Infection, Genetics and Evolution 12 (2012) 789–797

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

# TB-Lineage: An online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex

Amina Shabbeer [a], Lauren S. Cowan [c], Cagri Ozcaglar [a], Nalin Rastogi [d], Scott L. Vandenberg [e], Bülent Yener [a], Kristin P. Bennett [a,b,*]

[a] Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY, USA
[b] Mathematical Sciences Dept., Rensselaer Polytechnic Institute, Troy, NY, USA
[c] Centers for Disease Control and Prevention, Atlanta, GA 30333, USA
[d] Institut Pasteur de Guadeloupe, Abymes, Guadeloupe, France
[e] Computer Science Dept., Siena College, Loudonville, NY 12211, USA

ARTICLE INFO

*Article history:*
Available online 3 March 2012

*Keywords:*
Tuberculosis
Classification
Spoligotype
MIRU-VNTR
Spoligoforest
Lineage

ABSTRACT

This paper formulates a set of rules to classify genotypes of the *Mycobacterium tuberculosis* complex (MTBC) into major lineages using spoligotypes and MIRU-VNTR results. The rules synthesize prior literature that characterizes lineages by spacer deletions and variations in the number of repeats seen at locus MIRU24 (alias VNTR2687). A tool that efficiently and accurately implements this rule base is now freely available at http://tbinsight.cs.rpi.edu/run_tb_lineage.html. When MIRU24 data is not available, the system utilizes predictions made by a Naïve Bayes classifier based on spoligotype data. This website also provides a tool to generate spoligoforests in order to visualize the genetic diversity and relatedness of genotypes and their associated lineages. A detailed analysis of the application of these tools on a dataset collected by the CDC consisting of 3198 distinct spoligotypes and 5430 distinct MIRU-VNTR types from 37,066 clinical isolates is presented. The tools were also tested on four other independent datasets. The accuracy of automated classification using both spoligotypes and MIRU24 is >99%, and using spoligotypes alone is >95%. This online rule-based classification technique in conjunction with genotype visualization provides a practical tool that supports surveillance of TB transmission trends and molecular epidemiological studies.

© 2012 Elsevier B.V. All rights reserved.

## 1. Introduction

The classification of *Mycobacterium tuberculosis* complex (MTBC) strains into a phylogenetic framework has a long history. The analysis of spoligotype patterns led by Sola and Rastogi revealed the presence of related spoligotypes which have since been grouped into 62 spoligotype families or clades (Brudey et al., 2006). Early studies of sequence diversity in selected genes described three principal genetic groups which correlated well with spoligotype families (Sreevatsan et al., 1997). The sequencing of several MTBC complex genomes revealed additional single nucleotide polymorphisms (SNPs) defining additional genetic groupings (Filliol et al., 2006; Gutacker et al., 2006). Gagneux and colleagues used comparative genomics to identify large sequence polymorphisms (LSPs) defining major lineages in the MTBC (Gagneux et al., 2006) which were later supported by the sequencing of 89 genes from 108 strains (Hershberg et al., 2008). These studies have culminated in the description of a phylogenetic framework for strain classification along with a proposal for standardized nomenclature (Coscolla and Gagneux, 2010) which is summarized in Table 1.

Strains can be classified as ancient (also known as ancestral) or modern. The ancient strains belong to principal genetic group 1 and include *Mycobacterium africanum* (West African 1 and 2) and *Mycobacterium bovis* as well as the strains forming the Indo-Oceanic lineage (lineage 1). The modern strains include principal genetic group 1 strains belonging to the East Asian lineage (lineage 2) or to the East-African Indian lineage (lineage 3) and principal genetic group 2/3 strains belonging to the Euro-American lineage (lineage 4) (Gagneux and Small, 2007). Although SNPs and LSPs remain the gold standard for strain classification, the congruence between spoligotypes and LSPs has been reported (Coscolla and Gagneux, 2010; Kato-Maeda et al., 2011). This framework has allowed investigators to begin to answer questions regarding the impact of the genotypic diversity of MTBC on TB disease (reviewed in Coscolla and Gagneux, 2010).

