Available online at www.sciencedirect.com

SCIENCE DIRECT®

Infection, Genetics and Evolution 4 (2004) 5–14

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

Phylogenetic reconstruction of *Mycobacterium tuberculosis* within four settings of the Caribbean region: tree comparative analyse and first appraisal on their phylogeography

Véronique Duchêne[a], Séverine Ferdinand[a], Ingrid Filliol[a], Jean François Guégan[b], Nalin Rastogi[a,1], Christophe Sola[a,*]

[a] *Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, BP 484, F-97165 Pointe à Pitre Cedex, Guadeloupe, France*
[b] *Unité de Génétique des Maladies Infectieuses, UMR CNRS/IRD 9926, IRD, BP 64501, 34392 Montpellier Cedex 5, France*

Received 24 March 2003; received in revised form 28 August 2003; accepted 1 September 2003

Abstract

In order to compare phylogenetic methods and to reconstruct the evolutionary history of the tubercle bacilli, a set of macro-array-based genotyping data of *Mycobacterium tuberculosis* clinical isolates (called spoligotyping for spacer oligonucleotide *typing,* which assays the variability of the Direct Repeat -DR- locus), was analyzed in four settings of the Caribbean region (Guadeloupe, Martinique, Cuba and Haiti). A set of 47 alleles, split into 26 shared and 21 unique alleles) representative of 321 individual *M. tuberculosis* clinical isolates from patients residing in the above regions was studied. The following methods (and software in brackets) were investigated: numerical taxonomy distance methods (TAXOTRON), maximum parsimony procedure (PAUP), median-joining networks (NETWORK), and nested clade analysis (GEODIS). Results using these methods were analyzed, compared and discussed. The latter method (GEODIS) was investigated in detail by introducing geographical data together with genetic variability results to detect a link between population structure and population history, and to test the null hypothesis of no association between geography and genotypes. Irrespective of the methods used, our findings demonstrate that a core structure of four families (or clades) of *M. tuberculosis* strains is highly prevalent within the islands studied, indirectly reflecting passed colonization history of these different settings. Specificity of *M. tuberculosis* genotypes in each of the islands is discussed in the light of their respective colonial and contemporary histories.

Keywords: Mycobacterium tuberculosis; Evolutionary genetics; Bacterial genotyping techniques; Anthropology; Caribbean

1. Introduction

Directly responsible for more than two millions deaths worldwide annually, tuberculosis (TB) is a disease that may be as old as civilization (Grmek, 1994). It is a paradigm of a threatening re-emerging infectious disease and of all the accompanying changes in host-pathogen interactions that may be studied together with changes in epidemic dynamics. Phenomena such as Multi-drug-resistance and association with AIDS pandemic are also of increasing importance, hence a renewed interest for this disease (WHO, 1994). TB also happens to be a potential model to study paleodemography and paleoepidemiology of infectious diseases (Grmek, 1994). In the Caribbean region, human settlement goes back to 4000 years B.C., stemming from South America (Adelaïde-Merlande, 1994). Traditional Amerindian societies disappeared following Spanish, English and French successive settlements in the various islands for which European settlers fought for three centuries. The history of the "sugar islands" (Hispaniola now divided into Haiti and the Dominican Republic, St Kitts and others), is also made up of massive West African slave trade. This trade which started at the beginning of the XVIth century, was followed by Indian and to a lesser extent Chinese immigrations during the XIXth century after the slavery was abolished (Adelaïde-Merlande, 1994). Although it may be speculated from history that major ways of entry of TB in the Caribbean included sea-borne immigration of both humans and cattle, the passed incidence of this disease is poorly documented in the literature.

This study attempts to reconstruct passed natural history of TB spreading in a Caribbean setting and to link it to today's population structure of tubercle bacilli. For this

[*] Corresponding author. Tel.: +590-590-897-665; fax: +590-590-893-880.
E-mail addresses: nrastogi@pasteur-guadeloupe.fr (N. Rastogi), csola@pasteur-guadeloupe.fr (C. Sola).
[1] Co-corresponding author.

1567-1348/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2003.09.001

purpose, the available genetic variability data on *M. tuberculosis* clinical isolates from Cuba, Haiti, Guadeloupe and Martinique (Diaz et al., 2001; Ferdinand et al., 2003; Sola et al., 1999, 2001a), was subjected to a detailed phylogenetic investigation. Phylogenetic analysis has been often termed as a black box into which data are fed and out of which "the tree" springs (Swofford and Olsen, 1990). Nonetheless, it does permit the study of the evolution of an infectious disease with a complex interplay of demographic and epidemic dynamics, and population migrations. We have previously used a classical numerical taxonomy approach to investigate the population structure of *M. tuberculosis* in a set of 90 clinical isolates representative of the worldwide diversity of the tubercle bacilli (Sola et al., 2001b). The congruence obtained between independent genetic markers in this study argued in favor of a clonal structure of *M. tuberculosis* population and the existence of geographically structured clades or families of tubercle bacilli (Sola et al., 2001b). As additional information on the frequencies and locations of the different variants may be useful when studying a collection of haplotypes from a single species (Excoffier and Mouse, 1994), we decided to focus on the population structure of different Caribbean *M. tuberculosis* strains in the present study. Using distinct phylogenetic procedures, we show that a core of four major clades or families of *M. tuberculosis* present today reflects a combination of TB paleoepidemiology, human settlement and demographic history in this region of the world.

2. Materials and methods

2.1. Data collection, nomenclature

Spoligotyping data on 321 isolates from 321 individual patients originating from four locations within the Caribbean region (Cuba, $n = 126$; Haiti, $n = 56$; Guadeloupe, $n = 110$; and Martinique, $n = 29$), were collected from recently published studies (Diaz et al., 2001; Ferdinand et al., 2003; Sola et al., 1999, 2001a). In each case, spoligotyping was performed following the published procedure (Kamerbeek et al., 1997). Data from the 321 isolates corresponded to 47 distinct spoligotype patterns or alleles (Table 1); 26 of these were shared-types (ST), i.e. a common profile observed in isolates from two or more patients, and 21 were unique profile, i.e. found only once within this setting (also designated as "orphan").

The family and sub-family designation was performed according to the latest available classification and nomenclature, both for shared-types (ST) and families (Filliol et al., 2002, 2003).

