JOURNAL OF CLINICAL MICROBIOLOGY, May 2003, p. 1963–1970 Vol. 41, No. 5
0095-1137/03/$08.00+0 DOI: 10.1128/JCM.41.5.1963–1970.2003

Snapshot of Moving and Expanding Clones of *Mycobacterium tuberculosis* and Their Global Distribution Assessed by Spoligotyping in an International Study†

Ingrid Filliol,[1] Jeffrey R. Driscoll,[2] Dick van Soolingen,[3] Barry N. Kreiswirth,[4] Kristin Kremer,[3] Georges Valétudie,[1] Dang Duc Anh,[5] Rachael Barlow,[6] Dilip Banerjee,[7] Pablo J. Bifani,[4] Karine Brudey,[1] Angel Cataldi,[8] Robert C. Cooksey,[9] Debby V. Cousins,[10] Jeremy W. Dale,[11] Odir A. Dellagostin,[12] Francis Drobniewski,[13] Guido Engelmann,[14] Séverine Ferdinand,[1] Deborah Gascoyne-Binzi,[6] Max Gordon,[1] M. Cristina Gutierrez,[15] Walter H. Haas,[16] Herre Heersma,[3] Eric Kassa-Kelembho,[17] Ho Minh Ly,[5] Athanasios Makristathis,[18] Caterina Mammina,[19] Gerald Martin,[20] Peter Moström,[1] Igor Mokrousov,[21] Valérie Narbonne,[22] Olga Narvskaya,[21] Antonino Nastasi,[23] Sara Ngo Niobe-Eyangoh,[15] Jean W. Pape,[24] Voahangy Rasolofo-Razanamparany,[25] Malin Ridell,[26] M. Lucia Rossetti,[27] Fritz Stauffer,[28] Philip N. Suffys,[29] Howard Takiff,[30] Jeanne Texier-Maugein,[31] Véronique Vincent,[15] Jacobus H. de Waard,[32] Christophe Sola,[1]* and Nalin Rastogi[1]*

Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Pointe-à-Pitre, Guadeloupe[1]; Wadsworth Center, New York State Department Of Health, Albany, New York[2]; Public Health Research Institute, Tuberculosis Center, Newark, New Jersey[4]; Diagnosis Laboratory for Infectious Diseases and Perinatal Screening RIVM, Bilthoven, The Netherlands[3]; National Institute of Hygiene and Epidemiology, Hanoi, Vietnam[5]; Department of Microbiology, Leeds General Infirmary, Leeds,[6] Medical School, St Georges Hospital,[7] and Mycobacterium Reference Unit, PHLS, Dulwich Hospital,[13] London, and University of Surrey, Guildford, Surrey,[11] United Kingdom; Instituto de Biotecnologia INTA, Moron, Argentina[8]; Tuberculosis Mycobacteriology Branch, Centers for Disease Control and Prevention, Atlanta, Georgia[9]; Australian Reference Laboratory for Bovine Tuberculosis, Department of Agriculture, South Perth, Australia[10]; Centro de Biotecnologia, Universidade Federal de Pelotas, Pelotas,[12] Universidade Federal do Rio Grande do Sul, Porto Alegre,[27] and Department of Biochemistry and Molecular Biology, Fiocruz, Oswaldo Cruz Institute, Rio de Janeiro,[29] Brazil; University Childrens Hospital, Heidelberg,[14] Infektions Epidemiology, Robert Koch Institute, Berlin,[16] and Bundesinstitut für Gesundheitlichen Verbraucherschutz und Veterinärmedizin, Jena,[20] Germany; Centre National de Référence des Mycobactéries, Institut Pasteur, Paris,[15] Laboratoire de Bactériologie, CHU de Brest, Brest,[22] and Laboratoire de Bactériologie, CHU de Bordeaux, Bordeaux,[31] France; Institut Pasteur de Bangui, Bangui, Central African Republic[17]; Klinische Mikrobiologie, Hygiene-Institut der Universität,[18] and Bundesstaatliche Bakteriologisch-Serologische Untersuchungsanstalt Wien,[28] Vienna, Austria; Department of Hygiene and Microbiology, University of Palermo, Palermo,[19] and Department of Public Health, University of Florence, Florence,[23] Italy; Laboratory of Molecular Microbiology, Saint Petersburg Pasteur Institute, Saint Petersburg, Russia[21]; Les Centres Gheskio, INLR, Port au Prince, Haïti[24]; Institut Pasteur de Madagascar, Tananarive, Madagascar[25]; Institute of Medical Microbiology and Immunology, Göteborg University, Göteborg, Sweden[26]; and IVIC, Centro de Microbiología y Biología Celular, Laboratorio de Genética Molecular,[30] and Tuberculosis Laboratory, Instituto de Biomedicina,[32] Caracas, Venezuela

Received 21 October 2002/Returned for modification 6 January 2003/Accepted 31 January 2003

The present update on the global distribution of *Mycobacterium tuberculosis* complex spoligotypes provides both the octal and binary descriptions of the spoligotypes for *M. tuberculosis* complex, including *Mycobacterium bovis*, from >90 countries (13,008 patterns grouped into 813 shared types containing 11,708 isolates and 1,300 orphan patterns). A number of potential indices were developed to summarize the information on the biogeographical specificity of a given shared type, as well as its geographical spreading (matching code and spreading index, respectively). To facilitate the analysis of hundreds of spoligotypes each made up of a binary succession of 43 bits of information, a number of major and minor visual rules were also defined. A total of six major rules (A to F) with the precise description of the extra missing spacers (minor rules) were used to define 36 major clades (or families) of *M. tuberculosis*. Some major clades identified were the East African-Indian (EAI) clade, the Beijing clade, the Haarlem clade, the Latin American and Mediterranean (LAM) clade, the Central Asian (CAS) clade, a European clade of IS*6110* low banders (X; highly prevalent in the United States and United

* Corresponding author. Mailing address: Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Morne Jolivière, BP 484, 97165 Pointe à Pitre-Cedex, Guadeloupe. Fax: 590 (590) 893 880. E-mail for Christophe Sola: csola@pasteur-guadeloupe.fr. E-mail for Nalin Rastogi: nrastogi@pasteur-guadeloupe.fr.