* Corresponding author at: Mathematical Sciences Dept., Rensselaer Polytechnic Institute, Troy, NY, USA.
*E-mail addresses:* shabba@cs.rpi.edu (A. Shabbeer), los4@cdc.gov (L.S. Cowan), ozcagc2@cs.rpi.edu (C. Ozcaglar), nrastogi@pasteur-guadeloupe.fr (N. Rastogi), vandenberg@siena.edu (S.L. Vandenberg), yener@cs.rpi.edu (Bülent Yener), bennek@rpi.edu (K.P. Bennett).

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.02.010

**Table 1**
Mapping of lineage names from conventions used in prior literature.

| TB-Lineage family | LSP-based lineages (Gagneux et al., 2006) | SpolDB4 family (Brudey et al., 2006) | Principal genetic group (Sreevatsan et al., 1997) | TbD1 assignment (Brosch et al., 2002) |
|---|---|---|---|---|
| *M. bovis* | *M. bovis* | BOV1, BOV2, BOV3, BOV4, BOV1-variant1, BOV2-variant1, BOV1-variant2, BOV2-variant2 | 1 | Ancient |
| West African 1 | West African 1 (lineage 5) | Afri2, Afri3 | 1 | Ancient |
| West African 2 | West African 2 (lineage 6) | Afri1 | 1 | Ancient |
| Indo-Oceanic | Indo-Oceanic (lineage 1) | EAI-5, EAI1-SOM, EAI2-Manila, EAI2-Nonthaburi, EAI3-IND, EAI4-VNM, EAI6-BGD1, EAI6-BGD2, EAI8-MDG | 1 | Ancient |
| East Asian | East Asian (lineage 2) | Beijing, Beijing-like | 1 | Modern |
| East-African Indian | East-African Indian (lineage 3) | CAS1-Delhi, CAS1-Kili, CAS1-variant CAS2 | 1 | Modern |
| Euro-American | Euro-American (lineage 4) | T1, T1-RUS2, T2,T2-Uganda, T3, T3-ETH, T4, T4-CEU1, T5-Madrid2, T5-RUS1, Tuscany, S,X1, X2, X3, X2-variant1, X3-variant1, X3-variant2, H1, H1-variant1, H2, H3, LAM01, LAM02, LAM03, LAM04, LAM05, LAM06, LAM07, LAM08, LAM09, LAM10, LAM07-TUR, LAM10-CAM, LAM11-ZWE, LAM12-Madrid | 2 and 3 | Modern |

Spoligotypes and mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR) types are now routinely collected as part of TB surveillance in the United States as well as many other countries creating large databases rich with information regarding risk factors, clinical presentations and strain data. The genotyping of strains for the purpose of detecting chains of transmission requires standardized techniques that are fast, efficient, reproducible and highly discriminative (Kremer et al., 1999, 2005). Spoligotyping is the typing of MTBC strains based on the variability found in the direct repeat (DR) locus, which contains a variable number of short direct repeats interspersed with non-repeating spacers (Kamerbeek et al., 1997). MIRU-VNTR typing is based on the number of tandem repeats present at up to 24 identified loci distributed across the MTBC genome (Supply et al., 2000, 2006). However, LSP and SNP data are not appropriate tools for monitoring transmission and therefore are not collected by health care organizations.

To mine these large surveillance databases there is a need to use readily available genotyping data (spoligotype and MIRU-VNTR) to classify strains. In this study, we have defined an ordered set of rules to determine lineage using both spoligotype and MIRU-VNTR data or using spoligotype data alone. We also present a freely available, efficient and easy to use online tool that implements this classification method, TB-Lineage (http://tbinsight.cs.rpi.edu/run_tb_lineage.html). TB-Lineage is part of a suite of tools offered by the TB-Insight project for the epidemiology of TB including classification and visualization of MTBC strains and the investigation of host-pathogen relationships (Shabbeer et al., 2011). TB-Lineage is integrated with the TB-Vis tool to visualize the classified strains as a spoligoforest.

## 2. Material and methods

### 2.1. Training set

The CDC dataset contains the genotypes of 37,066 clinical isolates typed from January 2004 to September 2008. These records were assigned major lineage labels by CDC experts based on the spoligotype pattern and MIRU24 locus. A database was generated using Oracle Database 10g Express Edition to store and maintain the data. It comprises of 3198 distinct spoligotypes, 5430 distinct MIRU types and 10828 distinct genotypes (spoligotype-MIRU type pairs). The lineage-wise distribution of strains is described in Table 2. The training set was used to develop the clauses for the rule-based system and to decide the precedence of the rules. Additionally, it was used to determine the priors for the modern/ancestral Naïve Bayes classifier.

