2

Genomic Variability of *Mycobacterium tuberculosis*

María Mercedes Zambrano, Ginna Hernández-Neuta,
Iván Hernández-Neuta, Andrea Sandoval, Andrés Cubillos-Ruiz,
Alejandro Reyes and Patricia Del Portillo
Corporación CorpoGen, Bogotá D.C.,
Colombia

1. Introduction

Genomic variability provides the basis for adaptation and evolution and constitutes a fascinating aspect of the metabolically and phylogenetically diverse microbial world. Variability in bacteria has been extensively studied both because it enables inferring evolutionary relationships and because it plays an important role in host-pathogen interactions. Microbiologists, who have long struggled with species classification, have gained a more recent appreciation of the level of genetic diversity in microorganisms that has led to new awareness of what may constitute a bacterial "species" (Doolittle & Zhaxybayeva, 2009). In the clinical setting, genomic variability can represent a significant barrier to treatment. Many pathogens can acquire mutations or foreign genetic material through horizontal gene transfer (HGT) in response to the selective pressure imposed by the host immune system and by chemotherapy (Hawkey & Jones, 2009, Sampson, 2011), resulting in strains that are difficult to eradicate in hospitals as well as during long-term infection. Understanding the extent of genomic variability and its effects on disease in the case of pathogens that display genetic homogeneity and low variability, as is the case for the causative agent of tuberculosis, *Mycobacterium tuberculosis*, is particularly fascinating. The success of *M. tuberculosis* is intimately tied to the infectious process and its interaction with the human host, which is believed to have resulted from a long process of co-evolution (Donoghue, 2009, Gutierrez *et al*., 2005). As a result, *M. tuberculosis* is capable of subverting the immune response and persisting as a latent form within an individual and for millennia within the human population.

Despite the availability of chemotherapy and the continued efforts to control the disease, tuberculosis continues to be one of the top ten causes of morbidity and mortality worldwide, with approximately 9 million cases per year, according to the World Health Organization (Lawn & Zumla, 2011). In spite of the growing interest and continued efforts, there are still significant gaps in our knowledge regarding both the pathogen and its interaction with the host that hamper control strategies. The appearance and spread of multi-drug (MDR) as well as extensively drug resistant (XDR) strains of *M. tuberculosis* represent a growing threat worldwide and underscore the importance of effective diagnosis and treatment. Given the

burden to public health and the complexity of the disease, an effective control of tuberculosis must involve diverse approaches and will require a better understanding of the host as well as of the environmental and bacterial factors that govern disease outcome.

M. tuberculosis belongs to the *Mycobacterium tuberculosis* Complex (MTBC), a group of slow-growing mycobacteria that are closely related at the DNA level and share identical 16S rRNA gene sequences but that differ in terms of phenotypes and host preference (Brosch *et al*., 2001, Sreevatsan *et al*., 1997). The MTBC includes the human-adapted strains *M. tuberculosis*, *Mycobacterium africanum* and *Mycobacterium canneti*, being *M. cannetti* the most divergent within the MTBC complex (Gutierrez et al., 2005). The MTBC also includes animal-adapted strains. *M. bovis* has a wider host range and is the main cause of tuberculosis in other animal species. *M. microti* is a pathogen of rodents, *M. pinnipedii* causes disease in sea lions and seals, *M. caprae* is a pathogen of goats and, recently, "*M. mungi*" was isolated from mongoose (Alexander *et al*., 2010, Mostowy & Behr, 2005). The high similarity at the DNA level suggests that this group could have resulted from a bottleneck event that led to the expansion of a successful clone that then gave rise to different host-adapted ecotypes of the same species (Smith *et al*., 2006).

Understanding the differences that underlie the biology and evolution of the MTBC has been the focus of considerable work (Smith *et al*., 2009, Comas & Gagneux, 2009). Members of the MTBC have a highly clonal population structure where recent events of HGT are essentially absent (Supply et al., 2003, Gutierrez et al., 2005). This contrasts with many other microorganisms where horizontally acquired genetic material can play important roles in acquisition of novel virulence determinants and properties such as antibiotic resistance and the capacity to exploit different environmental niches. Recent surveys using MTBC strains that are more representative of global isolates, as well as advances in genome sequence analysis, have indicated, however, that there is more variation than previously anticipated and that this variation can be used to both distinguish isolates as well as to trace phylogenetic lineages (Hershberg *et al*., 2008).

A greater knowledge of the diversity present in *M. tuberculosis* and MTBC strains can also lead to deeper understanding of the biological consequences associated with strain variability. The variation in circulating *M. tuberculosis* isolates has been critical for identification of strains, outbreaks and changes within the population. It has also in some cases been associated with phenotypic properties that are relevant in terms of the disease, such as transmission potential, immunological response and manifestation of the disease (Nicol & Wilkinson, 2008). However, the link between genotypic and phenotypic properties is not necessarily evident given the complexity of the host-pathogen interaction and the effect of environmental factors. In this context, a deeper understanding of the population structure and dynamics of new clonal lineages, with mutations that contribute to a particular lineage's success, can provide great insight regarding the appearance and spread of strain variants relevant to public health and to the control, treatment and eradication of tuberculosis.

This chapter will provide an overview of recent studies regarding genetic variability in *M. tuberculosis*. This will include a brief description of the importance of variability for the study of the evolution of the MTBC. Also, we will address the mechanisms of genomic variation in a pathogen characterized by genetic homogeneity and inappreciable HGT by

illustrating how genomic variability can emerge as a consequence of mutations that result in Single Nucleotide Polymorphisms (SNPs) and Large Sequence Polymorphisms (LSPs), namely insertions and deletions. We will then discuss the importance of variability in disease outcome.

2. Genetic diversity and phylogeny of *M. tuberculosis*

The availability of the complete *M. tuberculosis* genome sequence (Cole *et al.*, 1998) opened new ways to conduct studies and to understand the evolution of the closely related MTBC strains. By using Bacterial Artificial Chromosomes (BAC) libraries it was shown that seven loci were deleted in *M. bovis* with respect to *M. tuberculosis,* reinforcing previous studies indicating that these strains probably originated from a common ancestor (Gordon *et al.*, 1999, Sreevatsan et al., 1997). This was more fully appreciated by comparative genomics studies (Brosch *et al.*, 2002) that also divided the *M. tuberculosis* strains into "ancient" and "modern" based on a deletion known as TbD1 in the modern strains. Several molecular markers have been developed to type strains and infer phylogenetic relationships. Some of these are considered more useful for epidemiological studies, such as transmission, re-infection and/or reactivation, while others are considered more robust phylogenetic markers that can help to decipher the evolution of *M. tuberculosis*. The methods used for epidemiology include restriction fragment length polymorphism (RFLP) of IS*6110* sites (van Embden *et al.*, 1993, van Soolingen *et al.*, 1993), spoligotyping to identify unique spacers within the Clustered Regulatory Short Palindromic Repeats (CRISPR) or Direct Repeat (DR) region (van Embden *et al.*, 2000, Brudey *et al.*, 2006, Kamerbeek *et al.*, 1997), and the identification of Variable Number of Tandem Repeats-Mycobacterial Interspersed Repetitive Units (MIRUs-VNTR) that are strain-specific repeats of short DNA sequences at different positions of the chromosome (Supply et al., 2003). Molecular markers that provide more robust phylogenetic information and have helped to shape the evolutionary scenario of *M. tuberculosis* include LSP, SNPs and Multilocus Sequence Analysis (MLSA) (Filliol *et al.*, 2006, Gagneux *et al.*, 2006, Gutacker *et al.*, 2006, Comas *et al.*, 2009) (Figure 1). Although it has been argued that the use of RFLPs, spoligotyping and VNTR markers is highly prone to convergent evolution and thus to homoplasies (i.e., the same spoligotyping can be observed in strains belonging to different lineages), recent studies show that, at least for the main lineages, this does not seem to be the case (Kato-Maeda *et al.*, 2011). However, more studies are required to clarify this issue.

