J Mol Evol (2001) 53:680–689
DOI: 10.1007/s002390010255

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 2001

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

Christophe Sola, Ingrid Filliol, Eric Legrand, Igor Mokrousov, Nalin Rastogi

Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Morne Jolivière, BP 484, F-97165 Pointe à Pitre-Cedex, Guadeloupe

Abstract. This paper deals with phylogenetic relationships among a set of 90 clinical strains representative of the worldwide diversity of the *Mycobacterium tuberculosis* complex (Kremer et al. 1999) using eight independent genetic markers: IS*6110,* IS*1081,* the direct repeat (DR) locus, and five variable number of tandem DNA repeat loci (VNTR). In a preliminary experiment, phylogenetic trees based on single markers were constructed that led to the detection of some similarities between the VNTR-based and the spoligotyping-based phylogenetic trees. In the second step, a more global phenetic approach based on pairwise comparison of strains within each typing system was used, followed by calculations of mean genetic distances based on all the eight loci and the use of the neighbor-joining algorithm for tree reconstruction. This analysis confirmed our preliminary observations and suggested the existence of at least two new phylogeographical clades of *M. tuberculosis,* one defined as the "East African–Indian family" (EA-I), which may find its origin on the African or Asian continents, and the other as the "Latin American and Mediterranean" (LA-M) family. The existence of these two families was also validated by an independent phylogenetic analysis of spoligotyping on a larger set of shared types ($n = 252$) and further corroborated by VNTR and *katG–gyrA* results. The potential origin of these families of bacilli is discussed based on cattle domestication and human migration history. In conclusion, the information contained in insertion sequence and repetitive DNAs may serve as a model for the phylogenetic reconstruction of the *M. tuberculosis* complex.

Key words: *Mycobacterium tuberculosis* complex — Molecular phylogeny — Multilocus genotyping

Introduction

Phylogenetic inferences are premised on the inheritance of ancestral characters and on the existence of an evolutionary history defined by changes in these characters according to the concept of "descent with modification" (Swofford and Olsen 1990; Darlu and Tassy 1993). In other words, it postulates that the observation and description of current OTUs (operational taxonomic unit, or terminal taxon) may help to draw phylogenetic scenarios. The study of similarities observed between OTUs, either through global analysis (phenetic approach) or after a split between plesiomorphic (primitive or ancestral) and apomorphic (derived) character states, allows conflictory approaches of phylogenetics to be developed, i.e., the phenetic, cladistic, and probabilistic approachs (Darlu and Tassy 1993). The methods of phylogenetics may also be classified into major groups according to heuristics: distance methods, likelihood methods, and parsimony methods [for a review of methods for estimating phylogenies see Nei (1996)].

The chronology of evolution and phylogenetic relationships among the *Mycobacterium tuberculosis* complex remains to be elucidated. Although a very homogeneous species at the genetic level (Frothingham et al.

Correspondence to: C. Sola; *email:* csola@pasteur.gp

1994), the *M. tuberculosis* complex, the agent of bovine and human tuberculosis, one of the oldest known diseases, contains five members (*M. tuberculosis, M. africanum, M. bovis, M. microti,* and *M. canettii*) that are characterized by different but overlapping epidemiologies (Tsukamura 1976). For epidemiological purposes, the *M. tuberculosis* complex has been subdivided previously into five variants (Collins and Yates 1982): classical human, Asian human, bovine, African I (West African), and African II (East African). Recently phenotypic and genotypic characteristics allowed further subdivision of *M. bovis* and *M. africanum* into three and two groups, respectively (Niemann et al. 2000). However, traditional genotyping methods are not fruitful for studying bacterial populations of the *M. tuberculosis* complex because of the unusually low structural gene variation found among *M. tuberculosis* complex strains, even when originating from very diverse geographic areas (Sreevatsan et al. 1997; Kremer et al. 1999). Repetitive DNA markers such as insertion sequences (IS*1081* or IS*6110*), variable number of tandem DNA repeat loci (VNTR), and the direct repeat locus (DR), provide markers of choice for phylogenetic reconstruction since they show extensive polymorphism (Groenen et al. 1993; Sreevatsan et al. 1997; Frothingham et al. 1999). Among the events that generate diversity, IS*6110*-mediated deletions or insertions, homologous recombination, and replication slippage are instrumental to explain the genome plasticity of tubercle bacilli (Brosch et al. 1999; Fang et al. 1999; Filliol et al. 2000). There are other, yet poorly understood, external selective forces that drive tubercle bacillus evolution such as passed human demography, migratory flows, passed tuberculosis epidemiology, and history of livestock domestication (Bates and Stead 1993).

Two single-nucleotide polymorphisms (SNPs) on the *katG* and *gyrA* genes were recently used to subdivide the *M. tuberculosis* complex into three major groups: ancestral group 1, which contains strains belonging to any of the five members of the complex, and groups 2 and 3, which were restricted to *M. tuberculosis* in the strict sense (Sreevatsan et al. 1997). Recently, Frothingham et al. (1999) further subdivided *M. africanum* from group 1 into two subgroups, 1A and 1B, based on *katG* 203, a distinction that corroborates earlier taxonomical studies on this group of strains (Clavel 1975; David et al. 1978). Finally, three subpopulations of *M. tuberculosis* genotypes based on IS*6110* RFLP were linked to specific ethnic groups in a recent multicenter study (Kremer et al. 1999). Consequently, it can be reasonably assumed that geographically specific clades of *M. tuberculosis* may have emerged in distinct regions by divergent evolution. Recently, we and others, used spacer-oligonucleotide typing (spoligotyping) for phylogenetic reconstruction of *M. tuberculosis* (Fang et al. 1998; Sola et al. 1999, 2001). In the present investigation, our aim was to use multilocus genotyping obtained on 90 strains representative of the worldwide diversity of the *M. tuberculosis* complex to evaluate global similarities between strains and to search for new phylogeographical clades or families of *M. tuberculosis* by neighbor-joining (NJ) tree building. Our results demonstrates the existence of two new clades of *M. tuberculosis,* one that is specific from Africa and the Indian subcontinent (EA-I) and another that may find its origin in the Mediterranean basin and whose descent is widely spread in Latin America (LA-M). The existence of these families was confirmed independently by spoligotyping on 252 shared types representative of 3319 strains. These results are congruent to those obtained by VNTR and *katG* and *gyrA* genotyping. Some hypotheses on the origin of these two clades of tubercle bacilli are discussed.