### 2.2. Test sets

This study utilized four different independent datasets to test the performance of the rules. A summarized description of these sets is presented in Table 2. It includes the total number of isolates in the datasets (the number of genotypes weighted by their number of occurrences), the number of distinct genotypes, the number of distinct spoligotypes and MIRU types and the distribution of the isolates by lineage. The CDC2011 dataset contains genotype results collected between September 2008 and February 2011 that have distinct genotypes (spoligotype and MIRU-VNTR type pairs) from those in the CDC dataset. The CDC2011 dataset corresponds to out-of-sample test data since it did not exist when the rules were

**Table 2**
Summary description of the datasets used in the development of the online tool TB-Lineage.

| Dataset | Total labeled | No. of distinct spoligotypes | No. of distinct MIRU-VNTR types | No. of distinct genotypes | *M. bovis* | West African 1 | West African 2 | Indo-Oceanic | East Asian | East-African Indian | Euro-American |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CDC | 37,066 | 3198 | 5430 | 10828 | 685 | 67 | 81 | 5177 | 4829 | 1446 | 24,781 |
| CDC2011 | 3138 | 1548 | 2363 | 3138 | 56 | 19 | 22 | 421 | 269 | 403 | 1948 |
| MIRU-VNTR*plus* | 165 | 84 | 109 | 128 | 11 | 20 | 11 | 12 | 10 | 10 | 91 |
| Brussels | 442 | 198 | 301 | 378 | 17 | 4 | 9 | 27 | 16 | 30 | 339 |
| SpolDB4 | 31,212 | 1589 | – | 1589 | 4731 | 228 | 89 | 2716 | 3973 | 425 | 19,050 |

generated The MIRU-VNTR*plus* dataset used comprises of the strains (described by spoligotype and 12 loci of MIRU) belonging to the major lineages from the highly curated MIRU-VNTR*plus* database. These strains were assigned a lineage label based on the SNP, LSP spoligotype and MIRU-VNTR profiles of the strains (Allix-Beguec et al., 2008b). The Brussels dataset comprises of 442 labeled data points, with 378 distinct genotypes (described by spoligotype and 12 loci of MIRU). The authors (Allix-Beguec et al., 2008a) assigned lineage to the genotypes in this dataset by matching with spoligotype signatures of classical prototypes, tree-based analysis using 24 loci MIRU-VNTR profiles and also by consistent best matches with the reference database in MIRU-VNTR*plus*. The SpolDB4 dataset comprises of a diverse set of strains assigned to spoligotype families based on visual rules and classical prototypes by the authors (Brudey et al., 2006). The spoligotype family labels and LSP-based lineage labels assigned to isolates in these datasets were converted to major lineage labels using the conventions outlined in Table 1. These datasets contain a large number of distinct genotypes (spoligotype and MIRU type pairs) that challenge the rule base and hence serve as good independent test sets.

Two sets of experiments were conducted with these datasets. First, the rules were applied on the spoligotype and MIRU type data for datasets CDC, CDC2011, MIRU-VNTR*plus* and Brussels. Second, the Naïve Bayes classifier was applied to only the spoligotype data from all five sets, CDC, CDC2011, MIRU-VNTR*plus*, Brussels and SpolDB4. The results from the predictions now serve as a surrogate for MIRU24 and are added to the spoligotype data. The rules were then applied on these augmented datasets.

### 2.3. Lineage classification

Prior literature describes spoligotype-based visual rules for classification, as well as bioinformatics approaches that identify specific spacers and loci that can be used to differentiate strains into groups (Borile et al., 2011; Brudey et al., 2006; de Jong et al., 2010; Ferdinand et al., 2004; Filliol et al., 2002, 2003; Sebban et al., 2002; Sola et al., 2001). This literature including descriptions of visual rules for SpolDB4 sub-families was synthesized to generate a set of observations for major lineage classification:

- Spacers 29–32 and 33–36 in the direct repeat locus may be used to determine if a strain belongs to principal genetic group 1 (PGG1) or groups 2 and 3 (PGG2/3). If spacers 33–36 are absent and at least one of spacers 29–32 is present, the strain belongs to PGG2/3. If at least one of spacers 33–36 is present, then the strain belong to PGG1. If spacers 29–36 are absent, then the principal genetic group cannot be determined on the basis of spoligotype alone.
- PGG 1 strains with spacer 3, 9, 16 and 39–43 absent are *M. bovis.*
- PGG1 strains with spacers 8, 9 and 39 absent are *M. africanum. M. africanum* strains that have spacers 8–12 and 37–39 absent belong to the West African 1 lineage and strains that have spacers 7–9 and 39 absent belong to West African 2.
- PGG 1 strains with spacers 29–32 absent and spacer 34 absent belong to the Indo-Oceanic lineage.
- PGG1 strains with spacers 1–34 absent belong to the East Asian lineage.
- PGG 1 strains with spacers 4–7 and 23–24 absent belong to the East-African Indian lineage.
- PGG 2 and 3 strains belong to the Euro-American lineage.
- The number of repeated units at locus MIRU24 correlates with the TbD1 deletion. Strains with the sequence intact (ancient) will have more than one repeated unit at MIRU24 while strains with the sequence deleted will have a single repeated unit in this locus.

These observations have been used to assign genotypes in CDC datasets to a lineage. An important note is that these observations are not exhaustive and are not meant to define a signature for all members of a lineage.

In order to evaluate the performance of the TB-Lineage tool, two forms of validation were performed as described in the next Sections 2.3.1 and 2.3.2.

#### 2.3.1. Validation of expert labels

First, since the rules were based on the approach used by the CDC experts, an independent tool provided by www.MIRU-VNTRplus.org was used to validate the labels assigned by the experts themselves. The tool finds the nearest neighbor in a highly curated reference database of strains by calculating the distance of the spoligotype and MIRU-VNTR profiles of a strain (Allix-Beguec et al., 2008b) to all the strains in the reference database. The strain is assigned a label corresponding to the nearest reference strain found by this best-match approach. A greater weight is assigned to MIRU when labels could not be ascertained using the default settings. The spoligotype family output was converted to lineage using the conventions outlined in Table 1.

#### 2.3.2. TB-Lineage performance evaluation

Classification accuracy was evaluated using the *f*-measure, the harmonic mean of the precision and recall. Precision is defined as the probability of a strain actually belonging to a lineage given that it is predicted to belong to that lineage as calculated by precision = true positives/(true positives + false negatives). Recall is the probability of correctly classifying a strain as belonging to a certain lineage given that the strain actually belongs to that lineage, and is calculated by recall = true positives/(true positives + false positives). The *f*-measure, given by $2 * \text{precision} * \text{recall}/(\text{precision} + \text{recall})$, is used as a performance metric for the classification task and takes into account both precision and recall of classes assigned by TB-Lineage as compared with expert-assigned labels.

### 2.4. Naïve Bayes classifer to predict modern or ancestral

Some rules rely on the number of repeats observed at MIRU24 locus to predict whether a strain is ancestral or modern (rules described in detail in Section 3). Modern strains typically contain one repeat at MIRU24 while ancestral strains contain greater than one repeat (Ferdinand et al., 2004). However, since MIRU-VNTR typing was adopted as a national standard only in 2004, MIRU-VNTR type data may not be available for earlier records. In order to overcome this limitation in data availability, a Naïve Bayes classifier was developed to classify spoligotypes as ancestral or modern. The model was trained using labeled spoligotype data with spoligotypes belonging to the East Asian, East-African Indian and Euro-American lineages assigned label modern and spoligotypes belonging to the West African 1 and 2, *M. bovis*, and Indo-Oceanic lineages assigned label ancestral. Previously, MTBC strains have been successfully classified at varying levels of granularity on the basis of spoligotype data using Bayesian Networks (Aminian et al., 2010; Vitol et al., 2006). Instead of the MIRU24 value that originally served as a surrogate for modern or ancestral, the modern or ancestral prediction made by the Naïve Bayes classifier is substituted into the rules.

