Palomino • Leão • Ritacco

TUBERCULOSIS 2007

From Basic Science to Patient Care

www.TuberculosisTextbook.com

Chapter 2: Molecular Evolution of the *Mycobacterium tuberculosis* Complex

Nalin Rastogi and Christophe Sola

2.1. A basic evolutionary scheme of mycobacteria

Mycobacteria are likely to represent a very ancient genus of bacteria. Probably, the *Mycobacterium* genus originates from a common ancestor whose offspring specialized in the process of colonizing very different ecological niches. The evolutionary relationships between organisms of the genus *Mycobacterium* have been investigated on the basis of the analysis of derived similarities ("shared derived traits", synapomorphies).

Since no contemporary living species may directly stem from another contemporary species, it is advisable to speak of «common ancestors», by building cladograms rather than genealogical trees when comparing a monophyletic group. Such cladistic analysis (the word clade is derived from the ancient Greek κλάδος, klados, meaning branch) forms an ideal basis for modern systems of biological classification. Cladograms so generated are invariably dependent on the amount of information selected by the researcher.

An ideal approach takes into account a wide variety of information in order to form a natural group of organisms (**clade**) which share a unique ancestor that is not shared with other organisms on the tree, i.e., each clade comprises a series of characteristics specific to its members (**synapomorphies**), and absent from the group of organisms from which it diverged. Such distinction involves the notion of **outgroups** (organisms that are closely related to the group but not part of it). The choice of an outgroup constitutes an essential step, since it can profoundly change the topology of a tree. Similarly, much attention is needed to distinguish between **characters** and **character states** prior to such analysis (e.g., "blue eyes" and "black eyes" are two character states of the character "eye-color"). A character state of a determined clade which is also present in its outgroups and its ancestor is designated as **plesiomorphy** (meaning "close form", also called ancestral state). The character state which occurs only in later descendants is called an **apomorphy** (meaning "separate form", also called the "derived" state). As only synapomorphies are used to characterize clades, the distinction between plesiomorphic and synapomorphic character states is made by considering one or more outgroups.

A collective set of plesiomorphies is commonly referred to as a **ground plan** for the clade or clades they refer to; and one clade is considered **basal** to another if it

holds more plesiomorphic characters than the other clade. Usually, a basal group is very species-poor in comparison to a more derived group. Thus, conservative (apomorphic) branches, defined as **anagenetic branches** represent species whose characteristics are closer to those of the ancestor than others.

Possibly, the founder of the genus *Mycobacterium* was a free-living organism and today's free-living mycobacterial species (and also some saprophytic species?) represent the conservative branches of founding mycobacteria. The more distant organisms are probably the ones that live in association with various multicellular organisms. It has been suggested that the mycobacteria that created a long-lasting association with marine animals (probably placoderms) are at the root of this phylogenetic branch. Thus, *Mycobacterium marinum* would stem from the conservative branch, whereas other vertebrate-associated mycobacteria would build the anagenetic branch. Grmek speculates that the association of a mycobacterial species with a marine vertebrate may have occurred during the superior Devonian (300 million years ago) (Grmek 1994). Figure 2-1 shows the phylogenetic position of the *Mycobacterium tuberculosis* complex species within the genus *Mycobacterium* based on a tree of the gene coding for the 16S ribosomal ribonucleic acid (rRNA).

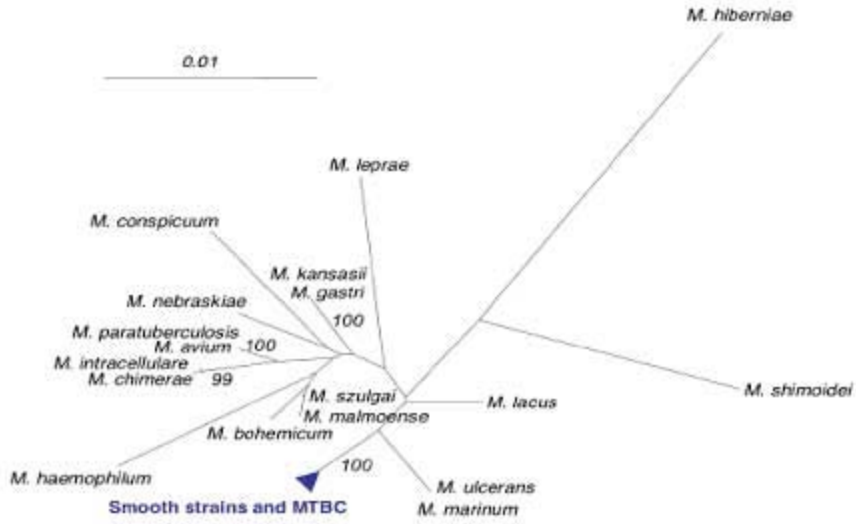


Figure 2-1: Phylogenetic position of the tubercle bacilli within the genus *Mycobacterium* (reproduced with permission from Gutierrez et al. 2005)

In the past, mycobacterial systematics used to rely on phenotypic characters; more recently, however, genetic techniques have boosted taxonomic studies (Tortoli 2003). The first natural characters used to distinguish between mycobacterial species were growth rate and pigmentation. Rapid growers (< 7 days) are free, environmental, saprophytic species, whereas slow growers are usually obligate intracellular, pathogenic species. The slow-fast grower division, which virtually always relies on the possession of one or two rRNA operons (*rrn* operon) (Jy 1994), was shown to be phylogenetically coherent (Stahl 1990, Devulder 2005).

In the ’50s, the hypothesis of co-evolution, or parallel evolution, between hosts and mycobacteria looked no more likely than the alternative hypothesis of «multiple, casual (furtive) introductions» of various saprophytes into different hosts. The traditional epidemiological belief for tuberculosis (TB) is that the anthropozoonosis due to *M. tuberculosis* may find its origin in a zoonotic agent, i.e., *Mycobacterium bovis* (Cockburn 1963). This view is still sustained by some authors (Smith 2006a). However, genetics brought some new clues into the debate (Brosch 2002). For example, the sequencing of the *Mycobacterium leprae* genome, by its defective nature, confirmed the previous history-driven hypothesis that *M. leprae* was a younger pathogen than *M. tuberculosis* (Cole 1998, Cole 2001). In the case of the *M. tuberculosis* complex, comparative genomics has also shown that the *M. bovis* genome is smaller than the *M. tuberculosis* genome, opening the way to a new scenario for the evolution of the tubercle bacillus (Brosch 2002). *M. bovis* genomic reduction (loss of genes) indeed suggests that it could be a younger pathogen than *M. tuberculosis* or, in other words, that human TB disease preceded bovine disease (Brosch 2002, Cockburn 1963). Figure 2-2 shows that the common ancestor of members of the *M. tuberculosis* complex is close to three of its branches: “*Mycobacterium canettii*”, *Mycobacterium africanum* and the ancestral East-African-Indian (EAI) clade. However, according to Smith et al., “*until it is demonstrated that strains of M. africanum subtype I can be maintained in immunocompetent cells, the host-association of the most recent common ancestor of the M. tuberculosis complex remains unsolved”* (Smith 2006b).

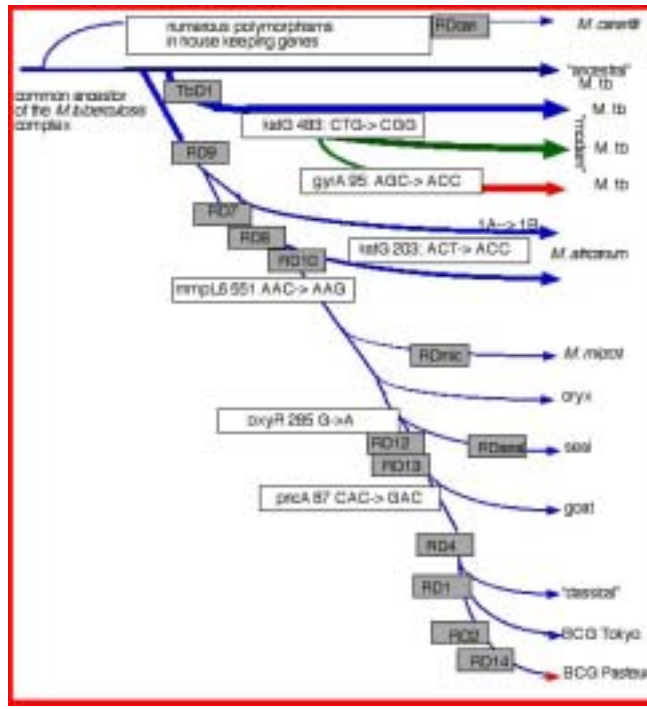


Figure 2-2: Scheme of the proposed evolutionary pathway of the *M. tuberculosis* bacilli illustrating successive loss of DNA in certain lineages (reproduced with permission from Brosch *et al.* 2002)

Ancient humans, bovids and mastodons experienced erosive diseases caused by *M. tuberculosis*. As an alternative to the classical hypothesis of TB spread being driven by human migration, bovids, mastodons, or simply diet might well be considered to be the natural epidemiological vehicle of TB. In this way, a poorly pathogenic environmental or animal *Mycobacterium* spp. would have progressively acquired some human-specific virulence traits (Rotschild 2001, Rotschild 2006a). The association of hyperdisease and endemic stability may have promoted a smooth and long-term transition from zoonosis to anthropozoonosis (Coleman 2001, Rotschild 2006b). Other complex anthropological parameters, such as the history of agriculture and livestock domestication, may also have been mediators of TB spread (Smith 1995, Bruford 2003). In this sense, it is also logical to compare the pathogenicity of the various *M. tuberculosis* complex members in various host species. Interestingly, it has been observed that *M. africanum* apparently elicits a more attenuated T cell response to the 6 kiloDalton (kDa) early secreted antigen

(ESAT-6) than *M. tuberculosis* in patients with TB. *M. africanum* could thus be considered to be an opportunistic human pathogen. If confirmed, these findings are new evidence that strain differences affect human interferon-based T cell responses (de Jong 2006). Strain-related differences in lymphokine (including interferon-gamma) response in mice with experimental infection were also reported in 2003 (Lopez 2003).

