Infection, Genetics and Evolution 12 (2012) 798–806

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

# *Mycobacterium tuberculosis* spoligotypes that may derive from mixed strain infections are revealed by a novel computational approach

Luiz Claudio O. Lazzarini [a,b], Jeffrey Rosenfeld [c], Richard C. Huard [d], Véronique Hill [e], José Roberto Lapa e Silva [a,b], Rob DeSalle [c], Nalin Rastogi [e,*], John L. Ho [b,g,*]

[a] Institute of Thoracic Disease, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[b] Division of International Medicine and Infectious Diseases, Department of Medicine, Weill Medical College of Cornell University, New York, NY, USA
[c] The Sackler Institute of Comparative Genomics, American Museum of Natural History, New York, NY, USA
[d] Clinical Microbiology Service, New York – Presbyterian Hospital, Columbia University Medical Center, New York, NY, USA
[e] WHO Supranational TB Reference Laboratory, Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Abymes Cedex, Guadeloupe, France
[g] Hosptialists Service, Southern Maine Medical Center, Biddeford, ME, USA

ARTICLE INFO

*Article history:*
Available online 5 September 2011

*Keywords:*
Spoligotyping
Tuberculosis
Mycobacterium
Genotyping, Mixed-infection
SpolDB4

ABSTRACT

Global control of tuberculosis is increasingly dependent on rapid and accurate genetic typing of *Mycobacterium tuberculosis*. Spoligotyping is a first-line genotypic fingerprinting method for *M. tuberculosis* isolates. An international online database (SpolDB4) of spoligotype patterns has been established wherein a clustered pattern (shared by ≥2 isolates) is designated a shared international type (SIT). Dual infections of single patients by distinct strains of *M. tuberculosis* is increasingly reported in high tuberculosis incidence areas, raising the possibility of false composite spoligotype patterns if performed upon mixed strain samples. A computational approach was applied to SpolDB4 and found that of the reported 1939 SITs, 54% could be a composite of two other SITs. Although many of the spoligotypes listed in SpolDB4 may be the product of admixing, the majority of patterns were reported with a corresponding low case frequency and so the effect of misclassification upon database integrity with these is likely minimal. Phylogenetic analysis of the five SITs most prone to be a composite demonstrated that these patterns designate nodes from which the ramifications of large families T, MANU, LAM, and EAI emerged. We illustrate how geographic context may indicate when an observed pattern could be the product of mixed infection. Importantly, when one of the most composite-prone SITs is obtained, further genetic testing by alternate methods is prudent to rule-out mixed infection, especially in high tuberculosis prevalence areas. These findings have broad practical implications for tuberculosis control and surveillance, as well as highlight the utility of a computational approach in providing solutions to biological questions in which the information can be digitalized.



## 1. Introduction

*Mycobacterium tuberculosis* is the etiologic agent of tuberculosis (TB) and is estimated to have infected one-third of the world's population, annually causing ~9 million new TB cases and ~1.3 million deaths worldwide (WHO, 2010). TB imposes a significant health burden, especially upon developing regions where resources are scarce and HIV coinfection is prevalent. An important component of TB control is rapid case detection and contact tracing, hereafter contacts may be offered preventative therapy. Molecular genotyping of *M. tuberculosis* strains can help establish transmission links between patients over time and help target public health interventions. Molecular epidemiologic investigations are also important to gauge the effectiveness of TB control efforts, to predict future epidemiological trends, and to advance our knowledge of *M. tuberculosis* strain-specific pathobiological characteristics, evolution, phylogeny, and global distribution.

Several molecular genotyping methods have been used to classify *M. tuberculosis* strain populations into families or clades based upon shared genetic markers. Such distinctions have provided support for lineage-specific differences in particular clinical manifestations of TB and have been used to establish the global phylogeography of the major *M. tuberculosis* lineages (Filliol et al., 2006; Gagneux et al., 2006). The most widely used fingerprinting techniques include spoligotyping, IS*6110*- restriction

* Corresponding authors. Present address: Centers for Disease Control in Haiti, American Embassy, 3400 Port-au-Prince Place, Dulles, VA, USA. Tel.: +1 404 553 8525 (J. L. Ho). Address: Institut Pasteur de Guadeloupe, Morne Jolivière, BP 484, 97183 Abymes Cedex, Guadeloupe, France. Tel.: +590 590 897661; fax: +590 590 893880 (N. Rastogi).

*E-mail addresses:* nrastogi@pasteur-guadeloupe.fr (N. Rastogi), millennium.john@gmail.com, HoJ@HT.CDC.GOV (J.L. Ho).

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.08.028

fragment length polymorphism (RFLP) analysis, and mycobacterial interspersed repetitive unit-variable numbers of tandem repeat (MIRU-VNTR) typing.

Spoligotyping is a standardized amplification method targeting the polymorphic *M. tuberculosis* complex-specific direct repeat (DR) locus and is visualized as a binary 43 digit pattern or reduced to an octal form of 15 digits (Kamerbeek et al., 1997). As compared to IS*6110*-RFLP and MIRU-VNTR typing, spoligotyping is usually unable to achieve the same level of discrimination of *M. tuberculosis* transmission clusters. However, the assay is rapid, inexpensive, robust, and the data can be easily exchanged between laboratories. Spoligotyping is therefore often used as a first-line genotyping method. It is according to shared patterns observed with spoligotyping that most of the major *M. tuberculosis* lineages have been named. These include Beijing (alternatively termed as East-Asian based on long sequence polymorphism [LSP] classification Gagneux et al., 2006), Haarlem, T, East African-Indian (EAI, also known as Indo-Oceanic Gagneux et al., 2006), and Latin American-Mediterranean (LAM), among several others. Each of these patterns is included in the SpolDB4, an international database developed and maintained by the Institute Pasteur de la Guadeloupe that includes spoligotypes submitted by investigators from all over the world (Brudey et al., 2006; available online: http://www.pasteur-guadeloupe.fr:8081/SITVITDemo).

