of special interest in the post genomic era, helping to understand the manifestation of disease produced by different strains involved in the current TB epidemic.

4.4.3. Environmental proteomics

The information obtained by genomic studies is static, because DNA is not essentially affected by the environment. In contrast, the proteomic profile of an organism in a particular physiological situation complements and helps to decipher its interaction with the environment. The study of the *M. tuberculosis* proteome in different physiological states is one of the most fascinating fields of research. Being an intracellular pathogen, the bacillus is challenged by a variety of environmental

changes. Inside the mammalian macrophage, the microorganism is subjected to a series of different weapons. Inside the granuloma, it faces low oxygen tension, starvation, low pH, reactive nitrogen, and reactive oxygen species, among other offenses (Schnappinger, 2006). The bacillus has developed adaptive mechanisms for survival and persistence in these hostile environments. The identification of proteins expressed under such conditions is a matter of demanding research.

M. tuberculosis has another outstanding characteristic: it is able to persist for years in its host causing latent infection, in a state known as dormancy. The mechanisms governing this state are still not fully understood and the protein expression profile in models mimicking the dormant state is an issue of intense research. Different *in vitro* models have been developed, aimed at simulating the *in vivo* conditions inducing dormancy (Wayne 1996, Betts 2002, Voskuil 2003). In the Wayne's model, which has been applied in proteomic studies, the level of oxygen is gradually depleted due to bacterial growth, defining two non-replicating stages: a microaerophilic stage NRP-1 (non-replicating persistence-1), followed by an anaerobic stage NRP-2. Still, the evidence linking human *M. tuberculosis* latent infection with *M. tuberculosis* dormant stages attained *in vitro* remains merely circumstantial.

Hypoxia is among the most conspicuous conditions encountered by the tubercle bacilli in the central part of the granuloma, where bacilli are considered to remain dormant. The hypoxic response of *M. tuberculosis* in the Wayne's model was investigated by performing a proteomic study of cell lysates and culture filtrates (Rosendkrands, 2002). The comparison of the protein content between aerobic and anaerobic cultures identified up to seven proteins that were more abundant in hypoxic conditions. The main proteins characterized were fructose biphosphate aldolase (in culture filtrate only), hypothetical protein *Rv0569*, and alpha-crystallin protein, also known as HspX. Other proteins identified included hypothetical proteins *Rv2623* and *Rv2626c*, L-alanine dehydrogenase (only in culture filtrates), and BfrB, a bacterioferritin involved in iron uptake and storage.

Using a modified Wayne's model, Stark *et al.* (Stark 2004) visualized 13 unique and 37 more abundant spots under hypoxia, revealed by 2-DE and MALDI-TOF/MS. Of these 50 spots, 16 proteins were identified, including some that had not been previously detected such as GroEL2, KasB, Ef-Tu, ScoB, TrxB2, and CmaA2. Among the hypothetical proteins found were *Rv2005c*, with similarity to universal stress proteins, *Rv0560c*, *Rv2185c*, and *Rv3866*.

Applying the ICAT technology to the comparison of the physiological NRP-1/NRP-2 states *versus* active growth (logarithmic phase), the number of newly identified proteins increased up to 875 (Cho 2006). A total of 586 proteins were

identified and quantified in the microaerobic stage of nonreplicating persistence stage (NRP-1), 628 others were detected in the anaerobic dormancy state (NRP-2), and 339 were common to both non-replicating persistence stages. Proteomic comparison between the NRP-1 and the logarithmic phase of growth showed that 6.5 % of the proteins were up-regulated in NRP-1, while 20.4 % of the proteins were up-regulated in NRP-2. The analysis of the proteomic profile showed that the NRP-1 state displayed a significant increase in proteins involved in small molecule degradation and the NRP-2 state a significant increase in energy metabolism, altogether suggesting an adaptive mechanism of *M. tuberculosis* to enter into the anaerobic environment.

M. tuberculosis pathogenicity is directly associated with its ability to establish invasion and division inside the host's macrophages, despite the antimicrobial properties of these cells. The study of the *M. tuberculosis* proteome in this physiological environment is a crucial step towards understanding the mechanisms involved in its pathogenicity. In spite of the enormous advances in biochemical analytical techniques, the purification and identification of proteins is not always an easy task. In a recent paper, Mattow *et al.* (Mattow 2006) applied subcellular fractionation of infected murine bone marrow-derived macrophages in combination with high-resolution 2-DE and MS/MS to analyze the proteome of *M. tuberculosis* inside the phagosome. The proteome was compared with that derived from broth cultures. Using this approach, 121 unique spots were detected in intra-phagosomal mycobacteria. Only 11 of them were identified as *M. tuberculosis* proteins, the remaining 110 were identified as murine proteins. The 11 identified proteins were: *Rv1240* (Malate dehydrogenase, MdH), *Rv1077* (Cystathione (beta)-synthase, CysM2), *Rv3396c* (GMP synthase, GuaA), *Rv0489* (Phosphoglycerate mutase I, Gpm), *Rv2773c* (Dihydrodipicolinate reductase, DapB), *Rv0009* (Peptidyl-prolyl cis-trans isomerase, PpiA), *Rv1627c* (lipid carrier protein), *Rv2961* (putative potassium uptake protein, TrkA), and hypothetical protein *Rv1130*, which was detected in two separate spots, *Rv0428c*, and *Rv1191*. Some of these proteins (CysM2, DapB, GuaA, MdH and PpiA) had previously been detected using 2-DE patterns of whole cell lysates and/or culture supernatants of *M. tuberculosis* H37Rv, indicating that they are not exclusive of the phagosomal milieu.

Proteomic studies seem to be a successful way to discover new virulence factors, drug target molecules and proteins involved in pathogenic mechanisms. Thousands of proteins have now been identified and many more await identification.