† This work is dedicated to the memories of Anne Devallois, who initiated the spoligotyping project in Guadeloupe in 1996 but died tragically at the age of 30 years, and Gerald Martin, coinvestigator, who also died tragically during the course of the present study.

Kingdom), and a widespread yet poorly defined clade (T). When the visual rules defined above were used for an automated labeling of the 813 shared types to define nine superfamilies of strains (*Mycobacterium africanum*, Beijing, *M. bovis*, EAI, CAS, T, Haarlem, X, and LAM), 96.9% of the shared types received a label, showing the potential for automated labeling of *M. tuberculosis* families in well-defined phylogeographical families. Intercontinental matches of shared types among eight continents and subcontinents (Africa, North America, Central America, South America, Europe, the Middle East and Central Asia, and the Far East) are analyzed and discussed.

Tuberculosis (TB) remains a major killer, with >8 million new cases and >2 million deaths each year. TB control essentially relies on improvement of expanded and reliable local microbiological diagnostic capacities, the availability of drugs on a worldwide basis, and compliance through adequate treatment strategy. Other related problems include the AIDS epidemic, less access to health services in countries needing it badly, and increasing multidrug resistance. Due to increased human migration from high-prevalence areas, there may also be a danger of the spread of multidrug-resistant TB and consequently a need for earlier detection of new outbreaks (8). A better knowledge of moving and expanding clones, such as Beijing (2, 3), that may harbor various degrees of virulence (25) is also urgently needed. In this context, the genotyping of tubercle bacilli and the study of the virulences of various clones from different settings may help to better define TB control measures.

Complementary to traditional epidemiology, molecular epidemiology based on PCR fingerprinting methods, such as spacer oligonucleotide typing (spoligotyping) (13), has emerged as a fast, reliable, and cost-effective alternative to traditional IS*6110* restriction fragment length polymorphism (RFLP) fingerprinting. Based on the variability of the direct-repeat (DR) locus (9, 10, 13, 14, 27, 29), spoligotyping is useful both for tracking epidemics (2, 7, 12, 13, 15, 32) and to detect new outbreaks and better define high-risk populations in order to focus prevention strategies on the subpopulations that need them most (28). In this context, the construction of polymorphism databases constitutes a powerful tool for studying the epidemiology and evolutionary genetics of tubercle bacilli (23) and, ultimately, the mechanisms of genetic variability by using data-mining methods (19). However, previous databases were only poorly representative of the worldwide diversity of *Mycobacterium tuberculosis* genomes (20, 21, 22), e.g., the previously published 259 shared types (identical spoligotypes shared by two or more patient isolates; available online at http://www.cdc.gov/ncidod/EID/vol7no3/sola_data.htm) contained more than two-thirds of the isolates from Europe and the United States. Nonetheless, these initial studies supported the fact that a significant number of *M. tuberculosis* isolates are confined to specific geographic locations (20, 21, 22).

In this article, we describe an update on the global distribution of *M. tuberculosis* complex spoligotypes (SpolDB3.0). With a total of 13,008 patterns from >90 countries, the SpolDB3.0 database contains both the octal (4) and binary (13) descriptions of the spoligotypes for all the current *M. tuberculosis* complex members (*M. tuberculosis*, *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium microti*, *Mycobacterium canetti*, and *Mycobacterium caprae*). It also better describes the diversity of TB genotypes, since the 24 most prevalent alleles now represent only 53% of the total instead of 65%, as in a previous study (22). Similarly, the computation of the genetic diversity index (*H* [24]) gives an *H* of 97.4% with the current version of the database compared to 93% with the previous version of the database.

MATERIALS AND METHODS

Database. Spoligotyping was performed using a methodology reported earlier (13). It is assumed that most of the new shared types represent true polymorphism and highlight both past and present population genetics of the *M. tuberculosis* complex. A total of 558 new alleles, observed at least twice, were added to the previously published data (22). The database (SpolDB3.0), containing shared types ST1 to 817, is available upon request or at http://www.pasteur-guadeloupe.fr/tb/spoldb3.htm (note that ST 633, 714, 729, and 770 were removed from the database due to artifacts and will be attributed at a later stage). SpolDB3.0 contains a total of 11,708 entries in an Excel spreadsheet, representative of various continents as follows: Africa, 1,303 (Burundi, *n* = 12; Burkina Faso, *n* = 1; Central African Republic, *n* = 122; Cameroon, *n* = 380; East Africa [undefined], *n* = 15; Egypt, *n* = 25; Ethiopia, *n* = 2; Guinea-Bissau, *n* = 189; Ivory Coast, *n* = 25; Kenya, *n* = 8; Mozambique, *n* = 4; Mauritania, *n* = 2; Namibia, *n* = 76; Rwanda, *n* = 2; Senegal, *n* = 64; Somalia, *n* = 3; Tunisia, *n* = 3; Tanzania, *n* = 1; Uganda, *n* = 5; South Africa, *n* = 123; Zimbabwe, *n* = 241); Asia, 1,048 (Middle East and Central Asia, *n* = 291; Far East Asia, *n* = 757); Middle East and Central Asia (Comoro Islands, *n* = 1; India, *n* = 44; Iran, *n* = 108; Sri Lanka, *n* = 3; Mauritius, *n* = 2; Madagascar, *n* = 62; Pakistan, *n* = 53; Reunion Island, *n* = 12; Saudi Arabia, *n* = 6); Far East Asia (China, *n* = 50; Indonesia, *n* = 29; Japan, *n* = 6; South Korea, *n* = 1; Malaysia, *n* = 27; Mongolia, *n* = 18; Philippines, *n* = 45; Thailand, *n* = 11; Vietnam, *n* = 570); Europe, 3,927 (Austria, *n* = 455; Belgium, *n* = 71; Switzerland, *n* = 1; Czech Republic, *n* = 8; Germany, *n* = 51; Denmark, *n* = 214; Spain, *n* = 103; France, *n* = 727; United Kingdom, *n* = 828; Ireland, *n* = 96; Italy, *n* = 203; Netherlands, *n* = 929; Portugal, *n* = 2; Romania, *n* = 10; Russia, *n* = 160; Sweden, *n* = 69); Americas, 5,157 (North America, *n* = 3,860; Central America, *n* = 530; South America, *n* = 767); and Oceania, 273 (Australia, *n* = 194; New Zealand, *n* = 32; French Polynesia, *n* = 1; United States-Hawaii, *n* = 46). The patterns from the Americas could be further split: North America (United States, *n* = 3,850; Canada, *n* = 10); Central America (Barbados, *n* = 6; Cuba, *n* = 219; Guadeloupe, *n* = 171; Haiti, *n* = 86; Honduras, *n* = 1; Martinique, *n* = 44; Mexico, *n* = 3); and South America (Argentina, *n* = 192; Bolivia, *n* = 4; Brazil, *n* = 248; Curaçao, *n* = 1; Chile, *n* = 2; Ecuador, *n* = 2; French Guiana, *n* = 191; Peru, *n* = 3; Surinam, *n* = 6; Venezuela, *n* = 118). It should be emphasized that the exact country designation was not available for some isolates from East Africa and that Madagascar (MDG) was included in subcontinent 6 (Middle East and Central Asia) for historical and anthropological reasons. Finally, in SpolDB3.0, the *M. tuberculosis* type strain, H37Rv (ST 451); the vaccinal strain *M. bovis* BCG (ST 482); and the rare species *M. canettii* (ST 592), *M. microti* (ST 639, 640, 641, and 642), and *M. caprae* (ST 644, 645, 646, 647, and 648) have been mentioned under the column geographic specificity. *M. africanum* and *M. bovis* isolates have not been marked specifically but are easily recognizable on the basis of their specific spoligotype signatures, i.e., the absence of spacers sp8, -9, and -39 in *M. africanum* (30) and sp39 to -43 in *M. bovis* (13).