2.2. Methods of analysis and softwares

The following softwares and methods used for phylogenetic reconstruction were as follows: TAXOTRON (Grimont, 1994), PAUP 4.0 (Swofford, 1998), NETWORK (Bandelt et al., 1999), and GEODIS (Posada et al., 2000).

2.2.1. TAXOTRON (numerical taxonomy)

The Pairwise distance between clinical isolates was computed using the 1-Jaccard ($1 - S_j$) index. The Jaccard Index is defined as $a/a + c$, where a is the number of simultaneously positive characters and c is the number of discrepancies (Jaccard, 1908). The neighbor-joining algorithm was used for clustering (Saitou and Nei, 1987). Average linkage (UPGMA, unweighted pair-group method using arithmetic averages) or neighbor joining are indeed fairly popular methods that have proved to be very useful to define some major phylogeographical clades within the *Mycobacterium tuberculosis* complex (van Soolingen et al., 1995; Kremer et al., 1999; Sola et al., 2001b; Filliol et al., 2002).

2.2.2. PAUP 4.0 (parsimony methods, see Swofford, 1988)

Given the high number of taxa, a unique parsimony approach was used. This procedure was performed using a heuristic search based on a branch swapping option, the tree bisection and reconnection algorithm (TBR), which is known to give good results in general. The most parsimonious tree among 1000 trees generated was kept. However, given the quantity of options available with this software, an exhaustive search of the best option was not systematically investigated.

2.2.3. NETWORK (phylogenetic networks)

Interconnected trees are often called networks. Geneticists working on the analysis of human mitochondrial DNA (MtDNA) variations originally developed this method (Bandelt et al., 1995; Bandelt et al., 1999). Here, the fingerprinting data (most of the time a DNA sequence) are considered as a set of binary characters. Absence of recombination, an hypothesis that may not always be true for spoligotyping, and a limited amount of ambiguous characters are two prerequisite assumptions. According to the author (Bandelt et al., 1995, 1999), a network that displays alternative potential evolutionary paths in the form of cycles may best express the multitude of plausible intraspecific trees. The method ("median joining" or MJ) for constructing networks from recombination-free population data combines features of Kruskal's algorithm for finding minimum spanning trees by favoring short connections, and Farris's maximum-parsimony (MP) heuristic algorithm, which sequentially adds new vertices called "median vectors". The MJ method is closely related to an earlier approach described for estimating MP trees but it can be adjusted to the level of homoplasy by setting a parameter epsilon (Foulds et al., 1979).

2.2.4. GEODIS (cladistic nested analysis)

This program implements the "cladistic nested analysis" (Posada et al., 2000). This methodology was first described by Templeton (Templeton et al., 1992, 1995). Using this

Table 1
Spoligotyping data from 321 *Mycobacterium tuberculosis* clinical isolates from four different areas of the Caribbean Islands

Type n°	Spoligotype description (binary)	Total n°	G	M	C	H
12	■□■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
13	■□■□□□□□□□□□■■□■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
14	■□■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	22	21	1	0	0
92	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	0	0	2	0
70	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■□■■■■	6	2	0	0	4
91	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	0	5
71	■■■■■■■■□□■■□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	5	0
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	44	20	4	16	4
119	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	4	0	3	0	1
51	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□□□	5	5	0	0	0
60	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■	3	0	0	3	0
81	■■■■■■■■■□■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	0	0	10	0
20	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	3	2	4	1
17	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	19	9	2	3	5
93	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	8	3	0	0	5
33	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	11	0	0	11	0
42	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	28	9	0	14	5
5	□□□□□□□□□■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	4	1	2	0	1
80	■■■■■□■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■■■	6	0	0	6	0
45	■■■■■■■■■■■■■■■■■■■■■■■□■□□□□□□■□□□□■■■■■■■	10	4	6	0	0
47	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	8	1	0	7	0
2	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■	35	7	2	15	11
62	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	5	0	4	1	0
103	■■■■■■■■■■■■■■■■■■■■■■□■■□□□□□□■□□□□■■■□■■■	3	3	0	0	0
50	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	38	11	2	16	9
58	■■■■■■■■■■■■■■■■■■■□■■□■■■■■■■■■□□□□■■■■■■■	5	0	0	5	0
C231	■■■□□■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	1	0	0	1	0
C260	■■■■■□■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■□□	1	0	0	1	0
C308	■■■■■■■■□□□■■■■□■■■■□□□□■■■■■■■■□□□□■■■■■■■	1	0	0	1	0
C309	■■■■■■■■□□□■■■■■■■□■□□□□■■■■■■■■□□□□■■■■■■■	1	0	0	1	0
C391	■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■□■□□□□■■■■■■■	1	0	0	1	0
C425	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■□□■■	1	0	0	1	0
C5	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■□■■■■■	1	0	0	1	0
C7	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■□	1	0	0	1	0
G1	□□□□□□□□■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	1	1	0	0	0
G2	■■■■■■□□□□□□□□□□□□□■□□□□■■■■■■■■□□□□■■■■■■■	1	1	0	0	0
G3	■■■■■■■■■■■■■■□■■■■■■■□■■□□□□□□■□□□□■■■□■■■	1	1	0	0	0
G4	■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□□□■■■■■	1	1	0	0	0
G5	■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■□□□□	1	1	0	0	0
G6	■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■□□□□■■■■■■■	1	1	0	0	0
G7	■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■□■□□□□■■■■■■■	1	1	0	0	0
H2	■■□■■■■■■■■■■■■■■■■■□□□□■■■□□□■■□□□□■■■■■■■	1	0	0	0	1
H4	■■□□■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■	1	0	0	0	1
H5	■■■■■■■■□□□■■■■■■■■■□□□□□□□□□■■■□□□□■■■■■■■	1	0	0	0	1
H6	■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	1	0	0	0	1
M1	■■■■■□□□■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	1	0	1	0	0
H7	□□□□□□□□□■■□■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	1	0	0	0	1
		321	**110**	**29**	**126**	**56**

G: Guadeloupe; M: Martinique; H: Haiti and C: Cuba. Type no.: nomenclature of the shared-types, types present at least twice, according to Sola et al., 2001a, p 745. Clinical isolates present only once (orphan isolates) are designated according to their country of origin (e.g. C231, isolate no. 231 from Cuba).