4.5. An insight into *M. tuberculosis* metabolomics

4.5.1. Metabolomics state-of-the-art

The term **metabolomics** was first coined in 1998 (Oliver 1998) to describe the “*change in the relative concentrations of metabolites as the result of deletion or over-expression of a gene*”. At the same time, the term **metabolome analysis** referred to the analysis of metabolites in the phenotypic profile of *E. coli* (Tweeddale 1998). Later on, metabolomics was considered the detection and measurement, under defined conditions, of cellular metabolites such as low molecular weight molecules present in an organism or biological sample. The field also includes information on the level of metabolite activities in the cell. Metabolites are in general defined as those small molecules, usually intermediate and final products of metabolism, but the definition also applies to high molecular weight molecules such as lipids, peptides and carbohydrates (sometimes referred to as “lipidomics”, etc).

Metabolomic approaches are now feasible due to the rapid improvements that have taken place during the last decade in two areas: analytical chemistry and bioinformatics. Metabolomic methodologies include the combination of classical technologies, such as gas chromatography-mass spectrometry (GC-MS), or nuclear magnetic resonance (NMR), with new developments to achieve improved sensitivity and discriminative power. Sophisticated informatic analysis and data mining are an important part of the methodology. The complete analysis involves *in silico* models on metabolite-protein interactions. This analysis can be qualitative or quantitative. In the latter case, all the conditions required for an accurate quantification should be considered, such as the use of appropriate data standards, etc (Nielsen 2005). Metabolomics can also help to validate *in silico* pathways prepared on the basis of available genome sequences and established databases (Park 2005).

Metabolomic analysis has mainly been used in studies on plants and human pathology; in this latter case, the attention was focused on searching for metabolites associated with disease, in other words, “*metabolites as biomarkers of disease*” (Weckwerth 2005). Microbial metabolomics has initially been devoted to explore bacterial or fungal strains carrying improved phenotypes with a certain biotechnology usefulness value (Wang 2006). Metabolomic approaches have also been directed to the development of new drugs addressed against novel microbial targets. A review on the basics and applications of microbial metabolomics can be read in van der Werf *et al.* (Werf 2005).

4.5.2. Has the metabolomic analysis of tuberculosis actually started?

From the beginning, a unique property of mycobacterial cells called the attention of scientists: the remarkably high lipid content of the cell envelope, which accounts for the most conspicuous mycobacterial features, including physiology and pathogenicity (Asselineau 1998, Barry 2001) (see Chapter 3). Many studies have been published on identification, characterization, and even practical applications (e.g. diagnostics) of several mycobacterial lipids. Almost all books on TB or mycobacteria have at least one chapter dedicated to lipids. Older books refer in more depth to the structural and chemical characterization of the envelope, as well as to the biosynthesis of lipids (Ratledge 1982, Kubica 1984); more recent books lay stress on the genetics and genes related to lipid metabolism (Cole 2005). Thus, the analysis of the lipid metabolic profiling cannot be regarded as a new field in mycobacteria, at least when considering all lipids as metabolites (Ortalo-Magne 1996). The main objection to doing so is the size of mycobacterial lipids. Indeed, mycobacterial lipids are rather big and complex. Most of them have a fatty acid backbone covalently linked to other kind of molecules, most frequently several types of saccharides (Asselineau 1998). Many lipids also belong to the molecular structure of bacterial lipoproteins (Sutcliffe 2004).

A detailed revision of mycobacterial lipids has been published relatively recently (Kremer 2005). The most representative lipids in mycobacteria are the mycolic acids. These molecules are larger in mycobacteria, compared to those of other related bacteria, such as *Corynebacterium* or *Nocardia*. Mycolic acids are the lipid component in the structure of complex glycolipids, including Mycolyl-ArabinoGalactan (mAG) and Trehalose-6-6'-Dimycolate (TDM). Another important group of lipids also contains trehalose as the glycosyl radical molecules and their fatty acids chains are multi-methylated. This group includes Di- Tri- and Pentaacyl Trehalose (DAT, TAT and PAT) and Sulfolipids (SL); the Phthiocerol Dimycocerosates (PDIMs) are also very important. These compounds and the closely related Phenolic Glycolipids (PGL) participate in the integrity of the cell envelope of *M. tuberculosis*. A last group of glycolipids contain D-mannan and D-arabinan in their molecules and have been considered of relevance in bacterial pathogenicity for a long time: Lipomannans (LM) and Lipoarabinomannans (LAM) (see Chapter 3).

Many lipids are unique to mycobacteria and therefore their metabolic analysis cannot be addressed by comparative lipidomic studies with other bacteria. Such specific metabolic pathways are viewed, in turn, as excellent targets for the design of new specific drugs (Draper 2000). Renewed efforts have been applied to the detection of metabolic routes and genes that participate in the biosynthesis and

degradation of complex lipids (Brennan 2003, Reed 2004, Portevin 2004, Veyron-Churlet 2004, Trivedi 2005, Kaur 2006), as well as genes involved in lipid transportation (Domenech 2004, Jain 2005). A review has recently been published on the biosynthesis, regulation and transport of long-chain multiple methyl-branched fatty acids, such as PDIM, PGL, and SL (Jackson 2006). This paper updates the knowledge on these complex topics, indicating that mycobacterial lipids share mechanisms in their metabolic routes, and that changes in a pathway could influence another pathway; in fact, some small molecules, namely metabolites, could be precursors of the more complex synthesis of lipids, and also be synthesized themselves during the lipids' metabolic pathway, thus being by-products or secondary products of the lipid's metabolism.

Few studies deal with the small metabolites from mycobacterial. A recent study on *Rv2221*, coding for a highly efficient *M. tuberculosis* adenylyl-cyclase, indicates that its catalytic activity is regulated by fatty acids. Thus, *M. tuberculosis* lipids seem to be involved in signal transduction through the main metabolite cAMP (Abdel-Motaal 2006). In fact, small metabolites are often involved in signaling transmission in many bacteria. The relevant PhoP/PhoR two-component system was demonstrated to be related to lipid metabolism in *M. tuberculosis* (Gonzalo Asensio 2006). Although the specific signal sensed by PhoR is still unknown (Jackson 2006), some small molecules (metabolites) might behave as its signaling effectors. In fact, the homologous two-component system (PhoP/PhoQ) is sensed by magnesium in other bacteria (Martin-Orozco 2006).