Based on our current view of its evolutionary history, *M. tuberculosis* can be divided into six phylogeographical lineages, which have been adapted to their local human populations (Figure 1). The use of different molecular makers, such as spoligotyping, LSPs and SNPs, can also classify the global population of MTB into comparable groups. For instance, Lineage 1 (Indo-Oceanic lineage) corresponds to the East African-Indian (EAI) family; Lineage 2 (East Asian Lineage) corresponds to the Beijing family; Lineage 3 or East African-Indian corresponds to the Central Asia (CAS) family; Lineage 4 is the Euro American Lineage that includes the Haarlem, LAM, X, T, S and Tuscany families; Lineage 5 (West African Lineage 1) and Lineage 6 (West African Lineage 2) correspond to AFRI 2 and AFRI 1, respectively, by spoligotyping (Sola *et al.*, 2001, Brudey et al., 2006). Based on the evidence accumulated from these studies, it has been suggested that *M. tuberculosis* evolved as a human pathogen in Africa, which is also the continent where all main *M. tuberculosis*

lineages have been isolated (Hershberg et al., 2008). Moreover, the "ancient" lineages described by Brosh (2002) are present in West Africa and the spread of the "modern" lineages are associated with the human migration out of the African continent (Wirth *et al.*, 2008, Hershberg et al., 2008).

Phylogenetic studies have also shown that clinical strains of *M. tuberculosis* are more genetically variable than originally expected (Hirsh *et al.*, 2004, Gagneux et al., 2006, Hershberg et al., 2008). Moreover, genetic variability can be translated into phenotypic differences, such as transmission capacity, virulence and pathogenicity that can have epidemiological consequences and affect the outcome of the disease. Although there are few studies showing a clear association between lineage and transmission capacity it is now clear that Lineage 2 (Beijing family) *M. tuberculosis* has spread globally more than any other lineage (Parwati *et al.*, 2010). The use of spoligotyping to type paraffin-embedded strains obtained from tuberculosis patients in different time periods has also shown an increase of this genotype over time in Africa. Its isolation from children, which is a measure of recent transmission, increased from 13% in 2000 to 33% in 2003 in South Africa (Cowley *et al.*, 2008). The capacity of the Beijing genotype to spread more than other lineages is not completely understood but it has been suggested that factors contributing to its expansion could involve the selective pressure imposed by BCG vaccination and drug treatment (Parwati et al., 2010).

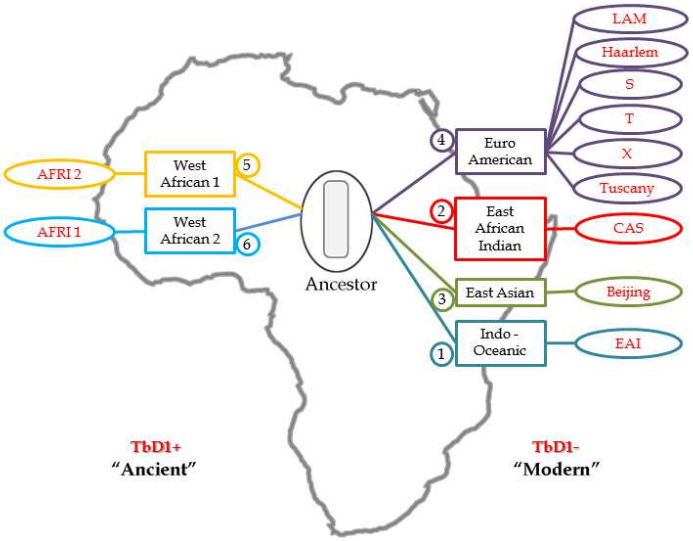


Fig. 1. Schematic representation of the phylogeography of *M. tuberculosis*. squares indicate the 6 main Lineages and circles are representative of the spoligotype families

The successful transmission of particular strains is not limited to the Beijing genotype. In a recent study, where guinea pigs were exposed to air from a HIV-tuberculosis ward, one non-Beijing strain was shown to be responsible for most of the secondary infections observed (Escombe *et al.*, 2008). In addition to transmission capacity, it is also currently accepted that genetically different *M. tuberculosis* strains produce markedly different

immuno-pathological events and affect disease manifestation. For example, in a study conducted in Vietnamese patients, a clear association between the Euro American Lineages of *M. tuberculosis* and pulmonary rather than meningeal tuberculosis was observed, suggesting these strains are less capable of extra-pulmonary dissemination than other strains in the study population (Caws *et al.*, 2008). In a study using a cohort of patients and household contacts in Gambia, both *M. africanum* and *M. tuberculosis* were equally transmitted to the household contacts but *M. tuberculosis* Beijing strains were most likely to progress to disease (de Jong *et al.*, 2008). Another source of evidence came from a recent study associating Lineages 1, 5 and 6, with a higher pro-inflammatory cytokine response when compared with the modern Lineages 2, 3 and 4 (Portevin *et al.*, 2011).

3. SNPs in *M. tuberculosis*

Genetic diversity within bacterial species is usually generated by mutations and by the exchange of genetic material. The process of HGT is thought to be an important driver of bacterial evolution in both pathogenic and non-pathogenic bacteria (Becq *et al.*, 2007). Horizontally transferred genes can be acquired in clusters known as genomic islands or pathogenicity islands that can be identified by characteristics that distinguish them from the host genome, such as GC content, flanking nucleotide repeats and insertion elements. In the case of *M. tuberculosis*, there is evidence of ancient gene transfer events that could have taken place in a progenitor tubercle bacilli pool before the clonal expansion that gave rise to the MTBC (Gutierrez et al., 2005). One of these events involved the Rv0986-8 virulence operon (Rosas-Magallanes *et al.*, 2006) that could have originated from genetic exchange between an environmental bacillus ancestor and other bacterial species (Nicol & Wilkinson, 2008). In the absence of recent events of HGT, modern *M. tuberculosis* lineages evolve essentially by mutations that alter its genome, resulting in SNPs and LSPs, such as deletions and insertions, the latter mainly mediated by transposition of the IS*6110* insertion element.

Although allelic variation in MTBC organisms is quite restricted when compared with other pathogenic bacteria (Sreevatsan et al., 1997), there is a growing recognition that there is substantial genetic diversity among isolates. At the level of SNPs changes can be either synonymous (sSNP) or non-synonymous (nsSNP) and this diversity has been undeniably useful for typing and defining evolutionary relationships among strains. SNPs provide many advantages for the analysis of phylogenetic relationships among microorganisms, especially among closely related clonal organisms such as the MTBC. Initial descriptions of the *M. tuberculosis* population structure involved analysis of SNPs in the *katG* and *gyrA* genes and defined three major genetic groups (Sreevatsan et al., 1997). Later surveys have extended this strategy to include more than 100 sSNPs identified in 112 *M. tuberculosis* isolates (Gutacker et al., 2002). In more recent work using 159 sSNPs identified by whole-genome comparison of sequenced strains, it was possible to classify 212 isolates into 56 haplotypes that grouped strains into six *M. tuberculosis* SNP Cluster Groups (SCG) and one SCG that grouped all the *M. bovis* strains (Filliol et al., 2006). A re-evaluation of the SNP phylogeny was obtained by using *de novo* sequencing of 89 randomly distributed genes in 108 global strains (Comas et al., 2009). This study suggested that initial classification could be done using a subset of discriminatory SNPs and then, if further molecular characterization were needed, a MIRU-VNTR typing technique could be applied to differentiate individual strains. However, the choice of discriminatory SNPs is not an easy

task. For example, SNPs comparison in 32 fully sequenced strains that caused an outbreak in a community in Canada, allowed the identification of two co-circulating "lineages" with the same MIRU-VNTR profile (Gardy *et al*., 2011), which would not have been evident if only the discriminatory SNPs used previously had been included. The study allowed tracking the transmission and demonstrated the power of coupling comparative genomics with social epidemiological studies.

Genetic variation at the SNP level can also have profound implications in strain fitness and disease outcome. One such case applies to the Esx protein family that has been implicated in host-pathogen interactions. To survey genetic diversity in the Esx family, and its potential for antigenic variation, all *esx* genes were sequenced from 108 clinical isolates of *M. tuberculosis* belonging to different clades. The SNP distribution affecting Esx proteins indicated high genetic variability and a total of 109 unique SNPs, 59 of which were non-synonymous. Some of the resultant amino acid substitutions affected known Esx epitopes likely to result in immune variation, thus revealing a dynamic *esx* gene family (Vasilyeva *et al*., 2009).