methodology, population structure can be separated from population history through rigourous statistical testing upon an estimated nested cladrogram. The input file consists of a description of the nested cladogram. The input file of Fig. 4 is available upon request from the corresponding author. The principles of the construction of the nested cladrogram have been described previously (Templeton et al., 1992). Spoligotypes are first clustered according to their common characteristics, i.e. the presence or absence of a given spacer. The starting hypothesis is that all the types are derived from a single ancestor whose spoligotype was made up of all 43 spacers. The hypothesis of no recombination between the order of the spacers (corroborated by recent findings on the structure of the DR locus; van Embden et al., 2000), is also assumed. The network is built by successive deletions of one to many consecutive spacers. These deletions are assumed to proceed in time by the accumulation of genetic elementary events such as replication slippage, insertion-mediated transposition, and homologous recombination (van Embden et al., 2000). The strains harboring lesser differences are linked to each other in a parsimonious approach. Once the tree has been built, the process used to create clades and to organize into a hierarchy the tree is followed as previously described (Templeton et al., 1992, 1995; Templeton and Sing, 1993). In this method, one always starts from the tips. If one or many tip-haplotypes are directly linked to an inner-haplotype, this group constitutes a step-one clade. It is designated as 1-x. The same is done for clades, i.e. if one or more tip-clades (or haplotypes) are directly connected to an inner-clade (or an inner haplotype), then this group constitutes a super-clade of step two, and so on, until all haplotypes are nested. No clades are created when no geographical or genetic variations are observed. Finally, the probability H is computed to estimate the validity of the parsimonious network created with the dataset (Templeton et al., 1992). On a given sample of n haplotypes, H is the probability by which two random types differ by more than one step (non-parsimonious state). Sample locations are treated as categorical variables and an exact permutational contingency test is performed for any clade at each nesting level. A χ^2 statistics is calculated from contingency tables in which rows represent genetic clades, and columns are geographical locations (Roff and Bentzen, 1989).

All softwares were run on a first generation iMac (Apple computer, Cupertino, California, USA), except for

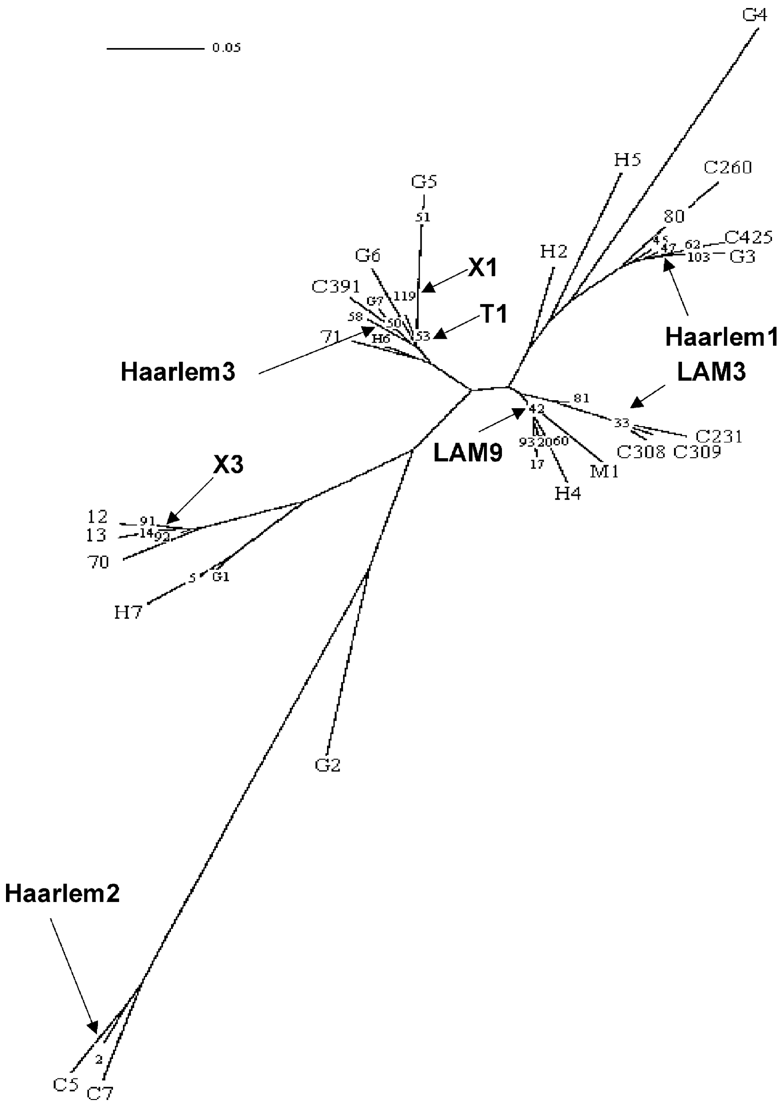

Fig. 1. Unrooted spoligotyping-based neighbor-joining phylogenetic tree. Results are based on pairwise calculation of the Jaccard similarity index $(1 - S_J)$ across the 47 alleles representative of 321 different *Mycobacterium tuberculosis* caribbean clinical isolates used in the present study. Note that four major groups are clearly defined by the method, and that four strains, i.e. strains G2, 2, C5, C7, are at a wrong place on this tree due to the peculiar structure of their DR locus.

NETWORK which was run on a Hewlett-Packard PC "HP brio", (Palo Alto, California, USA).

3. Results

3.1. Results of the TAXOTRON analysis

Fig. 1 depicts the unrooted phylogram constructed with TAXOTRON using the neighbor-joining (NJ) algorithm. This tree is made up of four main branches. These four distinct groups of *M. tuberculosis* strains are made up of three previously defined (Haarlem, LAM, and X) plus a poorly defined (T1) large clade of tubercle bacilli. This branch also clusters X1 and Haarlem 3, subfamilies of, respectively, the X and Haarlem (Filliol et al., 2002). Two minor artefact branches are observed, which include four patterns (G2, 2, C5 and C7) that appear unrelated to any of the other families.