Mixed infection is characterized by the simultaneous presence of more than one distinct *M. tuberculosis* strain in a given TB episode. Although previously considered a rare event, its occurrence is now shown with increasing frequency and it may be particularly common in areas of high TB prevalence (Baldeviano-Vidalón et al., 2005; Das et al., 2004; Warren et al., 2004; Yeh et al., 1999). The majority of papers that note dual infection used IS*6110*-RFLP or MIRU-VNTR typing as the principal methods to identify cases and to distinguish the individual strains causing the infection. Spoligotyping, in contrast, may not easily distinguish specimens harboring multiple strains. Indeed, in a previous study, we investigated the distribution of *M. tuberculosis* LSP designated as $RD^{Rio}$ by PCR-based methods (Lazzarini et al., 2007). Through the course of these studies, we identified a subset of isolates that were both positive and negative for the $RD^{Rio}$ LSP. We showed that these data were the result of a mixed infection by isolating two unique individual *M. tuberculosis* clones from each of two separate specimens derived from a single patient and repeating the molecular analyses. Of note, the spoligotype of the mixed clinical isolate was classified as shared international type (SIT)-291, a member of the T family and distributed to all continents (Brudey et al., 2006). However, SIT291 is a perfect composite of the individual spoligotypes from the two isolated clones (Fig. 1). Therefore if the primary isolate was not identified to be an admixture, the initial spoligotyping result would have been erroneously assigned. It is currently unknown how often such misclassifications could occur or which spoligotypes are more or least likely to be derived from a combination of others.

To begin to address these issues, we developed a computer program to evaluate the potential that each spoligotype present in SpolDB4 database could be generated from a composite of two other spoligotypes. As an example of the methodology, we additionally scrutinized the geographical localization of SITs that could potentially combine to produce a specific admixture pattern that had all the spoligotype spacers except 1 present (SIT100 which is characterized by the absence of spacer 34). This SIT was chosen based on the generally accepted principle that spoligotype patterns evolve by loss of spacers (Flores et al., 2007). The purpose behind this experimental approach was to come out with a strategy to classify which spoligotype patterns have a high probability to exist *per se*, as opposed to those that have a significant probability of existing due to admixture of two individual profiles.

## 2. Materials and methods

### 2.1. Computer program

A computer program was designed at the Sackler Institute of Comparative Genomics, American Museum of Natural History, New York, to generate relative occurrences of one spoligotype being the combination of two known spoligotypes. Briefly, the spoligotypes were converted into 0's (absence) and 1's (presence) (Supplemental Table 1). The program sorted through the list of 1939 different spoligotypes and systematically paired all combinations of two spoligotypes together and determined the pattern of 0's and 1' from the combined profile. The resulting combined profile was then compared to the list of spoligotypes to determine whether it resembled any individual spoligotypes. Such a match of a combined spoligotype to an individual spoligotype would indicate that the individual spoligotype might be a result of dual infection. The number of matches between two spoligotypes that exactly resemble a third spoligotype was recorded. The number of dual spoligotype profiles that match a particular known spoligotype (of the 1939 in the database) is provided in Supplemental Table 2. In addition, a list of spoligotypes (1 through 1939) can be generated with this program with the possible dual combinations that exactly resemble the initial spoligotype.

### 2.2. SpolDB4 database

The SpolDB4 database created and maintained by the Pasteur Institute of Guadeloupe (Brudey et al., 2006) is available online (http://www.pasteur-guadeloupe.fr:8081/SITVITDemo); the publicly available version contains 1939 SITs corresponding to 35,925 strains, as well as 3370 orphan patterns, from 121 countries. The five most common *M. tuberculosis* spoligotypes in this database were SIT53 (T1), SIT1 (Beijing), SIT50 (Haarlem3), SIT42 (LAM9) and SIT47 (Haarlem1).

The international distribution of each SIT in the SpolDB4 database shows that some lineages have preferential phylogeographical distributions (Beijing and EAI in East Asia, LAM in Europe, Africa, and Central and South America) while others are distributed more evenly around the globe (T, Haarlem, X, and S) (Brudey et al.,

| Mtb isolate | Genotype | SIT[a] | Spoligotype |
|---|---|---|---|
| Original | Mix | 291 | ■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□□□□■■■■■■■ |
| Strain #1 | $RD^{Rio}$ | 42 | ■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■ |
| Strain #2 | Non-$RD^{Rio}$ | N/A[b] | ■■■■■■■■■■■■■□□□□□□□■■■■■■■■□□■□□□□■■■■■■■ |

[a] SIT: Shared International Type.
[b] Orphan pattern.

**Fig. 1.** Illustrated are the spoligotype patterns for two strains derived from single colonies by plating the initial mixed sample. Strain # 1, was genotyped as $RD^{Rio}$ (SIT-42) and strain # 2 as Non-$RD^{Rio}$ (not SITed). The spoligotype from the mixed sample was a combination of both spoligotypes and erroneous assigned as SIT-291.