Another important area of research focuses on variability associated with specific phenotypes of clinical importance, such as antibiotic resistance. In *M. tuberculosis,* resistance to antibiotics results essentially from mutations, such as SNPs, that can be acquired during treatment and can spread within the population. The mutations conferring antibiotic resistance can have a variable effect on strain fitness and bacteria can develop compensatory mechanisms to recover fitness capacity (Borrell & Gagneux, 2011). Isoniazid (INH) resistance in *M. tuberculosis* is associated with mutations in the genes *katG*, *inhA* and *ahpC*. Most identified mutations map to *katG*, which encodes the catalase-peroxidase required to activate INH (Ramaswamy & Musser, 1998) and to protect *M. tuberculosis* from the oxidative free radicals in the macrophage. Thus *M. tuberculosis* INH resistant strains are less virulent (Pym et al, 2002). However, the katG^{S315T} mutation, the most common mutation for INH resistance (Sandgren *et al*., 2009), results in reduced INH activation while maintaining KatG activity and virulence in mice (Pym *et al*., 2002) suggesting compensatory evolution as has been suggested in other bacteria (Maisnier-Patin & Andersson, 2004). If compensatory evolution occurs in MDR and XDR strains it will have deep impacts in the control of tuberculosis (Borrell & Gagneux, 2011), an area that must be further investigated.

The identification of SNPs associated with resistance has also indicated the existence of multiple gene determinants for resistance, not all of which have been fully identified. Streptomycin (Sm) resistance, for example, is associated in the majority of cases with mutations in *rpsL* and *rrs* (Sreevatsan et al., 1997). However, 27% of Sm-resistant strains lack mutations in these genes. There is evidence that in some cases mutations in *gidB,* a gene coding for a 7-methylguanosine (m7G) methyltransferase specific for 16S rDNA, are associated with low level of Sm resistance (Donoghue, 2011). However, some susceptible strains also contain such mutations, thus requiring sequence analysis of more *M. tuberculosis* clinical isolates to better understand the role of *gid*B gene mutations in Sm resistance.

A longstanding question in tuberculosis has been the precise mechanisms by which mycobacteria can acquire resistant mutations, especially during latent infections. The mutation rates that confer antibiotic resistance have been determined *in vitro*, yet the slow growth and different metabolic states of *M. tuberculosis* during infection make it difficult to

assess the *in vivo* rates. This was achieved in a recent report, however, using whole genome sequencing and identification of SNPs generated during different disease states in macaque monkeys (Ford *et al*., 2011). Similar mutation rates were observed during latency and during active disease, and these were also consistent with *in vitro* rates. Based on these results and on the types of SNPs observed, it was suggested that *M. tuberculosis* can acquire mutations during latency and that these mutations are the result of oxidative DNA damage rather than errors in replication. This could be explained by increased oxidative damage during latency, as a result of the immune response, or by diminished DNA repair in metabolically quiescent bacilli (Ford *et al*., 2011).

The identification of SNPs in *M. tuberculosis* has provided important insight regarding genetic variability and evolution of this pathogen. SNPs can also impact strain fitness, as is evident by the acquisition of antibiotic resistance markers. It remains to be seen if many of the identified SNPs have an effect on the biology of *M. tuberculosis* and the host-pathogen interaction. Whole genome sequencing will undoubtedly allow more extensive SNP identification and analysis on a genome-wide scale. As more sequence data becomes available, comparative genomics studies may help to identify markers that can contribute to our understanding of the molecular mechanisms underlying phenotypes such as drug resistance and persistence.

4. Large Sequence Polymorphisms

LSPs can include both insertions and deletions (indels) and have been identified as one of the main sources of genomic variability in *M. tuberculosis*. The effect of LSPs can vary and may provide insights into the biology of *M. tuberculosis* strains. Large deletions have been shown to group closely related strains and have been associated with phylogeographical lineages, suggesting that a deletion event is specific to a particular lineage (Tsolaki *et al*., 2004). Some LSPs occur rarely in the population and could have arisen from random genomic events and then become associated with a particular phylogenetic lineage (Alland *et al*., 2007). In contrast, other LSPs are present in multiple strains from different lineages, as a result of selective pressure, and are not necessarily associated with particular groups (Alland et al., 2007).

Soon after completing the genome sequence of the laboratory strain H37Rv (Cole et al., 1998), the clinical isolate, strain CDC1551 that had caused an outbreak in the United States, was sequenced (Fleischmann *et al*., 2002). A whole genome comparative study carried out using these two genomic sequences identified 1,075 SNPs and 86 LSPs larger than 10 bp. The analysis of these LSPs using a panel of 169 clinical isolates, showed that clinical strains were genetically more variable than expected from a clonal bacterial population (Fleischmann et al., 2002).

The continued advances in methods for high-throughput nucleic acid sequencing now allow more rapid generation of sequence data and thus access to information from a growing number of sequenced clinical *M. tuberculosis* genomes. Up to now, there are more than 200 on-going sequencing projects of *M. tuberculosis* strains with different characteristics, such as strains with epidemic potential and strains characterized by multidrug resistance, as well as isolates obtained before and after a passage through an immunocompetent animal model, among others (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). This information,

together with new bioinformatics algorithms will be an invaluable resource for probing these genomes in an effort to further understand the evolution, epidemiology, emergence of drug resistance and phenotypic variability associated with tuberculosis disease.

4.1 Comparative genomics to assess variability

With the growing number of sequenced strains becoming available, the comparison of complete genomes from different clinical isolates of *M. tuberculosis* becomes an attractive and powerful tool to explore genotypic similarities and differences. This approach can also provide important insights regarding the genotype-phenotype relationship in *M. tuberculosis,* and can therefore contribute to the development of control measures for tuberculosis. In this respect, we carried out whole genome comparisons of six fully sequenced *M. tuberculosis* strains, four clinical isolates and two laboratory strains which showed high synteny, as expected, and no large rearrangements (Cubillos-Ruiz *et al.*, 2008), except for a large inversion seen in the KZN strain that could be due to sequencing errors and incomplete data (Figure 2). Most of the 1,428 LSPs identified were indels involving 120 genes that affected primarily 1) mobile genetic elements such as insertion sequences and prophages, 2) non-coding regions, and 3) the PE/PPE family of genes. The LSPs identified in this work differed among strains, were distributed along the entire genome and were used to identify strain-specific insertions and deletions. When fitted to an exponential decay function these data indicated a tendency towards accumulation of more deletions than insertions, consistent with the notion of genome decay in *M. tuberculosis* (Cubillos-Ruiz et al., 2008). One other remarkable finding was that laboratory strains contained less strain specific polymorphisms than the clinical isolates, suggesting that the selective pressure imposed by the human immune system could be driving variability. The existence of strain-specific polymorphisms also opened the possibility that specific indels could be associated with particular lineages and thus could also be used as markers for strain typing and surveillance.

Taking into account the growing evidence of the phylogeographical origin of *M. tuberculosis* (Gagneux et al., 2006, Wirth et al., 2008), we speculated that the strain-specific polymorphisms could be common to strains of a particular lineage rather than being an exclusive property of one particular isolate. To test this hypothesis, we evaluated strain-specific indels and previously identified SNPs associated with strains of the Haarlem lineage using a large panel of well-characterized *M. tuberculosis* strains (Olano et al, 2008, Cubillos-Ruiz et al., 2008). Six large deletions, two specific IS*6110* insertions and two SNPs were significantly associated with the Haarlem family and thus proposed as genomic signatures of this lineage (Cubillos-Ruiz *et al.*, 2010). These results were completely congruent with spoligotyping and with RFLP data, as well as with the new assignation of a URAL family instead of the Haarlem 4 sublineage (Abadia *et al.*, 2010). One particularly interesting result was the identification of deletions that affected previously proposed drug targets. These include the gene Rv1354c, which encodes a diguanylate cyclase (DGC) enzyme involved in regulating the levels of c-di-GMP, a bacterial second messenger implicated in survival and adaptation to different environmental conditions (Gupta *et al.*, 2010), and gene Rv2275 that codes for a cytochrome P450, Cyp121 (McLean & Munro, 2008). Both of these genes were deleted in the Haarlem strains analyzed, indicating that they would not be adequate targets for antimicrobials. Although this particular study was limited to Haarlem strains, it raises the possibility that other lineage-specific genomic differences might impact treatment and

control. More studies will be needed to address this issue and to verify the presence of specific gene targets.