3.2. Results of the PAUP analysis

The PAUP analysis corroborated the previous analysis (Fig. 2) as the global architecture into four main groups was conserved. In this tree, a new phylogenetic link between ST50 (Haarlem3) and ST47 (Haarlem1) through the G7 haplotype is suggested. As already seen in Fig. 1, The LAM family is split into at least two sub-branches ST42 (LAM9) and ST33 (LAM3), two prevalent alleles world-wide, and a geographic association between ST33 and Cuba may be noticed, which argues in favor of a Spanish origin of these alleles. In agreement with Fig. 1, another likely phylogenetical link between G5 and ST51 was also suggested. Lastly, strains G2 and the group 2, C7, C5 once again clearly cluster apart from the rest of the tree due to their peculiar structure (see Table 1). As with Taxotron, some major alleles (ST50, prototype allele of Haarlem 3 and ST119, prototype allele X1) are lost among the large undefined group T1.

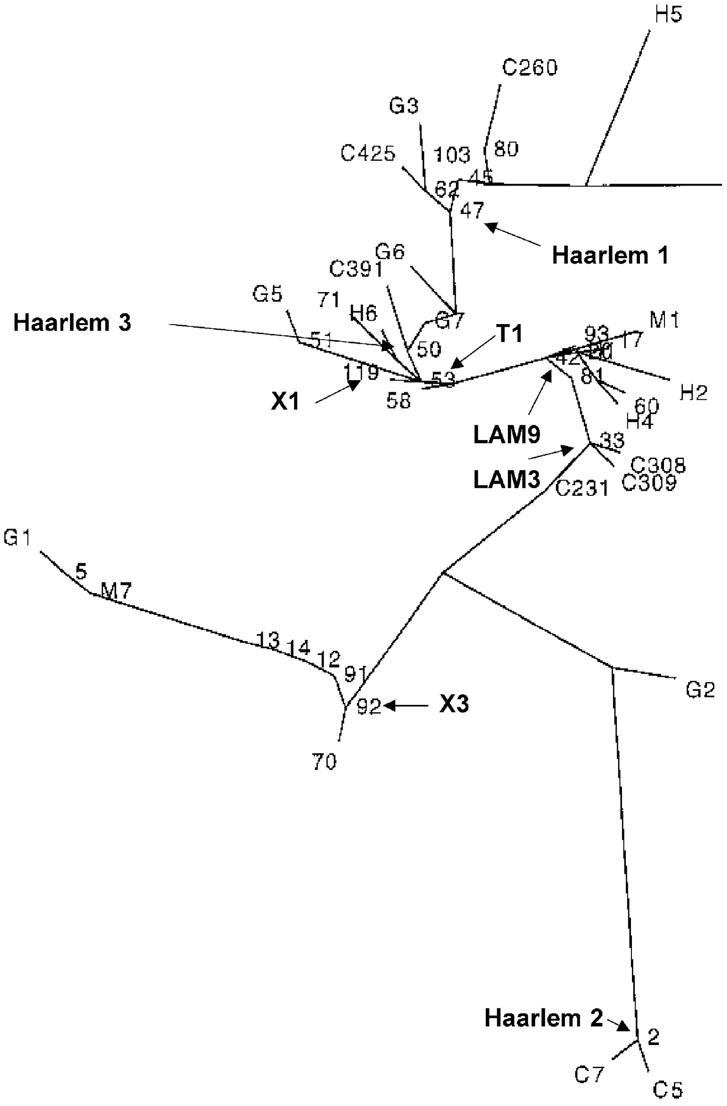

Fig. 2. Unrooted parsimony phylogenetic tree based on similar data as presented for Fig. 1. Fig. 2 gives similar results as those illustrated on Fig. 1 in that (i) the same four phylogenetic groups of strains are clearly identified, and (ii) strains G2, 2, C2 and C7 cluster apart but cannot be analyzed.

3.3. Results of the NETWORK analysis

The results obtained using the NETWORK software which implements the “median joining” (MJ) algorithm, were partly similar to those obtained by Taxotron or PAUP (Fig. 3). One important advantage of this last method is that the frequency of each haplotype is used in the visual display of the clusters since the diameter of each clade is proportional to the sample size. Consequently, a new population genetic dimension related to the sampling representation of strains within the studied dataset is included within Fig. 3. One major disadvantage is that this type of graphic, built in three dimensions, may be too complicated. Population genetic shows that *M. tuberculosis* most represented alleles are located within the central core of the phylogenetic network, e.g. ST53, ST50, and ST42 in the Fig. 3. Unfortunately, the three softwares used above (TAXOTRON, PAUP and NETWORK) do not give an exhaustive and true picture of all potential phylogenetical links, e.g. ST2, G2, C5 and C7 positions on the tree remains unresolved. Indeed, given the structure of these alleles (absence of contiguous blocks of spacers), the calculation of the pairwise distance between these spoligotypes and ST47 (for ST2, C5 and C7) or with ST42 (for G2) leads to an artifact in distance calculations, hence a wrong position in the tree. The congruence observed on some Variable Number of Tandem DNA Repeats and Mycobacterial Interpersed Repetitive Units (VNTR-MIRU; Sola et al., 2001a, 2003) results between the Haarlem family and ST2, C5 and C7 (VNTR allele = 32 333 in all cases) and the similarities between ST47 and ST2, led us to include this group of strains within the Haarlem family with the designation “Haarlem 2” (Filliol et al., 2002). This classification is not displayed using either of the three first methods.

3.4. Results of the GEODIS analysis

With GEODIS, a new concept—the nested clade concept and a new dimension—the geographical location of the data, is introduced in the analysis (Fig. 4 and Table 2). This new phylogenetic network (Fig. 4) suggests a new scenario of evolution between all the alleles reported in the present study. The display of the results is both highly structured and highly likely from an evolutionary standpoint. The Haarlem 2 strains are now located close to their likely ancestor, the Haarlem 3 family (Fig. 4). Although the overall test of