The association of mycobacterial lipids to *M. tuberculosis* pathogenicity is a matter of renewed interest (Riley 2006). A role in the establishment and progress of the pathology caused by the tubercle bacilli has been classically assigned for years to many of those lipids (Bloom 1994). However, most of the studies were conducted using lipids as isolated molecules, overlooking the interactions with other molecules within the bacterial cell and the environment. In fact, the lipid contents of the bacillus change according to the environmental conditions. Garton *et al.* described an increase of lipophilic inclusions according to the lipid content in the *in vitro* culture medium where bacteria were grown (Garton 2002). Lipid availability is probably not low inside man, the natural host, and *in vivo* bacilli could be lipolytic rather than lipogenic (Wheeler 1994). Trafficking of mycobacterial lipids from bacterial vacuoles to the endosomes of macrophages was demonstrated in *M. tuberculosis* infected macrophages, and mycobacterial lipids were detected even in uninfected cells. These findings indicate that through its own lipids, *M. tuberculosis* exerts a wide influence on its environment that extends beyond truly infected cells (Beatty 2000). Altogether, these data underline the great importance of the

metabolomic analysis for the interpretation of the biology of the tubercle bacillus and its relation with the host.

4.6. Concluding remarks

The availability of the first *M. tuberculosis* genome sequences triggered an overwhelming amount of knowledge on the genetics and the biology of *M. tuberculosis*. The way was opened for comparative and functional genomics. Scientists and medical doctors started to appreciate the potential coding capacity of this extraordinary organism. A broad picture of the *M. tuberculosis* gene content and coding capacity has been revealed. Sequencing and comparison with other genomes have shown the close relations that exist among the members of the *M. tuberculosis* complex, and have allowed the identification of a core mycobacterial genome, a minimal set of genes conserved in the different mycobacterial species. These advances were generated in a little more than seven years by no more than five publicly available genomes. Now, with the advent of new technologies and 21 genome projects in process, the study of mycobacteria and comparative genomics seems not only promising but very exciting. The application of new technologies, such as DNA microarray for the comparison of *M. tuberculosis* wild isolates, is a promising approach towards understanding its natural biology and adaptative evolution in the human population.

As research on the biology of TB expands, new and more accurate information is generated. In the coming years, knowledge about the real coding capacity of the tubercle bacillus will increase exponentially, and genome sequences will feed back from transcriptome and proteome analysis, filling old gaps and opening new ones in the understanding of *M. tuberculosis* biology. Functional genomics has become a key tool in the understanding of the biology of *M. tuberculosis*. By providing information about the pathogenesis of the disease, it is expected to promote the discovery of vaccine candidates and the investigation of novel drug targets. Investigations on complex biological systems can be now envisaged under a metabolomic perspective (Forst 2006). Metabolomics is a newborn methodology in microbiology and is even younger in mycobacteriology, therefore, almost everything remains to be learned in TB concerning that discipline.

It is clear that a long way still remains to be walked to understand how the tubercle bacillus behaves inside the host, its unique known environment. A more comprehensive integration of the knowledge generated by genomics, transcriptomics, proteomics and various molecular tools will surely provide a clearer picture of the amazing pathogen *M. tuberculosis* and the illness that it causes.

References

1. Abdel Motaal A, Tews I, Schultz JE, Linder JU. Fatty acid regulation of adenylyl cyclase *Rv2212* from *Mycobacterium tuberculosis* H37Rv. *FEBS J* 2006; 273: 4219-28.
2. Agarwal N, Raghunand TR, Bishai WR. Regulation of the expression of *whiB1* in *Mycobacterium tuberculosis*: role of cAMP receptor protein. *Microbiology* 2006; 152: 2749-56.
3. Asselineau J, Laneelle G. Mycobacterial lipids: a historical perspective. *Front Biosci* 1998; 3: 164-74.
4. Bagchi G, Chauhan S, Sharma D, Tyagi JS. Transcription and autoregulation of the *Rv3134c-devR-devS* operon of *Mycobacterium tuberculosis*. *Microbiology* 2005; 151: 4045-53.
5. Banaiee N, Jacobs WR Jr, Ernst JD. Regulation of *Mycobacterium tuberculosis whiB3* in the mouse lung and macrophages. *Infect Immun* 2006; 74: 6449-57.
6. Barnard A, Wolfe A, Busby S. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol* 2004; 7: 102-8.
7. Barry CE 3rd. Interpreting cell wall 'virulence factors' of *Mycobacterium tuberculosis*. *Trends Microbiol* 2001; 9: 237-41.
8. Bashyam MD, Hasnain SE. The extracytoplasmic function sigma factors: role in bacterial pathogenesis. *Infect Genet Evol* 2004; 4: 301-8.
9. Beaucher J, Rodrigue S, Jacques P-E, Smith I, Brzezinski R, Gaudreau L. Novel *Mycobacterium tuberculosis* anti- σ factor antagonists control σ^F activity by distinct mechanisms. *Mol Microbiol* 2002; 45: 1527-40.
10. Beatty WL, Rhoades ER, Ullrich HJ, Chatterjee D, Heuser JE, Russell DG. Trafficking and release of mycobacterial lipids from infected macrophages. *Traffic* 2000; 1: 235-47.
11. Behr MA, Wilson MA, Gill WP, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 1999; 284: 1520-3.
12. Betts JC, Dodson P, Quan S, et al. Comparison of the proteome of *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC 1551. *Microbiology* 2000; 146: 3205-16.
13. Betts JC. Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis. *IUBMB Life* 2002; 53: 239-42.
14. Bloom BR (Ed.) *Tuberculosis. Pathogenesis, protection and control.* ASM Press. Washington DC. 1994.
15. Brennan PJ. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 2003; 83: 91-7.
16. Brodin P, Eiglmeier K, Marmiesse M, et al. Bacterial artificial chromosome-based comparative genomic analysis identifies *Mycobacterium microti* as a natural ESAT-6 deletion mutant. *Infect Immun* 2002; 70: 5568-78.
17. Brosch R, Philipp WJ, Stavropoulos E, Colston MJ, Cole ST, Gordon SV. Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infect Immun* 1999; 67: 5768-74.
18. Brosch R, Gordon SV, Pym A, Eiglmeier K, Garnier T, Cole ST. Comparative genomics of the mycobacteria. *Int J Med Microbiol* 2000a; 290: 143-52.
19. Brosch R, Gordon SV, Buchrieser C, Pym AS, Garnier T, Cole ST. Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. *Yeast* 2000b; 17: 111-23.