Methods

Typing Methods

IS*6110* and IS*1081* RFLPs have been described previously (van Embden et al. 1993; van Soolingen et al. 1992).

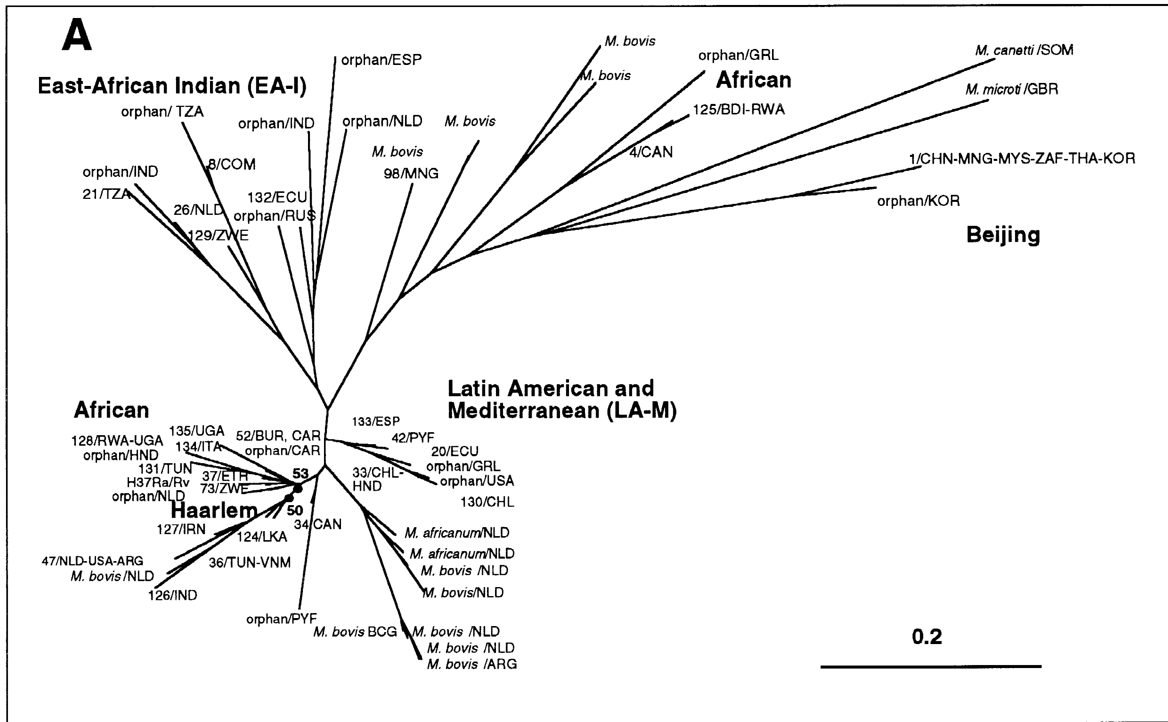
Spoligotyping is a fingerprinting technique based on the polymorphism of the direct repeat (DR) locus, which contains 36-bp direct repeats interspersed by specific 35- to 41-bp inter-DR sequences (Groenen et al. 1993; Kamerbeek et al. 1997). Some spoligotyping signatures specific for members of the *M. tuberculosis* complex have been published (Kamerbeek et al. 1997; van Soolingen et al. 1998; Aranaz et al. 1999). Variable number of tandem DNA repeats (VNTR) typing is a method based on the study of five polymorphic loci containing five to seven alleles for each locus, which has been shown to be useful for *M. tuberculosis* phylogenetic reconstruction (Frothingham and Meeker-O'Connell 1998; Frothingham et al. 1999).

Data

A recent multicenter study aimed at comparing fingerprinting methods for molecular epidemiology of tuberculosis described a set of 90 strains of *M. tuberculosis* originating from 38 countries (Kremer et al. 1999). This study constituted the first data set that may be assumed to be representative of the world wide diversity of *M. tuberculosis.* In preliminary experiments (Figs. 1A and B) we included all the strains that were differentiated either by spoligotyping or by VNTR typing (n = 71), which were subjected to a multilocus combined numerical analysis using IS*6110* RFLP, IS*1081* RFLP, VNTR, and spoligotyping (Figs. 2 and 3). The second data set, used to construct the trees shown in Figs. 4 and 5 concerned spoligotyping data extracted from a proprietary database described elsewhere (Sola et al. 2001) and contained 3319 spoligotypes from 47 countries with 252 shared types (identical spoligotypes shared by two or more patient isolates).

Phylogeny Reconstruction

The Taxotron package software (PAD Grimont, Taxolab, Institut Pasteur, Paris), which includes Recognizer, Restrictotyper, Adanson, and Dendrograf, was used throughout this study. Spoligo and VNTR data files were built into Recognizer files. For spoligotyping, the results were entered into the first 43 cupules under a binary format: 0 or empty cupule, no hybridization; 1 or full cupule, hybridization with the corresponding spacer (cupule number 0 was always kept empty). For VNTR, the same format was applied, but a maximum of 10 cupules


A
East-African Indian (EA-I)
orphan/ TZA
orphan/IND
8/COM
orphan/IND
21/TZA
26/NLD
129/ZWE
132/ECU
orphan/RUS
orphan/ESP
orphan/NLD
M. bovis
M. bovis
98/MNG
M. bovis
M. bovis
orphan/GRL
African
125/BDI-RWA
4/CAN
M. canetti /SOM
M. microti /GBR
1/CHN-MNG-MYS-ZAF-THA-KOR
orphan/KOR
Beijing
Latin American and
Mediterranean (LA-M)
African
128/RWA-UGA
orphan/HND
135/UGA
134/ITA
52/BUR, CAR
orphan/CAR
131/TUN
H37Ra/Rv
orphan/NLD
37/ETH
73/ZWE
53
Haarlem
50
127/IRN
124/LKA
34/CAN
47/NLD-USA-ARG
M. bovis /NLD
36/TUN-VNM
126/IND
orphan/PYF
133/ESP
42/PYF
20/ECU
orphan/GRL
orphan/USA
33/CHL-HND
130/CHL
M. africanum/NLD
M. africanum/NLD
M. bovis /NLD
M. bovis/NLD
M. bovis BCG
M. bovis /NLD
M. bovis /NLD
M. bovis /ARG
0.2