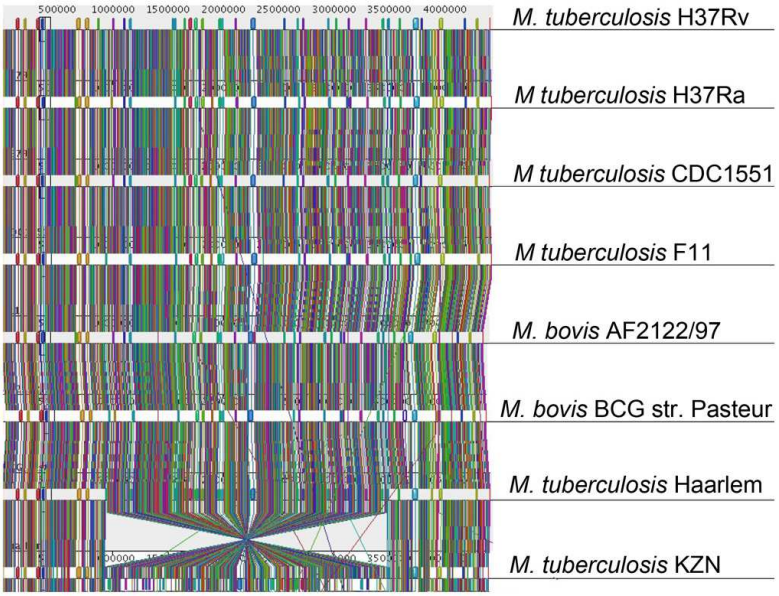


Fig. 2. Whole genome alignment generated with the MAUVE software, showing synteny and, in the case of strain KZN, the presence of a genomic inversion.

5. Insertions

The presence of insertion elements in different bacteria has been well appreciated for some time, especially because of the impact they can have on the host genome (Siguier *et al.*, 2006). Insertion elements can not only re-shape the genome but can also cause mutations and alter gene expression. In the case of pathogens such as *M. tuberculosis* the presence of insertion elements can generate genotypic variation and mediate changes that can affect gene function. This variability can therefore alter properties such as strain fitness and transmissibility and even play a role in the evolution of *M. tuberculosis*.

The insertion elements present in the *M. tuberculosis* genome were described in detail upon completion of the *M. tuberculosis* H37Rv whole genome sequence (Gordon et al., 1999). *M. tuberculosis* harbors four main insertion elements, IS*6110*, IS*1081*, IS*1547* and the IS-like element, all of them present in multiple copies. The best studied of these is the 1.36Kb IS*6110*, originally described by Thierry et al. in 1990 (Thierry *et al.*, 1990), which belongs to the group of IS*3* elements and is characterized by having two partially overlapping open reading frames that allow production of a transposase by translational frameshifting (McEvoy *et al.*, 2007). It also has 28 bp imperfect terminal inverted repeats and generates 3- to 4 bp direct repeats upon insertion (McAdam *et al.*, 1990, Thierry *et al.*, 1990, Mendiola *et al.*, 1992). The IS*6110* element is present exclusively in strains of the MTBC that can harbor

from zero to 25 copies per genome (Brosch *et al*., 2000). For this reason and due to its high degree of copy number and insertion site variation, IS*6110*- RFLP has been widely used for epidemiological purposes and is considered the "gold standard" to study the transmission dynamics of *M. tuberculosi*s (van Embden *et al*., 1993, Small *et al*., 1994, Safi *et al*., 1997). The discriminatory power of the IS*6110*-RFLP method depends on there being sufficient variation and copy number to differentiate between unlinked isolates while allowing identification of specimens that are related (Wall *et al*., 1999). Thus IS*6110*-RFLP is used to distinguish between epidemiological events but its use as marker for strain evolution is still under debate (McEvoy et al., 2007).

The consequence of IS*6110* transposition can differ depending on the position of integration, with phenotypic outcomes ranging from lethality to its bacterial host due to gene inactivation to possible benefits. There are four mutational events that can be generated by IS*6110* transposition: 1) Integration in intragenic regions; 2) Alteration of IS*6110* flanking regions; 3) Recombination/gene deletion; 4) Alteration of IS*6110* promoter activity (McEvoy et al., 2007). Intragenic insertions interrupt open reading frames and can inactivate genes; this is the most frequently described event in certain clinical isolates. For these interruptions to be observed they must occur in genes that are dispensable for survival of the bacterium or are redundant in function. These insertion events can also alter immune recognition or virulence properties, as has been suggested for insertions in members of the PPE gene family or in the phospholipase C gene region (McEvoy *et al*., 2009a, Vera-Cabrera *et al*., 2001). It has also been observed that the regions flanking an IS*6110* insertion contain additional mutations, suggesting that this element can have a disruptive effect on the DNA region of insertion that results in mutations (Warren *et al*., 2000).

Insertion elements can mediate deletions, as has also been shown for *M. tuberculosis,* where gene deletion can occur by homologous recombination between two flanking copies of IS*6110*. For example, deletion of the *plcA* gene in clinical *M. tuberculosis* strains displays a decrease capacity to cause pulmonary cavitation, clearly showing the phenotypic effects of transposition in a clinical setting (Kato-Maeda *et al*., 2001a). Not all the insertions described in clinical isolates have deleterious or silent effects on the mycobacterial cell; some studies have reported that IS*6110* can up-regulate expression of downstream genes from an outward-directed promoter at its 3′ end, conferring selective advantages. In particular, an insertion found within the *phoP* promoter region in an MDR *M.bovis* strain, which had produced outbreaks in the United States and Spain (Rivero *et al*., 2001), was shown to increase *phoP* expression 10-fold in *M. smegmatis* and was proposed to be responsible for the high transmissibility levels of the original *M. bovis* isolate (Soto *et al*., 2004).

Given the high variability of IS*6110* elements in the genomes of MTBC strains and the possible consequences of insertion on strain phenotype, there has been an interest in identifying the precise insertion locations in *M. tuberculosis* clinical isolates. Different methodologies developed, based on PCR, sequencing and cloning, have suggested that the IS*6110* element inserts preferentially into non-coding regions (Otal *et al*., 2008, Warren et al., 2000, Thorne *et al*., 2011, Kim *et al*., 2010, McEvoy *et al*., 2009b, Wall et al., 1999). This can be explained by the fact that insertions in functional genes essential for strain growth, maintenance and pathogen integrity would be harmful to the cell and thus, not maintained in the population. Preferential insertion loci or hotspots have also been identified, some of

which include the phospholipase C region (Vera-Cabrera et al., 2001), members of the PPE gene family (McEvoy et al., 2009a), the *dnaA - dnaN* intergenic region (Turcios *et al.*, 2009), the RD724 gene (Kim *et al.*, 2010) and insertion into the IS*1547* element (Fang *et al.*, 1999). PPE genes are considered to be important antigens during the host-pathogen interaction and have been proposed to play a role in evasion of the immune response (Sampson, 2011). Thus variability in the PPE genes generated through IS*6110* transposition could help to evade the immune system during infection and confer advantage to strains. In contrast to these hotspot regions, some loci where integration is rare or not observed have also been identified and these represent sites where *in vivo* transposition events can be harmful to strain fitness and growth (Yesilkaya *et al.*, 2005).

We recently developed a novel high-throughput method using next-generation sequencing to identify the flanking regions of the IS*6110* insertion element in over 500 *M. tuberculosis* isolates mainly from Latin-America and Europe. In this study we identified previously reported hotspot regions of insertion as well as novel sites (Table 1) (Reyes et al., submitted).

Locus	Gene ID	Description	# strains	# independent sites	Hotspots
MT3426:MT3427 (RvD5)	MT3426: moaA-3	Probable conserved membrane protein	195	1	A
Rv0403c	mmpS1		225	6	A*, H
Rv0835:Rv0836c	lpqQ: Rv0836c		227	5	A
Rv1754c	-	Conserved hypothetical protein	394	4	A*
Rv2336	-	Hypothetical protein	188	2	A*
Rv2814c-Rv2815c	-	IS6110, transposase	485	2	A*
Rv3113	-	Possible phosphatase	218	2	A

Table 1. Hotspot insertion sites in *M. tuberculosis* identified by high-throughput sequencing (Reyes et al., submitted). Hotspots (H) or Ancestral (A) insertions for a given lineage; A* indicates an ancestral insertion in a locus in that more than one lineage.