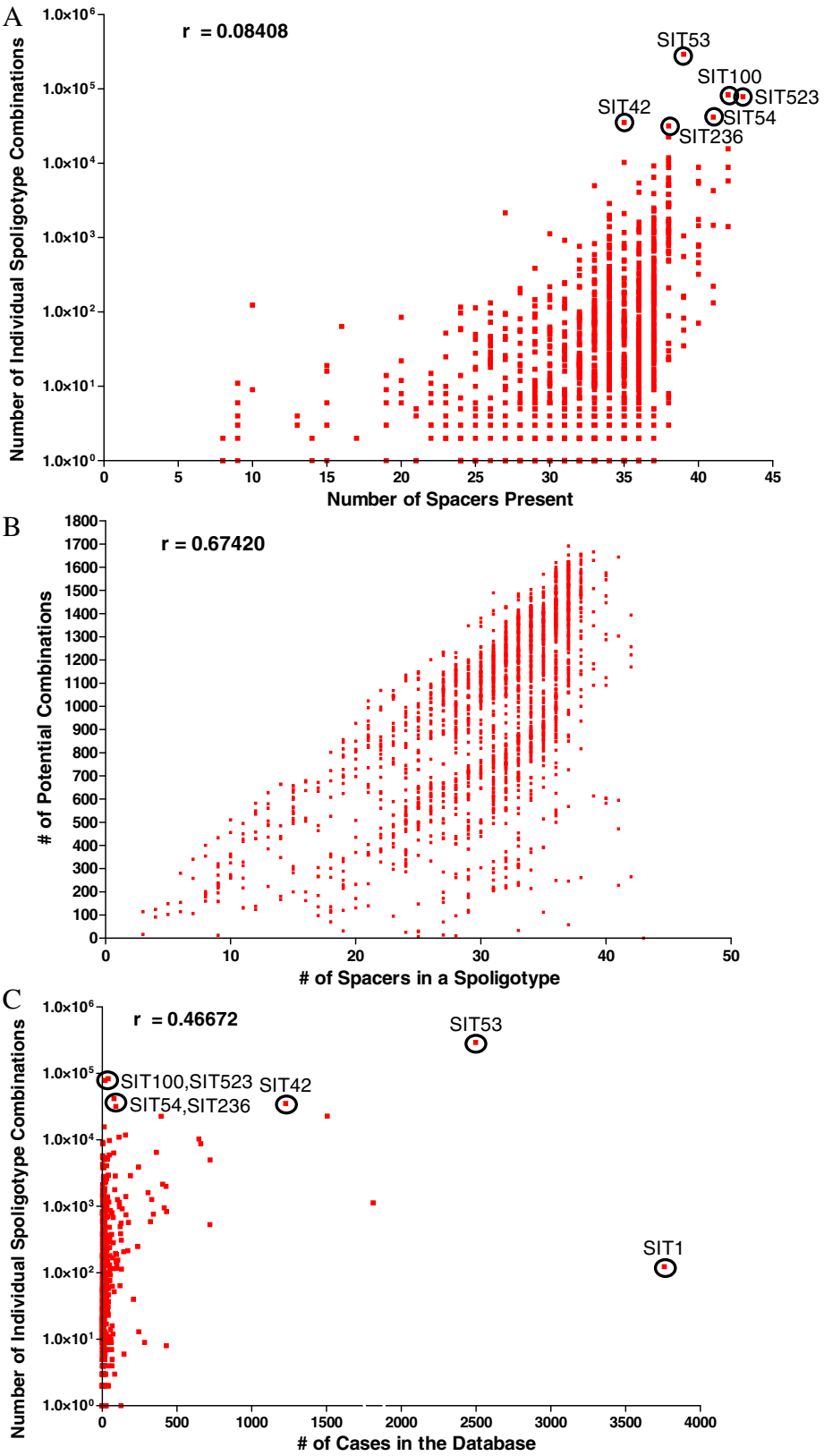


**Fig. 2.** Computational approach to analyze *M. tuberculosis* spoligotype clustered pattern (shared by $\geqslant 2$ isolates) designated a Spoligotype-International-Type (SIT). (A) Illustrated are the six shared-types with the highest chance of being a mixture of others types. The chance of combination increases in spoligotypes with 35 or more out of 43 spacers present. In this figure the $\log_{10}$ number of potential dual combinations that produce a known spoligotype is plotted versus the number of spacers in the known spoligotype. (B) Increasing the numbers of spaces present in a specific spoligotype is strongly correlated with the chance for combinations to form a composite pattern of two independent spoligotypes. (C) Comparison of the frequency of the spoligotypes in the database and the chance that these spoligotypes be a composite of other patterns. In this plot the $\log_{10}$ number of potential dual combinations for a given spoligotype in the database (of the 1939 known spoligotypes) is plotted versus the number of clinical cases of infection of the spoligotype. The small black circles indicate particularly important spoligotypes that are discussed in the text.

**Table 1**
International distribution of each SIT in SpolDB4.

| SIT | Lineage | % of SIT strains in SpolDB4 | Distribution in regions with ≥5% of a given SITs |
|---|---|---|---|
| 53 | T1 | 8.39 | Northern America 38.5, Western Europe 18.3, South America 8.0, Northern Europe 6.5, Southern Europe 6.3 |
| 42 | LAM9 | 3.42 | Northern America 23.2, South America 19.2, Southern Europe 12.4, Western Europe 10.8, Northern Asia 5.5, Northern Europe 5.4 |
| 236 | EAI5 | 0.25 | Northern America 35.2, Southeastern Asia 27.5, Northern Europe 11.0, Australasia 7.7, Southern Asia 7.7, Western Europe 5.5 |
| 54 | MANU2 | 0.22 | Southern Asia 24.4, Northern America 23.1, Northern Asia 12.8, Northern Europe 5.1, Western Europe 5.1 |
| 100 | MANU1 | 0.11 | Southern Asia 59.0, Northern Europe 15.4, Northern America 15.4, Southeastern Asia 7.7 |