B
Latin American and
Mediterranean (LA-M)
50/CZE/32343
orp/PYF/31333
73/ZWE/33432
50/CZE-ARG-IND-BOL-NLD/32333
36/TUN-VNM/32333
134/ITA/32333
124/LKA/32333
47/NLD/32333
H37Ra/Rv/33433
53/CHN/31432
127/IRN/31533
Haarlem
127/IRN/31433
133/ESP/21433
132/ECU/21433
37/ETH/22433
33/HND/22532
M. africanum/NLD/22442
42/PYF/22432
98/MNG/22232
orphan/HND/22232
53/CUR/22332
130/CHL/22233
20/ECU/22233
orp/USA/22233
33/CHL/22433
orphan/ESP/22433
orphan/NLD/22433
26/NLD/32235
orphan/IND42235
1/MNG/42436
Beijing
orphan/KOR/42435
1/CHN-MYS-ZAF-THA-KOR/42435
52/BUR42535
21/TZA/42234
orphan/CAR/42335
52/CAR42434
135/UGA/41634
African
128/RWA/31534
128/UGA/41424
125/RWA/32534
M. bovis/NLD/32544
M. bovis /SAU/22544
African
125/BUR/42533
Beijing
1/KOR/42432
orphan/RUS/42532
4/CAN/42443
34/CAN/42443
M. bovis /NLD/51543
M. africanum/NLD/53434
orphan/GRL/42473
orphan/GRL/42473
131/TNS/42473
M. bovis /NLD/62535
M. bovis/ NLD/64525
M. bovis/ NLD/35543
M. bovis BCG/Japan/55533
M. bovis BCG/Russia/55523
M. bovis BCG/Pasteur/55623
M. bovis/ ARG/75543
M. bovis/ ARG/75543
M. bovis/ARG/94463
129/ZWE/95466
M. bovis /NLD/83265
M. microti/GBR/83571
M. bovis /ARG/94423
M. canetti/SOM/104634
126/IND/64455
orphan/IND/61465
orphan/NLD/44466
orphan/TZA/54466
8/COM/64466
East-African Indian (EA-I)
0.2

was attributed to each exact tandem repeat (ETR) (Frothingham et al. 1999) and the number of alleles was entered as indicated previously; e.g., when ETR-C was equal to 3, the first three cupules only were blackened (Frothingham and Meeker-O'Connell 1998). IS*6110* RFLPs were transferred and analyzed into Gel Compar (Applied Maths, Kortrijk, Belgium) and molecular weights were transferred to Restrictotyper files. IS*1081* RFLPs were analyzed with Restrictoscan and Restrictotyper; RFLP comparison was done with a 4% error tolerance rate. Individual pairwise distance matrices were computed using the following mathematical indexes: the Jaccard (1908) index was used for spoligotyping and VNTR results, whereas the Dice (1945) index was used for IS*6110* and IS*1081* RFLP. The phylogenetic trees were built using the neighbor-joining (NJ) algorithm contained in the Adanson software; this algorithm is well suited for tree-buiding when a large number of OTUs is to be compared (Nei 1996). Among the distance matrix-based approach to phylogeny reconstruction, the NJ method produces a unique final tree under the principle of minimum evolution. Although it does not necessarily produce the minimum-evolution tree, computer simulations have shown that it is quite efficient in obtaining the correct tree topology and it is applicable to any type of evolutionary distance data (Saitou and Nei 1987). Combined numerical analysis was performed by averaging individual distance files, all strains being in the same order in each file, and a final NJ tree was built with this averaged distance file. Preliminary analyses (Figs. 1A and B) were also performed with the PHYLIP package (Felsenstein 1993) using the NEIGHBOR and DRAWTREE softwares for tree representation and gave similar results (results not shown).

Nomenclature

To simplify the handling of spoligotypes, an arbitrary numbering system (shared-type designation), which attributes a chronological number as soon as a new pattern leading to a shared-type is published, was used throughout the study in the figures (for a full description see Sola et al. 1999, 2001). All the spoligotyping results from the multicenter study by Kremer et al. (1999) were initially compared with our database and designated a number according to the above scheme. Patterns already present were assigned to a shared-type number, whereas new and unique strains were designated orphan. The VNTR results were expressed in their original format as described previously for alleles A to E (Frothingham and Meeker-O'Connell 1998). Country numbers are given according to ISO code 3166, available on the Internet at www.un.org/Depts/unsd/methods/m49.htm.

Results

A Preliminary Phylogenetic Reconstruction of the M. tuberculosis *Complex Suggests the Existence of at Least Two New Clades of the* M. tuberculosis *Complex: The East African–Indian (EA-I) and the Latin American and Mediterranean (LA-M) Types*

From the data published by Kremer et al. (1999), a set of 71 strains that contained all strains harboring a unique combination of spoligotyping and VNTR patterns was selected and two NJ trees based on either the spoligotyping results (Fig. 1A) or the VNTR results (Fig. 1B) were constructed. Each of these trees showed a specific spatial clustering of most of the strains according to their geographical origin. More than the five classical major *branches* defined previously by phenotypic methods may be seen on these trees (Collins and Yates 1982). For example, spoligotypes from strains originating from Tanzania, Zimbabwe, Comoro Islands, and India (all coming from the EA-I region) clustered distinctly into a separate branch (Fig. 1A). Strains from LA-M were distinctly located on a new and separate branch on this tree (Fig. 1A). As supported by a recent study suggesting the existence of various subtypes of *M. bovis* (Niemann et al. 2000), *M. bovis* tended to cluster separately into at least two *branches* in this NJ tree (Fig. 1A). The Beijing type was homogeneous and very distant from all other *M. tuberculosis* strains, whereas the "Haarlem" family strains aggregated together in a branch apparently close to some of the African strains, to a branch of *M. bovis,* and to the LA-M family of strains (Fig. 1A). Finally, the only two *M. africanum* strains in this study clustered with the *M. bovis* family, however their paucity (only two strains) precludes any interpretation of evolutionary branching in this study.