Description of SpolDB3.0 and definition of indices. In SpolDB3.0, the first column (type) attributes a number to each spoligotype in our database, the second column (full spoligotype description) shows the patterns obtained, the third column (octal nomenclature) shows the representation of the binary patterns according to the octal nomenclature described previously (4), the fourth column (geographic specificity) shows the source of the data (provider country) recorded as a three-digit ISO3166 code (available at http://www.din.de/gremien/nas/nabd/iso3166ma), the fifth column (total) shows the total number of isolates for each of the shared types described, the sixth column shows the percentage of a given spoligotype in the database, and the seventh column (area) shows the number of provider countries reporting the particular shared type. The eighth

Name	Abbreviation	Type of data	Rules for Definition of Qualifiers (C1 and C2)
Matching Code	MC	1-8 digits	If 1 digit, then C1 = Endemic If 2 digits, then C1 = Localized If ≥ 3 digits, go to area section below for further interpretation
Area	Ar	Number of countries	If MC= ≥ 3 digits and Areas ≤ 5 ; C1= Localized If MC= ≥ 3 digits and Areas ≥ 6 ; C1= Ubiquitous
Spreading Index	SI	Quantitative indicator (Minimum value: 1 Maximum value: 87)	If SI ≥ 30 ; C2 = Epidemic If SI between 10 and 29 ; C2 = Common If SI between 3 to 9 ; C2 = Recurrent If SI ≤ 2 ; C2 = Rare

Some examples of C1 and C2 calculation and interpretation:

Type	Full Spoligotype description	Total	Percentage	Area	SI	MC	C1	C2
1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	1282	10.95	29	44.21	12345678	Ubiquitous	Epidemic
6	□□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■	14	0.12	3	4.67	25	Localized	Recurrent
30	■■■■■■□■■■■■■■■■■□■■□□□□■■■■■■■■□□□□■■■■■■■	2	0.02	1	2.00	3	Endemic	Rare
17	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	92	0.79	14	6.57	12345	Ubiquitous	Recurrent
298	■□□■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■□■■□□□■■■■	17	0.15	1	17.00	5	Endemic	Common

FIG. 1. Definition of qualifiers C1 and C2 according to geographic and quantitative distributions of spoligotypes in SpolDB3.0 with examples.

column shows the matching code (MC), which summarizes the information recorded on the geographical specificity of a given shared type (1, Africa; 2, North America; 3, Central America; 4, South America; 5, Europe; 6, Middle East and Central Asia; 7, Far East Asia; and 8, Oceania). A one-digit number (value, 1 to 8) means that the shared type is observed in a single continent, whereas a two (or more)-digit number suggests an intercontinental match for a given shared type. The number of digits increases with the geographical spreading of a given shared type. Although the MC describes both old (in extinction) and new (epidemic) alleles at this stage, its significance will increase when the relative phylogenetic position of each spoligotype allele is known. Indeed, the exact dynamics behind the relative contribution of each of the given spoligotypes may not be easy to assess, e.g., rare and localized clones may be either undergoing extinction or emerging. The ninth column in SpolDB3.0 shows the spreading index (SI), which is obtained by dividing the total number of isolates for a given shared type by the number of areas where it has been observed. As opposed to the MC index, which gives an idea of the geographical specificity (the number of continents where similar shared types are found), the SI provides a quantitative indicator. Thus, for a given spoligotype, the correlation among the MC, SI, and areas (Ar) of distribution, according to ISO3166 codes, and the spoligotyping structure may help us to infer if a clone is undergoing extinction or emerging. The 10th and 11th columns, respectively, show the qualifiers C1 and C2 that tentatively define a shared type as endemic, localized, or ubiquitous (C1) and as rare, recurrent, common, or epidemic (C2). The qualifiers C1 and C2 and some typical examples extracted from SpolDB3.0 are described in the algorithm illustrated in Fig. 1. Indeed, the epidemic history of the disease in a given setting may provide important clues about the spreading of a given spoligotype, e.g., in low-prevalence countries, identical spoligotypes are more likely than in high-prevalence countries to represent past transmission events. It should be mentioned, however, that these qualifiers are not definitive, and they may be revised when the database grows further.