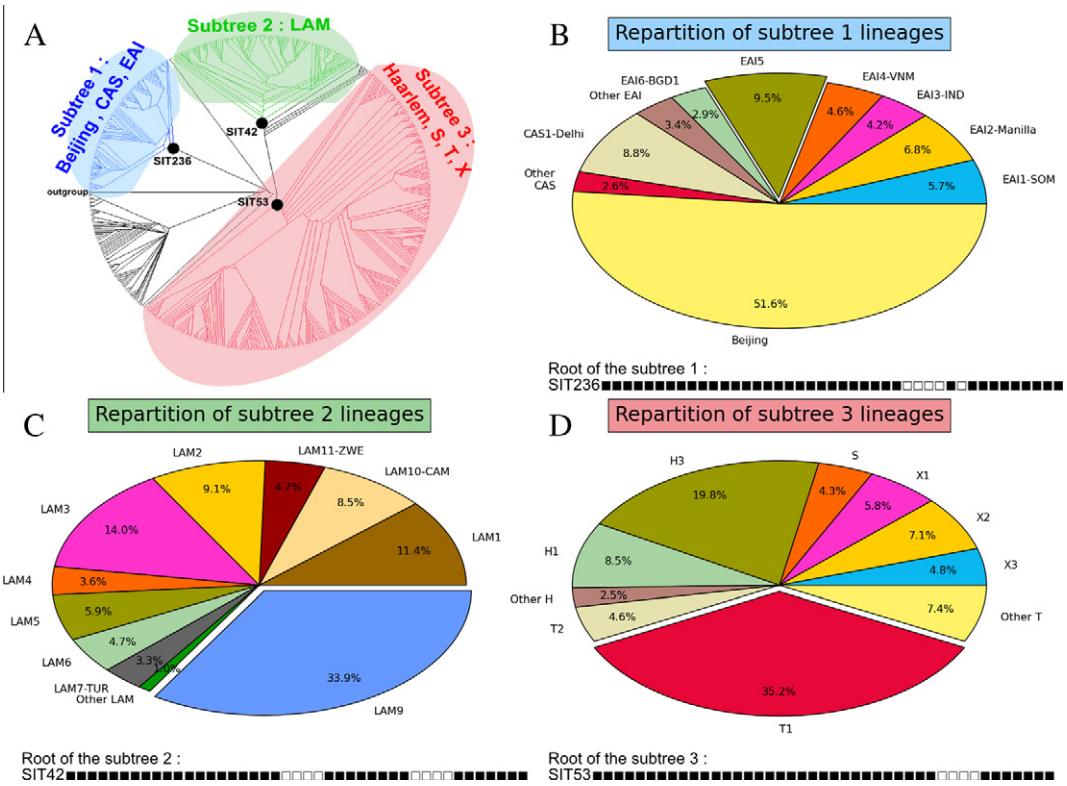


**Fig. 3.** Phylogenetic analysis and distribution of SITs defined by tree-nodes. (A) Phylogenetic analysis using the Camin–Sokal algorithm showing the ranking of 521 parsimonious phylogenetic patterns (those present ≥6 times in the database) classified assuming only one type of mutational event. The internal nodes corresponding to the appearance of patterns SIT236 (EAI5 prototype), SIT42 (LAM9 prototype), and SIT53 (T1 prototype) are highlighted by black dots, and the subtrees from these nodes are highlighted in different colors; SIT236/EAI5 for subtree 1 (blue), SIT42/LAM9 for subtree 2 (green), and SIT53/T1 for subtree 3 (red). (B) Distribution of lineages present in the subtree 1. (C) Distribution of lineages present in the subtree 2. (D) Distribution of lineages present in the subtree 3. The pie charts (for B, C and D) were constructed with a python 2D plotting library "matplotlib" available at http://matplotlib.sourceforge.net. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2006). For the spoligotypes most suspected of being involved in spoligotype pattern admixtures, their worldwide distribution was investigated for regions representing ≥5% of a given SIT as compared to their total number in the SpolDB4 database. The various macro-geographical regions and sub-regions were defined according to the specifications provided by the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm) as follows: Africa, Americas, Asia, Europe, and Oceania, subdivided in: Eastern, Middle, Central, Northern, Southern, South-Eastern, and Western. Note that in this classification the Caribbean belongs to Americas, while Oceania is subdivided in 4 sub-regions, Australasia, Melanesia, Micronesia, and Polynesia. Furthermore, Russia is attributed a new sub-region by itself (Northern Asia) instead of including it among the rest of Eastern Europe (which reflects both its geographical localization and the fact that it shares its predominant *M. tuberculosis* genotypes with those prevalent in Central, Eastern and South-Eastern Asia).

A phylogenetic tree was constructed with all the spoligotype patterns representing 6 or more clinical isolates in SpolDB4 database (for which a genotypic lineage was established, $n = 521$) using the Camin–Sokal algorithm available with PHYLIP software. Based on the principle that evolution is parsimonious, the Camin–Sokal algorithm retains trees requiring the smallest number of evolutionary changes (excluding reversions). This method is particularly suitable for phylogenetic studies based on a spoligotype43 marker, since it fits to the unidirectional evolution of the direct-repeat locus by loss of spacers. Furthermore, it also allows the generation of output files exportable to all other interactive phylogenetic software packages, which is highly useful for finding the optimal representation given the large sample size of this study.

## 3. Results

Of the reported 1939 spoligotypes, 1053 (54%) could possibly be the product or composite of two other spoligotypes. The potential that a spoligotype may be a composite of others increases with the number of spacers present in the spoligotype, as spoligotypes with more spacers are more prone to be a composite, but this chance is only relevant for spoligotypes with more than 35 spacers present (Fig. 2A). In total, the number of times that a given spoligotype present in the SpolDB4 database could be generated by the combination of two other spoligotypes in SpolDB4 varies from zero (i.e., no spoligotype pairs could falsely combine) to 292,266 possible combinations. The median of this occurrence, however, is only 2, indicating that the majority of spoligotypes in the database have a low likelihood of being a composite (data not shown).

The potential that a spoligotype be a mixture increases with the number of spacers present, but this is not an absolutely straightforward correlation (Fig. 2A). For example, SIT523 has all the 43 spacers present and could be generated by 78,292 possible combinations whereas SIT53, which has 39 out of the 43 spacers present, could be generated by the largest number of possible combinations (292,266). The reason for this is that spacers 33–36 were lost in conjunction with an evolutionary bottleneck that gave rise to the Euro-American *M. tuberculosis* superfamily, a clonally-derived radiation that now accounts for 58% of non-orphan individual spoligotypes listed in SpolDB4. The potential that a specific spoligotype could combine with another to form a composite is also related to the number of spacers present ($r = 0.674$) (Fig. 2B). For example, a spoligotype with 10 spacers present could combine up to 500 different ways whereas a spoligotype with 37 spacers present could combine up to 1692 times (Fig. 2B).

Of note, SIT1 (Beijing family), the most common spoligotype in the database with almost 3800 case-isolates submitted (herein, called cases), is unlikely to be an admixture of other spoligotypes, with only 124 possible combinations resulting in this particular pattern (Fig. 2C). In striking contrast, for SIT53 (T1 family), the second most described spoligotype in the SpolDB4 database, with almost 2500 cases, there are 292,266 possible spoligotype combinations that may result in this particular pattern (Fig. 2C). The potential of a SIT53 SpolDB4 strain submission to be a misclassification is therefore greater. There are 886 spoligotypes in the database that have no chance of being a composite and they are distributed among the different families, including the T family. Thus, our analysis indicates no specific family is spared, nor appears to have an outright predilection, for being misclassified. The five SITs that have the highest chance for having been observed due to dual infections are: SIT53 (T1) with 292,266 possible pair-wise combinations, SIT100 (MANU1) with 83,400 possible pair-wise combinations, SIT54 (MANU2) with 41,688 possible pair-wise combinations, SIT42 (LAM9) with 35,351 possible pair-wise combinations and SIT236 (EAI-5) with 31,863 possible pair-wise combinations. The total numbers of cases for each of these SITs (T1, MANU1, MANU2, LAM9, and EAI-5) in the database are: 2497, 39, 78, 1227, and 91, respectively, representing 10% of the total cases in the SpolDB4 (Supplemental Table 1). Although SIT523 (U) also has a high chance of being an admixture (78,657 possible combinations), with only 19 cases described in the database, misclassifications are less likely to impact database integrity and no further analyses were performed. SIT523 (U) was recently termed as "MANU-ancestor" (Helal et al., 2009).