The NJ-VNTR tree suggested similar results as those obtained by spoligotyping for the LA-M and the EA-I branches, which again and independently cluster on distinct branches of this new tree (Fig. 1B). The "Haarlem" family is homogeneous with a unique 32333 VNTR allele combination if we except a single strain from Czech Republic (32334). Except for one strain from Russia (42452), the Beijing type is very homogeneous when studied by this method (42432). A first combined numerical analysis of spoligotyping and VNTR was drawn and underlined the neighborhood between strains inside the LA-M as well as inside the EA-I clade (results not shown).

*Confirmation of the EA-I and LA-M Branches by a Combined Numerical Analysis Using IS*1081, *IS*6110, *Spoligotyping, and VNTR*

Insertion sequences such as IS*1081* and IS*6110* may be informative markers for biogeographical studies as well

←

Fig. 1. Neighbor-joining phylogenetical trees based on spoligotyping (**A**) and VNTR data (**B**) on 71 strains from the study by Kremer et al. (1999) showing the presence of two new phylogeographical clades, designated "East African–Indian" (EA-I) and "Latin American and Mediterranean" (LA-M). Each branch shows the following levels of information for *M. tuberculosis:* strain description, with "orphan" (for unique strains) or, alternatively, with a one- to three-digit number (spoligotype designation for shared types as shown in column G in Fig. 2), and a three-letter ISO code description for the country of origin. *M. bovis, M. africanum, M. microti,* and *M. canetti* are indicated, together with their geographical origin according to ISO code 3166. Potential clades as well as two highly prevalent shared types (types 53 and 50) are indicated in *boldface.* In addition, the VNTR results are shown as a five-digit number (e.g., 32333) in B.