2.2. *M. tuberculosis* complex population molecular genetics

Until recently, the question of individual genetic variation within the *M. tuberculosis* complex gained little attention and most research on *M. tuberculosis* was organism- rather than population-centered. The advent of molecular methods, and their widespread use in population studies, introduced both new conceptual and new technological developments. The inference of phylogenies from molecular data goes back to the early ’90s with the development of software such as PHYLIP and PAUP (Felsenstein 1993, Swofford 1990, Swofford 1998). In particular, the study of the *M. tuberculosis* complex phylogeny closely followed the development of increasing numbers of sophisticated genotyping methods. The way was opened by *M. tuberculosis* fingerprinting by restriction fragment length polymorphism based on insertion sequence IS*6110* (IS*6110* RFLP) (van Embden 1993). However, the use of IS*6110* RFLP in evolutionary genetics discovery was of limited value for many reasons:

- fast variation rate of this evolutionary marker (de Boer 1999)
- complexity of forces driving its transposition and risk of genetic convergence (Fang 2001)
- nature of experimental data produced which requires sophisticated software for analysis
- difficulty to build large sets of data (Heersma 1998, Salamon 1998)

The discovery in 1993 of the polymorphic nature of the Direct Repeat (DR) locus, and the subsequent development of the spoligotyping method based on DR locus variability, introduced more modern concepts and tools for *M. tuberculosis* complex genotyping (Groenen 1993, Kamerbeek 1997). Our research group bet that the highly diverse signature patterns observed by spoligotyping could indeed contain phylogenetical signals, and the construction of a diversity database was started *de novo* (Sola 1999). Today, a total of 62 *M. tuberculosis* complex clades/lineages are detailed in the Fourth International Spoligotyping Database (SpolDB4) which de-

scribes 1,939 shared-types representing a total of 39,295 *M. tuberculosis* strains from 122 countries (Brudey 2006). This database is available on the internet at SITVIT (http://www.pasteur-guadeloupe.fr:8081/SITVITDemo). Some of the major *M. tuberculosis* complex clades and their spoligotype signatures are described below under section 2.9. The assumption that the DR locus was neutral still remains speculative; however, the finding of other clustered regularly interspersed palindromic repeats (CRISPR) loci in both Archae and Bacteriae has become a hot issue (Jansen 2002, Pourcel 2005, Makarova 2006). Spoligotyping was immediately followed by the discovery of tandem repeat loci in the *M. tuberculosis* complex and the Variable Number of Tandem Repeats (VNTR) genotyping technique (Frothingham 1998). Later, the Mycobacterial Interspersed Repetitive Units (MIRU) technique (Supply 2001) was developed, which is also designated as Multiple Locus VNTR analysis (MLVA). Multi-Locus Sequence Typing (MLST) was introduced as an alternative method (Baker 2004). More recently, systematic Single Nucleotide Polymorphism (SNP) genotyping (Filliol 2006, Gutacker 2006) was described followed by Large Sequence Polymorphism (LSP), the latter performed either by microarray or real-time Polymerase Chain Reaction (PCR) (Mostowy 2002, Tsolaki 2005).

2.3. Co-evolution of *M. tuberculosis* with its hosts

Simulation models reported in 1988 suggested that a social network with a size of 180 to 440 persons is required for TB to occur with endemicity. In such conditions, host-pathogen coexistence would be maintained in populations (McGrath 1988). The concept of endemic stability, already mentioned above, suggests that an infectious disease may reach an epidemiological state in which the clinical disease is scarce, despite high levels of infection in the population (Coleman 2001). Clearly, this concept may apply to TB since it is most likely to have been a vertically transmitted disease before being responsible for large outbreaks.

The question of how many isolated communities of between 180 to 440 persons may have experienced, sequentially or concomitantly, the introduction of one or more founding genotypes of *M. tuberculosis* complex (each one with its own specific virulence), in other words, how TB was “seeded” is of paramount importance. To provide the initial conditions of a dynamic epidemic system we must understand how these early founding genotypes spread in low demographic conditions. Today, we can observe a phylogeographically structured global epidemic, built as a result of millennia of evolution. Some clones are extinct, others have an increased risk of emergence (Tanaka 2006). The evolution rate of TB is likely to have been succes-

sively slow (human and cattle migration and low endemicity or hyperendemicity but little or no disease), then moderate (five centuries of post-Columbus sail-based migration) with important anthropological changes that may have created bursting conditions linked to demographic growth and migration, and lastly, fast (since the introduction of air transportation), i.e. within the five decades of increasing movements of strains and people, concomitantly to new outbreaks in demographically active and resource-poor countries where the great majority of cases is now present.

Consequently, the worldwide bacterial genetic snapshot of the TB epidemic is the result of a combination of slow, medium, and fast evolving superimposition pictures of various outbreak histories. Such a jigsaw puzzle will be difficult, if not impossible, to reconstruct. However, looking for rare and isolated genotypes, which may have undergone a slower evolution, as well as searching for ancient desoxyribonucleic acid (DNA) may constitute two complementary scientific strategies in attempting to reach this goal.

One recent success of the first strategy is exemplified by the finding of a peculiar highly genetically diverse "*M. canettii*" in the Horn of Africa. "*M. canettii*" was likely to be the most probable source species of the *M. tuberculosis* complex, rather than just another branch of it (Fabre 2004). Further results confirm that, despite its apparent homogeneity, the "*M. canettii*" or "*M. prototuberculosis*" genome is a composite assembly resulting from horizontal gene transfer events predating clonal expansion. The large amount of synonymous single nucleotide polymorphism (sSNP) variation in housekeeping genes found in these smooth strains of "*M. prototuberculosis"* suggests that the tubercle bacilli were contemporaneous with early hominids in East Africa, and may have thus been evolving with their human host much longer than previously thought. These results open new perspectives for unraveling the molecular bases of *M. tuberculosis* evolutionary success (Gutierrez 2005).

The second strategy has also provided interesting results that support the notion of TB's ancient origin. The isolation and characterization of ancient *M. tuberculosis* DNA from an extinct bison, dated 17,000 years B.C., suggest the presence of TB in America in the late Pleiostocene (Rotschild 2001). The extensive infection of many individuals of the *Mammut americanum* species with the *M. tuberculosis* agent also suggests that, apart from *Homo sapiens,* mastodons and bovids may have spread the disease during the Pleistocene (Rotschild 2006a, Rotschild 2006b). When looking at human remains, several DNA studies served to trace back the presence of TB to Egyptian mummies, where *M. tuberculosis* and also *M. africanum* genotypes were identified (Zink 2003). Figure 2-3 shows an ancient Egyptian clay arte-

fact with a traditional kyphosis suggestive of Pott's disease. The presence of TB in America before the arrival of the Spanish settlers is also well demonstrated both by paleopathological evidence and studies on ancient DNA (Salo 1994, Arriaza 1995). Recent paleopathological evidence also suggests the presence of leprosy and TB in South East Asian human remains from the Iron Age (Tayles 2004). Taken together, these results may argue that the limited number of different genogroups that we observe today are likely to stem from those that were seeded in the past, have remained isolated by distance during millennia, and have had time to co-evolve independently before gaining reasonable statistical chances to meet.

Figure 2-3: Egyptian clay artefact of an emaciated man with a characteristic angular kyphosis suggestive of Pott's disease (reproduced from TB, Past, Present, 1999, TB Foundation)

2.4. *M. tuberculosis* through space and time

The concept of phylogeography was originally introduced by Avise (Avise 1987), as "*the history of processes that control the geographic distribution of genes and lineages by constructing the genealogies of populations and genes"*. The term was introduced as a way to bridge population genetics and molecular ecology and to describe geographically structured signals within species. This concept might well be applied to studies on the global spread of *M. tuberculosis* through time. If the ancestor of *M. tuberculosis* adapted specifically and slowly to human beings, it may have had the time to develop, via an extreme clonality, a deeply rooted and peculiar phylogeographical structure reflecting both the demographic history and the history of TB spread.

The geographic distribution of bacteriophage types was the only method to detect the geographic subdivision of the *M. tuberculosis* complex species during the '70s and the '80s (Bates 1969, Sula 1973); however, no phylogenetic relationships could be inferred at that time using mycobacteriophages. A numerical analysis of *M. africanum* taxonomy also suggested differences between isolates from West and East Africa (David 1978). The naming of two *M. africanum* variants (subtype I and II) created confusion and the status of *M. africanum* as a homogeneous sub-species of *M. tuberculosis* complex is still uncertain. The existence of some major geographical and epidemiological significant genetic variants of the *M. tuberculosis* complex was also recognized as early as 1982 (Collins 1982). Among these were the Asian, the bovine and the classical variants, in addition to *africanum* I and *africanum* II variants.

Lateral genetic transfer was presumed to be minor in *M. tuberculosis*, and the clonal structure of the *M. tuberculosis* complex was formally demonstrated by the finding of strong linkage disequilibrium within MIRU loci (Supply 2003). Only recently has the issue of *M. tuberculosis* complex lateral genetic transfer gained interest, particularly in regard to its links to genetic diversity and to potential acquisition of virulence (Kinsella 2003, Rosas-Magallanes 2006, Alix 2006). The importance of lateral genetic transfer in one species' history is of primary importance to better understand its specificity. As for the members of the *M. tuberculosis* complex, with the exception of *M. canettii*, there is no evidence for this kind of transfer or for housekeeping gene recombination (Smith 2006a). Indeed, recent evidence argues in favor of the existence of lateral genetic transfer in the precursor of the *M. tuberculosis* complex, and in favor of environmental mycobacteria being the source of certain genetic components in the *M. tuberculosis* complex. These findings reinforce the idea that the ancestor of the *M. tuberculosis* complex was an environ-

mental *Mycobacterium* (Rosas-Magallenes 2006). Another source of exogenous DNA may be plasmids that have been shown to be present in modern species of mycobacteria, and sometimes to carry virulence genes (Le Dantec 2001, Stinear 2000, Stinear 2004). The mosaic nature of the genome of ancestral "*M. prototuberculosis*" species also argues in favor of numerous gene transfer events and/or homologous recombination within ancient species of the *M. tuberculosis* complex (Gutierrez 2005).

2.5. Looking for robust evolutionary markers

When looking for robust evolutionary markers, the evolutionist will first choose markers that are assumedly neutral in order to avoid debates on function or potential selection, whether positive or stabilizing. For the *M. tuberculosis* complex, the very existence of an obligate intracellular life, which provides a stable chemical and metabolic environment, suggests that a classical metabolic selection scheme must have played a minimal role in the evolution of the *M. tuberculosis* complex genome (Musser 2000). Host specialization and niche adaptation may have been more important. Changes towards acquisition of an intracellular life style may also be responsible for loss of function and hence, loss of genes.

Silent mutations in housekeeping genes were the first candidates to be selected as evolutionary markers. However, the amount of genetic diversity found in the genes selected in that original study was unexpectedly low, which led to the hypothesis that TB had spread only recently from a unique precursor. Indeed, the rate of genetically neutral synonymous mutations (sSNP) was shown to be as low as 1/10,000 whereas the rate of non-synonymous mutations (nsSNP) outnumbered sSNPs by almost 2 to 1 (Sreevatsan 1997).

As for spoligo- and MIRU typing, at first glance it seems reasonable to consider these markers as neutral. No evident role for the DR locus, a member of CRISPR sequences, has been proven yet; however, there is an increased interest in CRISPR and the CRISPR-associated genes *cas*, which may mean to the bacterial world what silencing RNAs means for the eukaryotic world (Makarova 2006). Apart from the *sen*X3-*reg*X3 double component system, which was presumably involved in virulence, the function of MIRU loci remains poorly investigated (Parish 2003). In all cases, the phylogenetical information content obtained by studying the DR and the VNTR loci was previously shown to be rich (D. Falush 2003 - Prague, European Concerted Action Meeting, unpublished data).