Another Excel spreadsheet (not shown in the link to SpolDB3.0 provided above) contains the precise source of all information processed, such as the countries and names of all investigators and key identification numbers for strain identification. In the following analysis, matching of shared types was done between geographic areas of isolation and not between nationalities.

Visual rules for defining major spoligotyping families. Simultaneous analysis of a visual pattern made up of a binary succession of information of 43 bits is not an easy task for the human brain. Octal numbering (4) is an improvement for database storage but has not yet proven useful for taxonomic and phylogenetic analysis. Moreover, previous work using mathematical modeling has shown that not all spacer positions in a spoligotype carry the same amount of information (19). Consequently, in order to better recognize patterns visually, we found it convenient to define six major visual rules as follows: rule A, absence of sp29 to -32, presence of sp33, and absence of sp34; rule B, absence of sp21 to -24 and sp33 to -36; rule C, absence of sp18 and sp33 to -36; rule D, absence of sp39 to -43; rule E, absence of sp31 and sp33 to -36; rule F, absence of sp33 to -36. These six major rules were used together with the precise description of the extra missing spacers (minor rules) to define a total of 36 major clades of circulating *M. tuberculosis* isolates (5; http://www.cdc.gov/ncidod/EID/vol8no11/02-0125-Table.htm). Some major clades identified are the Beijing clade; the East African-Indian (EAI) clade; the Haarlem clade; the Latin American and Mediterranean (LAM) clade; the Central Asian (CAS) clade; a European clade of IS*6110* low banders, i.e., isolates containing ≤4 copies of the IS*6110* element (X; highly prevalent in the United States and United Kingdom); and a widespread yet poorly defined clade (T) characterized by the absence of sp33 to -36.

RESULTS

Global distribution of the shared types. The 13,008 spoligotype patterns were grouped into 813 shared types containing 11,708 (90%) of the isolates and 1,300 (10%) orphan patterns (clinical isolates showing unique spoligotypes). Since the publication of the previous database (22), the number of clustered isolates (shared types) has increased from 84 (2,779 of 3,319) to 90% (11,708 of 13,008). An identical clustering rate was found in the largest spoligotyping study published so far, which was performed in Texas and totaled 1,283 patients (20).

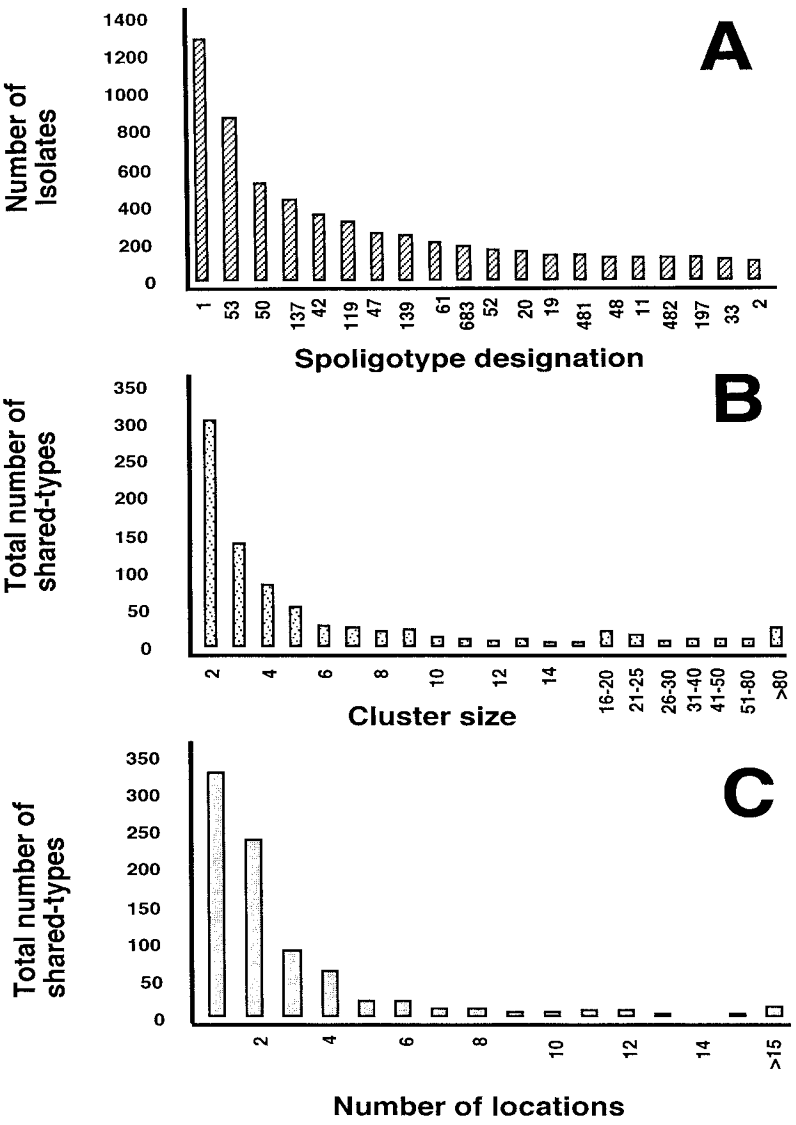

FIG. 2. Histograms derived from the database summarizing the distribution of shared types (A), their sizes (B), and their relative distributions in different locations (C).

The distribution of the shared types, their respective sizes, and their relative distributions in different locations are summarized in Fig. 2. The 20 most frequent types among the 813 shared types totaled 5,865 clinical isolates, i.e., 50% of all the clustered isolates. Three of these profiles correspond to *M. bovis*: types 683 and 481 for *M. bovis* and type 482 for *M. bovis* BCG. The addition of the next 30 most frequent spoligotypes slightly increased the total number of shared types assessed (65% instead of the initial 50%). SpolDB3.0 better describes the diversity of TB genotypes, since the 24 most prevalent alleles now represent only 53% of the total instead of 65% as in the previous study (22). The computation of the genetic diversity index (24) is done using the formula $H = 2n(1 - \Sigma x_i^2)/2n - 1$, where H is the genetic diversity index, x is the frequency of the allele i, and n is the population size, giving an H value of 97.4% with the current version of the database. An identical calculation with the previous version gave an H value of 93% (22).