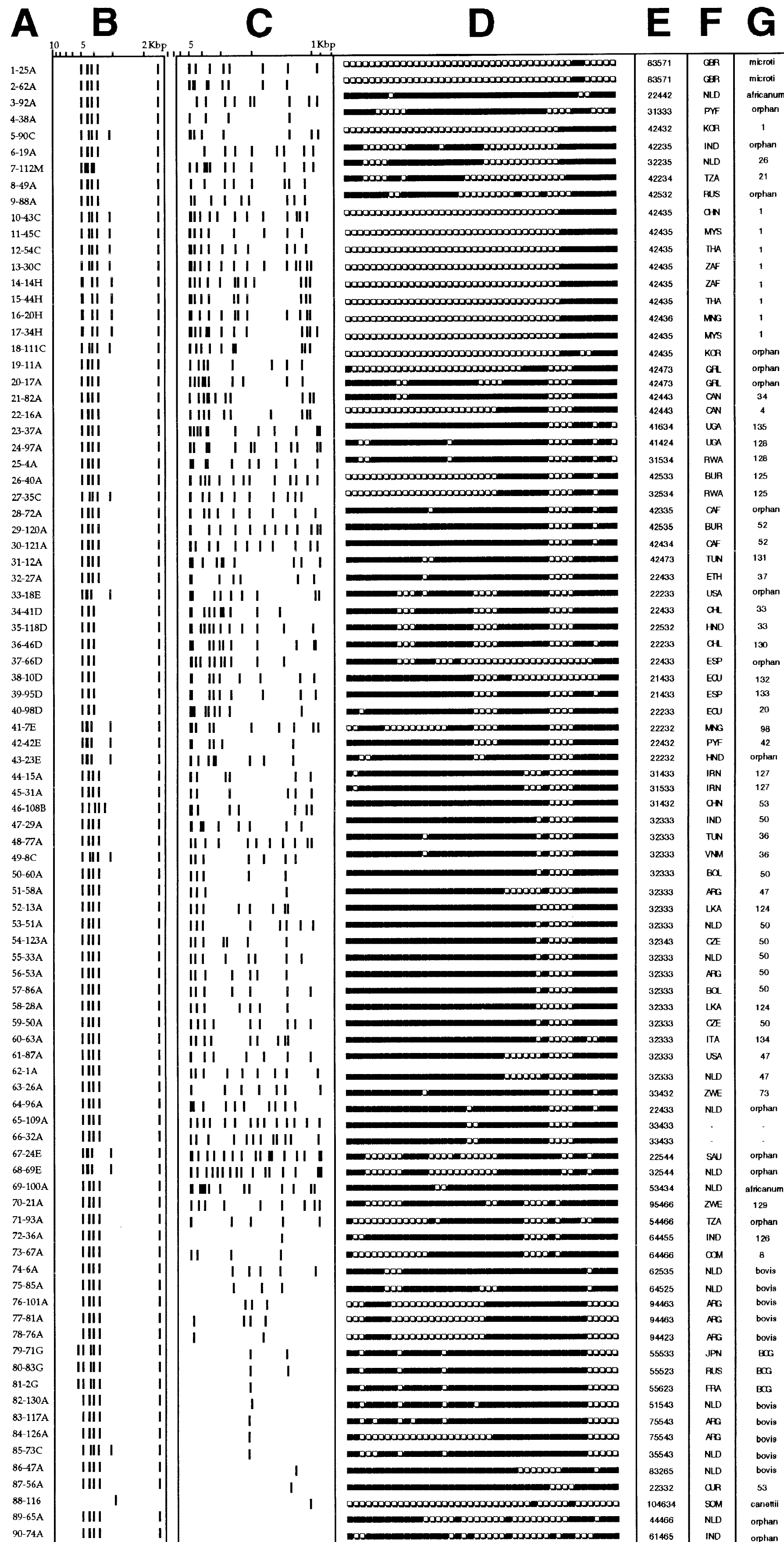


Fig. 2. Matrix of the repetitive DNA markers used for phylogenetical reconstruction of the *M. tuberculosis* complex extracted from Kremer et al. (1999). **A** Serial number (1 to 90) followed by strain number as in the original publication and IS*1081* type, designated A to M according to Park et al. (2000), e.g., strain 1 is isolate 25 harboring IS*1081* type A. **B** IS*1081* pattern. **C** IS*6110* pattern. **D** Spoligotype pattern. **E** VNTR allele combination. **F** Geographical origin expressed as ISO code 3166 (see also the legend to Fig. 3). **G** Arbitrary nomenclature of shared types of *M. tuberculosis* according to Sola et al. (2001) or designation of the *M. tuberculosis* complex subspecies (*M. bovis, M. africanum, M. microti, M. canettii*). Unique spoligotype patterns are labeled as "orphan." *Dashes* in column G for isolates 65 and 66 refer to type strains *M. tuberculosis* H37Rv and H37Ra. Due to different softwares, the IS*6110* and IS*1081* patterns do not appear exactly as in the original paper.

Fig. 3. Neighbor-joining phylogenetical tree based on the combined numerical analysis of spoligotyping, VNTR, IS*6110* RFLP, and IS*1081* RFLP results of 90 strains from the study by Kremer et al. (1999). Each branch shows the following levels of information: spoligotype designation for shared types as shown in column G in Fig. 2; ISO code 3166; isolate number; five-digit VNTR allele description; and IS*1081* type, designated A to M. The EA-I and LA-M clades are *circled.* Please note that the LA-M clade appears to be well separated from the Haarlem clade.

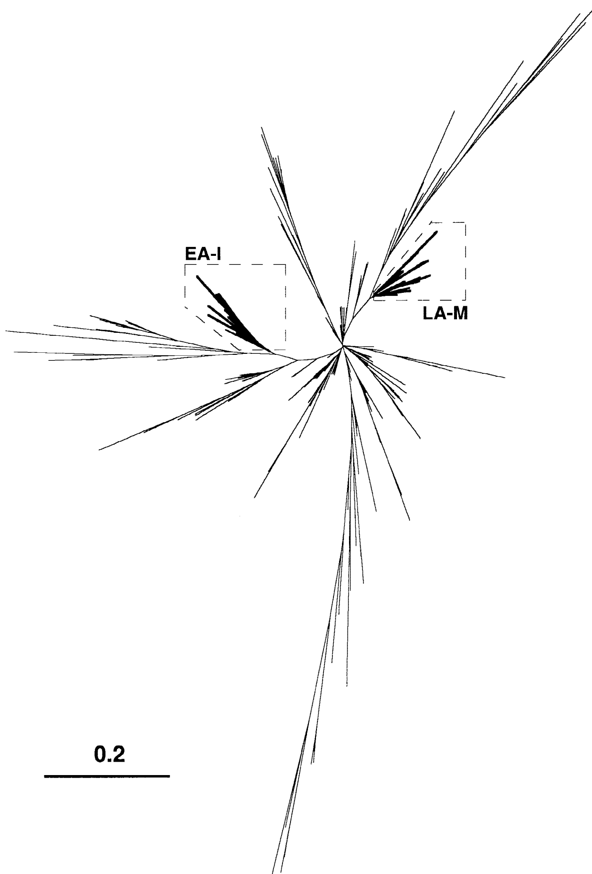


Fig. 4. Neighbor-joining phylogenetical tree based on 252 shared types from a worldwide spoligotyping database containing 3319 spoligotypes from the study by Sola et al. (2001). Please note that the EA-I and LA-M clades are well separated.

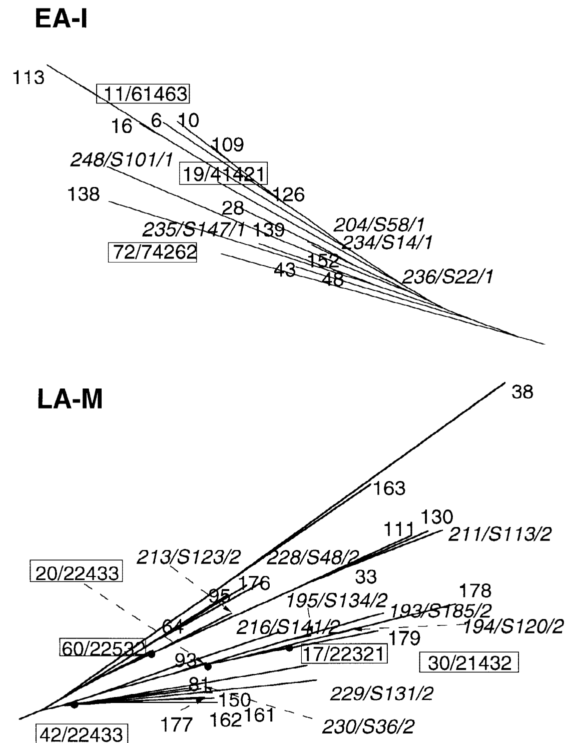


Fig. 5. Enlargement of the EA-I and LA-M clades from Fig. 3. Three levels of information are shown here: spoligotype designation for shared types according to the study by Sola et al. (2001); spoligotype designation according to Soini et al. (2000) (numbers starting with S); and, when given, a number corresponding to the major genetic groups 1 to 3 according to Sreevatsan et al. (1997). Numbers in *boxes* show two levels of information: the first number represents the shared-type spoligotype designation according to Sola et al. (2001), whereas the second number represents the five-digit VNTR description. Nodes shown on the branches (*black circles*) most likely represent ancestral and/or intermediate strains during the evolutionary process. Please note that all data from our laboratory are shown as *roman* characters, whereas those from the study by Soini et al. (2000) are in *italics*.