2.6. Why repeated sequences were so useful at the beginning

The description of repeated sequences goes back to the early age of molecular biology (Britten 1968). Their role in the selection of new vital functions in life is indeed of paramount importance for genetic evolution (Britten 2005). In the *M. tuberculosis* complex, repetitive DNA sequences were used as probes and showed to be useful for fingerprinting strains in epidemiological studies (Eisenach 1988). Shortly after the characterization of the insertion sequence IS*6110* (Thierry 1990), an international consensus method IS*6110* RFLP was adopted almost concomitantly to the World Health Organization declaration of TB as a public health emergency (van Embden 1993). IS*6110* RFLP changed the traditional belief that no more than 10 % of TB cases were due to recent transmission, and sparked a new hope for disease eradication by contributing to the adequate surveillance and prevention of TB transmission (Alland 1994, Small 1994). For diverse reasons, however, the use of IS*6110* was of little help in solving the phylogenetic structure of the *M. tuberculosis* complex because it turned out to be a poor phylogenetic marker (Fleischmann 2002). A rapidly emerging issue was that IS*6110* was ineffective in a large part of the world, including South-East Asia (Fomukong 1994). Another insertion sequence, IS*1081*, was also suggested as an interesting potential phylogenetic marker; however, its generalized use in *M. tuberculosis* complex population genetics was also hampered, among other reasons, by the RFLP format (van Soolingen 1997, Park 2000).

2.7. Regions of differences (RDs) and SNPs in *M. tuberculosis*

One approach to understanding the molecular evolution of the *M. tuberculosis* complex and looking for virulence genes is to identify regions of difference (RD) between *M. tuberculosis* complex genomes (Inwald 2003) or to look for Single Nucleotide Polymorphisms (SNPs). Substractive genomic hybridization was initially used to identify three distinct genomic regions between virulent *M. bovis*, *M. tuberculosis,* and the avirulent *M. bovis* bacille Calmette-Guérin (BCG) strain, designated respectively as RD1, RD2, and RD3 (Mahairas 1996). One of these regions, RD1, was shown to contain important virulence genes including the two immunodominant T-cell antigens ESAT6 and culture filtrate protein 10 (CFP10) (Pym 2002). In another study (Gordon 1999), restriction-digested bacterial artificial chromosome (BAC) arrays of H37Rv strain were used to reveal the presence of 10 regions of difference between *M. tuberculosis* and *M. bovis* (RD1 to 10); 7 of which (RD4-RD10) were deleted in *M. bovis*. The deletion pattern of *M. africanum* is closer to that of *M. tuberculosis* than to the pattern of *M. bovis* (Gordon 1999).

Brosch *et al*. analyzed the distribution of 20 variable regions resulting from insertion-deletion events in the genome of the tubercle bacilli in one hundred strains belonging to all sub-species of the *M. tuberculosis* complex (Brosch 2002). The authors showed that the majority of these polymorphisms resulted from ancient irreversible genetic events in common progenitor cells, the so-called Unique Event Polymorphisms (UEP). Based on the presence or absence of an *M. tuberculosis* specific deletion 1 (TbD1, a 2 kb sequence), *M. tuberculosis* can be divided into "ancient" TbD1 positive and "modern" TbD1 negative strains. This classification superimposes well with the previous principal genetic group (PGG) classification (Sreevatsan 1997); however, only two groups of strains, the EAI and the *M. africanum* strains are TbD1 positive. The RD9 deletion identifies an evolutionary lineage represented by *M. africanum*, *M. microti* and *M. bovis* that diverged from the progenitor of the present *M. tuberculosis* strains before TbD1 occurred (Brosch 2002). These findings contradict the long-held belief that *M. tuberculosis* evolved from a precursor *of M. bovis*, suggesting a new evolutionary scenario of the *M. tuberculosis* complex. Since *M. canettii* and other ancestral *M. tuberculosis* complex strains lack none of these regions, they are supposed to be direct descendants of the tubercle bacilli that existed before the *M. africanum-M. bovis* lineage separated from the *M. tuberculosis* lineage (Brosch 2002). This scenario was confirmed in a follow-up study in which *in silico* and macroarray based hybridization experiments confirmed the existence of a core set of 219 conserved genes shared by *M. leprae* and *M. tuberculosis.* Among these new phylogenetical markers is the *pks* 15/1 gene, which encodes one of the polyketide synthase enzymes required for the lipid metabolism of cell wall building. All modern strains show a 7-base pair (bp) frameshift deletion in this gene that induces a knock-out of the enzyme. *M. canettii,* most PGG1 ancestral EAI, and Beijing strains add two amino acids that do not interfere with *pks* function, whereas strains in the *M. bovis* lineage bear a 6-bp DNA deletion that involves deletion of these two extra amino acids (Constant 2002).

Three recent studies provide landmarks in TB molecular and phylogenetic population studies. The first one suggests the existence of six phylogeographical lineages, each associated with specific sympatric human populations (Gagneux 2006). These observations show that mycobacterial lineages are adapted to particular human populations. Whether these results are considered from either a "splitter" or from a "gatherer" perspective, they endorse the idea that there are probably just a small number of founding genogroups of the *M. tuberculosis* complex. Also, these results support previous results on *M. tuberculosis* complex genetic diversity and our hypothesis that *M. tuberculosis* complex is an ancient pathogen that co-evolved with its hosts (Sola 2001a, 2001b, Sebban 2002).

Two SNP-population-based phylogenies also provided similar results, i.e. a limited number of *M. tuberculosis* complex phylogeographical genogroups (Figure 2-4). According to a study led by Musser's group, eight deeply branching genetic groups (I to VIII) were found; however, this was still not representative of the worldwide genetic diversity of *M. tuberculosis* because of a biased sampling, e.g., lack of Central Asian (CAS) strains (Gutacker 2002). A second study corrected this bias by creating one new subgroup for the CAS lineage (Gutacker 2006). This lineage is close to the root, which suggests that the Indian subcontinent played a major role in TB evolution and expansion.

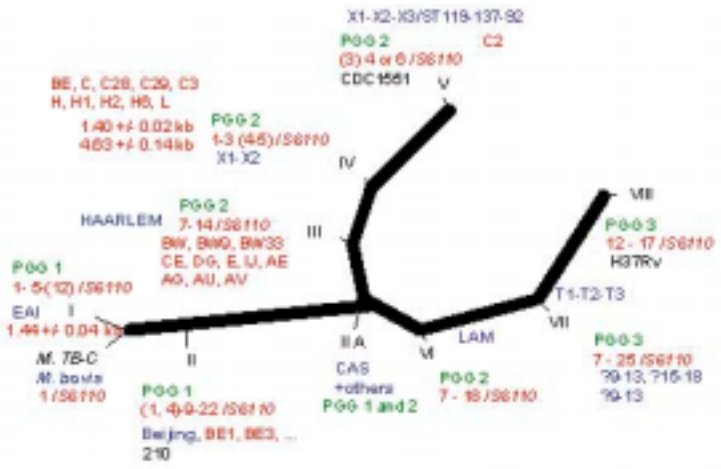
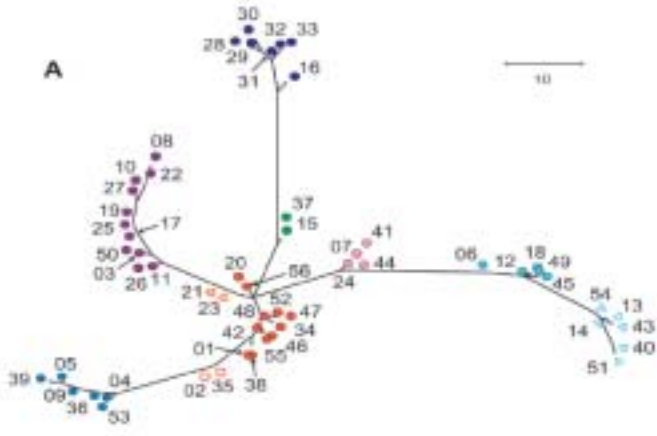


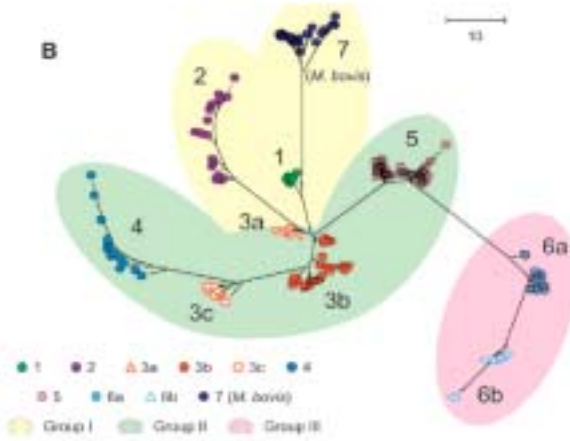
Figure 2-4 Phylogenetic tree obtained on SNPs, adapted from Gutacker *et al*. 2006 and supplemental data. In blue: spoligotyping-based nomenclature or characteristics. In red: IS*6110*-based clade nomenclature with some characteristics IS*6110* copy number or molecular weight data. In green: Musser's principal genetic group (Sreevatsan 1997). In black: SNP-based designation of clades with some characteristics strains (CDC1551, H37Rv, strain 210).

Similar results were obtained independently by Alland *et al*., reinforcing the idea that unrelated lineages may acquire the same number of IS*6110* by homoplasia (Alland 2003). The same group recently analyzed 212 SNPs in correlation with MIRU and spoligotyping on a worldwide representative collection of clinical isolates. Their results are illustrated in Figures 2-5 (A to C). The *M. tuberculosis* complex tree presented four main branches containing six SNP cluster groups (SCG1 to SCG6) and five subgroups as depicted in Figure 2-5 B (Filliol 2006). These results provide good congruence with spoligotyping and, to a lesser extent, with MIRU12,

endorsing the latest genetic diversity studies on spoligotyping (Brudey 2006). Still, it can be argued that in both SNP-based studies, identical bias could have been introduced since the SNPs analyzed in both cases were selected based on the four *M. tuberculosis* complex genome sequences available to date: *M. tuberculosis* strains 210, CDC1551, H37Rv and *M. bovis* strain AF2122.

Figure 2-5, A to C: (From Filliol *et al.* 2006 J. Bacteriol., reproduced with permission). A: a distance-based neighbor-joining tree on 159 sSNPs resolves the 219 *M. tuberculosis* complex isolates in 56 sequence types (ST). STs are indicated by a dot with numerical value and color code for SNP Cluster Group (SCG) belonging. B: Model-based neighbor-joining tree based on a data set with 212 SNPs, which resolves 327 *M. tuberculosis* complex isolates into 182 ST with identical cluster (compare with A). SNP Cluster Groups are indicated by colors. Principal Genetic Groups (1 to 3) are also highlighted. C: distribution of the spoligotype clades on the SNP-based phylogeny.