The copy number of IS*6110* elements in the genomes of circulating *M. tuberculosis* strains can be highly variable and is ultimately limited by the deleterious effects of IS*6110* transposition (McEvoy et al., 2007). Although most *M. tuberculosis* isolates have multiple copies of the IS*6110* element, the presence of a copy in the DR region of the MTBC strains suggests that this could be an ancestral insertion site. It has also been observed that some successful *M. tuberculosis* strains tend to have a high copy number of the IS*6110* element and that this might correlate with phenotypic properties (Alonso *et al.*, 2011). In a recent report, a Beijing family strain considered to have a high transmissibility rate was found to have 19 copies of the IS*6110* element, four of which were shown to up-regulate downstream gene expression. One of these was in the gene Rv2179c, which is normally expressed inside macrophages, suggesting that this gene could influence the infectious process and that the strain's high degree of transmissibility could be due to the up-regulation caused by the IS*6110* insertion (Alonso *et al.*, 2011). However, some clinical strains and MTBC members with a low number

of copies of the IS*6110* are also epidemiologically successful. In general, though, there is still insufficient information regarding the factors that influence the frequency of transposition, such as the genomic context of the insertion element within a particular strain background. The variation in the number of IS*6110* elements among *M. tuberculosis* isolates also raises the possibility that copy number is the result of the evolution of particular lineages as strains cope with IS*6110* transposition and its resulting genetic variability, and in some cases even selecting for phenotypically favorable events, while keeping genome integrity and avoiding deleterious effects.

6. Implication of variability on disease control

From the pool of individuals that come in contact with and are infected with *M. tuberculosis*, only about 10% will develop disease. The manifestation of the disease in these individuals, however, can vary greatly from a self-limited infection in the lungs to extra-pulmonary and disseminated cases (Nicol & Wilkinson, 2008). The outcome of infection must therefore be influenced both by host factors that may predispose to infection, and to genetic variation in the tubercle bacillus itself. Several host factors have been associated with risk for disease, such as malnutrition, vitamin D deficiency, NRAMP1 polymorphisms, diabetes and co-infection with HIV (Malik & Godfrey-Faussett, 2005). A recent study analysing the impact of pathogen variability in recombinant congenic mice indicated that host control of the infection varied depending on the infecting strain and the stage of infection. The dynamic response to disease suggests that in addition to host genetic determinants, the pathogen background also influences the outcome of infection (Di Pietrantonio *et al*., 2010). Studies with both laboratory and clinical strains have also suggested a correlation between strain genotype and the infectious process. This correlation, however, has been difficult to resolve in great part due to the difficulty associated with working with this slow-growing pathogen and to problems associated with extrapolation from animal models (Nicol & Wilkinson, 2008). The integration of genomics and epidemiological data has been able to link some cases of genetic variability with strain phenotypic characteristics. By analysing deletions in clinical isolates it was suggested, for example, that strains causing cavitary disease had fewer deletions, indicating that the accumulation of mutations affected pathogenesis (Kato-Maeda *et al*., 2001b). Mutations that altered the PE_PGRS33 protein, which may be involved in cell-cell interactions and antigenic variation, have also been connected with clustering and pathogenesis and thus with clinical and epidemiological characteristics of *M. tuberculosis* isolates (Talarico *et al*., 2007). Similarly, an analysis of the genetic variation at the *plcD* locus indicated that variability in this region was possibly associated with pathogenesis and disease manifestation (Yang *et al*., 2005). Studies involving strains that cause pulmonary and extra-pulmonary infections have also indicated that extra-respiratory strains were more efficient at infecting macrophages and could also have higher infectivity *in vivo* (Garcia de Viedma et al., 2005).

The growing consensus that the main MTBC lineages are associated with geographic origin suggests co-evolution of lineages with their hosts and thus adaptation that must involve events of strain variation. More recent evidence of the restricted geographical niche of certain lineages came from an Ibero-America MDR *M. tuberculosis* survey showing that circulation of Latin American MDR strains was restricted to particular areas and also that transnational transmission was scarce (Ritacco *et al*., 2011). The Beijing lineage, one of the most extensively studied families, has been responsible for several epidemic outbreaks and

in some cases has been associated with multidrug resistance (Hanekom *et al*., 2011). Its capacity to spread within a population is evident from epidemiological studies and emphasizes the possibility that certain strain properties could contribute to this lineage's expansion in the population (Nicol & Wilkinson, 2008). The increased virulence of these isolates was associated with the production of a phenolic glycolipid (PGL) that affects the host immune response, and which is absent in many other *M. tuberculosis* families (Ordway *et al*., 2007, Hanekom et al., 2011). More recent work suggests that although PGL can contribute to *M. tuberculosis* virulence, it probably requires additional bacterial factors (Sinsimer *et al*., 2008). Other examples stem from studies of strains that have caused outbreaks, such as strains CDC1551 and HN878, the latter also a member of the Beijing family. In these and other studied cases, it appears that some of the effects observed have to do with the capacity of these strains to induce variable inflammatory responses (Coscolla & Gagneux, 2010). Despite these studies, many of the clinical outcomes associated with strain variability still need to be further examined, particularly in other *M. tuberculosis* lineages before precise genotypic variability can be associated with phenotypic differences.

The emergence and spread of drug-resistant strains is particularly disturbing and provides additional examples where strain variability can have a profound effect on disease outcome. One particularly alarming case was the epidemic caused by an XDR strain in the KwaZulu-Natal region of South Africa that resulted in high mortality, causing the death of 52 of the 53 patients co-infected with HIV in the course of 16 days (Gandhi *et al*., 2006). To understand more about the dynamics of appearance and dispersion of this highly virulent KZN strain, whole genome sequence analysis was carried out for XDR, MDR and drug sensitive KZN strains. The results indicated that the outbreak was most probably due to clonal expansion of a single strain and that a particular strain genetic background did not necessarily contribute to acquisition of antibiotic resistance (Ioerger *et al*., 2009). Further work will be needed to better understand this strain's virulence and transmissibility in the community.

Part of the success of *M. tuberculosis* as a human pathogen is due to its capacity to be efficiently transmitted between hosts and to persist for long periods of time despite the host's immune response. A recent study involving whole genome sequencing of 21 strains from the six main *M. tuberculosis* lineages indicated that human T cell epitopes had very little sequence variation and were highly conserved relative to the rest of the genome. It was suggested that these antigens, contrary to expectations, might be under purifying selection and be benefitting from host immune recognition (Comas *et al*., 2010). This differs from the classical view of immune evasion due to the selective pressure imposed by the immune response and may indicate that new approaches should be considered for vaccine development and control of *M. tuberculosis*.

The genetic variability evident in strains of the MTBC bears relevance to control of tuberculosis since treatment must work against all circulating strains. Rapid and accessible diagnostics for both *M. tuberculosis* and drug resistant isolates are still required, as is the availability of a vaccine that can be universally effective, given the variable efficacy of the currently used BCG vaccine. There are now more that 10 vaccines under phase I trial and the hope is that in the near future at least one of these will prove to be safe and protective by containing *M. tuberculosis* and preventing reactivation. However, future strategies will need to address the need to prevent or eradicate latent infections, especially in view of additional factors affecting disease and the host immune response, such as co-infection with HIV

(Kaufmann, 2010). New and alternative drugs are also required to shorten the current duration of chemotherapy, to act against persistent bacilli and to counteract the spread of drug-resistant strains that frustrate global eradication programs. Due to renewed efforts in recent years, several novel drugs have been identified and are under clinical evaluation or being developed, many of which involve novel targets and mechanisms (Coxon & Dover, 2011). The discovery of novel drugs has involved different approaches that include the use of genomics to identify targets, whole-cell screening and re-engineering of known chemical molecules (Koul *et al*., 2011). Given the observed strain variability it is nonetheless possible that some of these drugs might vary in efficiency in different strain backgrounds, as was made evident for DGC and Cyp121 in the Haarlem lineage (Cubillos-Ruiz et al., 2010). Thus the heterogeneity among different strains and lineages, as well as of the host-pathogen interaction, must be taken into account when developing novel diagnostics and therapeutic strategies. Extensive analysis of circulating *M. tuberculosis* populations will be required to address the efficacy of treatment and vaccination in different genetic backgrounds. The advent of novel massive sequencing techniques to generate genomic data for multiple strains will undoubtedly allow examination of whole genomes and make such analyses more feasible (Lin & Ottenhoff, 2008).