20. Brosch R, Gordon SV, Eiglmeier K, Garnier T, Cole ST. Comparative genomics of the leprosy and tubercle bacilli. *Res Microbiol* 2000c; 151: 135-42.
21. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* 2001; 9: 452-8.
22. Brosch R, Gordon SV, Marmiesse M, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A* 2002; 99: 3684-9.
23. Camus JC, Pryor MJ, Medigue C, Cole ST. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 2002; 148: 2967-73.
24. Canneva F, Branzoni M, Riccardi G, Provvesi R, Milano A. Rv2358 and FurB: two transcriptional regulators from *Mycobacterium tuberculosis* which respond to zinc. *J Bacteriol* 2005; 187: 5837-40.
25. Capelli G, Volpe E, Grassi M, Liseo B, Colizzi V, Mariani F. Profiling of *Mycobacterium tuberculosis* gene expression during human macrophage infection: upregulation of the alternative sigma factor G, a group of transcriptional regulators, and proteins with unknown function. *Res Microbiol* 2006; 157: 445-55.
26. Chandler M, Mahillon J. Insertion sequences revisited. In: *Mobile DNA II* Craigi NL, Craigie R, Gellert M, Lambowitz AM (Eds.) ASM Press. 2002.
27. Cho SH, Goodlett D, Franzblau S. ICAT-based comparative proteomic analysis of non-replicating persistent *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 2006; 86: 445-60.
28. Clark-Curtiss JE, Haydel SE. Molecular genetics of *Mycobacterium tuberculosis* pathogenesis. *Annu Rev Microbiol* 2003; 57: 517-49.
29. Cole ST, Saint Girons I. Bacterial genomics. *FEMS Microbiol Rev* 1994; 14: 139-60.
30. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998a; 393: 537-44.
31. Cole ST. Comparative mycobacterial genomics. *Curr Opin Microbiol* 1998b; 1: 567-71.
32. Cole ST. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* 1999; 452: 7-10.
33. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. *Nature* 2001; 409: 1007-11.
34. Cole ST. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *Microbiology* 2002a; 148: 2919-28.
35. Cole ST. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl* 2002b; 36: 78s-86s.
36. Cole ST, Eisenack KD, McMurray DN, Jacobs WR. (Eds.) *Tuberculosis and the tubercle bacillus*. ASM Press. Washington DC. 2005.
37. Dainese E, Rodrigue S, Delogu G, et al. Posttranslational regulation of *Mycobacterium tuberculosis* extracytoplasmic-function sigma factor σ^L and roles in virulence and in global regulation of gene expression. *Infect Immun* 2006; 74: 2457-60.
38. Delogu G, Sanguinetti M, Posteraro B, Rocca S, Zanetti S, Fadda G. The hbbA gene of *Mycobacterium tuberculosis* is specifically upregulated in the lungs but not in the spleens of aerogenically infected mice. *Infect Immun* 2006; 74: 3006-11.
39. Dheenadhayalan V, Delogu G, Sanguinetti M, Fadda G, Brennan MJ. Variable expression patterns of *Mycobacterium tuberculosis* PE_PGRS genes: evidence that PE_PGRS16 and PE_PGRS26 are inversely regulated in vivo. *J Bacteriol* 2006; 188: 3721-5.

40. Domenech P, Barry CE 3rd, Cole ST. *Mycobacterium tuberculosis* in the post-genomic age. *Curr Opin Microbiol* 2001; 4: 28-34.
41. Domenech P, Reed MB, Dowd CS, Manca C, Kaplan G, Barry CE 3rd. The role of MmpL8 in sulfatide biogenesis and virulence of *Mycobacterium tuberculosis*. *J Biol Chem* 2004; 279: 21257-65.
42. Draper P. Lipid biochemistry takes a stand against tuberculosis. *Nat Med* 2000; 6: 977-8.
43. Ewann F, Jackson M, Pethe K, et al. Transient requirement of the PrrA-PrrB two-component system for early intracellular multiplication of *Mycobacterium tuberculosis*. *Infect Immun* 2002; 70: 2256-63.
44. Ewann F, Loch C, Supply P. Intracellular autoregulation of the *Mycobacterium tuberculosis* PrrA response regulator. *Microbiology* 2004; 150: 241-6.
45. Fiehn O, Weckwerth W. Deciphering metabolic networks. *Eur J Biochem* 2003; 270: 579-88.
46. Filliol I, Motiwala AS, Cavatore M, et al. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* 2006; 188: 759-72.
47. Finlay BB, Falkow S. Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev* 1997; 61: 136-69.
48. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995; 269: 496-512.
49. Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 2002; 184: 5479-90.
50. Forst CV. Host-pathogen systems biology. *Drug Discov Today* 2006; 11: 220-7.
51. Gagneux S, DeRiemer K, Van T, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2006; 103: 2869-73.
52. Garbe TR, Hibler NS, Deretic V. Isoniazid induces expression of the antigen 85 complex in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 1996; 40: 1754-6.
53. Garnier T, Eiglmeier K, Camus JC, et al. The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A* 2003; 100: 7877-82.
54. Garton NJ, Christensen H, Minnikin DE, Adegbola RA, Barer MR. Intracellular lipophilic inclusions of mycobacteria in vitro and in sputum. *Microbiology* 2002; 148: 2951-8.
55. Geiman DE, Kaushal D, Ko C, et al. Attenuation of late-stage disease in mice infected by the *Mycobacterium tuberculosis* mutant lacking the *sigF* alternate sigma factor and identification of *sigF*-dependent genes by microarrays analysis. *Infect Immun* 2004; 72: 1733-45.
56. Geiman DE, Raghunan TR, Agarwal N, Bishai WR. Differential gene expression in response to exposure to antimycobacterial agents and other stress conditions among seven *Mycobacterium tuberculosis* *whiB*-like genes. *Antimicrob Agents Chemother* 2006; 50: 2836-41.
57. Gey Van Pittius NC, Gamielien J, Hide W, Brown GD, Siezen RJ, et al. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. *Genome Biol* 2001; 2: research0044.1-0044.18.
58. Gonzalo Asensio J, Maia C, Ferrer NL, et al. The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in *Mycobacterium tuberculosis*. *J Biol Chem* 2006; 281: 1313-6.