B
7
(M. bovis)
2
1
5
4
3a
6a
3c
3b
6b
Group I
Group II
Group III

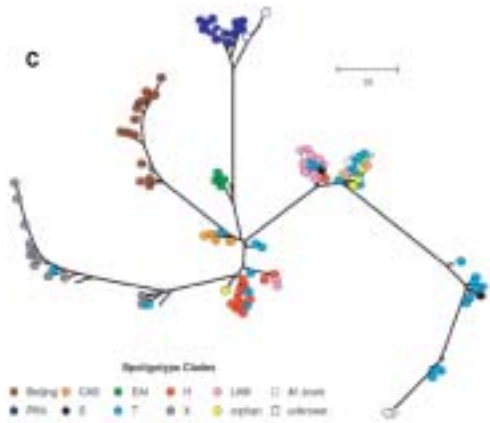

C

Table 2-1 provides a nomenclature correlation between *M. tuberculosis* complex groups defined by spoligotyping and those defined by sSNPs. As shown in this table, the most ancient clade, EAI defines SCG 1 or sSNP-I according to Alland's or to Musser's designation, respectively. SCG 2 and sSNP-II define the Beijing lineage. SCG 3a or sSNP-IIa defines the CAS or Delhi genogroup. SCG 3b or sSNP-III defines the Haarlem family of strains. SCG 3c and SCG 4, or sSNP-IV and sSNP-V, define the "IS*6110* European low-banders" or X genogroup (Sebban 2002, Dale 2003, Warren 2004). SCG 5 or sSNP-VI is mainly constituted by the Latin American and Mediterranean (LAM) genogroup (Sola 2001a). SCG 6a and SCG 6b (sSNP-VII and sSNP-VIII) define the poorly characterized Principal Genetic group 3 lineage that also includes some ill-defined T genotypes (Filliol 2002). Last but not least, SCG 7 defines the bovine and seal *M. tuberculosis* complex subspecies whereas no counterpart is provided in Musser's classification (Filliol 2006).

Table 2-1: Comparison of spoligotype and SNP terminology

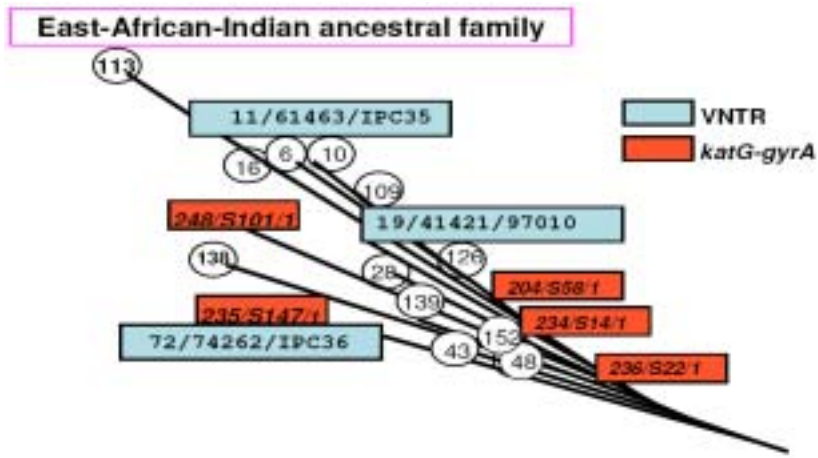
PGG (Sreevatsan 1997)	**Spoligotyping-based (Filliol 2003)**	**SCG-based (Filliol 2006)**	**SNP-based (Gutacker 2006)**
PGG	EAI	SCG 1	sSNP-I
PGG1	Beijing	SCG 2	sSNP-II
PGG1	CAS	SCG 3a	sSNP-IIA
PGG 1	Bovis	SCG 7	*M. tuberculosis* complex
PGG2	Haarlem	SCG 3b	sSNP-III
PGG2	X1	SCG 3c	sSNP-IV
PGG2	X1,X2,X3	SCG 4	sSNP-V
PGG2	LAM	SCG 5	sSNP-VI
PGG3	T (Miscellaneous)	SCG 6	sSNP-VII
			sSNP-VIII

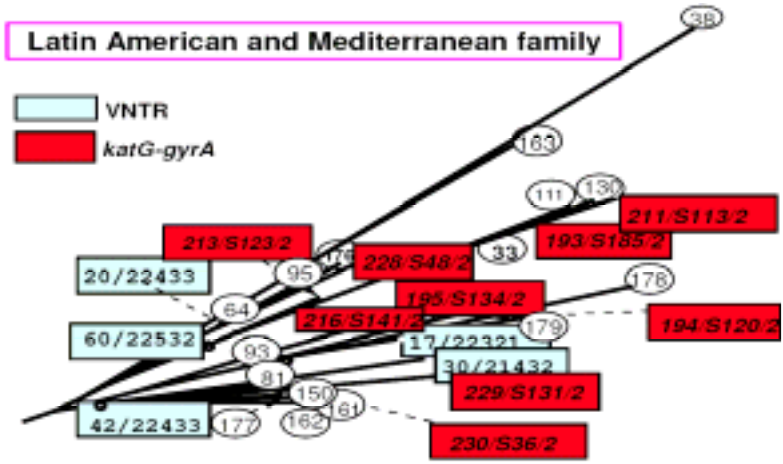
PGG = Principal Genetic Group
SCG = SNP cluster group
SNP = Single nucleotide polymorphism
EAI = East African Indian
CAS = Central Asian (or Delhi)l

2.8. Looking for congruence between polymorphic markers

The concept of molecular clock, attributed to Zuckerkandl and Pauling in 1962, was originally based on hemoglobin evolution and later generalized to DNA evolution (Zuckerkandl 1987). As for *M. tuberculosis*, we are dealing with polymorphic markers, i.e. repeated sequences, which are physically linked to the chromosome and therefore transmitted together with it. Concomitantly, these sequences are evolving at their own pace and hence possess more than one molecular clock. Although the combination of various molecular clocks of different paces in a single analysis may be criticized (Wilson 2003), this approach was used successfully in the past to detect the EAI and the LAM clades by observing congruence between spoligotyping and VNTR data (Figure 2-6, extracted from Sola 2001b).

Figure 2-6 : Close-up on a spoligotyping-based neighbor-joining (NJ) phylogenetical tree, built on SpolDB2 database showing the EAI and LAM branches. The superimposition of spoligotyping, VNTR and Principal Genetic grouping shows congruence between various markers (extracted from Sola 2001b) in blue boxes: Spoligotyping shared-type n°/VNTR allele: ETR-A to E from left to right). In red boxes: SpolDB-shared-type n°/Soini's spoligotyping number (see Soini et al. 2000)/Principal Genetic Group (see Sreevatsan 1997). In the blue boxes of the upper figure, the third number is the strain identification number. In circles: spoligotyping shared-type number.





IS*6110* RFLP was recognized very early to evolve faster than spoligotype since more RFLP than spoligo genotypes are present when a single set is analyzed (Kremer 1999). The mutation rate of IS*6110* was estimated recently to be 0.287 per genome per year for a strain with a typical number of 10 copies (Rosenberg 2003). Using the infinite allele model and the same set of data (Kremer 1999), the relative mutation rate of spoligotype is calculated to be 13.5 % of the rate of IS*6110* (Tanaka 2005). This corresponds to a spoligotype mutation rate of around 0.039 events per year. A more complex model was recently developed, which assumes that the mutation rate of a given spoligotype is proportional to the number of spacer units present in the DR region. This new model allows the detection of emerging strains of *M. tuberculosis* (Tanaka 2006).

Population bottlenecks are important in biology since they create genetic conditions that favor founder effect and speciation. Among many bottleneck hypotheses, the one ascribed to the late Pleistocene is very attractive. It involves volcanic winter and differentiation of modern humans at a time comprised between 50,000 and 15,000-25,000 years ago (Ambrose 1998). These events may have created environmental conditions favoring the spread of *M. tuberculosis*. We may hypothesize on the global spreading of a single clone (Kapur 1994), or of a limited number of clones, based on the expansion of the surviving re-founders, preserved in various small refuges located in tropical areas (Ambrose 1998). The ample human genetic diversity observed today in Africa (as well as the apparently ample *M. tuberculosis* genetic diversity) may be due either to a longer evolutionary period, or to the pres-

ervation of such ample diversity in this continent during the bottleneck event. Consequently, for the *M. tuberculosis* complex, we can hypothesize that the high genetic diversity observed in "*M. prototuberculosis*" could be a remnant of this bottleneck event, with a strong resilience and hence a high preservation of the previous genetic diversity inside these tropical refuges. This ecological perspective is also supported by data suggesting that human beings migrated back to Africa after the demographic expansion into the South-East Asian peninsula (Cruciani 2002). Thus, if demographic and epidemic factors are considered in addition to evolutionary and genetic factors, the modern tubercle bacilli are more likely to find their origin in India or South-East Asia rather than in Africa. The fact that the TbD1 positive East-African-Indian strains, which are likely to have disseminated when adequate demographical conditions were fulfilled, are genetically the closest to the *M. canetti*- "*M. prototuberculosis*" strains argues in favor of this hypothesis.

Given the astonishingly reduced SNP diversity observed initially in the *M. tuberculosis* complex (Sreevatsan 1997, Musser 2000), the bottleneck hypothesis is seducing. However, the 15,000- to 25,000-year time frame was calculated by computation of synonymous mutation rates based on *Escherichia coli* and *Salmonella* divergence, i.e. based on a uniform calibration rate for nucleotide substitution (the basic molecular clock). This choice of independency from growth rate (doubling time) and other parameters, such as mutation rate and population size, may be criticized. The doubling time of *E. coli* is 20 min and that of *M. tuberculosis* is 20 hours. If we logically assume that sSNPs acquisition is related to DNA metabolism, then, a ratio of 60x should be applied to the computation presented in Kapur's paper, thus providing a much larger time-frame (900,000-1,500,000 years) for the presence of *M. tuberculosis* complex bacilli on earth, an hypothesis that is consistent with the latest results obtained on "*M. prototuberculosis*", which shows an unusually high SNP diversity (Gutierrez 2005).

According to a recent multigenic phylogenetic approach, the speciation process in mycobacteria might have been progressive and relatively homogenous across the whole genome (Devulder 2005). When comparing substitution rates of fast and slow growing mycobacteria by means of a relative rate test, non-significant differences were observed. These findings suggest that the two groups evolved at the same rate. In other words, the evolutionary rate does not necessarily correlate to the number of generations. This framework fits with the strictly clonal evolution of *M. tuberculosis* and the co-evolution hypothesis that suggests adaptation between particular mycobacterial lineages and particular human populations (Supply 2003, Gagneux 2006). However more recent genetic studies using SNPs analysis suggests that some genes such as the ones coding for the PE-PGRS and PPE proteins that

have the motifs Pro-Glu (PE) and Pro-Pro-Glu (PPE), thought to be critical in host-pathogen interactions, are prone to recombination and gene conversion events (Karboul 2006, Liu 2006).

2.9. Main lineages within the *M. tuberculosis* species

Within the scope of this chapter is the description of the results of the molecular population approach that allowed the definition of genetically homogenous clusters of *M. tuberculosis* complex, which are now shown to be preferentially linked to some human hosts (Brudey 2006, Gagneux 2006). Table 2-2 provides the latest description of statistically, epidemiologically or phylogeographically relevant clonal complexes of the *M. tuberculosis* complex based on spoligotyping signatures described in the SpolDB4 database (a high resolution image can be downloaded at: http://www.biomedcentral.com/1471-2180/6/23/figure/F1?highres=y; from Brudey 2006).