For the purpose of the Naïve Bayes classifier, each spoligotype is represented as a binary twelve-dimensional feature vector. Each dimension represents the presence/absence of a contiguous deletion. Presence of a deletion means no spacers are present in the subsequence, while absence means at least one spacer is present in the subsequence. The evolution of the DR locus occurs via deletion of one or more contiguous Direct Variable Repeats (DVRs) with some non-negligible probability, whereas insertion of DVRs is

highly unlikely (Warren et al., 2002). The features selected to represent a spoligotype were single deletions of spacers 3, 16, 8, 9, 39 and contiguous deletions of spacers 1–34, 25–28, 29–32, 33–36, 39–43, 4–7 and 23–24. The selection of deletion sequences was made on the basis of the observations listed above about the sequences that are relevant for the lineage classification task and further validated based on the information gain as described in the Supplementary information. If at least one spacer is present in the deletion sequence it is represented as a 1, presence of the deletion is represented by a 0. These bits are concatenated to form a binary vector, e.g. considering the deletions in the order specified, if spacers 3,16,8,9,39 and 23–24 are absent, while there is at least one spacer present in sequences 1–34, 25–28, 29–32, 33–36, 39–43 and 4–7 the spoligotype is represented as 000001111110.

The Naïve Bayes model assumes that each feature $S_d$ with possible values 0 or 1, is independent given the class $C_i$ (modern or ancestral). $S_d$ takes value 1 if at least one spacer is present in deletion $d$ and 0 otherwise. The probability of the deletion $d$ being absent (at least one spacer being present) for a strain of class $i$ is given by $p_{id}$, and the probability of a deletion $d$ being absent (no spacer(s) being present) is given by $(1 - p_{id})$. We assume the conditional independence of the features of variable $S$. The probability of occurrence of a spoligotype $S$ given the class $C_i$ is therefore $P(S \mid C_i) = \prod_d (p_{id})^{S_d}(1 - p_{id})^{(1-S_d)}$.

In the test phase, strains were classified into the two classes – modern and ancestral. The probability of a strain $S$ belonging to class $C_i$ was computed as

$$P(C_i|S) \propto P(C_i)P(S|C_i)/P(S)$$
$$\propto P(C_i)P(S|C_i)$$

The values of $p_{id}$ and $P(C_i)$ were calculated by counting the appropriate proportion of values in the data with Laplace smoothing to deal with deletions observed 0 times. The strain S is assigned to the class i for which its conditional probability $P(S|C_i)$ is the highest. The value of the MIRU24 locus is predicted on the basis of the modern/ancestral classification and used in place of the MIRU24 in the rules.

### 2.5. TB-Vis spoligoforests

Spoligoforests represent evolutionary relationships between genotypes in the dataset. Nodes represent spoligotype-based clusters and directional relationships between nodes represent the deletion of one or more adjacent spacers (contiguous deletions) in the spoligotype. These relationships between spoligotype clusters are henceforth referred to as 'parent–child' relations. A set of candidate parents is generated for each spoligotype cluster based on the knowledge that spacers may be lost but are rarely gained. Based on the assumption that convergent evolution is very infrequent, the single most likely parent from amongst the candidates must be chosen. A mathematical analysis of convergent evolution of spoligotypes is provided in (Reyes et al., 2011), and a discussion of the consequences of convergent evolution in spoligotypes as well as MIRU on TB epidemiology is discussed in (Comas et al., 2009). The choice of parent is made based on the principle of maximum parsimony by choosing the genotype whose genetic distance from the genotype of concern is the minimum.

The choice of parent is made by eliminating unlikely candidates from the set of possible parents as described here. The first step is based on using the Hamming distance between MIRU-VNTR types, i.e. the number of loci in which the number of repeats differ. If there remain multiple candidate parents, the Hamming distance between spoligotypes, i.e. the number of bits in which the two pairs of spoligotypes differ is used to break the tie. This arises from the fact that short spacer deletions are more likely than longer deletions as indicated by the Zipf model shown to best-fit the frequencies of deletions of various lengths (Reyes et al., 2008). If ambiguity in the choice of parents persists, it is resolved using the Euclidean distance between MIRU-VNTR types which reflects the difference in the *number* of repeats observed at each of the MIRU loci. In the rare event that more than one parent remains, a random pick is made between the best possible parents. The step-wise algorithm for spoligoforest generation is presented here.

---

Input: Strain dataset with spoligotypes and (optionally) MIRU patterns.
Output: Spoligoforest G = (V,E), where each node ϵ V is a spoligotype cluster and each edgeϵ E represents an inferred spoligotype mutation.
*/*Initialize*/*
E={}
/*Determine spoligoforest adjacency list*/
**for** each node sϵ V **do**
  Find the set of candidate parents P for node s using the assumption that spacers are lost but rarely gained.
  */*Determine single most likely parent*/*
  **if** MIRU data present then

    From set P, find subset $P_0$, the set of strains with MIRU patterns that have minimum Hamming distance to MIRU pattern associated with s. Set $P=P_0$.