as for phylogeny reconstruction of *M. tuberculosis* (van Soolingen et al. 1995; Park et al. 2000; Warren et al. 2000). We therefore included IS*1081* and IS*6110* genotyping data in a new phylogeny reconstruction of *M. tuberculosis* based on the description of the full collection of 90 strains published by Kremer et al. (1999) by spoligotyping, VNTR, IS*1081,* and IS*6110.* We decided to limit our analysis to the most informative genetic markers, which could be easily analyzed and compared, using computer-assisted comparison between strains. Results of this analysis are shown in Fig. 3. We reasoned that, as each typing method identifies strains with atypical characters, the combination of these four data sets would reduce the impact of minor discrepancies and produce a much more robust tree. As expected, the two previously defined branches, EA-I and LA-M, remained distinct; the EA-I branch was related to a *M. bovis* branch, whereas the LA-M branch appeared to be part of a larger clade which may share an ancestor with the Haarlem family (Fig. 3). However, the LA-M family was clearly distinguished from the Haarlem family, as all but one strain in the LA-M branch harbored a typical IS*1081* profile, defined as "pattern D" by Park et al. (2000).

Independent Validation of the LA-M and EA-I Branches Using a Larger Spoligotyping Database and Hypothetical Scenarios of Their Evolution

Independently, we constructed another NJ tree based on spoligotyping on a much larger set of data (252 shared types representative of 3319 strains). The resulting tree (Fig. 4) is quite complex and is not discussed in detail here, however, the presence of branches LA-M and EA-I is clearly indicated, and a preliminary correlation among spoligotyping, *katG–gyrA,* and VNTR data for these branches is illustrated in the enlargement shown in Fig. 5. Some interesting conclusions may be drawn; all strains of the EA-I branch (Fig. 5A) harbor a *katG*463 CTG (Leu) and *gyrA*95 ACC (Thr) allele and are ancestral strains as defined previously by Sreevatsan et al. (1997). Furthermore, the VNTR data show a high ETR-A allele number (four or more copies) in strains of the EA-I branch (Fig. 5A), a feature also shared by the isolates of the *M. bovis* or *M. africanum* descent [Group IA or IB of

Frothingham et al. (1999)]. On the other hand, the strains of the LA-M branch harbor a *katG*463 C*G*G (Arg) and *gyrA*95 ACC (Thr) allele, and as defined by Sreevatsan et al. (1997), these are probably intermediate strains at the evolutionary level (Fig. 5B). The VNTR data show that, contrary to the EA-I branch, their ETR-A allele number is low (two instead of four or more copies).

Discussion

For a disease considered as old as the history of humanity (Bates and Stead 1993), its phylogeny remains highly speculative (Kapur et al. 1994; Stead et al. 1995; Frothingham 1999). *M. bovis* has been suspected as the ancestor of *M. tuberculosis,* as tuberculosis was presumed to be a zoonosis before becoming an anthropozoonosis (Stead et al. 1995), however, genetic evidence to this effect is missing. As repetitive DNAs of mycobacteria have been proposed to be useful for studying the evolution of *M. tuberculosis* (Fang et al. 1998; Sola et al. 1999, 2001), we focused on the use of spoligotyping, VNTR, IS*1081,* and IS*6110* to investigate phylogenetic links among two sets of *M. tuberculosis* complex strains in the present study. One set was composed of an exhaustively characterized collection of 90 strains, whereas the second set contained a larger but less characterized collection of 3319 spoligotypes for which other genotyping data (VNTR, *katG–gyrA*) were available for selected isolates (Sola et al. 2001). As shown under Results, instead of five *M. tuberculosis* families in the world defined previously by analysis of phenotypic character (Collins and Yates 1982), spoligotyping and VNTR analysis suggested the existence of more than five genotype families (Fig. 1), an observation corroborated by the combined numerical analysis based on spoligotyping, VNTR, IS*1081* RFLP, and IS*6110* RFLP (Fig. 3).

However, one may argue that combining markers with different molecular clocks may not be the best way to elucidate the different depths of phylogenetic divergence over different time scales, a problem that has received theoretical attention only recently (Wilson et al. 2001). Though not perfect, the phenetic approach has the merit of attempting to define global similarities using all available characters, which considerably reduces the impact of minor discrepancies and produces a much more robust tree. Concerning the choice of individual markers, IS*6110* dynamics may reportedly vary between clades since it may evolve too rapidly among high-copy isolates such as the Beijing strains (Warren et al. 2000). Nevertheless, it is suitable for slowly evolving clades of low-IS*6110* banders, and consequently, we decided to keep this marker along with slower molecular clock markers such as spoligotyping, VNTR, and IS*1081.* Finally, data should be interpreted with caution for *M. tuberculosis* complex subspecies for which only rare isolates are yet available such as *M. africanum, M. canettii,* and *M. microti.*

The results obtained showed the existence of new biogeographical subtypes of *M. tuberculosis,* and according to the most frequent origins of the strains which define these branches, two families were baptized the “East African-Indian” (EA-I) and “Latin American and Mediterranean” (LA-M) clades. However, it should be underlined that this designation does not necessarily signify that the ancestors of these families are to be found in these regions, e.g., strains of the EA-I branch were independently reported in Guinea–Bissau (Källenius et al. 1999). These authors suggested that this family of strains (designated group C in their study) may find its origin in West Africa. Concerning the origin and spread of the EA-I clade, it may be hypothesized that these isolates were historically endemic in Asia or Africa and spread with early human migration and/or later through Indian–Muslim trade from the Indian subcontinent toward East Africa or the Portuguese–Spanish colonization of the 15th–16th century. Alternatively a potential bovine origin of this family of strains may also be linked to cattle migratory routes of taurine or zebu east to west as evidenced by the existence of two independent cattle domestication events (Loftus et al. 1994).