150 Genomics and Proteomics

59. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* 1999; 32: 643-55.
60. Gupta S, Sinha A, Sarkar D. Transcriptional autoregulation by *Mycobacterium tuberculosis* PhoP involves recognition of novel direct repeat sequences in the regulatory region of the promoter. *FEBS Lett* 2006; 580: 5328-38.
61. Gutacker MM, Smoot JC, Migliaccio CA, et al. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002; 162: 1533-43.
62. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999; 17: 994-9.
63. Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, Chen X. Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. *Mol Cell Proteomics* 2003; 2: 1284-96.
64. Hahn MY, Raman S, Anaya M, Husson RN. The *Mycobacterium tuberculosis* extracytoplasmic-function sigma factor *SigL* regulates polyketide synthases and secreted or membrane proteins and is required for virulence. *J Bacteriol* 2005; 187: 7062-71.
65. Haydel SE, Benjamin Jr WH, Dunlap NE, Clark-Curtiss JE. Expression, autoregulation and DNA binding properties of the *Mycobacterium tuberculosis* TrcR response regulator. *J Bacteriol* 2002; 184: 2192-203.
66. Haydel SE, Clark-Curtiss. Global expression analysis of two-component system regulator genes during *Mycobacterium tuberculosis* growth in human macrophages. *FEMS Microbiol Lett* 2004; 236: 341-7.
67. He H, Zahrt TC. Identification and characterization of a regulatory sequence recognized by *Mycobacterium tuberculosis* persistence regulator MprA. *J Bacteriol* 2005; 187: 202-12.
68. He H, Hovey R, Kane J, Singh V, Zahrt TC. MprA is a stress-responsive two-component system that directly regulates expression of sigma factors SigB and SigE in *Mycobacterium tuberculosis*. *J Bacteriol* 2006; 188: 2134-43.
69. Himpens S, Loch C, Supply P. Molecular characterization of the mycobacterial SenX3-RegX3 two-component system: evidence for autoregulation. *Microbiology* 2000; 146: 3091-8.
70. Hirschfield GR, McNeil M, Brennan PJ. Peptidoglycan-associated polypeptides of *Mycobacterium tuberculosis*. *J Bacteriol* 1990; 172: 1005-13.
71. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci U S A* 2004; 101: 4871-6.
72. Isobe T, Uchida K, Taoka M, Shinkai F, Manabe T, Okuyama T. Automated two-dimensional liquid chromatographic system for mapping proteins in highly complex mixtures. *J Chromatogr* 1991; 588: 115-23.
73. Hughes AL, Friedman R, Murray M. Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. *Emerg Infect Dis* 2002; 8: 1342-6.
74. Jackson M, Stadthagen G, Gicquel B. Long-chain multiple methyl-branched fatty acid-containing lipids of *Mycobacterium tuberculosis*: Biosynthesis, transport, regulation and biological activities. *Tuberculosis (Edinb)* 2007; 87: 78-86.

75. Jain M, Cox JS. Interaction between polyketide synthase and transporter suggests coupled synthesis and export of virulence lipid in *M. tuberculosis*. *PLoS Pathog* 2005; 1: 12-9.
76. Johansen KA, Gill RE, Vasil ML. Biochemical and molecular analysis of phospholipase C and phospholipase D activity in mycobacteria. *Infect Immun* 1996; 64: 3259-66.
77. Jungblut PR, Schaible UE, Mollenkopf HJ, et al. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* 1999; 33: 1103-17.
78. Jungblut PR, Muller EC, Mattow J, Kaufmann SH. Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect Immun* 2001; 69: 5905-7.
79. Kanduma E, McHugh TD, Gillespie SH. Molecular methods for *Mycobacterium tuberculosis* strain typing: a users guide. *J Appl Microbiol* 2003; 94: 781-91.
80. Kaur D, Berg S, Dinadayala P, et al. Biosynthesis of mycobacterial lipoarabinomannan: role of a branching mannosyltransferase. *Proc Natl Acad Sci U S A* 2006; 103: 13664-9.
81. Kazmierczak MJ, Wiedmann M, Boor KJ. Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev* 2005; 69: 527-43.
82. Kempell KE, Ji YE, Estrada IC, Colston MJ, Cox RA. The nucleotide sequence of the promoter, 16S rRNA and spacer region of the ribosomal RNA operon of *Mycobacterium tuberculosis* and comparison with *Mycobacterium leprae* precursor rRNA. *J Gen Microbiol* 1992; 138: 1717-27.
83. Kendall SL, Movahedzadeh F, Rison SCG, Wernisch L, Paris T, Duncan K, Betts JC, Stoker NG. The *Mycobacterium tuberculosis* dosRS two-component system is induced by multiple stresses. *Tuberculosis* 2004; 84: 247-55.
84. KINHAIKAR AG, VARGAS D, LI H, et al. *Mycobacterium tuberculosis* malate synthase is a laminin-binding adhesin. *Mol Microbiol* 2006; 60: 999-1013.
85. Kremer L, Besra GS. A waxy tale, by *Mycobacterium tuberculosis*. Chapter 19, pgs:287-305. In *Tuberculosis and the tubercle bacillus*. Cole ST (Ed.) ASM Press. 2005.
86. Kubica GP, Wayne LG (Eds.) *The Mycobacteria: A sourcebook*. Part A. Dekker md. Microbiology series Vol 15. 1984.
87. Lee BY, Hefta SA, Brennan PJ. Characterization of the major membrane protein of virulent *Mycobacterium tuberculosis*. *Infect Immun* 1992; 60: 2066-74.
88. Li L, Bannantine JP, Zhang Q, et al. The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis. *Proc Natl Acad Sci U S A* 2005; 102: 12344-9.
89. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J Bacteriol* 1996; 178: 1274-82.
90. Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I. Differential expression of 10 sigma factor genes in *Mycobacterium tuberculosis*. *Mol Microbiol* 1999; 31: 715-24.
91. Manganelli R, Voskuil MI, Schoolnik GK, Smith I. The *Mycobacterium tuberculosis* ECF sigma factor σ^E : role in global gene expression and survival in macrophages. *Mol Microbiol* 2001; 41: 423-37.
92. Manganelli R, Voskuil MI, Schoolnik GK, Dubnau E, Gomez M, Smith I. Role of the extracytoplasmic-function σ^H in *Mycobacterium tuberculosis* global gene expression. *Mol Microbiol* 2002; 45: 365-74.
93. Manganelli R, Fattarini L, Tan D, et al. The extra cytoplasmic function sigma factor σ^E is essential for *Mycobacterium tuberculosis* virulence in mice. *Infect Immun* 2004; 72: 3038-41.