    From set P, find subset $P_0$, the set of strains with MIRU patterns that have minimum Hamming distance to spoligotype of s. Set $P=P_0$.

    From set P, find subset $P_0$, the set of strains with minimum L2 distance to MIRU pattern of s. Set $P=P_0$.
  **else**
    From set P, find subset $P_0$, the set of strains with MIRU patterns that have minimum Hamming distance to spoligotype of s. Set $P=P_0$.
  **end if**
  **if** |P|> l then
      Pick a node p ϵ P at random
  **else**
      Pick the last remaining node p ϵ P
  **end if**
  Assign node p as the unique parent of node s.
  Add the edge $e_{ps}$ from node p to node s to the set E.
  E=E U $e_{ps}$
**end for**

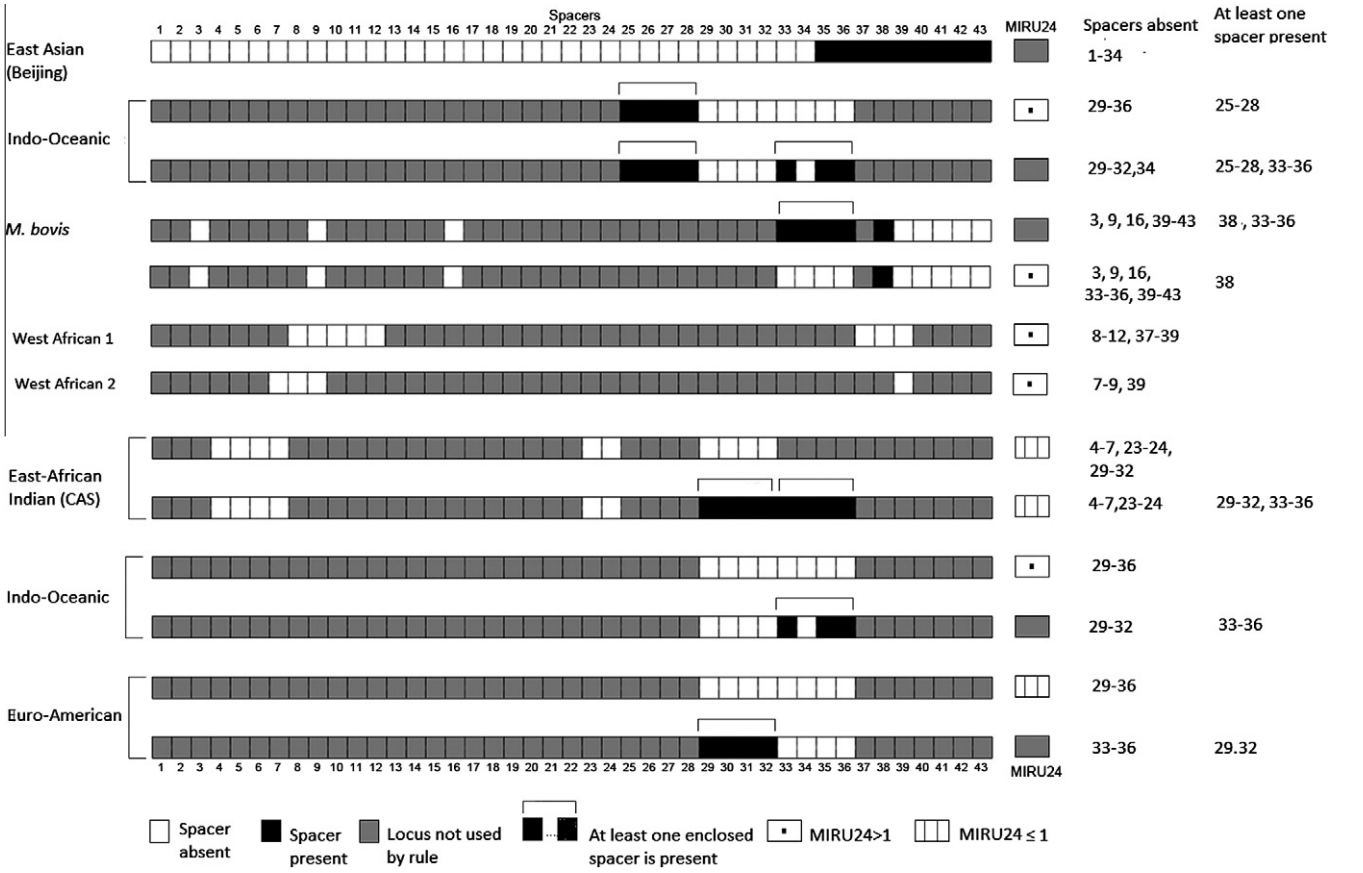


**Fig. 1.** Rules to classify *M. tuberculosis* strains into major lineages based on the presence or absence of spacer sequences and the number of repeats observed at the MIRU24 locus. The rules are applied in the order specified. The label assigned to a strain corresponds to the first rule that is satisfied.

### 2.6. Website development

This multifunctional website was implemented using Perl/CGI and Python scripts running on a Unix web server. The web-form accepts as input either a single genotype or multiple genotypes in a tab delimited text file in one of these formats: (1) spoligotype only, (2) spoligotype and MIRU24 locus, (3) spoligotype and 12 loci of MIRU-VNTR type, (4) spoligotype and 15 loci of MIRU-VNTR type, (5) spoligotype and 24 loci of MIRU-VNTR type. Sample data files available on the website provide templates of these file formats. The scripts parse the user-uploaded data and process it by testing against the defined logical clauses. Classification results are written to a spreadsheet that is available for immediate download. The option to generate and save an image of the spoligoforest based on the classification results is made available after classification is completed. The spoligoforest generating algorithm is implemented in Java, and the graph-drawing package Graphviz (Gansner and North, 2000) is used to render the graph.

## 3. Results and discussion

The rules developed in this study for the classification of MTBC into major lineages using spoligotype and MIRU24 comprise of precisely defined logical clauses based on contiguous deletions of one or more of the 43 spacers and the number of repeats at MIRU24. Prior literature describes spoligotype-based visual rules for classification, as well as bioinformatics approaches that identify specific spacers and loci that can be used to differentiate strains into groups (Brudey et al., 2006; Ferdinand et al., 2004; Filliol