2.9.1. Principal lineages of the Genetic group 1

2.9.1.1. The East African-Indian (EAI) lineage

This lineage was first described in Guinea-Bissau (Källenius 1999) and was shown to be frequent in South-East Asia, India, and East Africa (Kremer 1999). This group of strains is characterized by a low number of IS*6110* copies. A subgroup of these strains harboring a single copy of IS*6110* was shown to be widespread in Malaysia, Tanzania, and Oman (Fomukong 1994). In combined datasets (i.e. pooled datasets characterized by one or more methods), this lineage demonstrated congruence between spoligotypes (absence of spacers 29-32, presence of spacer 33, absence of spacer 34), VNTR [exact tandem repeat A (ETR-A) allele ≥4], *kat*G-*gyr*A grouping (Group 1), and later the presence of the TbD1 sequence (Soini 2000, Sola 2001b). More recently, the presence of an *oxyR* C37T transition was shown to be specific to the lineage (Baker 2004). This lineage was shown to belong to cluster group 1 or Cluster I (Filliol 2006, Gutacker 2006). It harbors a specific region of difference, RD239 and was renamed as Indo-Oceanic in the work of Gagneux *et al.* (Gagneux 2006). It is speculated that this lineage, which is endemic in South-East Asia, South-India, and East-Africa, may have originated in Asia, where TB could have historically found favorable spreading conditions. The Manila family was first identified by Douglas in 1997, and was later thoroughly characterized by the same group (Douglas 2003). This genotype was identified based on the prevalence of clustered strains isolated from Philippino immigrants in the

United States (US) and was only later shown to be prevalent in the Philippines. The Manila family bears ST19 as prototypic spoligo-signature and is actually identical to EAI-2 (Filliol 2002). ST89, which defines the Nonthaburi (Thailand) group of strains, is a derived clone (Namwat 1998). In this family, specific variants have been also described for Vietnam (ST139 or EAI-4), Bangladesh (ST591, ST1898 or EAI-6 and 7) and Madagascar (ST109, EAI-6).

We have no precise idea about the prevalence of the EAI lineage in India and China, although it is evident that this genotype is more specifically linked with South-East Asia and South India than with Northern China. This may be due to differences in civilization and agriculture histories between North and South China (Sola 2001b). It is also very difficult to analyze what links these clones may have with strains in the major genetic group 2, given the presence of the spacer 33 in this group of strains (a spacer that is absent in groups 2 and 3). A striking discovery related to these strains was made recently when analyzing medieval human remains discovered in an English parish. TB was confirmed by amplifying multiple *M. tuberculosis* loci and EAI genotypes were apparently identified by spoligotyping (Taylor 1999). Whether these spoligotyping results obtained on medieval remains are reliable or not should be confirmed independently; however, the possibility of the presence of EAI genotypes in 13th century England should not be excluded.

2.9.1.2. The Beijing lineage

The Beijing genotype belongs to the principal genetic group 1 of Sreevatsan, and its specific spoligotype signature (absence of spacer 1-33, presence of spacer 34-43) was discovered in 1995 (van Soolingen 1995). However, a notorious outbreak due to a multidrug resistant clone of one of its offspring (New York W strain) had been characterized earlier, at the beginning of the ’90s (Plikaytis 1994, Bifani 2002). The emergence of this family of related genotypes continues to pose a serious threat to TB control due to its high virulence and frequent association with multidrug resistance. It was hypothesized that this genotype emerged successfully in East Asia due to mass BCG vaccination during the 20th century (van Soolingen 1995, Abebe 2006). However, Beijing should also be considered as a group of variant clones that evolved from a common precursor at an undefined time, maybe during the Genghis Khan reign or before (Mokrousov 2005).

These strains are characterized by the presence of an inverted IS*6110* copy within the DR region, an IS*6110* element at a particular insertion site (within the origin of replication) and one or two IS*6110* copies in a DNA region called NTF (Plikaytis 1994, Kurepina 1998). A characteristic Beijing lineage-defining SNP (G81A in Rv3815c) has been reported by Filliol *et al*. According to SNP analysis, the Beijing

cluster was designated as SCG 2 or sSNP-II (Filliol 2006, Gutacker 2006). Other characteristic sSNPs of the Beijing lineage were described in putative DNA repair genes (Rad 2003).

More recently, new phylogenetically-informative specific LSP markers were found, such as RD105, which is present in all Beijing/W or RD142, RD150 and RD181. It allows a further division of the Beijing lineage into four monophyletic subgroups (Tsolaki 2005). The Beijing lineage was recently renamed as the East Asian Lineage by other authors (Gagneux 2006). Its most frequent VNTR signature is 42435 (Kremer 1999).

Recent evidence points to an early dispersal of the Beijing genotype in correlation to genetic haplotype diversity of the male Y chromosome (i.e. in correlation with human phylogeography). These results suggest that the spreading history of Beijing has a molecular evolutionary history that is much more intricate and more deeply rooted to human history than initially thought. Using the Beijing genotype as a model, and comparing its phylogeography to Y-chromosome-based phylogeography, Mokrousov *et al.* hypothesized that two events shaped the early history of this genotype: (1) its upper Paleolithic origin in the *Homo sapiens sapiens* K-M9 cluster in central Asia, and (2) a primary dispersal of the secondary Beijing NTF:: IS*6110* lineage by Proto-Sino-Tibetan farmers within East-Asia (human O-M214/M122 haplogroup) (Mokrousov 2005).

2.9.1.3. The Central-Asian (CAS) or Delhi lineage

The presence in India of a specific lineage of the *M. tuberculosis* complex was concomitantly and independently reported by two different groups using IS*6110* RFLP and spoligotyping, respectively (Bhanu 2002, Filliol 2003). This lineage was also shown to be endemic in Sudan, other sub-Saharan countries and Pakistan (Brudey 2006). Using IS*6110* RFLP, the Delhi lineage shows a characteristic band pair in the high molecular weight region (12.1 and 10.1 kilobase pairs) and its specific spoligotype signature is formed in the absence of spacers 4-27 and 23-34. This spoligo-signature shows numerous variants and several subgroups such as CAS1-Kili (for Kilimanjaro) and CAS1-Dar (for Dar-es-Salaam), which have already been defined on the basis of new spoligotype-signatures that are specific for each new clonal complex (Mc Hugh 2005, Eldholm 2006). Still, more results using other polymorphic markers should complement these data. VNTR signatures of *M. tuberculosis* complex clinical isolates from South-Asian immigrants in London and native patients in Rawalpindi, Pakistan, were identical (allele combination 42235) and correlated with the CAS spoligotype (Gascoyne-Binzi 2002, Brudey unpublished results).

This genotype family could be the ancestor of the Beijing family since it clusters close to Beijing when analyzed by a combination of MIRU, spoligotyping and VNTR (Sola 2003). In India, its frequency varies from one region to another: it is more prevalent in the North than in the South, where the EAI family predominates (Suresh 2006). An outbreak strain named CH was recently reported in Leicester, United Kingdom. It belongs to the CAS family and harbors a specific deletion (Rv1519). In broth media, this strain was found to grow more slowly and to be less tolerant to acid and H_2O_2 than two laboratory reference strains, CDC1551 and H37Rv. Nevertheless, its ability to grow in human monocyte-derived macrophages was not impaired. This strain induced more anti-inflammatory IL-10, more IL-6 gene transcription/secretion from monocyte-derived macrophages, and less protective IL-12p40 than CDC1551 and H37Rv strains. Thus, this strain seems to compensate the microbiological attenuation by skewing the innate response toward a phagocyte deactivation. The complementation of Rv1519 reversed its ability to elicit anti-inflammatory IL-10 production by macrophages. These results suggest that the Rv1519 polymorphism confers an immune subverting *M. tuberculosis* phenotype that might contribute to the persistence and outbreak potential of this lineage (Newton 2006).

2.9.2. Lineages belonging to the Principal Genetic groups 2 and 3

2.9.2.1. The Haarlem family

The Haarlem family was described in the Netherlands in 1999 (Kremer 1999). On IS*6110* RFLP, these strains harbor a double band at 1.4 kb. Their spoligotype is characterized by the absence of the spacer 31, which is due to the presence of a second copy of IS*6110* in the DR region (Groenen 1993). Due to an asymmetric insertion within the DR locus, this second IS*6110* copy hinders the detection of spacer 31 (Filliol 2000, Legrand 2001). Three main spoligotype-signatures define the variants H1 to H3 (Filliol 2002). However, many Haarlem clonal complexes may harbor other Haarlem-based spoligo-signatures that are, as yet, poorly characterized. Another characteristic of the Haarlem lineage is the frequent VNTR pattern 33233 (Kremer 1999). The Haarlem family is highly prevalent in Northern Europe. It is present in the Caribbean to a lesser extent and is also prevalent in Central Africa, where it is believed to have been introduced during the European colonization (Filliol 2003). This family, which is highly diverse, merits further studies to better understand its evolutionary history. A SNP in the *mgt* gene of the *M tuberculosis* Haarlem genotype was discovered recently (Alix 2006). More SNPs are expected to be specific of the Haarlem lineage.

2.9.2.2. The Latin American and Mediterranean (LAM) family

The LAM family was defined by the finding of linkage disequilibrium between the absence of spacers 21-24 in the spoligotyping and the presence of an ETR-A allele equal to 2 (Sola 2001b). However, this genotype family is more diverse and its study is more complicated than initially thought. Strains belonging to the LAM3/F11 family and the S/F28 family harbor identical spoligotypes of the shared type ST4, revealing the existence of genetic convergence between spoligotypes (Warren 2002). This phenomenon seems, however, to be rare and highly dependent on the structure of the observed spoligotype. The absence of spacers 21-24 may also have occurred more than once in tubercle bacilli evolution although no genetic evidence has suggested such a convergence event until now. Many sub-motifs - LAM1 to LAM12 - have been suggested according to the latest international spoligotype database project SpolDB4 (Brudey 2006). However, the phylogenetic significance of the common absence of spacers 23-24 has not been demonstrated in this lineage. In this sense, some genotypes that show strong geographical specificity (for example the LAM10-Cameroon or the LAM7-Turkey) were initially labeled as LAM, although there is no evidence of their phylogenetical relation to other LAM spoligo-signatures (Niobe-Eyangoh 2003, Zozio 2005). Recently, a specific deletion designated as RD^{RIO} was shown to be linked to certain LAM spoligo-signatures present in Rio de Janeiro, Brazil (L Lazzarini, R Huard, JL Ho personal communication).

The LAM clade is frequent in Mediterranean countries and its presence in Latin America is supposed to be linked to the Lusitanian-Hispanian colonization of the New World. Conversely, it may have been endemic in Africa and/or in South America, spreading to Europe later. At this stage, we must highlight that paleopathological and ancient DNA data support the existence of TB before the arrival of Spanish settlers to Latin America in the 15^{th} century (Arriaza 1995, Salo 2001).

2.9.2.3. The X family: the European IS*6110* low banders

The X family of strains is defined by two concomitant features, a low number of IS*6110* copies and the absence of spacer 18 in the spoligotyping (Sebban 2002). This latter is indeed an important characteristic common to at least three spoligotype shared types: ST119, ST137, and ST92. Both characteristics are present in the CDC1551 strain, which was once suggested to be highly virulent. The X family was also the first group identified in Guadeloupe (Sola 1997) and the French Polynesia (Torrea 1995). Specific epidemic variants of this genotype family were described in South Africa (Streicher 2004). The absence of spacer 18 bears phylogenetical significance because it is improbable that this spacer was deleted more than

once in the evolution of *M. tuberculosis*. The distribution of the X family appears to be linked to Anglo-Saxon countries (Dale 2003). It is also highly prevalent in South Africa and to a lesser extent in the Caribbean. Currently, it is only poorly documented in India. The strong presence of this genotype family in Mexico could be explained by its close proximity to the USA.