152 Genomics and Proteomics

94. Marmiesse M, Brodin P, Buchrieser C, et al. Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. *Microbiology* 2004; 150: 483-96.
95. Marri PR, Bannantine JP, Golding GB. Comparative genomics of metabolic pathways in *Mycobacterium* species: gene duplication, gene decay and lateral gene transfer. *FEMS Microbiol Rev* 2006; 30: 906-25.
96. Marsh IB, Whittington RJ. Deletion of an *mmpL* gene and multiple associated genes from the genome of the S strain of *Mycobacterium avium* subsp. *paratuberculosis* identified by representational difference analysis and in silico analysis. *Mol Cell Probes* 2005; 19: 371-84.
97. Martin-Orozco N, Touret N, Zaharik ML, et al. Visualization of vacuolar acidification-induced transcription of genes of pathogens inside macrophages. *Mol Biol Cell* 2006; 17: 498-510.
98. Mattow J, Jungblut PR, Schaible UE, et al. Identification of proteins from *Mycobacterium tuberculosis* missing in attenuated *Mycobacterium bovis* BCG strains. *Electrophoresis* 2001; 22: 2936-46.
99. Mattow J, Siejak F, Hagens K, et al. Proteins unique to intraphagosomally grown *Mycobacterium tuberculosis*. *Proteomics* 2006; 6: 2485-94.
100. Mawuenyega KG, Forst CV, Dobos KM, et al. *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling. *Mol Biol Cell* 2005; 16: 396-404.
101. Mizrahi V, Andersen SJ. DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? *Mol Microbiol* 1998; 29: 1331-9.
102. Molle V, Palframan WJ, Findlay KC, Buttner MJ. WhiD and WhiB, homologous proteins required for different stages of sporulation in *Streptomyces coelicolor* A3 (2). *J Bacteriol* 2000; 182: 1286-95.
103. Mollenkopf HJ, Jungblut PR, Raupach B, et al. A dynamic two-dimensional polyacrylamide gel electrophoresis database: the mycobacterial proteome via Internet. *Electrophoresis* 1999; 20: 2172-80.
104. Mooney RA, Darst SA, Landick R. Sigma and RNA polymerase: an on-again, off-again relationship? *Molecular Cell* 2005; 20: 335-45.
105. Nagai S, Wiker HG, Harboe M, Kinomoto M. Isolation and partial characterization of major protein antigens in the culture fluid of *Mycobacterium tuberculosis*. *Infect Immun* 1991; 59: 372-82.
106. Newton SM, Smith RJ, Wilkinson KA, et al. A deletion defining a common Asian lineage of *Mycobacterium tuberculosis* associates with immune subversion. *Proc Natl Acad Sci U S A* 2006; 103: 15594-8.
107. Nielsen J, Oliver S. The next wave in metabolome analysis. *Trends Biotechnol* 2005; 23: 544-6.
108. Ochman H, Moran NA. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 2001; 292: 1096-9.
109. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 1975; 250: 4007-21.
110. Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 1998; 16: 373-8.
111. Ortalo-Magne A, Lemassu A, Laneelle MA, et al. Identification of the surface-exposed lipids on the cell envelopes of *Mycobacterium tuberculosis* and other mycobacterial species. *J Bacteriol* 1996; 178: 456-61.