7. Concluding remarks

Over the last decades there has been a substantial increase in our understanding of the molecular bases of *M. tuberculosis* biology and its interaction with the host. However, the clinical and epidemiology consequences of *M. tuberculosis* infection are still poorly understood. Despite the restricted variability and clonality of the MTBC population, various studies make evident that circulating strains vary in terms of their genomic makeup and differ with respect to virulence and immunogenicity properties. The differences observed in the interactions between pathogen and host and in disease manifestation indicate that variation must play a role in disease and in clinical outcome, even though the extent of the impact of this strain diversity is still unclear (Coscolla & Gagneux, 2010). Thus, the precise role of bacterial factors and the importance of strain diversity in pathogenicity and tuberculosis disease remain elusive, partly due to the complex interplay between host and pathogen that is compounded by additional environmental factors. The strain-to-strain variation also has important consequences for the development of efficient control strategies. The development of new diagnostics tools, drugs and vaccines must somehow incorporate analysis of the differences that characterize host responses and strains, highlighting the importance of continued studies regarding the genetic makeup of circulating strains. The use of modern genetic and molecular tools, including the availability of massive sequencing techniques, can contribute significantly to our understanding of *M. tuberculosis* variability and its possible association with biological properties. Only by realizing the need to incorporate this added level of complexity to the study of tuberculosis, will we be able to tackle the intricacies of this disease and achieve an adequate level of control on a global scale.

8. Acknowledgment

Support was obtained from the CCITB, Colciencias and the StopLATENT-TB network (Collaborative Project) supported by the EC under the Health Cooperation

Work Programme of the 7th Framework Programme (G.A. no. 200999) (http://cordis.europa.eu/fp7/dc/index.cfm).

9. References

Abadia, E., J. Zhang, T. dos Vultos, V. Ritacco, K. Kremer, E. Aktas, T. Matsumoto, G. Refregier, D. van Soolingen, B. Gicquel & C. Sola, (2010) Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol* 10: 1066-1074.

Alexander, K. A., P. N. Laver, A. L. Michel, M. Williams, P. D. van Helden, R. M. Warren & N. C. Gey van Pittius, (2010) Novel *Mycobacterium tuberculosis* complex pathogen, M. mungi. *Emerg Infect Dis* 16: 1296-1299.

Alonso, H., J. I. Aguilo, S. Samper, J. A. Caminero, M. I. Campos-Herrero, B. Gicquel, R. Brosch, C. Martin & I. Otal, (2011) Deciphering the role of IS6110 in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. *Tuberculosis (Edinb)* 91: 117-126.

Alland, D., D. W. Lacher, M. H. Hazbon, A. S. Motiwala, W. Qi, R. D. Fleischmann & T. S. Whittam, (2007) Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. *J Clin Microbiol* 45: 39-46.

Becq, J., M. C. Gutierrez, V. Rosas-Magallanes, J. Rauzier, B. Gicquel, O. Neyrolles & P. Deschavanne, (2007) Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol* 24: 1861-1871.

Borrell, S. & S. Gagneux, (2011) Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. *Clin Microbiol Infect* 17: 815-820.

Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen & S. T. Cole, (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A* 99: 3684-3689.

Brosch, R., S. V. Gordon, A. Pym, K. Eiglmeier, T. Garnier & S. T. Cole, (2000) Comparative genomics of the mycobacteria. *Int J Med Microbiol* 290: 143-152.

Brosch, R., A. S. Pym, S. V. Gordon & S. T. Cole, (2001) The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* 9: 452-458.

Brudey, K., I. Filliol, S. Ferdinand, V. Guernier, P. Duval, B. Maubert, C. Sola & N. Rastogi, (2006) Long-term population-based genotyping study of *Mycobacterium tuberculosis* complex isolates in the French departments of the Americas. *J Clin Microbiol* 44: 183-191.

Caws, M., G. Thwaites, S. Dunstan, T. R. Hawn, N. T. Lan, N. T. Thuong, K. Stepniewska, M. N. Huyen, N. D. Bang, T. H. Loc, S. Gagneux, D. van Soolingen, K. Kremer, M. van der Sande, P. Small, P. T. Anh, N. T. Chinh, H. T. Quy, N. T. Duyen, D. Q. Tho, N. T. Hieu, E. Torok, T. T. Hien, N. H. Dung, N. T. Nhu, P. M. Duy, N. van Vinh Chau & J. Farrar, (2008) The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. *PLoS Pathog* 4: e1000034

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin,

S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead & B. G. Barrell, (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537-544.

Comas, I., J. Chakravartti, P. M. Small, J. Galagan, S. Niemann, K. Kremer, J. D. Ernst & S. Gagneux, (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42: 498-503.

Comas, I. & S. Gagneux, (2009) The past and future of tuberculosis research. *PLoS Pathog* 5: e1000600.

Comas, I., S. Homolka, S. Niemann & S. Gagneux, (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* 4: e7815.

Coscolla, M. & S. Gagneux, (2010) Does M. tuberculosis genomic diversity explain disease diversity? *Drug Discov Today Dis Mech* 7: e43-e59.

Cowley, D., D. Govender, B. February, M. Wolfe, L. Steyn, J. Evans, R. J. Wilkinson & M. P. Nicol, (2008) Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. *Clin Infect Dis* 47: 1252-1259.

Coxon, G. D. & L. G. Dover, (2011) Current Status and Research Strategies in Tuberculosis Drug Development. *J Med Chem*.

Cubillos-Ruiz, A., J. Morales & M. M. Zambrano, (2008) Analysis of the genetic variation in *Mycobacterium tuberculosis* strains by multiple genome alignments. *BMC Res Notes* 1: 110.

Cubillos-Ruiz, A., A. Sandoval, V. Ritacco, B. Lopez, J. Robledo, N. Correa, I. Hernandez-Neuta, M. M. Zambrano & P. Del Portillo, (2010) Genomic signatures of the haarlem lineage of *Mycobacterium tuberculosis*: implications of strain genetic variation in drug and vaccine development. *J Clin Microbiol* 48: 3614-3623.

de Jong, B. C., P. C. Hill, A. Aiken, T. Awine, M. Antonio, I. M. Adetifa, D. J. Jackson-Sillah, A. Fox, K. Deriemer, S. Gagneux, M. W. Borgdorff, K. P. McAdam, T. Corrah, P. M. Small & R. A. Adegbola, (2008) Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J Infect Dis* 198: 1037-1043.

Di Pietrantonio, T., C. Hernandez, M. Girard, A. Verville, M. Orlova, A. Belley, M. A. Behr, J. C. Loredo-Osti & E. Schurr, (2010) Strain-specific differences in the genetic control of two closely related mycobacteria. *PLoS Pathog* 6: e1001169.

Donoghue, H. D., (2009) Human tuberculosis--an ancient disease, as elucidated by ancient microbial biomolecules. *Microbes Infect* 11: 1156-1162.

Donoghue, H. D., (2011) Insights gained from palaeomicrobiology into ancient and modern tuberculosis. *Clin Microbiol Infect* 17: 821-829.

Doolittle, W. F. & O. Zhaxybayeva, (2009) On the origin of prokaryotic species. *Genome Res* 19: 744-756.

Escombe, A. R., D. A. Moore, R. H. Gilman, W. Pan, M. Navincopa, E. Ticona, C. Martinez, L. Caviedes, P. Sheen, A. Gonzalez, C. J. Noakes, J. S. Friedland & C. A. Evans, (2008) The infectiousness of tuberculosis patients coinfected with HIV. *PLoS Med* 5: e188.

Fang, Z., C. Doig, N. Morrison, B. Watt & K. J. Forbes, (1999) Characterization of IS1547, a new member of the IS900 family in the *Mycobacterium tuberculosis* complex, and its association with IS6110. *J Bacteriol* 181: 1021-1024.

Filliol, I., A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbon, M. Bobadilla del Valle, J. Fyfe, L. Garcia-Garcia, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. Leon, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendon, J. Sifuentes-Osornio, A. Ponce de Leon, M. D. Cave, R. Fleischmann, T. S. Whittam & D. Alland, (2006) 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* 188: 759-772.

Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs Jr, Jr., J. C. Venter & C. M. Fraser, (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 184: 5479-5490.