The phylogenetic analysis of the five most prevalent SITs demonstrated that these patterns designate nodes from which the ramifications of large families T, MANU, LAM, and EAI emerged. The differential geographical distribution of these five SITs as provided in SpolDB4 is described in Table 1. The tree in Fig. 3A drawn using the Camin–Sokal algorithm shows a ranking of 521 parsimonious phylogenetic patterns classified assuming only one type of mutational event, the loss of spacers. In this figure, the internal nodes corresponding to the appearance of patterns SIT236 (EAI5 prototype), SIT42 (LAM9 prototype), and SIT53 (T1 prototype) are indicated by black dots, and the subtrees from these nodes are highlighted in different colors; a blue subtree-1 rooted to SIT236, a green subtree-2 rooted to SIT42, and a red subtree-3 rooted to SIT53. An algorithm was written to analyze a dataset of given patterns to find all resulting composite (admixture) patterns. The logic was based on the supposition that each single pattern could act as a probable parent; and make the pair with another pattern to give the admixture profile. This algorithm was applied to all patterns present in each subtree. The patterns ($n = 83$) of the subtree-1 generated 214 composite patterns (Supplemental Table 3); 31% of which were indeed traced back to the tree (100% are positioned in subtree-1 shown in Fig. 3A). The most predominant composite pattern was SIT236 (EAI5) which corresponded to 1302 associated pairs or 38.2% of all pairs, followed by SIT26 (CAS1-Delhi prototype) which represented only 4.4% of all pairs. The patterns ($n = 96$) of the subtree-2 generated 200 composite patterns (Supplemental Table 4); 45% of which were traced back to the tree (96% are positioned in subtree-2 shown in Fig. 3A). The most predominant composite pattern was SIT42 (LAM9) which corresponded to 2634 associated pairs or 57.8% of all pairs, followed by SIT93 (LAM5 prototype) representing 5.1% of all pairs. Lastly, the patterns ($n = 264$) of the subtree-3 generated 613 composite patterns (Supplemental Table 5); 35% of which were traced back to the tree (96% are positioned in subtree-3 shown in Fig. 3A). The most predominant composite pattern was SIT53 (T1 prototype) which corresponded to 22,638 associated pairs or 65.2% of all pairs, followed by SIT50 (H3 prototype) representing 3.7% of all pairs.

These results clearly illustrate that combinations of distal patterns in each of the subtrees often trace back to the central nodes made up of SIT236/EAI5 (subtree 1), SIT42/LAM9 (subtree 2), and SIT53/T1 (subtree 3). These 3 central nodes therefore represent patterns which have a higher probability to arise from a combination of other spoligotypes since the combination of two patterns in the same subtree may more often (but not always) results in a pattern shown at the root (central node). Data are not shown for MANU1/SIT100 and MANU2/SIT54 since these SITs were present in too small numbers in SpolDB4. The pie charts in Fig. 3B–D illustrate various lineages defined by these prevalent SITs (corresponding to 6 or more clinical isolates) linked to each of the node in subtrees 1–3. In the analysis shown, LAM9 (Fig. 3C) and T1 (Fig. 3D) are clearly predominant in their respective pie charts; however in the subtree 1 (Fig. 3B), the extremely high number of Beijing strains in SpolDB4 database surpasses all other lineages that could also simultaneously lead to the admixture pattern SIT236/EAI5. Nonetheless, EAI5 is the second most predominant lineage in this group.

Considering the scarcity of SIT100/MANU1 (0.11% of the SITed strains in SpolDB4) and its high number of possible pair-wise combinations, we decided to study it specifically by investigating "all pairs" in the SpolDB4 database that could generate this pattern in four subregions: Southern Asia, Northern Europe, Northern America and Southeastern Asia, areas where it is mainly found (Fig. 4A, Table 1). Histograms in Fig. 4B show the combined incidence of most relevant pairs (for incidence $\geq$1/10,000) capable of giving a SIT100 pattern. Note that the histogram colors refer to the colors of the subregions on the map, and the real incidence of SIT100 in each of the subregions studied is shown on the main world map figure (Fig. 4A); all SITs implicated in the probability of giving the SIT100/MANU1 pattern are shown in Table 2. The pairs illustrated in the histograms (Fig. 4B) show that a MANU1 pattern could arise due to an admixture of an ancestral strain