In Fig. 2A, which depicts the 20 most frequent spoligotypes, the Beijing type (ST 1) is the most frequent (1,282 isolates, or 11% of all clustered isolates), followed by the Haarlem type (ST 47 and ST 50, representing ~6% of all clustered isolates). The newly designated X1 and X2 spoligotypes (ST 119 and ST 137), which tend to be highly prevalent in the United Kingdom and the United States, represent 6.4% of the clustered isolates. Figure 2B shows that one-third of all the shared types consist of two isolates only. This result suggests an important local diversity of spoligotyping. Nonetheless, a match of two identical but rare spoligotypes found in a single setting is different from a match found by database comparison of two widely separated isolates. The first case may be an early indicator of clonal expansion and ongoing transmission, whereas the second case, depending on the spoligotype, is likely to be due to homoplasy (independent acquisition of two similar structures without common ancestors). Alternatively, such matches, whether near extinction or not, may also reflect past epidemi-

TABLE 1. Total numbers of intercontinental matches of shared types among eight continents and subcontinents[a]

Locations and MC[b]	No. of matches							
	Africa	Americas			Europe	Asia		Oceania
		North America	Central America/ Caribbean	South America		Middle East and Central Asia	Far East	
Africa (1)	51	11	1	5	22	1	0	0
North America (2)		119	7	9	88	3	12	6
Central America (3)			8	4	13	2	2	2
South America (4)				27	17	0	0	2
Europe (5)					163	16	3	10
Middle East (6)						6	0	1
Far East (7)							9	3
Oceania (8)								2

[a] Africa, North America, Central America, South America, Europe, Middle East, Central Asia, and Far East. The matches found at two geographic locations were defined by MC analysis among the eight locations ($n = 625$).

[b] MCs summarize the geographical specificity of a given shared type in the database. A single-digit number (value, 1 to 8) means that the shared type is observed in a single continent, whereas a two (or more)-digit number suggests an intercontinental match for a given shared type. Only matches between two locations are shown.

ological events. Figure 2C shows that one-third of the shared types are repeatedly found within a single geographic area. This result corroborates the observations from Fig. 2B and suggests that spoligotyping performed as a single genotyping method in a new setting may be a good indicator of strain identity and helps to produce a precise picture of epidemiologically important clones.

Matching analysis of shared types found in one or two geographic locations. Table 1 shows the results of matching analysis of shared types which have been reported in one or two continental regions as defined in Materials and Methods ($n = 625$). This analysis demonstrates that the diversity of clustered spoligotypes is high within Europe ($n = 163$), the United States ($n = 119$), and Africa ($n = 53$). When the sample size is normalized, the diversity, limited to a specific continent (the number of shared types limited to a given continent divided by the total number of isolates within the continent), appears to be highest within Europe ($n = 163$ of 3,927, or 0.041), followed by Africa ($n = 53$ of 1,303, or 0.039) and the United States ($n = 119$ of 5,157, or 0.023). On the other hand, the lowest diversity is found in the Far East ($n = 9$ of 757, or 0.011). The greatest number of intercontinental matches is found between North America and Europe ($n = 88$). This class is made up of clones that are likely to represent, at least partly, historical TB transmission events between Europe and the United States, a phenomenon that may be explained either by the relatively old demographic links between these two continents or by recent transmission from identical high-prevalence countries. A significant number of these intercontinental matches are also found between Africa and Europe ($n = 22$), South America and Europe ($n = 17$), Central America and Europe ($n = 13$), the Middle East and Europe ($n = 13$), and, to a lesser extent, between Far East Asia and North America ($n = 12$). These data should be interpreted in the light of old, as well as recent, migratory flux and deserve further study. Among the matches between Europe and the Middle East and Central Asia, a majority concerned IS*6110* low banders from the United Kingdom that are known to be linked to the EAI clade of *M. tuberculosis* (14, 23). However, the recent finding of a spoligotype harboring a typical EAI signature (sp29 to -32 with sp34 missing) in a sample from 15th-century *M. tuberculosis* DNA found in the Wharram Percy medieval village in the United Kingdom is controversial as far as the precise origin of the EAI clade (16).

Snapshot of moving and expanding clones of *M. tuberculosis* and their global distribution. In a previous report on 259 shared types observed for 3,319 isolates from 47 countries, at least six major clades of tubercle bacilli were described (22). The present study permitted us to define a total of 36 potential superfamilies of spoligotypes using visual major rules A to F, defined in Materials and Methods, and a number of minor rules available online at http://www.cdc.gov/ncidod/EID/vol8no11/02-0125-Table.htm. Automated Excel labeling of the whole database ($n = 813$ shared types) following the visual rules defined above, and on a total of nine superfamilies of strains (*M. africanum*, Beijing, *M. bovis*, EAI, CAS, T group of families, Haarlem, X family, and LAM family), resulted in the labeling of 788 of 813 (96.9%) shared types. These results should be further assessed and generalized using data-mining methods (19).

The distribution of the most frequently observed spoligotypes, schematized in Fig. 3, underlined some major differences among the continental regions studied, e.g., the number of orphan types (or singletons) ranged from a low of 8% (North America) to a high of 21% (Middle East and Central Asia). Similarly, minor shared types ranged from 12% in the Far East to >50% in Europe and the Middle East and Central Asia. Among major clades, the heterogeneity of the distribution of the Beijing type (type 1 in the database) was noteworthy: it ranged from <2% in South America to 3 to 5% in Central America, Europe, Africa, and the Middle East and Central Asia, 13% in Oceania, 16% in North America, and as high as 45% in the Far East. Considering the multidrug resistance of the Beijing strains (2, 3, 29), the high prevalence of this clade in certain regions of the world is an important issue for effective TB control. Another interesting feature from Africa is the significantly high proportion of *M. africanum* strains (type 181), which represent 6% of all spoligotypes.