et al., 2002, 2003; Sebban et al., 2002; Sola et al., 2001). However, these previously defined methods (Brudey et al., 2006; Ferdinand et al., 2004; Sebban et al., 2002; Sola et al., 2001; Streicher et al., 2007) define identifying characteristics for spoligotype families that lie within the major lineages. Moreover, the precedence of these visual rules is not defined, leading to ambiguous assignment of labels in an automated system. In TB-Lineage, the rules are applied in a specified order as defined in Fig. 1; the families with the most stringent and well-defined signatures are tested first. Furthermore, precision and recall on the CDC dataset guided the assignment of precedence i.e. the order of the rules was varied, and the order that resulted in the highest precedence and recall on the training set was used. The genotype is classified by the first rule met in the following defined sequence. If no rules are satisfied, the lineage is indicated as being "not determined". These rules are not comprehensive and will not identify all strains within a lineage.

(1) *East Asian:* contiguous deletion of spacers 1–34.
(2) *Indo-Oceanic:*
(i) At least one spacer present in loci 25–28 AND contiguous deletion in loci 29–36 AND MIRU24 > 1[1], OR
(ii) Spacer 34 absent AND contiguous deletion in loci 29–32 AND at least one spacer present in loci 33–36 AND at least one spacer present in loci 25–28.
(3) *M. bovis:*
(i) Spacers 3, 9, 16 absent AND spacer 38 present AND contiguous deletion from loci 39 to 43 AND at least on spacer present in loci 33–36, OR

[1] Number of repeats at locus MIRU24 > 1.

(ii) Spacers 3, 9, 16 absent AND spacer 38 present AND contiguous deletion from loci 39 to 43 AND MIRU24 > 1 AND contiguous deletion in 33–36.
(4) *West African 1:* spacers 8–12, 37–39 absent AND MIRU24 > 1.
(5) *West African 2:* spacers 7–9, 39 absent AND MIRU24 > 1.
(6) *East-African Indian:*
(i) Contiguous deletion in spacers 4–7 AND contiguous deletion in spacers 23–24 AND contiguous deletion in loci 29–32 AND MIRU24 $\leqslant$ 1, OR
(ii) Contiguous deletion in spacers 4–7 AND contiguous deletion in spacers 23–24 AND at least one spacer in loci 29–32 AND at least one spacer in loci 33–36 AND MIRU24 $\leqslant$ 1.
(7) *Indo-Oceanic:*
(i) Contiguous deletion from 29 to 36 AND MIRU24 > 1, OR
(ii) Spacer 34 absent AND contiguous deletion from 29–32 AND at least one spacer present in loci 33–36.
(8) *Euro-American:*
(i) Contiguous deletion in loci 33–36 AND contiguous deletion in loci 29–32 AND MIRU24 $\leqslant$ 1, OR
(ii) Contiguous deletion from loci 33 to 36 AND at least one spacer present in loci 29–32.