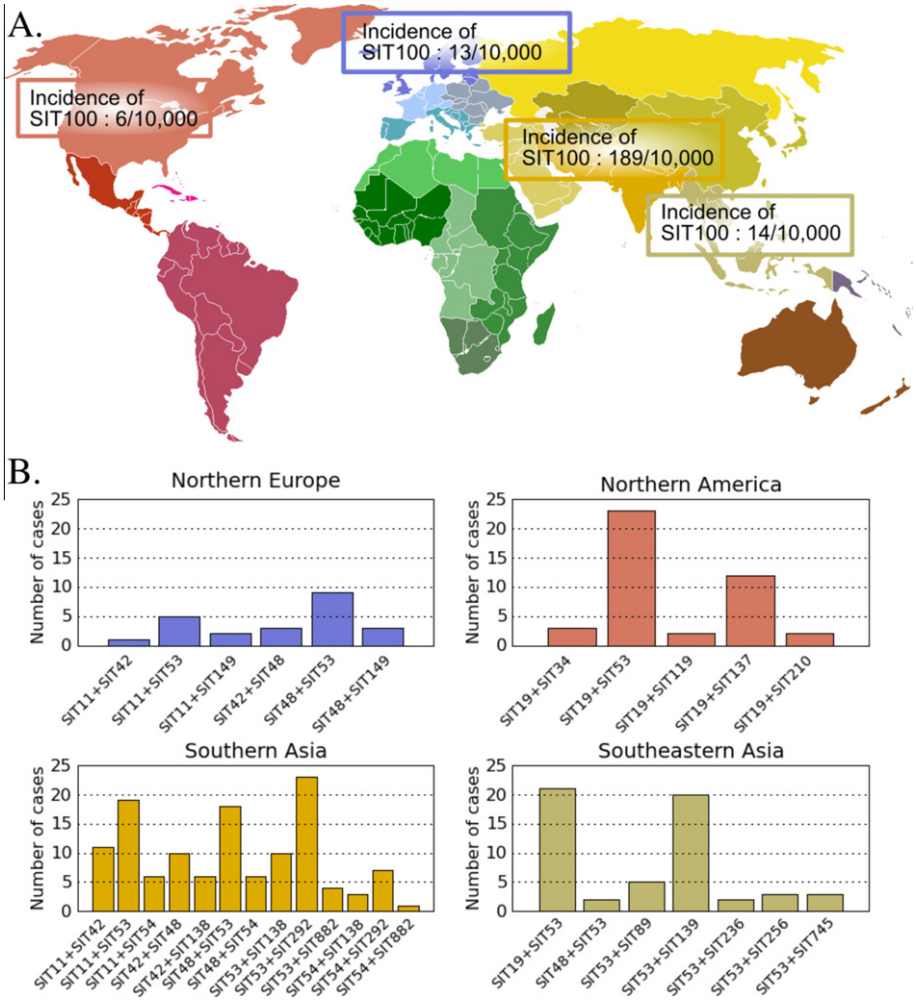


**Fig. 4.** (A) Incidence of SIT100 where MANU strains are observed. Description of "all pairs" in the SpolDB4 database that could generate the pattern of SIT100/MANU1 in four geographic subregions: Southern Asia, Northern Europe, Northern America and Southeastern Asia, where this pattern is mainly found. (B) Histograms show the incidence of most relevant pairs (incidence ≥1/10,000) capable of giving SIT100 pattern. Note that the histogram colors refer to the colors of the subregions on the map, and the real incidence of SIT100 in each of the subregions studied is shown on the main map figure. (For interpretation to colours in this figure, the reader is referred to the web version of this paper.)

**Table 2**
SITs in the SpolDB4 database potentially combining to give a SIT100/MANU1 pattern.

| SIT | Binary value of spoligotype 43 | Lineage |
|---|---|---|
| 11 | ■■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□□■■■■ | EAI3-IND |
| 19 | ■■□■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■■ | EAI2-Manilla |
| 34 | ■■■■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■ | S |
| 42 | ■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■■ | LAM9 |
| 48 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■ | EAI1-SOM |
| 53 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■ | T1 |
| 54 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■ | MANU2 |
| 89 | ■■□■■■■■□□□□□□□□□□□□□□□□□□■■■□□□□■□■■■■■■■■■ | EAI2-nonthaburi |
| 119 | ■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■■■■■■ | X1 |
| 137 | ■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■□□□□■ | X2 |
| 138 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□□□□ | EAI5 |
| 139 | ■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□□□■□■■■■■■■■■■ | EAI4-VNM |
| 149 | ■■■■■■■■■□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■■■■■■■■ | T3-ETH |
| 210 | ■■■■■■■■■□□□□□■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■ | T1 |
| 236 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■■ | EAI5 |
| 256 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■ | EAI5 |
| 292 | ■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■□□□□■□■■□■■■■■■ | EAI6-BGD1 |
| 745 | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□■□■■■ | EAI1-SOM |
| 882 | ■□■■■■■■■■■■■■■■■■■■■■■■■□■■■■■□□□□■□■■■■■■■■■ | EAI6-BGD1 |
| **100** | ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■ | **MANU1** |

SIT – shared international type.

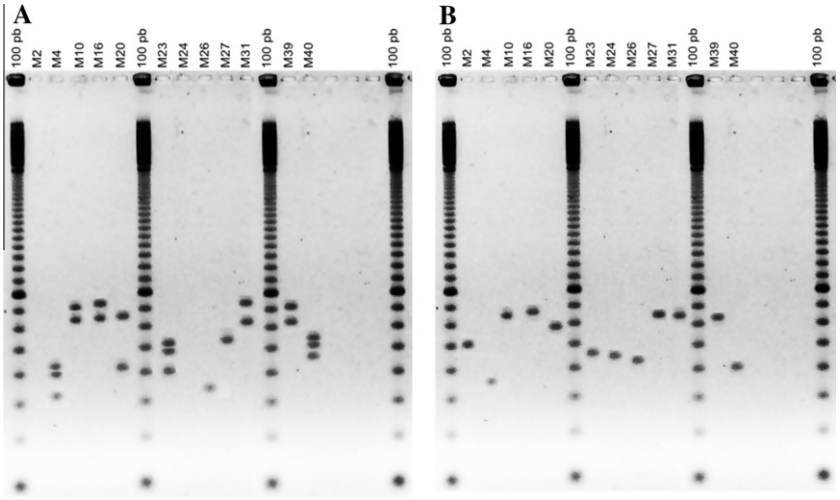


**Fig. 5.** Illustration of MIRU-VNTR gel electrophoresis for a suspected admixture pattern "MANU1 lineage" strain for each of the 12 MIRU-VNTR loci tested: (A) The MANU1 test strain and (B) *M. tuberculosis* H37Rv control. Symbols: M followed by a number shows the MIRU loci tested; 100 pb, 100 base-pair ladder every five lanes. The 12-loci MIRU VNTR-typing was performed at the Institut Pasteur de Guadeloupe as previously described (Millet et al., 2009). The presence of more than one band is indicative of admixture of two of more strains.

(e.g., EAI lineage characterized by loss of spacers 29–32 and 34) with an evolutionary-modern strain (characterized by loss of spacers 33–36). Knowing that modern strains are well represented throughout the world, the limiting factor for outbreaks of admixture patterns in this example is the occurrence of EAI strains in a given region. Most interestingly, if one looks at the proportion of EAI in the various subregions (Supplemental Fig. 1), the highest incidence of SIT100 strains is perfectly concordant with the highest percentages of EAI (>7%) in the same subregions: AMER-N, S-ASIA, ASIA-SE, EURO-N. Interestingly, the incidences for SIT100/MANU1 and the sum of paired patterns are more or less of the same order of magnitude in the subregions in question (Fig. 4A). These observations suggest that an investigator should take extra precaution to exclude the presence of two strains when a SIT100/MANU1 strain is observed in regions with high proportion of EAI lineage, a possibility that may be easily checked based on MIRU-VNTR typing of the isolates.