2.9.2.4. The T families and others

The «ill-defined» T group is characterized «by default». It includes strains that miss spacers 33-36 and can hardly be classified in other groups. This is a general characteristic of strains belonging to the principal genetic groups 2 and 3, together with the absence of an intact *pks* 15/1 gene (Marmiesse 2004). The presence of intact polyketide synthase genes, active in the synthesis of the specific lipid complex of the *M. tuberculosis* complex is now known to be linked to virulence (Constant 2002). Conversely, the 7 bp frameshift deletion in *pks*15/1 may be considered as a phylogenetical marker specific for the modern *M. tuberculosis* strains (Gagneux 2006) and may define the recently designated Euro-American lineage. It is expected that the combination of spoligotype and improved MIRU signatures will be the best way to precisely define epidemiological clonal complexes (Supply 2006). Alternatively, RDs and/or SNPs may also improve the taxonomic definition of these clones.

Table 2-2 shows the nomenclature correspondence between the main spoligotyping-based *M. tuberculosis* complex lineages and those recently described by MLST-SNPs (Baker 2004) and LSP (Gagneux 2006). As shown, spoligotyping appears to be more discriminative than the other two typing systems since it is able to resolve clinical isolates within the branch of the modern strains that are not solved by LSP. Specific RDs are described for many individual spoligotype-signatures; however, no Table is yet available for LSP and/or SNP synthetic correspondence with spoligotype.

Even if there is consensus in the fact that the main branches of the genetic tree of the *M. tuberculosis* complex have now been found, many uncertainties still remain with regard to the chronology of the evolution of the *M. tuberculosis* complex. For example, Gagneux *et al*. suggest that West African 2 diverged from an ancestral branch of *M. bovis,* whereas West African 1, characterized by a deletion of RD711, did not (Gagneux 2006).

Table 2-2 Comparison of spoligotype, Multi Locus Sequence Typing (MLST) and Large Sequence Polymorphism (LSP) nomenclature

Spoligotyping-based (Filliol 2003)	MLST (Baker 2004)	LSP (Gagneux 2006)	Comment
East-African-Indian (EAI)	IV	Indo-Oceanic	Prevalent in South East Asia, East Africa and South India
Beijing	I	East-Asian	Prevalent in China, Japan, South East Asia, Russia
Central-Asian (CAS)	III	East-African-Indian	Prevalent in North India, Pakistan, Libya, Sudan
X, Haarlem, LAM	II	Euro-American	Ubiquitous
M. africanum	NA	West African 1	Nigeria, Ghana
M. africanum	NA	West African 2	Senegal, Gambia

Recent results in our laboratory have shown that, in certain cases, it should be possible to reconstruct the past evolutionary history of some modern clones of the *M. tuberculosis* complex belonging to the principal groups 2 and 3. As an example, a striking identity was found recently between the MIRU typing results of the main LAM7-Turkey clonal complex (Zozio 2005) and the Japanese group T3-OSA (Ano 2006) (Millet *et al*. unpublished results). The meaning of this identity is under investigation and there is no reason to believe that it is due to convergence. Similarly, an endemic clone found in Nunavik (Nguyen 2003) was shown to be related to a clone found to be prevalent in central Europe (Poland and Germany) (Sola *et al.* unpublished results). Once again, we are trying to analyze how and when such movement of strains took place and whether they are representative of a deeply rooted anthropological structure or from modern outbreaks.

2.10. When did the bovine-human switch of *M. tuberculosis* take place?

The question of the molecular evolution of *M. bovis* provides an interesting framework for comparison with that of *M. tuberculosis* (Smith 2006a). In particular, Smith *et al*. discuss in detail how population bottlenecks and selective sweeps deeply affect the population structure of strictly clonal pathogens, such as members of the *M. tuberculosis* complex. Using the genetic diversity of *M. bovis* in the United Kingdom as a model, these authors demonstrate that all *M. bovis* genotypes derive from a single clonal complex that is likely to have emerged as a result of the actions of bovine TB control programs, which have been in force for the last 100 years. These authors also suggest that comparative genomics between two selected genomes that have gone through very different selection pressures (H37Rv and *M.*

bovis AF2122) may have wrongly suggested that *M. bovis* is an offspring clone of *M. tuberculosis*. As Brosch *et al*. identified deletions in *M. bovis* by comparing it with the only *M. tuberculosis* chromosome sequence available at that time, it was inevitable to conclude that *M. bovis* was the terminal group in the lineage (Smith 2006a). The assumption that the RD9-deleted lineage (including *M. bovis*) descended from an *M. tuberculosis*–like ancestor also implies, by parsimony, that the most recent common ancestor of these strains was adapted to humans. The exact host-association of *M. africanum* subtype I strains has not been examined so far. There is some evidence that *M. africanum*, which is less virulent than other *M. tuberculosis* complex genotypes, is currently extinct in settings where it was the most prevalent strain only three decades ago. Instead, it is being replaced by imported, more virulent genotypes (V. Vincent, unpublished results). The genetic susceptibility of the indigenous African population to TB during World War I is a well-known fact which supports the idea that TB caused by a more virulent genotype evokes a different, acute and even fatal disease, very different from that produced by *M. africanum*.

2.11. Comparative genomics and evolution of tubercle bacilli

The wealth of completed genome sequences, the development of microarray technology, and the decreasing cost of sequencing have enabled scientists to thoroughly study the significance of strain to strain variation in bacteria such as *Streptococcus agalactiae* and to define the "pan-genome" concept (Tettelin 2005). According to this concept, any species is made up of a common and a strain-specific genetic pool. Depending on the population structure of the studied organism and on the levels of lateral gene transfer, the relative part of these two pools may vary significantly. The **core genome** contains genes present in all strains, and the **dispensable genome** contains genes present in two or more strains as well as genes unique to single strains. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome (Medini 2005).

LSP analysis is of particular interest in the *M. tuberculosis* complex, given the low level of sSNPs (Sreevatsan 1997, Kato-Maeda 2001, Alland 2007). Figure 2-7 shows the non-randomness of deletions in the 16 clinical isolates that were tested by microarray against the H37Rv genome. Some isolates contained unique deletions whereas other deletions were shared by many isolates. This study was extended to 100 different and unique IS*6110* RFLP types representing the global genetic diversity of the *M. tuberculosis* complex observed in San Francisco over

12 years (Tsolaki 2004). LSP size varied between 105 and 11,985 bp, with eight deleted sequences larger than 5,000 bp. LSPs tend to occur in genomic regions that are prone to repeated insertion-deletion events and may be responsible of a high degree of genomic variation in the *M. tuberculosis* complex (Alland 2007). Chapter 4 provides an exhaustive review on the comparative genomics of members of the *M. tuberculosis* complex.

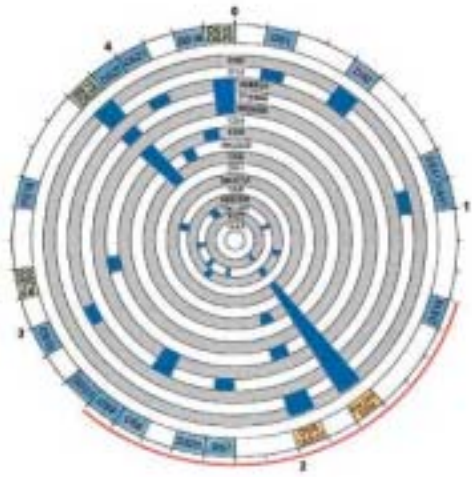


Figure 2-7: Circular map of genomic deletions among *M. tuberculosis* showing that the pattern of deletions differs between clones and is not spatially random. The outer numbers show the scale in mega base pairs (O=replication origin). In blue: genomic locations of deleted sequences. The outer circle summarizes the sum of all detected deletions. Color code (blue, orange, green) is linked to number of deletions (respectively 1, 2 and 3 deletions). The thin red line spans the genomic region of the genome where the number of deletions detected is greater than expected by chance alone. CDC1551 appears as the third ring on this picture. (Reproduced with permission from M. Kato-Maeda and P. Small)

2.12. Short-term evolutionary markers and database building

There are also ongoing debates about the true status of *"M. prototuberculosis"* (Gutierrez 2005). Whereas some consider *"M. prototuberculosis"* to be the most likely common ancestor to all *M. tuberculosis* complex members (Brisse 2006), others do not believe in the fact that these smooth variants of the tubercle bacilli are the true ancestors of today's tubercle bacilli (Smith 2006b). According to Smith, the computation providing a 3 million-year time frame is not reliable and there is no reason to believe that "*M. prototuberculosis*" is a more likely ancestor to the *M.*

tuberculosis complex than any animal pathogen still to be characterized. There is agreement, however, that the gene mosaicism found in *"M. prototuberculosis"* is real. Also, it is widely acknowledged that further studies on the genetic diversity of *"M. prototuberculosis"* will allow light to be shed on lateral genetic transfer and homologous recombination events in the *M. tuberculosis* complex.

Research on the molecular evolution of the *M. tuberculosis* complex is today addressed by exploiting multiple markers such as the DR locus, insertion sequences, deletion regions, mini-satellites, and SNPs, etc. However, in order to data-mine these large polymorphism databases better, newer methods of data analysis are needed in order to discover intelligible rules and to eliminate noisy data. Simplified decision rules are also needed to distinguish emerging pathogenic clones from those in extinction or from others reflecting ongoing TB transmission. A practical consequence of such studies would be a simplification of typing methods, which in turn, would result in a reduction of experimental constraints and an increase in the number of samples processed. At the Institut Pasteur of Guadeloupe, a new version of the spoligotyping database is currently incorporating MIRU-VNTR alleles and will be released for web-based consultation in 2007. In the future, similar websites will add new markers, allowing the performance of combined searches, including country of isolation, country of origin and ethnicity of the patient, multiple genotyping data, as well as a fine analysis of their geographical distribution. Further links of such databases to geographic information systems (GIS) for real-time map construction and clinical expression of the disease might help to shed new light on a stable association between populations of tubercle bacilli and their human hosts over time and across environments, as well as providing brand new tools to tackle the multifactorial nature of the variable clinical expression of the disease.

2.13. Conclusion and Perspectives

The description of the main branches of phylogeographically specific *M. tuberculosis* clonal complexes and the incipient unraveling of the molecular evolution of the *M. tuberculosis* complex took very long and there are reasons to believe that the task has just started. Some of the reasons are to be found within the complexity of the problem itself. A likely ancient TB pathogen may have had the time to create a large number of population-adapted genetic variants. Other challenges may lie in the slow development of efficient methods to characterize the intra-species genetic diversity of the *M. tuberculosis* complex. Also, we may invoke the recent introduction of new concepts, such as statistical phylogeography, whose application to TB will require the construction of an adequate dataset and even more time for the

requisite reconstruction (Knowles 2004). However, the increasing human mobility worldwide is expected to blur the picture of the history of spread of the *M. tuberculosis* complex.

Lastly, a more precise understanding of the evolutionary genetic network of all *M. tuberculosis* complex clonal complexes may also emerge thanks to new studies using the recently standardized MIRU format (Supply 2006). Figure 2-8 illustrates the minimum spanning tree approach, built on polymorphisms of 24 MIRUs, found in a cosmopolitan sample including "*M. prototuberculosis*" isolates. The dotted lines represent some doubtful links (for example, the ancestral position of Beijing, relatively to CAS and EAI is totally speculative since this type of graph does not represent phylogenetical links).