Along the same lines, the LA-M family of strains either were endemic in the Americas in the pre-Columbus era or originated in the Mediterranean basin and were introduced later into the Americas through infected cattle and human migration (Fanning 1994). Although it was hypothesized that tuberculosis may not have been a serious health problem in pre-Columbian America (Stead 1997), a study performed in South America, that reviewed 483 skeletons, suggested an estimated tuberculosis prevalence of between 1 and 2% between 2000 B.C. and 1500 A.D. (Arriaza et al. 1995). The endemic presence of tuberculosis in America before the settlement of Caucasians is also supported by the finding of *M tuberculosis* DNA in a Peruvian mummy (Salo et al. 1994). At the same time, a potential link to the Mediterranean basin of this clade is indicated by a hypothetical scenario of evolution among strains of types 42, 20, and 17 as discussed under Results (Fig. 5), and also by their current geographical distribution. Indeed no strain of the “ancestral” type 42 was reported among the 1283 strains described by Soini et al. (2000) in the USA although derived types were found (type 229, 230).

In conclusion, the precise relationship among the current genotype prevalence of the *M. tuberculosis* complex, human demographic expansion and migration, and tuberculosis transmission dynamics should be the focus of further investigations. Such an approach may help to elucidate if the *M. tuberculosis* diversity followed or preceded human demographic expansion; e.g., a recent study suggested that the dynamics of repeated elements depends more on IS*6110* transpositional events than epi-

demic parameters alone (Tanaka et al. 2000). Consequently, today's molecular diversity of tubercle bacilli may derive from a relatively recent spreading of tuberculosis linked to changes in human demography, lifestyle, and tuberculosis epidemiology essentially after 1500 A.D. Alternatively, it may be due to a relatively ancient spreading which may then have occurred either before or after the domestication of livestock around 7000 B.C. (Bradley et al. 1996). According to early hominid biogeography (Strait and Wood 1999), four to seven early dispersal events may have taken place independently between southern and eastern Africa and the Malawi River. These events may have led to the foundation of *M. tuberculosis* evolutionary history. Along similar lines, at least five independent events of livestock and two of bovine cattle domestication took place (Smith 1995; Loftus et al. 1994); in the latter case two types of aurochs (*Bos primigenius*) supposedly led to the current species, i.e., *Bos bos indicus,* on the Indian subcontinent and in East Africa, and another unknown ancestor which spread from Africa to Europe and gave rise to *Bos bos taurus* (Loftus et al. 1994). A more recent reintroduction of Asian *B. indicus* into Africa through the Arab invasion (670 A.D.) gave rise to the typical East African zebu, a fact supported by archeological evidence (Loftus et al. 1994). In this context, the recent discovery of *M. bovis* subsp. *caprae* suggests that a common *M. bovis* ancestor may also have diverged to give host-specific variants for goats (Aranaz et al. 1999). Thus a complex intermingling of humans and animals for the last 7000–15,000 years may be responsible for the diversity and evolution of major *M. tuberculosis* clades. Likewise, the hantavirus and the rodent coevolution models show that virus and hosts phylogeny may be superimposed (Vapalahti et al. 1999).

Data on human diversity reveal that differences among continents represent roughly one-tenth of human molecular diversity and that genetic variation is high even among small population groups (Barbujani et al. 1997), a model that may also hold true for tubercle bacillus diversity. Alternatively, bottleneck events such as climatic changes may have lowered ancestral human populations to small effective sizes (Zietkiewicz et al. 1998), limiting the current diversity of tubercle bacilli to an association between most fit clones of bacilli with most demographically—and/or migratorily—active human populations. This would obviously affect the selection of a limited and recent number of major families of *M. tuberculosis* genotypes.

Acknowledgments. The authors are grateful to "Délégation Générale au Réseau International des Instituts Pasteur et Instituts Associés," Institut Pasteur, Paris, Fondation Française Raoul Follereau, Paris, and Ministère des Affaires Etrangères, Paris, for financial support. We also thank Laurence Ballanger, Ariane S.A., Guadeloupe, for her advice on software utilization and Dr. F.-X. Weill for stimulating discussions. We warmly acknowledge Drs. K. Kremer, D. van Soolingen, and J. van Embden for publishing their study.

References

Aranaz A, Liebana E, Gomez-Mampaso E, Galan JC, Cousins D, Ortega A, Blazquez J, Baquero F, Mateos A, Suarez G, Dominguez L (1999) *Mycobacterium tuberculosis* subsp. caprae subsp. nov.: A taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. Int J Syst Bacteriol 49:1263–1273

Arriaza BT, Salo W, Aufderheide AC, Holcomb TA (1995) Pre-Columbian tuberculosis in northern Chile: Molecular and skeletal evidence. Am J Phys Anthropol 98:37–45

Barbujani G, Magnani A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–19

Bates JH, Stead WW (1993) The history of tuberculosis as a global epidemic. Med Clin North Am 77:1205–1217

Bradley DG, Hugh DEM, Cunningham P, Loftus RT (1996) Mitochondrial diversity and the origins of African and European cattle. Proc Natl Acad Sci USA 93:5131–5135

Brosch R, Philipp WJ, Stavropoulos E, Colston MJ, Cole ST, Gordon SV (1999) Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. Infect Immun 67:5768–5774

Clavel S (1975) Lysotypie du bacille tuberculeux dans quelques pays d'Afrique tropicale et en Guadeloupe. Comparison avec la France. Rapports avec la classification de "*Mycobacterium africanum.*" Ann Sclavo 17:571–577

Collins CH, Yates MD (1982) Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. J Hyg Cambr 89:235–242

Darlu P, Tassy P (1993) In: La reconstruction phylogénétique: Concepts et méthodes. Masson, Paris, pp 15–22

David HL, Jahan MT, Jumin A, Grandry J, Lehmann EH (1978) Numerical taxonomy of *Mycobacterium africanum.* Int J Syst Bacteriol 28:467–472

Dice LR, (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

Fang Z, Morrison N, Watt B, Doig C, Forbes KJ (1998) IS*6110* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. J Bacteriol 180:2102–2109

Fang Z, Doig C, Kenna DT, Smittipat N, Palittapongarnpim P, Watt B, Forbes KJ (1999) IS*6110*-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis.* J Bacteriol 181:1014–1020

Fanning EA (1994) *Mycobacterium bovis* infection in animals and humans. In: Davies PDO (ed) Clinical tuberculosis. Chapman & Hall, London, pp 361–365

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.57c for Power MacIntosh. Department of Genetics University of Washington, Seattle