DISCUSSION

Some recent papers have dealt with the construction of spoligotyping databases (20, 21, 22). Soini et al. (20) described a study of 1,429 *M. tuberculosis* isolates from 1,283 patients as

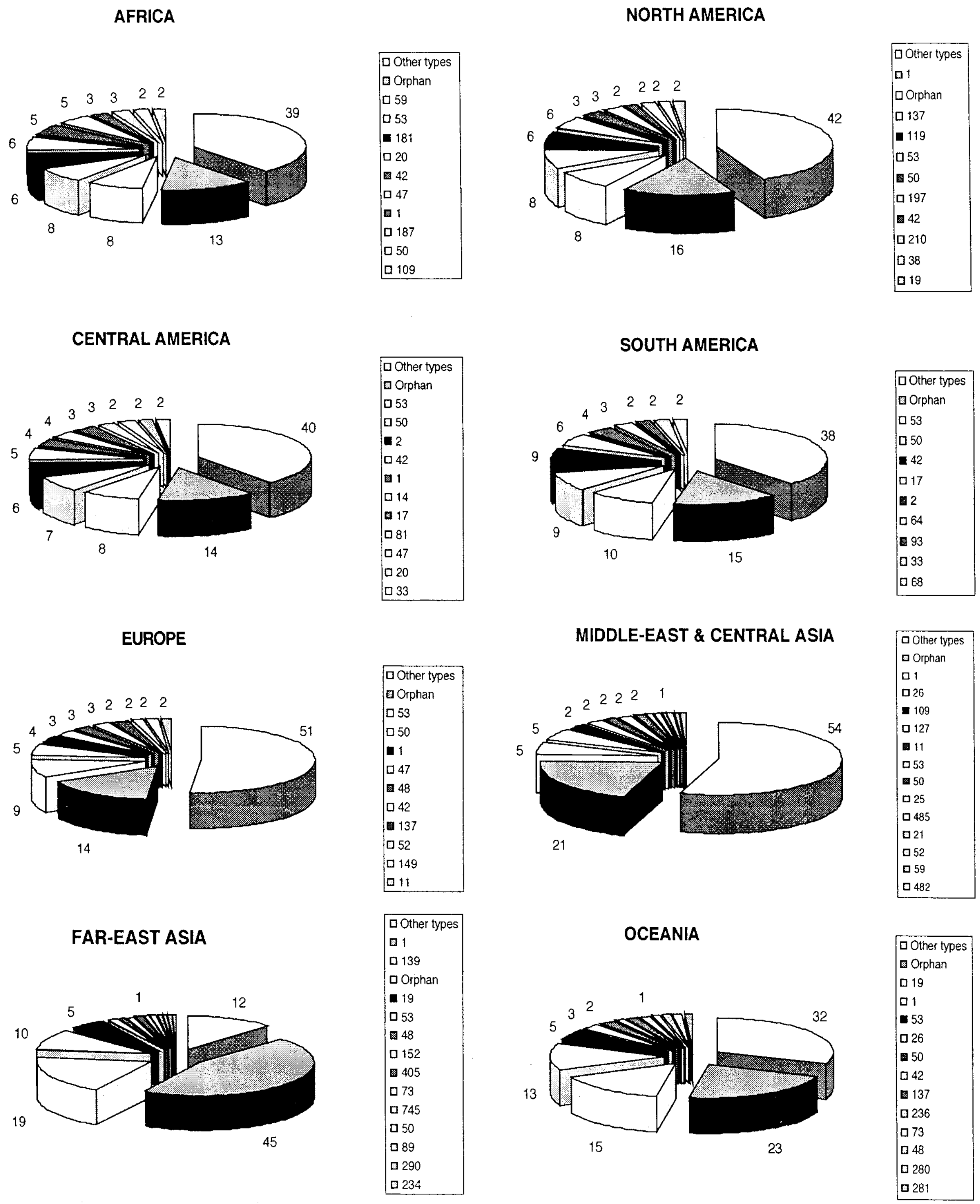

FIG. 3. Global distribution of *M. tuberculosis* complex spoligotypes assessed on 11,708 clustered isolates split into 813 shared types by continental regions, as defined in SpolDB3.0 (http://www.pasteur-guadeloupe.fr/tb/spoldb3).

part of an ongoing population-based TB epidemiology study in Houston, Tex. This paper was soon followed by a report of the biogeographical distribution of 3,319 spoligotype patterns and 259 shared types from 47 countries worldwide (22). The first study essentially focused on isolates from patients residing in a single state in the United States, whereas in the second study, >73% of the isolates described were from Europe and the United States. Despite these limitations, the studies underlined the fact that a significant number of *M. tuberculosis* isolates in circulation were essentially confined to specific geographic locations (20, 22). By including new spoligotyping data from all over the world, SpolDB3.0 has increased the overall representation; nonetheless, a more representative description of the worldwide diversity of tubercle bacilli should be possible through the acquisition of information from Asian and African countries.

The construction of global polymorphism databases constitutes a powerful tool, as it permits a quantitative estimation of the measure of DNA variations at the chromosomal level by the number of genetic structures observed so far. Similarly to what is done in *Drosophila melanogaster* population genetics, where inversions have been classified as "common ubiquitous, rare endemic, recurrent endemic and unique endemic" (24), we attempted to categorize most of the geographic variations in the DR loci observed so far by spoligotyping, so as to have a better knowledge of moving and expanding clones of *M. tuberculosis*. For this purpose, we introduced new indices (MC and SI) and qualifiers (C1 and C2) in order to better describe the spatiotemporal status of natural populations of the *M. tuberculosis* complex. For the spatial distribution, the populations studied were defined as endemic, localized, or ubiquitous. For the quantitative distribution, the populations were defined as epidemic, common, recurrent, or rare. These definitions synthetically define a spatiotemporal status for each shared type and, together with its genetic structure, may provide a global idea of its evolutionary history.

The results obtained also underline the well-known fact that casual contacts and sporadic cases, although difficult to detect, are responsible for most of the microepidemics and constitute an important means of TB transmission (6). Our next objective is to better describe the genetic diversity of the *M. tuberculosis* complex worldwide, which may be achieved by recruitment of adequate clinical isolates or DNA samples or inclusion of representative spoligotyping data in the database. Construction of new mathematical models that permit an interpretation based on the combination of DNA fingerprinting, epidemiological, and demographical data should further improve our knowledge of evolutionary processes that intervene in the development and spread of infectious diseases.