Ford, C. B., P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. Galagan, N. Mohaideen, T. R. Ioerger, J. C. Sacchettini, M. Lipsitch, J. L. Flynn & S. M. Fortune, (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* 43: 482-486.

Gagneux, S., K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell & P. M. Small, (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 103: 2869-2873.

Gandhi, N. R., A. Moll, A. W. Sturm, R. Pawinski, T. Govender, U. Lalloo, K. Zeller, J. Andrews & G. Friedland, (2006) Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet* 368: 1575-1580.

Garcia de Viedma, D., G. Lorenzo, P. J. Cardona, N. A. Rodriguez, S. Gordillo, M. J. Serrano & E. Bouza, (2005) Association between the infectivity of *Mycobacterium tuberculosis* strains and their efficiency for extrarespiratory infection. *J Infect Dis* 192: 2059-2065.

Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. Jones, F. S. Brinkman, R. C. Brunham & P. Tang, (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364: 730-739.

Gordon, S. V., B. Heym, J. Parkhill, B. Barrell & S. T. Cole, (1999) New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 145 (Pt 4): 881-892.

Gupta, K., P. Kumar & D. Chatterji, (2010) Identification, activity and disulfide connectivity of C-di-GMP regulating proteins in *Mycobacterium tuberculosis*. *PLoS One* 5: e15072.

Gutacker, M. M., B. Mathema, H. Soini, E. Shashkina, B. N. Kreiswirth, E. A. Graviss & J. M. Musser, (2006) Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis* 193: 121-128.

Gutacker, M. M., J. C. Smoot, C. A. Migliaccio, S. M. Ricklefs, S. Hua, D. V. Cousins, E. A. Graviss, E. Shashkina, B. N. Kreiswirth & J. M. Musser, (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 162: 1533-1543.

Gutierrez, M. C., S. Brisse, R. Brosch, M. Fabre, B. Omais, M. Marmiesse, P. Supply & V. Vincent, (2005) Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog* 1: e5.

Hanekom, M., N. C. Gey van Pittius, C. McEvoy, T. C. Victor, P. D. Van Helden & R. M. Warren, (2011) *Mycobacterium tuberculosis* Beijing genotype: A template for success. *Tuberculosis (Edinb)*.

Hawkey, P. M. & A. M. Jones, (2009) The changing epidemiology of resistance. *J Antimicrob Chemother* 64 Suppl 1: i3-10.

Hershberg, R., M. Lipatov, P. M. Small, H. Sheffer, S. Niemann, S. Homolka, J. C. Roach, K. Kremer, D. A. Petrov, M. W. Feldman & S. Gagneux, (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* 6: e311.

Hirsh, A. E., A. G. Tsolaki, K. DeRiemer, M. W. Feldman & P. M. Small, (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci U S A* 101: 4871-4876.

Inoue, J., P. C. Donoghue & Z. Yang, (2010) The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol* 59: 74-89.

Ioerger, T. R., S. Koo, E. G. No, X. Chen, M. H. Larsen, W. R. Jacobs, Jr., M. Pillay, A. W. Sturm & J. C. Sacchettini, (2009) Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS One* 4: e7778.

Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal & J. van Embden, (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 35: 907-914.

Kato-Maeda, M., P. J. Bifani, B. N. Kreiswirth & P. M. Small, (2001a) The nature and consequence of genetic variability within *Mycobacterium tuberculosis*. *J Clin Invest* 107: 533-537.

Kato-Maeda, M., S. Gagneux, L. L. Flores, E. Y. Kim, P. M. Small, E. P. Desmond & P. C. Hopewell, (2011) Strain classification of *Mycobacterium tuberculosis*: congruence between large sequence polymorphisms and spoligotypes. *Int J Tuberc Lung Dis* 15: 131-133.

Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat & P. M. Small, (2001b) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* 11: 547-554.

Kaufmann, S. H., (2010) Future vaccination strategies against tuberculosis: thinking outside the box. *Immunity* 33: 567-577.

Kim, E. Y., P. Nahid, P. C. Hopewell & M. Kato-Maeda, (2010) Novel hot spot of IS*6110* insertion in *Mycobacterium tuberculosis*. *J Clin Microbiol* 48: 1422-1424.

Koul, A., E. Arnoult, N. Lounis, J. Guillemont & K. Andries, (2011) The challenge of new drug discovery for tuberculosis. *Nature* 469: 483-490.

Lawn, S. D. & A. I. Zumla, (2011) Tuberculosis. *Lancet* 378: 57-72.

Lin, M. Y. & T. H. Ottenhoff, (2008) Host-pathogen interactions in latent *Mycobacterium tuberculosis* infection: identification of new targets for tuberculosis intervention. *Endocr Metab Immune Disord Drug Targets* 8: 15-29.

Maisnier-Patin, S. & D. I. Andersson, (2004) Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res Microbiol* 155: 360-369.

Malik, A. N. & P. Godfrey-Faussett, (2005) Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. *Lancet Infect Dis* 5: 174-183.

McAdam, R. A., P. W. Hermans, D. van Soolingen, Z. F. Zainuddin, D. Catty, J. D. van Embden & J. W. Dale, (1990) Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS3 family. *Mol Microbiol* 4: 1607-1613.

McEvoy, C. R., A. A. Falmer, N. C. Gey van Pittius, T. C. Victor, P. D. van Helden & R. M. Warren, (2007) The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 87: 393-404.

McEvoy, C. R., P. D. van Helden, R. M. Warren & N. C. Gey van Pittius, (2009a) Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. *BMC Evol Biol* 9: 237.

McEvoy, C. R., R. M. Warren, P. D. van Helden & N. C. Gey van Pittius, (2009b) Multiple, independent, identical IS*6110* insertions in *Mycobacterium tuberculosis* PPE genes. *Tuberculosis (Edinb)* 89: 439-442.

McLean, K. J. & A. W. Munro, (2008) Structural biology and biochemistry of cytochrome P450 systems in *Mycobacterium tuberculosis*. *Drug Metab Rev* 40: 427-446.

Mendiola, M. V., C. Martin, I. Otal & B. Gicquel, (1992) Analysis of the regions responsible for IS*6110* RFLP in a single *Mycobacterium tuberculosis* strain. *Res Microbiol* 143: 767-772.

Mostowy, S. & M. A. Behr, (2005) The origin and evolution of *Mycobacterium tuberculosis*. *Clin Chest Med* 26: 207-216, v-vi.

Nicol, M. P. & R. J. Wilkinson, (2008) The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. *Trans R Soc Trop Med Hyg* 102: 955-965.

Olano, J., B. Lopez, A. Reyes, M. P. Lemos, N. Correa, P. Del Portillo, L. Barrera, J. Robledo, V. Ritacco, and M. M. Zambrano. 2007. Mutations in DNA repair genes are associated with the Haarlem lineage of *Mycobacterium tuberculosis* independently of their antibiotic resistance. Tuberculosis (Edinb) 87:502-8.

Ordway, D., M. Henao-Tamayo, M. Harton, G. Palanisamy, J. Troudt, C. Shanley, R. J. Basaraba & I. M. Orme, (2007) The hypervirulent *Mycobacterium tuberculosis* strain HN878 induces a potent TH1 response followed by rapid down-regulation. *J Immunol* 179: 522-531.

Otal, I., A. B. Gomez, K. Kremer, P. de Haas, M. J. Garcia, C. Martin & D. van Soolingen, (2008) Mapping of IS*6110* insertion sites in Mycobacterium bovis isolates in relation to adaptation from the animal to human host. *Vet Microbiol* 129: 333-341.

Parwati, I., R. van Crevel & D. van Soolingen, (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* 10: 103-111.

Portevin, D., S. Gagneux, I. Comas & D. Young, (2011) Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog* 7: e1001307.

Pym, A. S., B. Saint-Joanis & S. T. Cole, (2002) Effect of katG mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect Immun* 70: 4955-4960.

Ramaswamy, S. & J. M. Musser, (1998) Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber Lung Dis* 79: 3-29.

Ritacco, V., M. J. Iglesias, L. Ferrazoli, J. Monteserin, E. R. Dalla Costa, A. Cebollada, N. Morcillo, J. Robledo, J. H. de Waard, P. Araya, L. Aristimuno, R. Diaz, P. Gavin, B. Imperiale, V. Simonsen, E. M. Zapata, M. S. Jimenez, M. L. Rossetti, C. Martin, L. Barrera & S. Samper, (2011) Conspicuous multidrug-resistant *Mycobacterium tuberculosis* cluster strains do not trespass country borders in Latin America and Spain. *Infect Genet Evol*.