112. Parida BK, Douglas T, Nino C, Dhandayuthapani. Interactions of anti-sigma factor antagonists of *Mycobacterium tuberculosis* in the yeast two-hybrid system. *Tuberculosis (Edinb)* 2005; 85: 347-55.
113. Parish T, Smith DA, Roberts G, Betts J, Stoker NG. The senX3-regX3 two-component regulatory system of *Mycobacterium tuberculosis* is required for virulence. *Microbiology* 2003a; 149: 1423-35.
114. Parish T, Smith DA, Kendall S, Casali N, Bancroft GJ, Stoker NG. Deletion of two-component regulatory systems increases the virulence of *Mycobacterium tuberculosis*. *Infect Immun* 2003b; 71: 1134-40.
115. Park H-D, Guinn KM, Harrell MI, Liao R, Voskuil MI, Tompa M, Schoolnik GK, Sherman DR. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis* *Mol Microbiol* 2003; 48: 833-43.
116. Park SJ, Lee SY, Cho J, et al. Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Appl Microbiol Biotechnol* 2005; 68: 567-79.
117. Patterson SD. Proteomics: the industrialization of protein chemistry. *Curr Opin Biotechnol* 2000; 11: 413-8.
118. Philipp WJ, Poulet S, Eiglmeier K, et al. An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*. *Proc Natl Acad Sci U S A* 1996; 93: 3132-7.
119. Portevin D, De Sousa-D'Auria C, Houssin C, et al. A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc Natl Acad Sci U S A* 2004; 101: 314-9.
120. Rajakumar K, Shafi J, Smith RJ, et al. Use of genome level-informed PCR as a new investigational approach for analysis of outbreak-associated *Mycobacterium tuberculosis* isolates. *J Clin Microbiol* 2004; 42: 1890-6.
121. Raman S, Song T, Puyang X, Bardarov S, Jacons Jr. WR, Husson RN. The alternative sigma factor SigH regulates major components of oxidative and heat stress response in *Mycobacterium tuberculosis*. *J Bacteriol* 2001; 183: 6119-25.
122. Raman S, Hazra R, Dascher CC, Hussin RN. Transcription regulation by the *Mycobacterium tuberculosis* alternative sigma factor SigD and its role in virulence. *J Bacteriol* 2004; 186: 6605-16.
123. Raman S, Puyang X, Cheng T-Y, Young DC, Moody DB, Husson RN. *Mycobacterium tuberculosis* SigM positively regulates Esx secreted proteins and nonribosomal peptide synthetase genes and down regulates virulence-associated surface lipid synthesis. *J Bacteriol* 2006; 188: 8460-8.
124. Ratledge C, Stanford J. (Eds.) *The Biology of the Mycobacteria*. Vol I. Academic Press. London/NewYork. 1982.
125. Reed MB, Domenech P, Manca C, et al. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 2004; 431: 84-7.
126. Rickman L, Scott C, Hunt DM, et al. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the *rpfA* gene coding for a resuscitation promoting factor. *Mol Microbiol* 2005; 56: 1274-86.
127. Riley LW. Of mice, men, and elephants: *Mycobacterium tuberculosis* cell envelope lipids and pathogenesis. *J Clin Invest* 2006; 116: 1475-8.
128. Rodrigue S, Provvedi R, Jacques P-E, Gaudreau L, Manganelli R. The σ factors of *Mycobacterium tuberculosis*. *FEMS Microbiol Rev* 2006; 30: 926-41.

154 Genomics and Proteomics

129. Rodriguez GM, Smith I. Mechanisms of iron regulation in mycobacteria: role in physiology and virulence. *Mol Microbiol* 2003; 1485-94.
130. Rosenkrands I, Weldingh K, Jacobsen S, et al. Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. *Electrophoresis* 2000a; 21: 935-48.
131. Rosenkrands I, King A, Weldingh K, Moniatte M, Moertz E, Andersen P. Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* 2000b; 21: 3740-56.
132. Rosenkrands I, Slayden RA, Crawford J, Aagaard C, Barry CE 3rd, Andersen P. Hypoxic response of *Mycobacterium tuberculosis* studied by metabolic labeling and proteome analysis of cellular and extracellular proteins. *J Bacteriol* 2002; 184: 3485-91.
133. Saïd-Salim B, Mostowy S, Kristof AS, Behr MA. Mutations in *Mycobacterium tuberculosis* Rv0444c, the gene encoding anti-SigK, explain high level expression of MPB70 and MPB83 in *Mycobacterium bovis*. *Mol Microbiol* 2006; 62: 1251-63.
134. Sassetti CM, Boyd DH, Rubin EJ. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci U S A* 2001; 98: 12712-7.
135. Schnappinger D, Ehrh S, Voskuil MI, et al. Transcriptional adaptation of *Mycobacterium tuberculosis* within macrophages: insights into the phagosomal environment. *J Exp Med* 2003; 198: 693-704.
136. Schnappinger D, Schoolnik GK, Ehrh S. Expression profiling of host pathogen interactions: how *Mycobacterium tuberculosis* and the macrophage adapt to one another. *Microbes Infect* 2006; 8: 1132-40.
137. Schmidt F, Donahoe S, Hagens K, et al. Complementary analysis of the *Mycobacterium tuberculosis* proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol Cell Proteomics* 2004; 3: 24-42.
138. Schoolnik GK. Functional and comparative genomics of pathogenic bacteria. *Curr Opin Microbiol* 2002; 5: 20-6.
139. Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrel MI, Schoolnik GK. Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding alpha-crystallin. *Proc Natl Acad Sci USA* 2001; 98: 7534-9.
140. Shires K, Steyn L. The cold-shock stress response in *Mycobacterium smegmatis* induces the expression of a histone-like protein. *Mol Microbiol* 2001; 39: 994-1009.
141. Sinha S, Kosaloi K, Arora S, et al. Immunogenic membrane-associated proteins of *Mycobacterium tuberculosis* revealed by proteomics. *Microbiology* 2005; 151: 2411-9.
142. Simpson RJ, Connolly LM, Eddes JS, Pereira JJ, Moritz RL, Reid GE. Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database. *Electrophoresis* 2000; 21: 1707-32.
143. Sirakova TD, Dubey VS, Kim HJ, Cynamon MH, Kolattukudy PE. The largest open reading frame (*pks12*) in the *Mycobacterium tuberculosis* genome is involved in pathogenesis and dimycocerosyl phthiocerol synthesis. *Infect Immun* 2003; 71: 3794-801.
144. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol* 2001; 53: 680-9.
145. Soliveri JA, Gomez J, Bishai WR, Chater KF. Multiple paralogous genes related to the *Streptomyces coelicolor* developmental regulatory gene *whiB* are present in *Streptomyces* and other actinomycetes. *Microbiology* 2000; 146: 333-43.