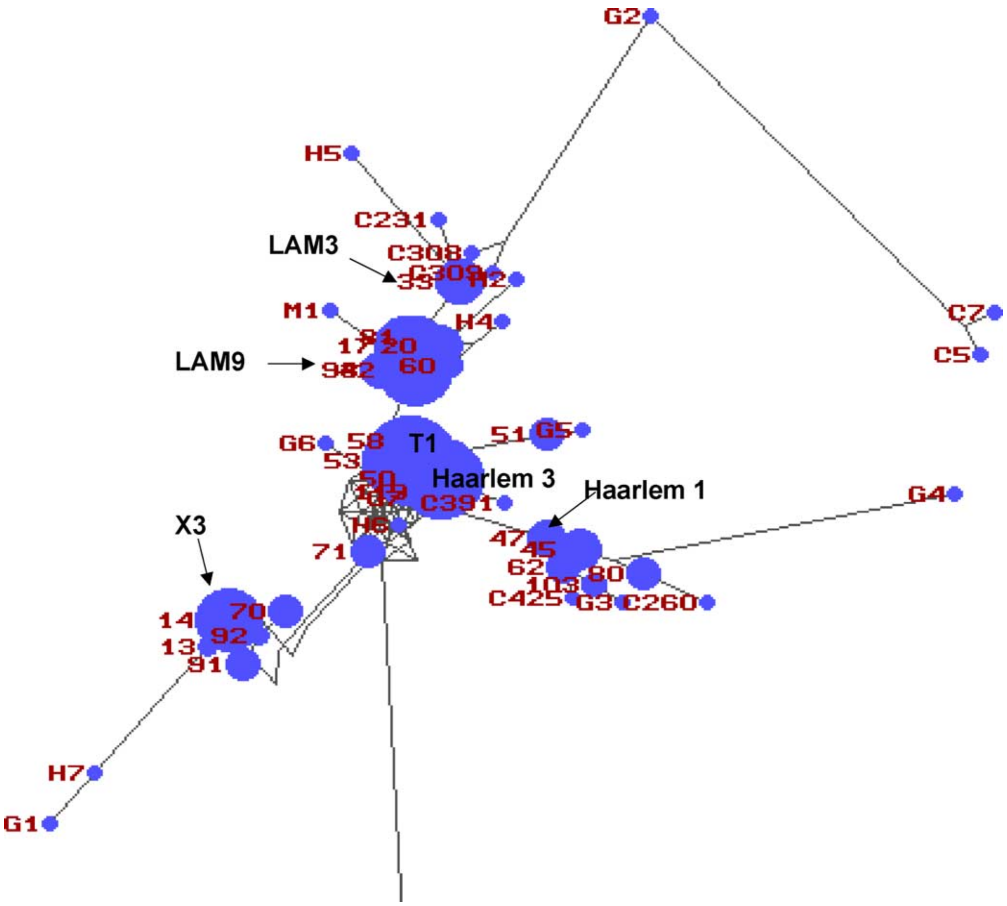

Fig. 3. Median-joining network constructed using the NETWORKsoftware®, and based on the same data-set as those used for reconstructing Figs. 1 and 2. The size of the circles is proportional to the number of strains harboring a given spoligotype. Note here that the hypercube in the center of the figure reflects the impossibility for identifying both the nature and the sense of genetic events at the early origin of the strain epidemics.

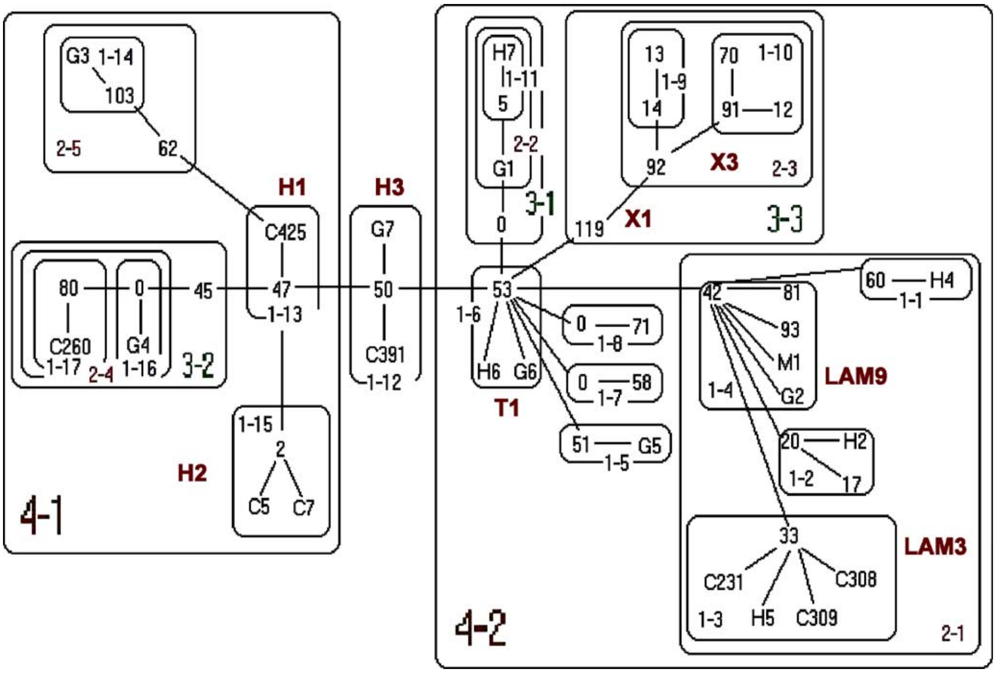

Fig. 4. Cladistic nested analysis (cladogram) based on 47 haplotypes—spoligotyping data of *M. tuberculosis* clinical isolates from Cuba, Haïti and the French Antilles. On the graph, “0” indicates intermediate haplotype states, which are missing in the sample. Step 1 clades are designated as 1-*X* (with *X* the code number of a given clade), and similarly Step 2 clades are labelled 2-*X*, and so on. The nomenclature of superfamilies is based on the analysis of a world-wide spoligotyping database (SpolDB3.0) which contained 817 shared-types representative of 11 708 clinical isolates from 90 countries (full nomenclature is described in (Filliol et al., 2002, 2003).