A laboratory example of such a dual infection resulting in a novel orphan "MANU1 lineage" spoligotype in which DNA was available is illustrated; Fig. 5A shows the PCR results for 12 classical MIRU-VNTR loci for the MANU1 (labeled as "Test admixture pattern" in Supplemental Table 6), as compared to the *M. tuberculosis* H37Rv control (Fig. 5B). Multiple bands for eight loci are observed in Fig. 5A for the suspected admixed pattern strain (M4, M10, M16, M20, M23, M31, M39, M40; no amplification was obtained for loci M2 and M24). A detailed list of all possible pairs that may lead to this proven false pattern is shown in Supplemental Table 6, and suggests the possible association of LAM10-CAM and EAI strains to give rise to the incriminated MANU1 lineage strain in question.

## 4. Discussion

Until recently it was believed that TB was always caused by a single infecting *M. tuberculosis* strain. This principle has been an underlying rationale in epidemiological studies to track outbreaks and to evaluate the susceptibility to anti-TB drugs. However, with the availability of genetic tools, it has become clear that concurrent infections with more than one distinct *M. tuberculosis* strain may be much more common than previously thought. It is epidemiologically and clinically relevant to identify infection with multiple strains for several reasons: (i) true outbreaks may be missed, (ii) epidemiological and molecular analyses may be distorted and (iii) by not identifying differential susceptibility to anti-TB drugs, treatment failures may arise due to the emergence of undetected drug-resistant strains. The likelihood of mixed infections is greatest in areas with a high prevalence of TB and therefore greater exposure risk (Richardson et al., 2002; Shamputa et al., 2004; Warren et al., 2004), with prevalence of mixed infection varying from 1% to 19%, depending on the geographic region (Lazzarini et al., 2007; Richardson et al., 2002; Shamputa et al., 2004; Wang et al., 2010; Warren et al., 2004). In our previous study of the $RD^{Rio}$ family in Rio de Janeiro, Brazil, we found a 1% rate of mixed infection ($RD^{Rio}$ and non-$RD^{Rio}$ strains in the same episode) but this is probably an underestimation as the $RD^{Rio}$ multiplex PCR method employed screens for a single genomic locus and is not able to identify mixed infections between two $RD^{Rio}$ strains or two non-$RD^{Rio}$ strains. The same applies to other molecular tests with single targets, such as those that segregate Beijing versus non-Beijing *M. tuberculosis* strains (Hillemann et al., 2006). The diagnosis of multiple infections is expensive and labor intensive for it is necessary to identify single colonies from the original culture and genotype each one individually. Because a major proportion of current molecular epidemiological data is based on spoligotyping, the true proportion of mixed infections remains unknown.

The SpolDB4 database is a fundamental instrument for understanding the *M. tuberculosis* population dynamics on both local and global scale. Nevertheless, its developers were aware that the presence of mixed infections could be a possible explanation for some mis-assignments and could possibly explain some undefined patterns (Brudey et al., 2006). The observation that infection with multiple strains could generate false spoligotypes has also been previously reported (Shamputa et al., 2004; Mokrousov et al., 2009). However, the current paper is the first to investigate the differential probability that some spoligotypes described in the international database may result from an admixture of one or more patterns. In addition, based on the number of cases described for each spoligotype, the weight of any misclassification could be accessed. For example, SIT523 (U Family) may be created by 78,657 possible combinations, but with only 19 cases described, the importance of any possible misclassification is relatively minimal. On the other hand, SIT53 (T family) may be produced by 292,266 possible combinations and is described in 2497 cases.

The risk of misclassification for ST53 (T family) therefore takes on greater relevance.

The T family has been considered an evolutionary modern clade and, different from other families, it seems to represent at least two divergent TB lines based on the fact that it can be classified in both Principal Genetic Groups 2 and 3 of *M. tuberculosis* by select SNP analyses (Lazzarini et al., 2007). T family phylogenetic divergence was also identified through whole-genome SNP analyses (Gutacker et al., 2006) and MIRU-VNTR (Allix-Béguec et al., 2008). Notably, the T family is described in all continents and, as a group, is the largest clade worldwide, representing more than 25% of the worldwide distribution of isolates (Brudey et al., 2006). Based on our results, it is possible that a proportion of these cases may be the result of misclassification In strikingly contrast, SIT1 (Beijing family) is less likely to be a composite and not a major concern for mis-assignment.

An experimental approach to confirm this interpretation would have been desirable but the Pasteur Institute of Guadeloupe does not hold isolates or the DNA from the submitted spoligotypes. However, our hypothesis was recently independently corroborated (Stavrum et al., 2009) by a group studying the diversity of *M. tuberculosis* genotypes in South Africa, a country with a high prevalence of TB. They found that the two most predominant spoligotypes were exactly SIT53/T family (11% of total) and SIT1/Beijing family (10%). Remarkably, over half of all SIT53/T family were proved to be of mixed *M. tuberculosis* populations by MIRU-typing whereas none of the SIT1/Beijing family samples were mixed - results that suggest that our computational results may be used in real life situations to identify strains requiring additional screening for dual infections based upon their derived spoligotype.

Globally, the striking prevalence of LAM9, T1, and to a lesser extent of EAI5, may be taken as a flag to suspect admixture patterns. Knowledge of the relative prevalence of co-localized spoligotype pairs that could combine to produce a separate frequently observed spoligotype, as we illustrated in the case of MANU1, may serve as an additional criterion to pinpoint SIT designations that should be ideally reconfirmed using MIRU-VNTR typing or other molecular methods to exclude polyclonal infections. Clearly, further studies systematically focusing on the identified "troublesome" spoligotype lineages are needed to ascertain whether routine secondary typing to exclude a dual infection is warranted in these cases. Nonetheless, based on current data, the impact of misclassification owing to individual submissions of mixed strain isolates in the SpolDB4 database is probably minimal, although a proportion of SIT53 entries may be suspect.

## 5. Conclusion

We describe the development and use of a novel computational approach to analyzing spoligotype data which, although speculative, does provide a framework for future experimental research to test the important theories and hypotheses raised. Over half of all spoligotypes in the SpolDB4 database could potentially be a composite of two other spoligotypes, however, the likelihood seems in general to be low. Exceptions primarily include SIT53 (T1), SIT100 (MANU1), SIT54 (MANU2), SIT42 (LAM9), and SIT236 (EAI-5). In contrast, SIT1 (Beijing family), the most common spoligotype in the database, is unlikely to be an admixture of other spoligotypes. Spoligotyping will therefore continue to be an invaluable resource for epidemiological purposes and for investigators around the globe to exchange and compare data. However, when one of the above suspect patterns is identified, especially in areas with a high TB incidence where mixed infections are more likely to occur, it is highly recommended that mixed infection be ruled-out by employing MIRU-VNTR or another molecular typing method.