146. Sonden A, Kocincova D, Deshayes C, et al. Gap, a mycobacterial specific integral membrane protein, is required for glycolipid transport to the cell surface. *Mol Microbiol* 2005; 58: 426-40.
147. Song T, Dove SL, Lee KH, Husson RN. Rsh, an anti-sigma factor that regulates the activity of the mycobacterial stress response sigma factor SigH. *Mol Microbiol* 2003; 50: 949-59.
148. Sonnenberg MG, Belisle JT. Definition of *Mycobacterium tuberculosis* culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing, and electrospray mass spectrometry. *Infect Immun* 1997; 65: 4515-24.
149. Sreevatsan S, Pan X, Stockbauer KE, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A* 1997; 94: 9869-74.
150. Stanley SA, Raghavan S, Hwang WW, Cox JS. Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc Natl Acad Sci U S A* 2003; 100: 13001-6.
151. Starck J, Kallenius G, Marklund BI, Andersson DI, Akerlund T. Comparative proteome analysis of *Mycobacterium tuberculosis* grown under aerobic and anaerobic conditions. *Microbiology* 2004; 150: 3821-9.
152. Stewart GR, Wernisch L, Stabler R, et al. Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. *Microbiology* 2002; 148: 3129-38.
153. Steyn AJ, Collins DM, Hondalus MK, Jacobs WR Jr, Kawakami RP, Bloom BR. *Mycobacterium tuberculosis* WhiB3 interacts with ProV to affect host survival but is dispensable for in vivo growth. *Proc Natl Acad Sci USA* 2002; 99: 3147-52.
154. Sun R, Converse PJ, Ko Ch, Tyagi S, Morrison NE, Bishai WR. *Mycobacterium tuberculosis* ECF sigma factor sigC is required for lethality in mice and for the conditional expression of a defined gene set. *Mol Microbiol* 2004; 52: 25-38.
155. Sutcliffe IC, Harrington DJ. Lipoproteins of *Mycobacterium tuberculosis*: an abundant and functionally diverse class of cell envelope components. *FEMS Microbiol Rev* 2004; 28: 645-59.
156. Taboada EN, Acedillo RR, Luebbert CC, Findlay WA, Nash JH. A new approach for the analysis of bacterial microarray-based Comparative Genomic Hybridization: insights from an empirical study. *BMC Genomics* 2005; 6: 78.
157. Talaat AM, Lyons R, Howard ST, Johnston SA. The temporal expression profile of *Mycobacterium tuberculosis* infection in mice. *Proc Natl Acad Sci USA* 2004; 101: 4602-7.
158. Tekai F, Gordon SV, Garnier T, Brosch R, Barrell BG, Cole ST. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber Lung Dis* 1999; 79: 329-42.
159. Timm J, Post FA, Bekker L-G, et al. Differential expression of iron-, carbon-, and oxygen-responsive mycobacterial genes in the lungs of chronically infected mice and tuberculosis patients. *Proc Natl Acad Sci USA* 2003; 100: 14321-6.
160. Trivedi OA, Arora P, Vats A, et al. Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. *Mol Cell* 2005; 17: 631-43.
161. Tsolaki AG, Hirsh AE, DeRiemer K, et al. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* 2004; 101: 4865-70.
162. Tsolaki AG, Gagneux S, Pym AS, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. *J Clin Microbiol* 2005; 43: 3185-91.

156 Genomics and Proteomics

163. Tweeddale H, Notley-McRobb L, Ferenci T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J Bacteriol* 1998; 180: 5109-16.
164. Urquhart BL, Cordwell SJ, Humphery-Smith I. Comparison of predicted and observed properties of proteins encoded in the genome of *Mycobacterium tuberculosis* H37Rv. *Biochem Biophys Res Commun* 1998; 253: 70-9.
165. van der Werf MJ, Jellema RH, Hankemeier T. Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *J Ind Microbiol Biotechnol* 2005; 32: 234-52.
166. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993; 31: 406-9.
167. Veyron-Churlet R, Guerrini O, Mourey L, Daffe M, Zerbib D. Protein-protein interactions within the Fatty Acid Synthase-II system of *Mycobacterium tuberculosis* are essential for mycobacterial viability. *Mol Microbiol* 2004; 54: 1161-72.
168. Volpe E, Cappelli G, Grassi M, et al. Gene expression profiling of human macrophages at late time of infection with *Mycobacterium tuberculosis*. *Immunology* 2006; 118: 449-60.
169. Voskuil MI, Schnappinger D, Visconti KC, et al. Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. *J Exp Med* 2003; 198: 705-13.
170. Walters SB, Dubnau E, Kolesnikova I, Laval F, Daffe M, Smith I. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. *Mol Microbiol* 2006; 60: 312-30.
171. Wang QZ, Wu CY, Chen T, Chen X, Zhao XM. Integrating metabolomics into a systems biology framework to exploit metabolic complexity: strategies and applications in microorganisms. *Appl Microbiol Biotechnol* 2006; 70: 151-61.
172. Wayne LG, Hayes LG. An in vitro model for sequential study of shutdown of *Mycobacterium tuberculosis* through two stages of nonreplicating persistence. *Infect Immun* 1996; 64: 2062-9.
173. Wayne LG, Sohaskey CD. Nonreplicating persistence of *Mycobacterium tuberculosis*. *Annu Rev Microbiol* 2001; 55: 139-63.
174. Weckwerth W, Morgenthal K. Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* 2005; 10: 1551-8.
175. Wheeler PR, Ratledge C. Metabolism of *Mycobacterium tuberculosis*. Chapter 23, pgs:353-385. In *Tuberculosis and the tubercle bacillus* Bloom BR (Ed.) ASM Press. Washington DC. 1994.
176. West AH, Stock AM. Histidine kinases and response regulator proteins in two-component signaling systems. *Trends Biochem Sci* 2001; 26: 369-76.
177. Zhang Y, Wallace RJ Jr, Mazurek GH. Genetic differences between BCG substrains. *Tuber Lung Dis* 1995; 76: 43-50.