Rivero, A., M. Marquez, J. Santos, A. Pinedo, M. A. Sanchez, A. Esteve, S. Samper & C. Martin, (2001) High rate of tuberculosis reinfection during a nosocomial outbreak of multidrug-resistant tuberculosis caused by Mycobacterium bovis strain B. *Clin Infect Dis* 32: 159-161.

Rosas-Magallanes, V., P. Deschavanne, L. Quintana-Murci, R. Brosch, B. Gicquel & O. Neyrolles, (2006) Horizontal transfer of a virulence operon to the ancestor of *Mycobacterium tuberculosis*. *Mol Biol Evol* 23: 1129-1135.

Safi, H., J. Aznar & J. C. Palomares, (1997) Molecular epidemiology of *Mycobacterium tuberculosis* strains isolated during a 3-year period (1993 to 1995) in Seville, Spain. *J Clin Microbiol* 35: 2472-2476.

Sampson, S. L., (2011) Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin Dev Immunol* 2011: 497203.

Sandgren, A., M. Strong, P. Muthukrishnan, B. K. Weiner, G. M. Church & M. B. Murray, (2009) Tuberculosis drug resistance mutation database. *PLoS Med* 6: e2.

Siguier, P., J. Filee & M. Chandler, (2006) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* 9: 526-531.

Sinsimer, D., G. Huet, C. Manca, L. Tsenova, M. S. Koo, N. Kurepina, B. Kana, B. Mathema, S. A. Marras, B. N. Kreiswirth, C. Guilhot & G. Kaplan, (2008) The phenolic glycolipid of *Mycobacterium tuberculosis* differentially modulates the early host cytokine response but does not in itself confer hypervirulence. *Infect Immun* 76: 3027-3036.

Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley & G. K. Schoolnik, (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 330: 1703-1709.

Smith, N. H., R. G. Hewinson, K. Kremer, R. Brosch & S. V. Gordon, (2009) Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 7: 537-544.

Smith, N. H., K. Kremer, J. Inwald, J. Dale, J. R. Driscoll, S. V. Gordon, D. van Soolingen, R. G. Hewinson & J. M. Smith, (2006) Ecotypes of the *Mycobacterium tuberculosis* complex. *J Theor Biol* 239: 220-225.

Sola, C., I. Filliol, M. C. Gutierrez, I. Mokrousov, V. Vincent & N. Rastogi, (2001) Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. *Emerg Infect Dis* 7: 390-396.

Soto, C. Y., M. C. Menendez, E. Perez, S. Samper, A. B. Gomez, M. J. Garcia & C. Martin, (2004) IS*6110* mediates increased transcription of the phoP virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J Clin Microbiol* 42: 212-219.

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

Supply, P., R. M. Warren, A. L. Banuls, S. Lesjean, G. D. Van Der Spuy, L. A. Lewis, M. Tibayrenc, P. D. Van Helden & C. Locht, (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol Microbiol* 47: 529-538.

Talarico, S., M. D. Cave, B. Foxman, C. F. Marrs, L. Zhang, J. H. Bates & Z. Yang, (2007) Association of *Mycobacterium tuberculosis* PE PGRS33 polymorphism with clinical and epidemiological characteristics. *Tuberculosis (Edinb)* 87: 338-346.

Thierry, D., M. D. Cave, K. D. Eisenach, J. T. Crawford, J. H. Bates, B. Gicquel & J. L. Guesdon, (1990) IS*6110*, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Res* 18: 188.

Thorne, N., S. Borrell, J. Evans, J. Magee, D. Garcia de Viedma, C. Bishop, J. Gonzalez-Martin, S. Gharbia & C. Arnold, (2011) IS*6110*-based global phylogeny of *Mycobacterium tuberculosis*. *Infect Genet Evol* 11: 132-138.

Tsolaki, A. G., A. E. Hirsh, K. DeRiemer, J. A. Enciso, M. Z. Wong, M. Hannan, Y. O. Goguet de la Salmoniere, K. Aman, M. Kato-Maeda & P. M. Small, (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* 101: 4865-4870.

Turcios, L., Y. Casart, I. Florez, J. de Waard & L. Salazar, (2009) Characterization of IS*6110* insertions in the dnaA-dnaN intergenic region of *Mycobacterium tuberculosis* clinical isolates. *Clin Microbiol Infect* 15: 200-203.

van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick & et al., (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 31: 406-409.

van Embden, J. D., T. van Gorkom, K. Kremer, R. Jansen, B. A. van Der Zeijst & L. M. Schouls, (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol* 182: 2393-2401.

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

Vasilyeva, S. V., E. Unur, R. M. Walczak, E. P. Donoghue, A. G. Rinzler & J. R. Reynolds, (2009) Color purity in polymer electrochromic window devices on indium-tin oxide and single-walled carbon nanotube electrodes. *ACS Appl Mater Interfaces* 1: 2288-2297.

Vera-Cabrera, L., M. A. Hernandez-Vera, O. Welsh, W. M. Johnson & J. Castro-Garza, (2001) Phospholipase region of *Mycobacterium tuberculosis* is a preferential locus for IS*6110* transposition. *J Clin Microbiol* 39: 3499-3504.

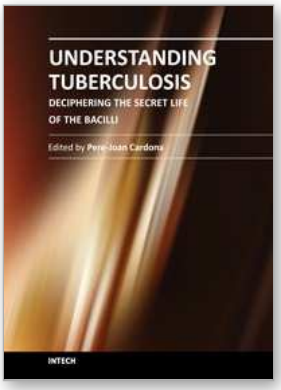
Wall, S., K. Ghanekar, J. McFadden & J. W. Dale, (1999) Context-sensitive transposition of IS*6110* in mycobacteria. *Microbiology* 145 (Pt 11): 3169-3176.

Warren, R. M., S. L. Sampson, M. Richardson, G. D. Van Der Spuy, C. J. Lombard, T. C. Victor & P. D. van Helden, (2000) Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol Microbiol* 37: 1405-1416.

Wirth, T., F. Hildebrand, C. Allix-Beguec, F. Wolbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rusch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply & S. Niemann, (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog* 4: e1000160.

Yang, Z., D. Yang, Y. Kong, L. Zhang, C. F. Marrs, B. Foxman, J. H. Bates, F. Wilson & M. D. Cave, (2005) Clinical relevance of *Mycobacterium tuberculosis* plcD gene mutations. *Am J Respir Crit Care Med* 171: 1436-1442.

Yesilkaya, H., J. W. Dale, N. J. Strachan & K. J. Forbes, (2005) Natural transposon mutagenesis of clinical isolates of *Mycobacterium tuberculosis*: how many genes does a pathogen need? *J Bacteriol* 187: 6726-6732.



Understanding Tuberculosis - Deciphering the Secret Life of the Bacilli

Edited by Dr. Pere-Joan Cardona

ISBN 978-953-307-946-2
Hard cover, 334 pages
Publisher InTech
Published online 17, February, 2012
Published in print edition February, 2012

Mycobacterium tuberculosis, as recent investigations demonstrate, has a complex signaling expression, which allows its close interaction with the environment and one of its most renowned properties: the ability to persist for long periods of time under a non-replicative status. Although this skill is well characterized in other bacteria, the intrinsically very slow growth rate of Mycobium tuberculosis, together with a very thick and complex cell wall, makes this pathogen specially adapted to the stress that could be generated by the host against them. In this book, different aspects of these properties are displayed by specialists in the field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

María Mercedes Zambrano, Ginna Hernández-Neuta, Iván Hernández-Neuta, Andrea Sandoval, Andrés Cubillos-Ruiz, Alejandro Reyes and Patricia Del Portillo (2012). Genomic Variability of Mycobacterium tuberculosis, Understanding Tuberculosis - Deciphering the Secret Life of the Bacilli, Dr. Pere-Joan Cardona (Ed.), ISBN: 978-953-307-946-2, InTech, Available from: http://www.intechopen.com/books/understanding-tuberculosis-deciphering-the-secret-life-of-the-bacilli/genomic-variability-of-mycobacterium-tuberculosis

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821