The longer a clone takes to evolve, the more extensive the observed genetic diversity will be. In view of the assumedly ancient origin of TB, much work remains to be done to unravel the true genealogy of the numerous clonal complexes of the *M. tuberculosis* complex that have been described so far. Many others remain to be discovered since the sampling is still very small compared to the extent of diversity that is likely to exist.

Most of the scientific contributions reviewed in this chapter find an echo best translated by Douglas Young's concluding remarks in the lecture "Ten years of research progress and what's to come" (Young 2003): "*Armed with powerful new molecular tools and renewed momentum, laboratory-based researchers are beginning to tackle the fundamental questions of persistence and pathogenesis of human TB that have frustrated previous generations. Progress in fundamental understanding of disease process poses the exciting challenge of translating new ideas into practical tools that will assist in the global control of TB*". It is quite satisfying to see that the research conducted in the last 12 years is clearly advancing towards a better understanding of the tubercle bacillus and its interaction with the host, the mechanisms of pathogenicity involved, and the co-evolution of the bacterium and its host through time and space.

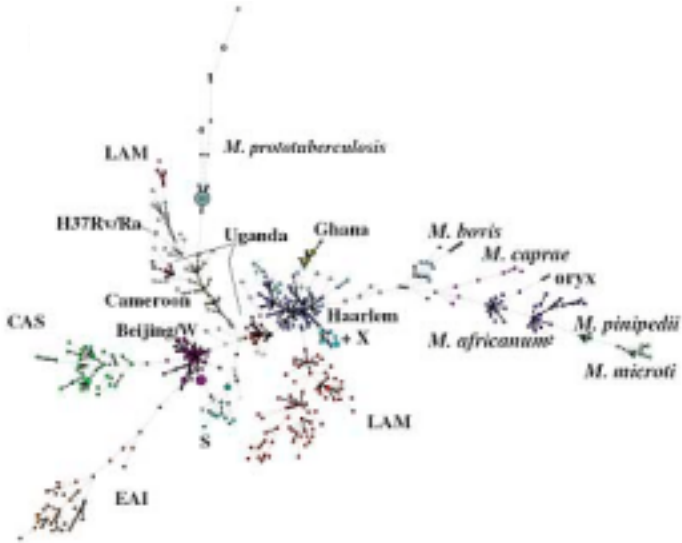


Figure 2-8: Minimum spanning tree based on MIRU-VNTR relationships among tubercle bacilli. Circles correspond to the different types identified by the set of 24 loci among the 494 *M. tuberculosis* isolates from cosmopolitan origins, and 35 *"M. prototuberculosis"*. The corresponding species names and spoligotype family designations (except T types) are indicated. Linkage by a single, double, or triple locus variation is boldfaced. EAI = East-African Indian (Indo-Oceanic in Gagneux's 2006 terminology), CAS = Central Asian (East-African-Indian in Gagneux's 2006 terminology), Beijing/W (East-Asian in Gagneux's 2006 terminology) LAM = Latino-American and Mediterranean, X = European IS*6110* low-banders, S = Sicily-Sardinia clade (all these clades are designated as Euro-American lineages in Gagneux's 2006 terminology) (Reproduced from Supply 2006 with authorization)

References

1. Abebe F, Bjune G. The emergence of Beijing family genotypes of *Mycobacterium tuberculosis* and low-level protection by bacille Calmette-Guerin (BCG) vaccines: is there a link? Clin Exp Immunol 2006; 145: 389-97.
2. Alix E, Godreuil S, Blanc-Potard AB. Identification of a Haarlem genotype-specific single nucleotide polymorphism in the mgtC virulence gene of *Mycobacterium tuberculosis.* J Clin Microbiol 2006; 44: 2093-8.
3. Alland D, Kalkut GE, Moss AR, et al. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med 1994; 330: 1710-6.
4. Alland D, Whittam TS, Murray MB, et al. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J Bacteriol 2003; 185: 3392-9.

5. Alland D, Lacher DW, Hazbon MH, et al. 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 2007; 45: 39-46.
6. Ambrose SH. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. J Hum Evol 1998; 34: 623-51.
7. Ano H, Matsumoto T, Yoshida H, et al. [Molecular epidemiology of tuberculosis by the use of IS*6110* restriction fragment length polymorphism: a study from 2001 to 2003]. Kekkaku 2006; 81: 321-8.
8. Arriaza BT, Salo W, Aufderheide AC, Holcomb TA. Pre-Columbian tuberculosis in northern Chile: molecular and skeletal evidence. Am J Phys Anthropol 1995; 98: 37-45.
9. Avise JC, Arnold J, Ball RM, et al. Intraspecific phylogeography: the mitochondrial DANN bridge between population genetics and systematics. Ann Rev Ecol Syst 1987; 18: 489-522.
10. Baker L, Brown T, Maiden MC, Drobniewski F. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerg Infect Dis 2004; 10: 1568-77.
11. Bates JH, Mitchison DA. Geographic distribution of bacteriophage types of *Mycobacterium tuberculosis*. Am Rev Respir Dis 1969; 100: 189-93.
12. Bhanu NV, van Soolingen D, van Embden JD, Dar L, Pandey RM, Seth P. Predominance of a novel *Mycobacterium tuberculosis* genotype in the Delhi region of India. Tuberculosis (Edinb) 2002; 82: 105-12.
13. Bifani P, Mathema BJ, Kurepina NE et al. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol 2002 ; 10: 45-52.
14. Brisse S, Supply P, Brosch R, Vincent V, Gutierrez MC. A re-evaluation of *"M. prototuberculosis"*: continuing the debate. PLoS Pathog 2006; 2.
15. Britten RJ, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 1968; 161: 529-40.
16. Britten RJ. The majority of human genes have regions repeated in other human genes. Proc Natl Acad Sci U S A 2005; 102: 5466-70.
17. 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.
18. Brudey K, Driscoll JR, Rigouts L, et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 2006; 6: 23.
19. Bruford MW, Bradley DG, Luikart G. DNA markers reveal the complexity of livestock domestication. Nat Rev Genet 2003; 4: 900-10.
20. Cockburn A. The evolution and Eradication of Infectious Diseases. 1963. John Hopkins Press, Baltimore.
21. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998; 393: 537-44.
22. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409: 1007-11.
23. Coleman PG, Perry BD, Woolhouse ME. Endemic stability - a veterinary idea applied to human public health. Lancet 2001; 357: 1284-6.
24. Collins CH, Yates MD, Grange JM. Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. J Hyg (Lond) 1982; 89: 235-42.

25. Constant P, Perez E, Malaga W, et al. Role of the *pks*15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methly esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the *pks*15/1 gene. J Biol Chem 2002; 277: 38148-58.
26. Cruciani F, Santolamazza P, Shen P, et al. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 2002; 70: 1197-214.
27. Dale JW, Al-Ghusein H, Al-Hashmi S, et al. Evolutionary relationships among strains of *Mycobacterium tuberculosis* with few copies of IS*6110*. J Bacteriol 2003; 185: 2555-62.
28. David HL, Jahan MT, Grandry J, Lehmann EH. Numerical taxonomy of *Mycobacterium africanum*. Int J System Bacteriol 1978; 28: 467-72.
29. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J Infect Dis 1999; 180: 1238-44.
30. de Jong BC, Hill PC, Brookes RH, et al. *Mycobacterium africanum* elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. J Infect Dis 2006; 193: 1279-86.
31. Devulder G, Perouse de Montclos M, Flandrois JP. A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. Int J Syst Evol Microbiol 2005; 55: 293-302.
32. Douglas JT, Qian L, Montoya JC, et al. Characterization of the Manila family of *Mycobacterium tuberculosis*. J Clin Microbiol 2003; 41: 2723-6.
33. Eisenach KD, Crawford JT, Bates JH. Repetitive DNA sequences as probes for *Mycobacterium tuberculosis*. J Clin Microbiol 1988; 26: 2240-5.
34. Eldholm V, Matee M, Mfinanga SG, Heun M, Dahle UR. A first insight into the genetic diversity of *Mycobacterium tuberculosis* in Dar es Salaam, Tanzania, assessed by spoligotyping. BMC Microbiol 2006; 6: 76.
35. Fabre M, Koeck JL, Le Fleche P, et al. High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of *hsp65* gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". J Clin Microbiol 2004; 42: 3248-55.
36. Fang Z, Kenna DT, Doig C, et al. Molecular evidence for independent occurrence of IS*6110* insertions at the same sites of the genome of *Mycobacterium tuberculosis* in different clinical isolates. J Bacteriol 2001; 183: 5279-84.
37. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.57c. 1993. Department of Genetics, University of Washington.
38. Filliol I, Sola C, Rastogi N. Detection of a previously unamplified spacer within the DR locus of *Mycobacterium tuberculosis*: epidemiological implications. J Clin Microbiol 2000; 38: 1231-4.
39. Filliol I, Driscoll JR, van Soolingen D, et al. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg Infect Dis 2002; 8: 1347-9.
40. Filliol I, Driscoll JR, van Soolingen D, et al. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol 2003; 41: 1963-70.
41. 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.

42. Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 2002; 184: 5479-90.

43. Fomukong NG, Tang TH, al-Maamary S, et al. Insertion sequence typing of *Mycobacterium tuberculosis*: characterization of a widespread subtype with a single copy of IS6110. Tuber Lung Dis 1994; 75: 435-40.

44. Frothingham R, Meeker-O´Connell WA. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 1998; 144: 1189-96.

45. 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.

46. Gascoyne-Binzi DM, Barlow RE, Essex A, et al. Predominant VNTR family of strains of *Mycobacterium tuberculosis* isolated from South Asian patients. Int J Tuberc Lung Dis 2002; 6: 492-6.

47. 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.

48. Grmek M. Les Maladies à l'aube de la civilisation occidentale. 1994. Payot, Paris.

49. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol Microbiol 1993; 10: 1057-65.

50. 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.

51. Gutacker MM, Mathema B, Soini H, et al. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J Infect Dis 2006; 193: 121-8.

52. Gutierrez MC, Brisse S, Brosch R, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog 2005; 1: e5.

53. Heersma HF, Kremer K, van Embden JD. Computer analysis of IS*6110* RFLP patterns of *Mycobacterium tuberculosis*. Methods Mol Biol 1998; 101: 395-422.

54. Inwald J, Jahans K, Hewinson RG, Gordon SV. Inactivation of the *Mycobacterium bovis* homologue of the polymorphic RD1 gene Rv3879c (Mb3909c) does not affect virulence. Tuberculosis (Edinb) 2003; 83: 387-93.

55. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 2002; 43: 1565-75.

56. Ji YE, Colston MJ, Cox RA. The ribosomal RNA (rrn) operons of fast-growing mycobacteria: primary and secondary structures and their relation to rrn operons of pathogenic slow-growers. Microbiology 1994; 140: 2829-40.

57. Kallenius G, Koivula T, Ghebremichael S, et al. Evolution and clonal traits of *Mycobacterium tuberculosis* complex in Guinea-Bissau. J Clin Microbiol 1999; 37: 3872-8.

58. Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997; 35: 907-14.