Table 2
Analysis of association between nested clades as defined on Fig. 4 and their respective geographic locations

Exact contigency test		
Clade	Chi-square observed	Probability
1-1	4.0000	0.2570
1-2	6.7274	0.3850
1-3	15.0000	0.1550
1-4	71.8132	0.0000
1-6	9.5584	0.3530
1-9	0.0949	1.0000
1-10	6.7407	0.0590
1-11	1.8750	1.0000
1-12	3.7307	0.6840
1-13	0.1406	1.0000
1-15	2.4874	1.0000
2-1	25.1300	0.0080
2-2	2.4000	1.0000
2-3	62.2407	0.0000
2-4	8.0000	0.1260
2-5	9.0000	0.0100
3-2	14.7600	0.0000
3-3	23.4433	0.0000
4-1	31.3806	0.0000
4-2	66.4122	0.0000
Cladrogram total	12.0038	0.0570

Statistics are based on Chi-square deviation at the 5% level. The overall test of association is marginally not significant ($P = 0.0570$), but eight groups show a strong genetic association with their geographical location: clade (1-4), clades (2-1, 2-3 and 2-5), clades (3-2 and 3-3) and clades (4-1 and 4-2). For each clade, results relies on permutation analysis based on 1000 resamples.

association between nested clades and geographic location is marginally non significant ($P = 0.0570$), for eight clades, this contigency test (permutation with 1000 resamplings) shows a strong correlation, suggesting that there is a strong geographical association between these clades and their location. This concerns one step-1 clade (1-4) or LAM9, three step-2 clades (2-1, 2-3 and 2-5, respectively LAM, X3 and the ST62 clade, two step-3 clades (3-2 and 3-3, respectively the ST45 clade and X), and two step-4 clades (4-1 and 4-2, broadly equivalent to Haarlem and T1). Clade 1-4 is based on the ST42 which is at the node of the LAM9 family. Clade 2-1 is the broader LAM superfamily enclosing two nodes, ST20 and ST33 (LAM3). These alleles have a high prevalence as shown in the spolDB3 database (Filliol et al., 2003). Clade 2-3 defines the X3 clade, likely to be of Caribbean, English or Indian origin. Clade 2-5 is a specific, yet potentially homoplasic, evolution of Haarlem1 (node ST62) since it is found both in Cuba, Martinique and Guadeloupe. Clade 3-2 looks like another specific evolution of Haarlem1 (node ST45), whereas Clade 3-3 defines the large X family as a whole ($X1 + X3$). Last but not least, clades 4-1 and 4-2 define, respectively, “Haarlem” and T1 families.

4. Discussion

In this study, we used four methods of phylogenetic reconstruction on a set of randomly selected isolates from the Caribbean region. Congruent results between the four methods were obtained. Geodis provides the best results in terms of likelihood and quality of display. We suggest that, based on the structure of the Direct Repeat locus, four main super-families of *M. tuberculosis* or super-haplogroups are historically prevalent in this region. All of these strains belong to the principal genetic groups 2 and 3 in Sreevatsan's classification (Sreevatsan et al., 1997), arguing in favor of a recent (after 1492 A.D.) spread of tuberculosis in the Caribbean. This feature is in sharp contrast to what is observed in Asia, where Beijing, East African-Indian (EAI) and Central-Asian (CAS) TB bacilli all belong to principal genetic group 1 organisms, and are likely to be of more ancient origin (Filliol et al., 2002, 2003).

The first major clade, the "Haarlem" superfamily, was initially found to be prevalent in European countries, and named according to the Dutch town, Haarlem, where it was first discovered (Kremer et al., 1999). It is characterized by the absence of spacers 26–31 and 33–36 (ST47). The absence of spacer 31 inside Haarlem strains is often linked to an asymetrically-inserted second copy of IS*6110* within the DR locus hindering its detection (Filliol et al., 2000; Legrand et al., 2001). The Haarlem superfamily is currently subdivided into at least three families; Haarlem 1 (ST47), Haarlem 2 (ST2), and Haarlem 3 (ST50) (Filliol et al., 2002, 2003). Assuming the hypothesis of evolution of the DR locus by loss of spacers (van Embden et al., 2000), ST50 would definitively be older than ST47. The presence of Haarlem may thus be related to European settlers in the Caribbean.

The second clade, the Latin American and Mediterranean or "LAM" superfamily (split into variants, of which LAM3 and LAM9, Fig. 1), has been recently defined by the simultaneous absence of spacers 21–24 and 33–36 (Sola et al., 2001b). This signature (ST42), may in some cases recover homoplasic events, however, the congruence observed between spoligotyping and VNTR-MIRU data (Sola et al., 2001a, 2003), as well as the geographical prevalence of the LAM family around the Mediterranean basin and Latin America justifies its existence as a broad superfamily of strains (Sola et al., 2001b). Based on their respective structures, ST33 and its sub-lineages are likely to be younger than ST42 and would belong to Sreevastan's group 3 organisms as already shown for some of these alleles (Soini et al., 2000). This family of strains (under the designation F11) has recently been shown to be prevalent in South Africa (Warren et al., 2002). The ST51 genotypes, are also prevalent in the Western Mediterranean islands of Sicily and Sardinia, both linked to a Spanish heritage, and are also likely to belong to group III organisms in Sreevatsan's classification (Soini et al., 2000). As shown for DR structures implying a left or right deletion (Warren et al., 2002), some convergence events on this specific DR structure (ST51) may not be excluded in this stage. In our setting, the LAM3, LAM9, and ST51 alleles would represent the traces of the pioneering Spanish influence on the Caribbean history.

The third clade designating the X superfamily was initially defined by the simultaneous absence of spacer 18 and 33–36, and reported as highly prevalent in English-speaking countries such as the United Kingdom and the USA (Sebban et al., 2002). Since this first definition, this group of strains has been shown to be part of a large group of European (largely English) and American IS*6110* low-copy number strains. Such isolates are highly prevalent today in the American State of Michigan and more generally in the USA (Cowan et al., 2002a,b), but may also have been prevalent in UK for decades or even centuries. We recently suggested that this superfamily could be subdivided into at least three distinct families, X1 to X3 (Filliol et al., 2002). Studies using other markers such as synonymous single-nucleotide polymorphisms agree with the definition of this superfamily of strains (Dale et al., 2003; Gutacker et al., 2002; Alland et al., 2003). These strains may reflect the long-lasting Anglo-Saxon colonial influences on Caribbean history, and tracking their current evolution and expansion may also help to trace transmission between Central and North America in certain settings (Ferdinand et al., 2003).

T1, the last clade ("ill-defined" or defined by default), is characterized by the absence of spacers 33–36. Its prototypic shared type is ST53. It belongs to groups II and III organisms according to Sreevatsan 's classification (Soini et al., 2000). This family clusters various strains whose evolution remains unresolved (e.g. 71, H6). ST53 alleles are highly prevalent in some countries in Africa such as the Ivory Coast (N. Guessend, unpublished data). We may suggest the African origin of these alleles.