## Acknowledgments

We are grateful to Thierry Zozio (Institut Pasteur de la Guadeloupe) for constructive criticism and suggestions and to Julie Millet (Institut Pasteur de la Guadeloupe) for helping with MIRU typing. VH was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe. RD and JR thank the Sackler Institute for comparative genomics and the Korein Family Foundation for their support. This work was supported by US National Institutes of Health (NIH) Grant R21 AI063147 (to J.L.H), and NIH Fogarty International Center Grant U2R TW006883 [J.R. Lapa e Silva]. L.C.O.L was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and was Fogarty International Center trainee. The work done at the Pasteur Institute of Guadeloupe was financed by the Regional Council of Guadeloupe (CR/08-1612: Biodiversité et Risque Infectieux dans les modèles insulaires).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.08.028.

## References

Allix-Béguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J. Clin. Microbiol. 46, 2692–2699.

Baldeviano-Vidalón, G.C., Quispe-Torres, N., Bonilla-Asalde, C., Gastiaburú-Rodriguez, D., Pro-Cuba, J.E., Llanos-Zavalaga, F., 2005. Multiple infection with resistant and sensitive *M. Tuberculosis* strains during treatment of pulmonary tuberculosis patients. Int. J. Tuberc. Lung Dis. 9, 1155–1160.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaïa, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rüsch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Das, S., Narayanan, S., Hari, L., Mohan, N.S., Somasundaram, S., Selvakumar, N., Narayanan, P.R., 2004. Simultaneous infection with multiple strains of *Mycobacterium tuberculosis* identified by restriction fragment length polymorphism analysis. Int. J. Tuberc. Lung Dis. 8, 267–270.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., León, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendón, A., Sifuentes-Osornio, J., Ponce de León, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 188, 759–772.

Flores, L., Van, T., Narayanan, S., DeRiemer, K., Kato-Maeda, M., Gagneux, S., 2007. Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. J. Clin. Microbiol. 45, 3393–3395.

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

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

Helal, Z.H., Ashour, M.S., Eissa, S.A., Abd-Elatef, G., Zozio, T., Babapoor, S., Rastogi, N., Khan, M.I., 2009. Unexpectedly high proportion of ancestral Manu genotype *Mycobacterium tuberculosis* strains cultured from tuberculosis patients in Egypt. J. Clin. Microbiol. 47, 2794–2801.

Hillemann, D., Warren, R., Kubica, T., Rusch-Gerdes, S., Niemann1, S., 2006. Rapid detection of *Mycobacterium tuberculosis* Beijing genotype strains by real-time PCR. J. Clin. Microbiol. 44, 302–306.

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., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Lazzarini, L.C., Huard, R.C., Boechat, N.L., Gomes, H.M., Oelemann, M.C., Kurepina, N., Shashkina, E., Mello, F.C., Gibson, A.L., Virginio, M.J., Marsico, A.G., Butler, W.R., Kreiswirth, B.N., Suffys, P.N., Lapa e Silva, J.R., Ho, J.L., 2007. Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro. Brazil J. Clin. Microbiol. 45, 3891–3902.

Millet, J., Baboolal, S., Akpaka, P.E., Ramoutar, D., Rastogi, N., 2009. Phylogeographical and molecular characterization of an emerging *Mycobacterium tuberculosis* clone in Trinidad and Tobago. Infect. Genet. Evol. 9, 1336–1344.

Mokrousov, I., Valcheva, V., Sovhozova, N., Aldashev, A., Rastogi, N., Isakova, J., 2009. Penitentiary population of *Mycobacterium tuberculosis* in Kyrgyzstan: exceptionally high prevalence of the Beijing genotype and its Russia-specific subtype. Infect. Genet. Evol. 9, 1400–1405.

Richardson, M., Carroll, N.M., Engelke, E., Van Der Spuy, G.D., Salker, F., Munch, Z., Gie, R.P., Warren, R.M., Beyers, N., Van Helden, P.D., 2002. Multiple *Mycobacterium tuberculosis* strains in early cultures from patients in a high-incidence community setting. J. Clin. Microbiol. 40, 2750–2754.

Shamputa, I.C., Rigouts, L., Eyongeta, L.A., El Aila, N.A., van Deun, A., Salim, A.H., Willery, E., Locht, C., Supply, P., Portaels, F., 2004. Genotypic and phenotypic heterogeneity among *Mycobacterium tuberculosis* isolates from pulmonary tuberculosis patients. J. Clin. Microbiol. 42, 5528–5536.

Stavrum, R., Mphahlele, M., Ovreås, K., Muthivhi, T., Fourie, P.B., Weyer, K., Grewal, H.M., 2009. High diversity of *Mycobacterium tuberculosis* Genotypes in South Africa and preponderance of mixed infections among ST53 isolates. J. Clin. Microbiol. 47, 1848–1856.

Wang, J.Y., Hsu, H.L., Yu, M.C., Chiang, C.Y., Yu, F.L., Yu, C.J., Lee, L.N., Yang, P.C.; the TAMI Group., 2010. Mixed infection with Beijing and non-Beijing strains in pulmonary tuberculosis in Taiwan: prevalence, risk factors, and dominant strain. Clin. Microbiol. Infect. Oct 14. doi:10.1111/j.1469-0691.2010.03401.x.

Warren, R.M., Victor, T.C., Streicher, E.M., Richardson, M., Beyers, N., Gey van Pittius, N.C., van Helden, P.D., 2004. Patients with active tuberculosis often have different strains in the same sputum specimen. Am. J. Respir. Crit. Care Med. 169, 610–614.

WHO, 2010. Tuberculosis fact sheet no. 104. World Health Organization, Geneva, Switzerland. http://www.who.int/mediacentre/factsheets/fs104/en/index.html.

Yeh, R.W., Hopewell, P.C., Daley, C.L., 1999. Simultaneous infection with two strains of *Mycobacterium tuberculosis* identified by restriction fragment length polymorphism analysis. Int. J. Tuberc. Lung Dis. 3, 537–539.