Filliol I, Sola C, Rastogi N (2000) Detection of a previously unamplified spacer within the DR locus of *Mycobacterium tuberculosis:* Epidemiological implications. J Clin Microbiol 38:1231–1234

Frothingham R (1999) Evolutionary bottlenecks in the agents of tuberculosis, leprosy, and paratuberculosis. Med Hypotheses 52:95–99

Frothingham R, Hills HG, Wilson KH (1994) Extensive DNA sequence conservation throughout the *Mycobacterium tuberculosis* complex. J Clin Microbiol 32:1639–1643

Frothingham R, Meeker-O'Connell WA (1998) Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 144:1189–1196

Frothingham R, Strickland PL, Bretzel G, Ramaswamy S, Musser JM, Williams DL (1999) Phenotypic and Genotypic characterization of *Mycobacterium africanum* isolates from West Africa. J Clin Microbiol 37:1921–1926

Groenen PMA, Bunschoten AE, van Soolingen D, van Embden JDA (1993) Nature of DNA polymorphism in the direct repeat cluster of

Mycobacterium tuberculosis; Application for strain differentiation by a novel typing method. Mol Microbiol 10:1057–1065

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44:223–270

Källenius G, Koivula T, Ghebremichael S, Hoffner SE, Norberg R, Svensson E, Dias F, Marklund B, Svenson SB (1999) Evolution and clonal traits of *Mycobacterium tuberculosis* in Guinea-Bissau. J Clin Microbiol 37:3872–3878

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

Kapur V, Whittam TS, Musser JM (1994) Is *Mycobacterium tuberculosis* 15000 old? J Infect Dis 170:1349–1350

Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PWM, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, Musser JM, van Embden JDA (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* strains: Interlaboratory study of discriminatory power and reproducibility. J Clin Microbiol 37: 2607–2618

Loftus RT, Hugh DEM, Bradley DG, Sharp PM, Cunningham P (1994) Evidence for two independent domestications of cattle. Proc Natl Acad Sci USA 91:2757–2761

Nei M (1996) Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet 30:371–403

Niemann S, Richter E, Rüsch-Gerdes S (2000) Differentiation among members of the *Mycobacterium tuberculosis* complex by molecular and biochemical features: Evidence for two pyrazinamide-susceptible subtypes of *M. bovis.* J Clin Microbiol 38:152–157

Park Y-K Bai G-H Kim S-J, (2000) Restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* from countries in the Western pacific region. J Clin Microbiol 38:191–197

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Salo WI, Aufderheide AC, Buikstra J, Holcomb TA (1994) Identification of *Mycobacterium tuberculosis* in a pre-Columbian Peruvian mummy. Proc Natl Acad Sci USA 91:2091–2094

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

Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM (2000) Characterization of Mycobacterium tuberculosis isolates from patients in Houston, Texas, by spoligotyping. J Clin Microbiol 38: 669–676

Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N (1999) Tuberculosis in the Caribbean: Using spacer oligonucleotide typing to understand strain origin and transmission. Emerg Infect Dis 5:404–414

Sola C, Filliol I, Guttierez C, Mokrousov I, Vincent V, Rastogi N (2001) An update of the spoligotype database of *Mycobacterium tuberculosis:* Analysis of the biogeographical distribution of shared-types and epidemiological and phylogenetical perspective. Emerg Infect Dis 7:390–396

Sreevatsan S, Pan X, Stockbauer K, Connell N, Kreiswirth B, Whittam T, Musser J (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA 97:9869–9874

Stead WM, Eisenach KD, Cave MD, Beggs ML, Templeton GL, Thoen CO, Bates JH (1995) When did *Mycobacterium tuberculosis* infection first occur in the new world? Am J Respir Crit Care Med 151:1267–1268

Stead WW (1997) The origin and erratic global spread of tuberculosis. Clin Chest Med 18:65–77

Strait DS, Wood BA (1999) Early hominid biogeography. Proc Natl Acad Sci USA 96:9196–9200

Swofford DL Olsen FJ, (1990), In: Hillis DM, Moritz C (eds) Molecular systematics. Sinauer Associates, Sunderland, MA, pp 411–515

Tanaka MM, Small PM, Salamon H, Feldman MW (2000) The dynamics of repeated elements: Applications to the epidemiology of tuberculosis. Proc Natl Acad Sci USA 97:3532–3537

Tsukamura M (1976) Numerical classification of slowly growing mycobacteria. Int J System Bacteriol 26:409–420

Van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small PM (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology. J Clin Microbiol 31:406–409

Van Soolingen D, Hermans PWM, de Haas PEW, van Embden JDA (1992) Insertion element IS*1081*-associated restriction fragment length polymorphisms in *Mycobacterium tuberculosis* complex species: A reliable tool for recognizing *Mycobacterium bovis* BCG. J Clin Microbiol 30:1772–1777

van Soolingen D, Qian L, de Haas PEW, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JDA (1995) Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J Clin Microbiol 33:3234–3238

van Soolingen D, van der Zanden AGM, de Haas PEW, Noordhoek GT, Kiers A, Foudraine NA, Portaels F, Kolk AHJ, Kremer K, van Embden JDA (1998) Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. J Clin Microbiol 36:1840–1845

Vapalahti O, Lundkvist A, Fedorov V, Conroy CJ, Hirvonen S, Plyusnina A, Nemirov K, Fredga K, Cook JA, Niemimaa J, Kaikusalo A, Henttonen H, Vaheri A, Plyusnin A (1999) Isolation and characterization of a hantavirus from *Lemmus sibiriscus:* Evidence for host switch during Hantavirus evolution. J Virol 73:667–675

Warren RM, Simpson SL, Richardson M, van der Spuy GD, Lombard P, Victor TC, van Helden PD (2000) Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. Mol Microbiol 37:1405–1416

Wilson IJ, Wade ME, Balding DJ (2001) Inferences from DNA data: Population histories, evolutionary processes, and forensic match probabilities. http://maths.abdn.ac.uk/~ijw/wb.ps

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, Labuda D (1998) Genetic structure of the ancestral population of modern humans. J Mol Evol 47:146–155