59. Kapur V, Whittam TS, Musser JM. Is *Mycobacterium tuberculosis* 15,000 years old? J Infect Dis 1994; 170: 1348-9.

60. Karboul A, Gey van Pittius NC, Namouchi A, et al. Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE-PGRS duplicated gene pair. BMC Evol Biol 2006; 6: 107.

61. Kato-Maeda M, Rhee JT, Gingeras TR, et al. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res 2001; 11: 547-54.

62. Kinsella RJ, Fitzpatrick DA, Creevey CJ, McInerney JO. Fatty acid biosynthesis in *Mycobacterium tuberculosis*: lateral gene transfer, adaptive evolution, and gene duplication. Proc Natl Acad Sci U S A 2003; 100: 10320-5.

63. Knowles LL. The burgeoning field of statistical phylogeography. J Evol Biol 2004; 17: 1-10.

64. Kremer K, van Soolingen D, Frothingham R, et al. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J Clin Microbiol 1999; 37: 2607-18.

65. Kurepina NE, Sreevatsan S, Plikaytis BB, et al. Characterization of the phylogenetic distribution and chromosomal insertion sites of five IS*6110* elements in *Mycobacterium tuberculosis*: non-random integration in the dnaA-dnaN region. Tuber Lung Dis 1998; 79: 31-42.

66. Le Dantec C, Winter N, Gicquel B, Vincent V, Picardeau M. Genomic sequence and transcriptional analysis of a 23-kilobase mycobacterial linear plasmid: evidence for horizontal transfer and identification of plasmid maintenance systems. J Bacteriol 2001; 183: 2157-64.

67. Legrand E, Filliol I, Sola C, Rastogi N. Use of spoligotyping to study the evolution of the direct repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. J Clin Microbiol 2001; 39: 1595-9.

68. Liu X, Gutacker MM, Musser JM, Fu YX. Evidence for recombination in *Mycobacterium tuberculosis*. J Bacteriol 2006; 188: 8169-77.

69. Lopez B, Aguilar D, Orozco H, et al. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. Clin Exp Immunol 2003; 133: 30-7.

70. 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.

71. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 2006; 1: 7.

72. 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(Pt 2): 483-96.

73. McGrath JW. Social networks of disease spread in the lower Illinois valley: a simulation approach. Am J Phys Anthropol 1988; 77: 483-96.

74. McHugh TD, Batt SL, Shorten RJ, Gosling RD, Uiso L, Gillespie SH. *Mycobacterium tuberculosis* lineage: a naming of the parts. Tuberculosis (Edinb) 2005; 85: 127-36.

75. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005; 15: 589-94.

76. Mokrousov I, Ly HM, Otten T, et al. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res 2005; 15: 1357-64.

77. Mostowy S, Behr MA. Comparative genomics in the fight against tuberculosis: diagnostics, epidemiology, and BCG vaccination. Am J Pharmacogenomics 2002; 2: 189-96.

78. Musser JM, Amin A, Ramaswamy S. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics 2000; 155: 7-16.

79. Namwat W, Luangsuk P, Palittapongarnpim P. The genetic diversity of *Mycobacterium tuberculosis* strains in Thailand studied by amplification of DNA segments containing direct repetitive sequences. Int J Tuberc Lung Dis 1998; 2: 153-9.

80. 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: 15594-98.

81. Nguyen D, Proulx JF, Westley J, Thibert L, Dery S, Behr MA. Tuberculosis in the Inuit community of Quebec, Canada. Am J Respir Crit Care Med 2003; 168: 1353-7.

82. Niobe-Eyangoh SN, Kuaban C, Sorlin P, et al. Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon. J Clin Microbiol 2003; 41: 2547-53.

83. 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 2003; 149: 1423-35.

84. Park YK, Bai GH, Kim SJ. Restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolated from countries in the western pacific region. J Clin Microbiol 2000; 38: 191-7.

85. Plikaytis BB, Marden JL, Crawford JT, Woodley CL, Butler WR, Shinnick TM. Multiplex PCR assay specific for the multidrug-resistant strain W of *Mycobacterium tuberculosis*. J Clin Microbiol 1994; 32: 1542-6.

86. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 2005; 151: 653-63.

87. Pym AS, Brodin P, Brosch R, Huerre M, Cole ST. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. Mol Microbiol 2002; 46: 709-17.

88. Rad ME, Bifani P, Martin C, et al. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. Emerg Infect Dis 2003; 9: 838-45.

89. Rosas-Magallanes V, Deschavanne P, Quintana-Murci L, Brosch R, Gicquel B, Neyrolles O. Horizontal transfer of a virulence operon to the ancestor of *Mycobacterium tuberculosis*. Mol Biol Evol 2006; 23: 1129-35.

90. Rosenberg NA, Tsolaki AG, Tanaka MM. Estimating change rates of genetic markers using serial samples: applications to the transposon IS*6110* in *Mycobacterium tuberculosis.* Theor Popul Biol 2003; 63: 347-63.

91. Rothschild BM, Martin LD, Lev G, et al. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. Clin Infect Dis 2001; 33: 305-11.

92. Rothschild BM, Martin LD. Did ice-age bovids spread tuberculosis? Naturwissenschaften 2006a; 93: 565-9.

93. Rothschild BM, Laub R. Hyperdisease in the late Pleistocene: validation of an early 20th century hypothesis. Naturwissenschaften 2006b; 93: 557-64.

94. Salamon H, Segal MR, Ponce de Leon A, Small PM. Accommodating error analysis in comparison and clustering of molecular fingerprints. Emerg Infect Dis 1998; 4: 159-68.

95. Salo WL, Aufderheide AC, Buikstra J, Holcomb TA. Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. Proc Natl Acad Sci U S A 1994; 91: 2091-4.

96. Sebban M, Mokrousov I, Rastogi N, Sola C. A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. Bioinformatics 2002; 18: 235-43.

97. Small PM, Hopewell PC, Singh SP, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med 1994; 330: 1703-9.

98. Smith BD. The emergence of agriculture. 1995. Scientific American Library, New York.

99. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nat Rev Microbiol 2006a; 4: 670-81.

100. Smith NH. A Re-Evaluation of *"M. prototuberculosis"*. PLoS Pathog 2006b; 2.

101. Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J Clin Microbiol 2000; 38: 669-76.

102. Sola C, Horgen L, Goh KS, Rastogi N. Molecular fingerprinting of *Mycobacterium tuberculosis* on a Caribbean island with IS*6110* and DRr probes. J Clin Microbiol 1997; 35: 843-6.

103. Sola C, Devallois A, Horgen L, et al. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. Emerg Infect Dis 1999; 5: 404-14.

104. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. Emerg Infect Dis 2001a; 7: 390-6.

105. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS*1081*, IS*6110*, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. J Mol Evol 2001b; 53: 680-9.

106. Sola C, Filliol I, Legrand E, et al. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect Genet Evol 2003; 3: 125-33.

107. 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.

108. Stahl DA, Urbance JW. The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. J Bacteriol 1990; 172: 116-24.

109. Stinear TP, Jenkin GA, Johnson PD, Davies JK. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. J Bacteriol 2000; 182: 6322-30.

110. Stinear TP, Mve-Obiang A, Small PL, et al. Giant plasmid-encoded polyketide synthases produce the macrolide toxin of *Mycobacterium ulcerans*. Proc Natl Acad Sci U S A 2004; 101: 1345-9.

111. Streicher EM, Warren RM, Kewley C, et al. Genotypic and phenotypic characterization of drug-resistant *Mycobacterium tuberculosis* isolates from rural districts of the Western Cape Province of South Africa. J Clin Microbiol 2004; 42: 891-4.

112. Sula L, Redmond WB, Coster JF, et al. WHO cooperative studies on the phage-typing of mycobacteria. 1. Phage lysis of Czechoslovak and Italian strains of *Mycobacterium tuberculosis*. Bull World Health Organ 1973; 48: 57-63.
113. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol 2001; 39: 3563-71.
114. Supply P, Warren RM, Banuls AL, et al. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol Microbiol 2003; 47: 529-38.
115. Supply P, Allix C, Lesjean S, et al. Proposal for standardization of optimized Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 2006; 44: 4498-510.
116. Suresh N, Singh UB, Arora J, et al. *rpoB* gene sequencing and spoligotyping of multidrug-resistant *Mycobacterium tuberculosis* isolates from India. Infect Genet Evol 2006; 6: 474-83.
117. Swofford DL, Olson FJ. Phylogeny Reconstruction. In: Molecular Systematics. 1990. Sinauer Associates, Sunderland, MA.
118. Swofford DL. PAUP (Phylogeny Analysis using Parsimony) v4.0. 1998. Sinauer Associates, Fitchburg, MA.
119. Tanaka MM, Francis AR. Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. Infect Genet Evol 2005; 5: 35-43.
120. Tanaka MM, Francis AR. Detecting emerging strains of tuberculosis by using spoligotypes. Proc Natl Acad Sci U S A 2006; 103: 15266-71.
121. Tayles N, Buckley HR. Leprosy and tuberculosis in Iron Age Southeast Asia? Am J Phys Anthropol 2004; 125: 239-56.
122. Taylor GM, Goyal M, Legge AJ, Shaw RJ, Young D. Genotypic analysis of *Mycobacterium tuberculosis* from medieval human remains. Microbiology 1999; 145: 899-904.
123. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 2005; 102: 13950-5.
124. Tortoli E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. Clin Microbiol Rev 2003; 16: 319-54.
125. Thierry D, Cave MD, Eisenach KD, et al. IS6110, an IS-like element of *Mycobacterium tuberculosis* complex. Nucleic Acids Res 1990; 18: 188.
126. Torrea G, Levee G, Grimont P, Martin C, Chanteau S, Gicquel B. Chromosomal DNA fingerprinting analysis using the insertion sequence IS*6110* and the repetitive element DR as strain-specific markers for epidemiological study of tuberculosis in French Polynesia. J Clin Microbiol 1995; 33: 1899-904.
127. 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.
128. 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.
129. 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.

130. van Soolingen D, Qian L, de Haas PE, et al. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J Clin Microbiol 1995; 33: 3234-8.

131. van Soolingen D, Hoogenboezem T, de Haas PE, et al. A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa. Int J Syst Bacteriol 1997; 47: 1236-45.

132. Warren RM, Streicher EM, Sampson SL, et al. Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J Clin Microbiol 2002; 40: 4457-65.

133. Warren RM, Victor TC, Streicher EM, et al. Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers. J Clin Microbiol 2004; 42: 5774-82.

134. Wilson IJ, Weale ME, Balding DJ. Inferences from DNA data : population histories, evolutionary processes, and forensic match probabilities. J R Stat Soc 2003; Ser A166: 155-201.

135. Young DB. Ten years of research progress and what's to come. Tuberculosis 2003; 83: 77-81.

136. Zink AR, Sola C, Reischl U, et al. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. J Clin Microbiol 2003; 41: 359-67.

137. Zozio T, Allix C, Gunal S, et al. Genotyping of *Mycobacterium tuberculosis* clinical isolates in two cities of Turkey: description of a new family of genotypes that is phylogeographically specific for Asia Minor. BMC Microbiol 2005; 5: 44.

138. Zuckerkandl E. On the molecular evolutionary clock. J Mol Evol 1987; 26: 34-46.