Understanding the genetic variability of *M. tuberculosis* through space and time is of recent interest (van Embden and van Soolingen, 2000). One may always argue the real value of using a single locus (the DR locus), the evolutionary mode of which is unclear, to derive phylogenetic relationships among strains. However, previous studies have demonstrated, (i) the usefulness of this sole locus to define some highly important families of tubercle bacilli, i.e. Beijing (van Soolingen et al. 1995), LAM (Sola et al., 2001a), EAI (Sola et al., 2001b), CAS (Filliol et al., 2002), X (Sebban et al., 2002) and, (ii) given the clonal structure of tubercle bacilli populations, the congruence of phylogenetical analysis whatever the markers used, i.e. MIRU-VNTR and spoligotyping has previously been demonstrated (Sola et al., 2001b, 2003).

Since Cockburn's hypothesis, the traditional historical belief in *M. tuberculosis* epidemiology was that the bovine tubercle bacilli (*M. bovis*) 'appeared before the human form' (Cockburn, 1963). Nevertheless, recent genetic evidence opposes this viewpoint, suggesting that human forms of tubercle bacilli (*M. tuberculosis*, *M. africanum*) preceded the bovine subspecies (Brosch et al., 2002; Zink et al., 2003). It has been suggested that a close human community of around 180–440 persons would be enough in most cases for endemic disease and vertical transmission of TB (from parents to children) to occur (McGrath, 1988). Consequently, the hypothesis that tuberculosis bacilli diversity may be highly structured on a geographical scale is likely, and the analysis of available genotyping TB databases agrees with this assumption (Cowan and Crawford, 2002b; Filliol et al., 2002, 2003). This fact prompted us to investigate the specific co-evolutionary history between human beings and tubercle bacilli in our setting. The Caribbean region, by its insular nature and history, is undoubtedly a setting where one may expect to unravel the TB spreading history using genetic fingerprinting. Indeed, the Caribbean region constitutes a heterogeneous human settlement with a rich *M. tuberculosis* genetic diversity (Sola et al., 1999; Diaz et al., 2001; Ferdinand et al., 2003). This diversity could either reflect an old (pre-Columbian history), a more recent (post-Columbian history), or an admixture of both. In the last 500 years, within all the Caribbean islands, European, African and later Indian settlers have rapidly replaced the original anthropological structure which stemmed from the South American continent. Consequently, it may not be surprising that our sampling does not seem to detect any evidence of an ancient and specific founding endemic strains of tubercle bacilli that would have been specific for the first Caribbean inhabitants, the Amerindians. Our results support the view that the history of tuberculosis spread in the Caribbean is linked to the recent (post-Columbus) human migratory and demographic events. Nevertheless, it should also be mentioned that recent debates argued in favor of the presence of tuberculosis in America before the Columbus era (Salo et al., 1994; Arriaza et al., 1995). Genotyping on ancient TB American DNA and comparison of the population structure of TB bacilli in South America and the Caribbean may solve this issue. Among other external influences not studied here, a specific historical Chinese immigration in Cuba, may explain the historical presence of the Beijing family of strains in this particular setting (Diaz et al., 2001). However, a particular population structure can be the result of distinct processes, e.g. adaptation, acting in different points through time and space, or genetic drift resulting from the process of disease transmission dynamics. It may also reflect historical rather than ongoing genetic population level processes (Excoffier and Mouse, 1994). New extended studies combining both anthropological, genetic and historical approaches are needed to shed further light on the natural history of the TB disease in Central and South America, as well as to comprehend the factors affecting its global transmission.

Acknowledgements

This work was made possible thanks to collaboration established between the Gheskio Centers in Port au Prince, Pr. J.W. Pape, and the Institut Pasteur, through a grant from the ANRS (Agence National de Recherche contre le SIDA). The "Unité de la Tuberculose et des Mycobactéries" is supported by grants from the "Réseau International des Instituts Pasteur et Instituts Associés" and "Fondation Raoul Follereau", Paris. The "Unité de Génétique des Maladies Infectieuses" is supported both by the I.R.D. (Institut de Recherche pour le Développement) and the C.N.R.S. (Centre National de la Recherche Scientifique). V. D. was a 3rd year summer student from the "Ecole de Biotechnology de Cergy-Pontoise". Lastly, we are grateful to the two anonymous reviewers who helped us to significantly improve this paper.

References

Adelaïde-Merlande, J., 1994. Histoire Générale des Antilles et des Guyanes. L'Harmattan, Paris.

Alland, D., Whittam, T.S., Murray, M.B., Cave, M.D., Hazbon, M.H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J.A., Fraser, C.M., Fleischmann, R.D., 2003. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J. Bacteriol. 185, 3392–3399.

Arriaza, B.T., Salo, W., Aufderheide, A.C., Holcomb, T.A., 1995. Pre-Columbian tuberculosis in northern Chile: molecular and skeletal evidence. Am. J. Phys. Anthropol. 98, 37–45.

Bandelt, H.J., Forster, P., Sykes, B.C., Richards, M.B., 1995. Mitochondrial portraits of human populations using median networks. Genetics 141, 743–753.

Bandelt, H.J., Forster, P., Rohl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16, 37–48.

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

Cockburn, A., 1963. The Evolution and Eradication of Infectious Diseases. John Hopkins Press, Baltimore.

Cowan, L.S., Crawford, J.T., 2002b. Genotype Analysis of *Mycobacterium tuberculosis* Isolates from a Sentinel Surveillance Population. Emerg. Infect. Dis. 8, 1294–1302.

Cowan, L.S., Mosher, L., Diem, L., Massey, J.P., Crawford, J.T., 2002a. Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis Isolates with Low Copy Numbers of IS*6110* by Using Mycobacterial Interspersed Repetitive Units. J. Clin. Microbiol. 40, 1592–1602.

Cowan, L.S., Mosher, L., Diem, L., Massey, J.P., Crawford, J.T., 2002b. Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110* by using mycobacterial interspersed repetitive units. J. Clin. Microbiol. 40, 1592–1602.

Dale, J.W., Al-Ghusein, H., Al-Hasmi, S., Butcher, P.D., Dickens, A., Drobniewski, F., Forbes, K.J., Gillespie, S., Lamprecht, D., McHugh, T.D., Pitman, R., Rastogi, N., Sola, C., Yesilkaya, H., 2003. Evolutionary relationships amongst isolates of *Mycobacterium tuberculosis* with few copies of IS*6110*. J. Bacteriol. 185, 2555–2562.