Regarding the genetic variability of the DR locus, it was recently shown to be a part of a larger family of sequence repeats among prokaryotes (11). Much remains to be done to precisely define the potential phylogenetic links within various alleles of this locus, as well as to investigate potential links that are found across individual studies targeting local epidemiological issues, particularly since TB does not respect man-made frontiers. Little is also known about the microevolutionary events associated with the DR locus and how they may influence the interpretation of both spoligotyping and IS*6110* RFLP data (31). Indeed, different isolates from the same strain family and isolates from different strain families may rarely converge to give the same spoligotype pattern (31). Though of limited importance, this bias may be investigated in detail in future by using second-generation spoligotyping based on a set of new spacer oligonucleotides (26) or by assessment of other genetic markers (18) in selected strains. The management of such projects will be facilitated by automation of data entry and data mining to further update SpolDB3.0 (1, 19). The data acquisition, similarity search, and matching process; labeling; and translation from binary to octal format and vice versa are already automated, and future data exchange and internet working of SpolDB3.0 with other databases (such as IS*6110* RFLP or mycobacterial interspersed repetitive units) should soon allow new queries to be screened against an updated version.

The facility by which detection of matches between potentially linked strains can be achieved may make SpolDB3.0 a new tool for international studies of TB transmission. Indeed, the detection of a match between two rare profiles in SpolDB3.0 may be a start to gathering complementary genotyping information, such as IS*6110* RFLP or polymorphic GC-rich-sequence RFLP in other international databases, to demonstrate clonality of the studied isolates (17) and to detect unsuspected epidemiological links. In conclusion, SpolDB3.0 constitutes a potential tool for global TB epidemiology and population genetics and *M. tuberculosis* complex taxonomy and phylogeny. It underlines major differences in the population structures of tubercle bacilli within the eight subcontinents studied, and by using new indices and qualifiers, it has led to better interpretation methods and the possibility of future comparison with other methods, such as mycobacterial interspersed repetitive units (18). Nevertheless, further work is still needed to get a more exhaustive global picture of worldwide tubercle bacillus genetic variability. Another major issue will be the ability to link this genetic diversity to virulence and/or fitness factors and ultimately to the genetic predisposition factors of the human or animal hosts.

ACKNOWLEDGMENTS

We are highly grateful to G. Källenius and T. Koivula (Sweden), P. Palittapongarnpim (Thailand), H. Kasai (Japan), G. Haase (Germany), and R. Frothingham (United States) for their collaboration. We also thank all investigators who transmitted their spoligotyping data to Institut Pasteur de Guadeloupe or the RIVM in Bilthoven, as well as other investigators who published studies with exhaustive descriptions of their spoligotyping patterns and the origins of the isolates, allowing such databases to be constructed.

This work was supported through grants by the Délégation Générale au Réseau International des Instituts Pasteur et Instituts Associés, Institut Pasteur, Paris, France, and the Fondation Française Raoul Follereau, Paris, France. It also benefited from the EU Project QLK2-CT-2000-630, entitled New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis.

REFERENCES

1. **Allen, J. F.** 2001. In silico veritas. Data-mining and automated discovery: the truth is in there. EMBO Rep. **2:**542–544.
2. **Caminero, J. A., M. J. Pena, M. I. Campos-Herrero, J. C. Rodriguez, I. Garcia, P. Cabrera, C. Lafoz, S. Samper, H. Takiff, O. Afonso, J. M. Pavon, M. J. Torres, D. van Soolingen, D. A. Enarson, and C. Martin.** 2001. Epidemiological evidence of the spread of a *Mycobacterium tuberculosis* strain of the Beijing genotype on Gran Canaria island. Am. J. Respir. Crit. Care. Med. **164:**1165–1170.
3. **Chan, M. Y., M. Borgdorff, C. W. Yip, P. E. de Haas, W. S. Wong, K. M. Kam, and D. van Soolingen.** 2001. Seventy percent of the *Mycobacterium tuberculosis* isolates in Hong Kong represent the Beijing genotype. Epidemiol. Infect. **127:**169–171.
4. **Dale, J. W., D. Brittain, A. A. Cataldi, D. Cousins, J. T. Crawford, J. Driscoll, H. Heersma, T. Lillebaek, T. Quitugua, N. Rastogi, R. Skuce, C. Sola, D. van Soolingen, and V. Vincent.** 2001. Spacer oligonucleotide typing of *Mycobacterium tuberculosis*: recommendations for standardized nomenclature. Int. J. Tuberc. Lung. Dis. **5:**216–219.
5. **Filliol, I., J. R. Driscoll, D. van Soolingen, B. N. Kreiswirth, K. Kremer, G. Valétudie, D. D. Anh, R. Barlow, D. Banerjee, P. J. Bifani, K. Brudey, A. Cataldi, R. C. Cooksey, D. V. Cousins, J. W. Dale, O. A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, M. Gordon,

M. C. Gutierrez, W. H. Haas, H. Heersma, G. Källenius, E. Kassa-Kelembho, T. Koivula, H. M. Ly, A. Makristathis, C. Mammina, G. Martin, P. Moström, I. Mokrousov, V. Narbonne, O. Narvskaya, A. Nastasi, S. N. Niobe-Eyangoh, J. W. Pape, V. Rasolofo-Razanamparany, M. Ridell, M. L. Rossetti, F. Stauffer, P. N. Suffys, H. Takiff, J. Texier-Maugein, V. Vincent, J. H. de Waard, C. Sola, and N. Rastogi. 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg. Infect. Dis. **8:**1347–1349.