Diaz, R., Gomez, R., Restrepo, Evol., Rumbaut, R., Sevy-Court, J., Valdivia, J.A., van Soolingen, D., 2001. Transmission of tuberculosis in Havana, Cuba: a molecular epidemiological study by IS*6110* restriction fragment length polymorphism typing. Mem. Inst. Oswaldo. Cruz 96, 437–443.

Excoffier, L., Smouse, P.E., 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136, 343–359.

Ferdinand, S., Sola, C., Verdol, B., Legrand, E., Goh, K.S., Berchel, M., Aubéry, A., Timothée, M., Joseph, P., Pape, J.W., Rastogi, N., 2003. Molecular characterization and Drug-resistance of *Mycobacterium tuberculosis* patients in an AIDS counseling center in Port-au-Prince, Haiti: a one-year study. J. Clin. Microbiol. 41, 694–702.

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.

Filliol, I., Driscoll, J.R., van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valétudie, G., Anh, D.D., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Källenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H.M., Makristathis, A., Mammina, C., Martin, G., Moström, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Eyangoh, S.N.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., de Waard, J., Sola, C., Rastogi, N., 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg. Inf. Dis. 8, 1341–1343.

Filliol, I., Driscoll, J.R., Van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valetudie, G., Anh, D.D., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Kassa-Kelembho, E., Ly, H.M., Makristathis, A., Mammina, C., Martin, G., Mostrom, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., De Waard, J.H., Sola, C., Rastogi, N., 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J. Clin. Microbiol. 41, 1963–1970.

Foulds, L.R., Hendy, M.D., Penny, D., 1979. A graph theoretic approach to the development of minimal phylogenetic trees. J. Mol. Evol. 13, 127–149.

Grimont P.A.D., 1994. TAXOTRON. Institut Pasteur, Paris.

Grmek, M., 1994. Chap. VII: Une grande Tueuse, la tuberculose. Les maladies à l'aube de la civilisation occidentale. Payot, Paris, pp. 261–290.

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

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

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

Kremer, K., van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W.M., Martin, C., Palittapongarnpim, P., Plikaytis, B.B., Riley, L.W., Yakrus, M.A., Musser, J.M., van Embden, J.D.A., 1999. Comparison of methods based on different molecular epidemiologial markers for typing of *Mycobacterium tuberculosis* strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. 37, 2607–2618.

Legrand, E., Filliol, I., Sola, C., Rastogi, N., 2001. Use of spoligotyping to study the evolution of the direct repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. J. Clin. Microbiol. 39, 1595–1599.

McGrath, J.W., 1988. Social networks of disease spread in the lower Illinois valley: a simulation approach. Am. J. Phys. Anthropol. 77, 483–496.

Posada, D., Crandall, K.A., Templeton, A.R., 2000. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Mol. Ecol. 9, 487–488.

Roff, D.A., Bentzen, P., 1989. The statistical analysis of mitochondrial DNA polymorphisms: chi-squared and the problem of small samples. Mol. Biol. Evol. 6, 539–545.

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

Salo, W.L., Aufderheide, A.C., Buikstra, J., Holcomb, T.A., 1994. Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. Proc. Natl. Acad. Sci. U.S.A. 91, 2091–2094.

Sebban, M., Mokrousov, I., Rastogi, N., Sola, C., 2002. A Data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. Bioinformatics 18, 235–243.

Soini, H., Pan, X., Amin, A., Graviss, E.A., Siddiqui, A., Musser, J.M., 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. Inf. Dis. 5, 404–414.

Sola, C., Filliol, I., Guttierez, C., Mokrousov, I., Vincent, V., Rastogi, N., 2001a. Spoligotype database of *Mycobacterium tuberculosis*: Biogeographical distribution of shared types and epidemiological and phylogenetic perspectives. Emerg. Inf. Dis. 7, 390–396.

Sola, C., Filliol, I., Legrand, E., Mokrousov, I., Rastogi, N., 2001b. *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. 53, 680–689.

Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P., Rastogi, N., 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect. Genet. Evol. 3, 125–133.

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. U.S.A. 97, 9869–9874.

Swofford, D.L., 1998. PAUP. Sinauer Associates, Fitchburg, MA.

Swofford, D.L., Olsen, F.J., 1990. Phylogeny reconstruction. In: Molecular Systematics. Sinauer Associates, Sunderland, MA, pp. 411–515.

Templeton, A.R., Sing, C.F., 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. Part IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134, 659–669.

Templeton, A.R., Crandall, K.A., Sing, C.F., 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. Part III. Cladogram estimation. Genetics 132, 619–633.

Templeton, A.R., Routman, E., Phillips, C.A., 1995. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics 140, 767–782.

van Embden, J., van Soolingen, D., 2000. Molecular epidemiology of tuberculosis: coming of age. Int. J. Tuberc. Lung. Dis. 4, 285–286.

van Embden, J.D.A., van Gorkom, T., Kremer, K., Jansen, R., van der Zeijst, B.A.M., Schouls, L.M., 2000. Genetic variation and evolutionary origin of the Direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. 182, 2393–2401.

van Soolingen, D., Qian, L., de Haas, P.E.W., Douglas, J.T., Traore, H., Portaels, F., Qing, H.Z., Enkhsaikan, D., Nymadawa, P., van Embden, J.D.A., 1995. Predominance of a Single Genotype of *Mycobacterium tuberculosis* in Countries of East Asia. J. Clin. Microbiol. 33, 3234–3238.

Warren, R.M., Streicher, E.M., Sampson, S.L., Van Der Spuy, G.D., Richardson, M., Nguyen, D., Behr, M.A., Victor, T.C., Van Helden, P.D., 2002. Microevolution of the direct repeat region *of Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J. Clin. Microbiol. 40, 4457–4465.

Zink, A., Sola, C., Reischl, U., Grabner, W., Rastogi, N., Wolf, H., Nerlich, A.G., 2003. Characterization of *Mycobacterium tuberculosis* complex findings from Egyptian mummies by spoligotyping. J. Clin. Microbiol. 41, 359–367.

World Health Organization, 1994. Tuberculosis—a global emergency. W.H.O. Report on the TB epidemic. World Health Organization, Geneva, Switzerland.