6. **Golub, J. E., W. A. Cronin, O. O. Obasanjo, W. Coggin, K. Moore, D. S. Pope, D. Thompson, T. R. Sterling, S. Harrington, W. R. Bishai, and R. E. Chaisson.** 2001. Transmission of *Mycobacterium tuberculosis* through casual contact with an infectious case. Arch. Intern. Med. **161:**2254–2258.
7. **Goyal, M., S. Lawn, B. Afful, J. W. Acheampong, G. Griffin, and R. Shaw.** 1999. Spoligotyping in molecular epidemiology of tuberculosis in Ghana. J. Infect. **38:**171–175.
8. **Grein, T. W., K. B. Kamara, G. Rodier, A. J. Plant, M. J. Ryan, T. Ohyama, and D. L. Heymann.** 2000. Rumors of disease in the global village: outbreak verification. Emerg. Infect. Dis. **6:**97–102.
9. **Groenen, P. M. A., A. E. Bunschoten, D. van Soolingen, and J. D. A. van Embden.** 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.
10. **Hermans, P. W. M., D. van Soolingen, E. M. Bik, P. E. W. de Haas, J. W. Dale, and J. D. A. van Embden.** 1991. Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect. Immun. **59:** 2695–2705.
11. **Jansen, R., J. D. van Embden, W. Gaastra, and L. M. Schouls.** 2002. Identification of a novel family of sequence repeats among prokaryotes. Genomics **6:**23–33.
12. **Källenius, G., T. Koivula, S. Ghebremichael, S. E. Hoffner, R. Norberg, E. Svensson, F. Dias, B. Marklund, and S. B. Svenson.** 1999. Evolution and clonal traits of *Mycobacterium tuberculosis* in Guinea-Bissau. J. Clin. Microbiol. **37:**3872–3878.
13. **Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. D. A. van Embden.** 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. **35:**907–914.
14. **Kremer, K., D. van Soolingen, R. Frothingham, W. H. Haas, P. W. M. Hermans, C. Martin, P. Palittapongarnpim, B. B. Plikaytis, L. W. Riley, M. A. Yakrus, J. M. Musser, and J. D. A. van Embden.** 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.
15. **Martinez, J. I., F. S. Chaves, A. A. Arce, M. S. Alonso, E. M. Palenque, F. H. Jaen, M. M. M. Teresa, M. G. Gor, and A. N. Rodriguez.** 2000. Recent transmission of tuberculosis in Madrid (Spain): usefulness of molecular techniques. Med. Clin. **115:**241–245.
16. **Mays, S., G. M. Taylor, A. J. Legge, D. B. Young, and G. Turner-Walker.** 2001. Paleopathological and biomolecular study of tuberculosis in a medieval skeletal collection from England. Am. J. Phys. Anthropol. **114:**298–311.
17. **Poulet, S., and S. T. Cole.** 1995. Characterization of the highly abundant polymorphic GC-rich repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. Arch. Microbiol. **163:**87–95.
18. **Savine, E., R. M. Warren, G. D. van der Spuy, N. Beyers, P. D. van Helden, C. Locht, and P. Supply.** 2002. Stability of variable-number tandem repeats of mycobacterial interspersed repetitive units from 12 loci in serial isolates of *Mycobacterium tuberculosis*. J. Clin. Microbiol. **40:**4561–4566.
19. **Sebban, M., I. Mokrousov, N. Rastogi, and C. Sola.** 2002. A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. Bioinformatics **18:**235–243.
20. **Soini, H., X. Pan, A. Amin, E. A. Graviss, A. Siddiqui, and J. M. Musser.** 2000. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J. Clin. Microbiol. **38:**669–676.
21. **Sola, C., A. Devallois, L. Horgen, J. Maïsetti, I. Filliol, E. Legrand, and N. Rastogi.** 1999. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. Emerg. Infect. Dis. **5:**404–414.
22. **Sola, C., I. Filliol, C. Guttierez, I. Mokrousov, V. Vincent, and N. Rastogi.** 2001. Spoligotype database of *Mycobacterium tuberculosis*: biogeographical distribution of shared types and epidemiological and phylogenetic perspectives. Emerg. Infect. Dis. **7:**390–396.
23. **Sola, C., I. Filliol, E. Legrand, I. Mokrousov, and N. Rastogi.** 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. J. Mol. Evol. **53:** 680–689.
24. **Solignac, M., G. Periquet, D. Anxolabéhère, and C. Petit.** 1995. Mesure de la variation, p. 81–101. *In* Génétique et evolution, vol. 1. Hermann, Paris, France.
25. **van Crevel, R., R. H. Nelwan, W. de Lenne, Y. Veeraragu, A. G. van der Zanden, Z. Amin, J. W. van der Meer, and D. van Soolingen.** 2001. *Mycobacterium tuberculosis* Beijing genotype strains associated with febrile response to treatment. Emerg. Infect. Dis. **7:**880–883.
26. **Van der Zanden, A. G. M., K. Kremer, L. M. Schouls, K. Caimi, A. Cataldi, A. Hulleman, N. J. D. Nagelkerke, and D. van Soolingen.** 2002. Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. J. Clin. Microbiol. **40:**4628–4639.
27. **van Embden, J. D. A., T. van Gorkom, K. Kremer, R. Jansen, B. A. M. van der Zeijst, and L. M. Schouls.** 2000. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. **182:**2393–2401.
28. **van Soolingen, D.** 2001. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. J. Intern. Med. **249:**1–26.
29. **van Soolingen, D., L. Qian, P. E. W. de Haas, J. T. Douglas, H. Traore, F. Portaels, H. Z. Qing, D. Enkhsaikan, P. Nymadawa, and J. D. A. van Embden.** 1995. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J. Clin. Microbiol. **33:**3234–3238.
30. **Viana-Niero, C., C. Gutierrez, C. Sola, I. Filliol, F. Boulahbal, V. Vincent, and N. Rastogi.** 2001. Genetic diversity of *Mycobacterium africanum* clinical isolates based on IS*6110* restriction fragment length polymorphism analysis, spoligotyping, and variable number of tandem DNA repeats. J. Clin. Microbiol. **39:**57–65.
31. **Warren, R. M., E. M. Streicher, S. L. Sampson, G. D. van der Spuy, M. Richardson, D. Nguyen, M. A. Behr, T. C. Victor, and P. D. van Helden.** 2002. Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J. Clin. Microbiol. **40:**4457–4465.
32. **Wilson, S. M., S. Gross, and F. Drobniewski.** 1998. Evaluation of strategies for molecular fingerprinting for use in the routine work of a *Mycobacterium* reference unit. J. Clin. Microbiol. **36:**3385–3388.