This coherent set of rules for lineage classification was generated on the basis of the observations listed in Section 2.3. The order of the rules is equivalent to first checking membership in PGG 1 and assigning to lineages within PGG 1. The rules are tested in the order of stringency of clauses in order to minimize the number of multiple label assignments. The MIRU24 is used to determine if a strain is ancestral and belongs to PGG 1, since assignment to principal genetic groups cannot be made on the basis of spoligotype alone when there is a deletion from 29 to 36. If the MIRU24 data is not available the Naïve Bayes classifier based on the spoligotype is used to predict if a strain is ancestral and belongs to PGG 1. Assignment to specific lineages within PGG 1 is also based on MIRU24 (or modern/ancestral prediction by the Naïve Bayes classifier) and pertinent clauses for the lineages. The Beijing lineage that requires spacers 1–34 to be absent is the first in the sequence of rules. The main conditions used to test for the Indo-Oceanic lineage are absence of spacers 29–32 and 34 in PGG 1 strains. An additional clause requiring the presence of at least one spacer in 25–28 is included in Rule 2 based on observations in the training dataset. Note, while this is not a required clause to identify Indo-Oceanic strains, it is incorporated in order to impose precedence of rules such that high accuracy (as measured by *f*-measure described in Section 2.3.2) is achieved on the training set. Additionally, strains that simply satisfy absence of 29–32, 34, and are determined to belong to PGG 1 by presence of at least one spacer in 33–36 or MIRU24 > 1, and that fail to satisfy any other lineage tests would be labeled Indo-Oceanic by Rule 6. Since some of the clauses for *M. africanum* (West African 1 and 2) are subsets of the clauses in *M. bovis*, the corresponding rules for West African 1 and 2 (Rules 4 and 5) appear after that of *M. bovis* (Rule 3). When there is a contiguous deletion in 4–7 and 23–24 and MIRU24 $\leqslant$ 1, Rule 5 is satisfied and the isolate is labeled as East-African Indian. The Euro-American lineage comprises a broad range of previously defined sub-families – the LAM, Haarlem, X and T clades. The T sub-group is ambiguously defined. However, all of these families correspond to the PGG 2 and 3 and have the spacers 33–36 deleted. Absence of spacers 33–36 is the main condition tested in Rule 7 for Euro-American identification. Thus a precise, ordered set of rules are defined for MTBC major lineage classification.

As a first validation step, in order to verify the expert-defined labels themselves, the expert labels were compared to those generated by the strain-identification tool available at www.MIRU-VNTRplus.org using spoligotype and MIRU (Allix-Beguec et al., 2008b) (as described in Section 2.3). Of the 10828 distinct genotypes in the CDC dataset, 90.3% of records were assigned concordant labels by the MIRU-VNTR*plus* tool based on the three nearest neighbors. Of these labeled records there is a 99.8% correspondence between the CDC expert-defined labels and the lineages assigned by the MIRU-VNTR*plus* strain-identification tool. Table 3 is a confusion matrix showing this high correspondence between expert labels and those assigned by the MIRU-VNTR*plus* tool using all the spoligotype and MIRU-VNTR data in the CDC dataset.

The accuracy of the rules was evaluated on two datasets from the CDC, with CDC2011 containing genotypes not analyzed during development, along with datasets from MIRU-VNTR*plus* (Allix-Beguec et al., 2008b), Brussels (Allix-Beguec et al., 2008b) and SpolDB4 (Filliol et al., 2002) described in prior publications. A high level of accuracy was reported on all datasets, with the labels assigned by the online tool matching labels determined by CDC experts, LSP-based analysis or SNP-based analysis in greater than 99% of cases. A comparison of the *f*-measure of the rules used with spoligotype and MIRU-VNTR pattern for each lineage applied to all the datasets is reported in Table 4. Near perfect *f*-measure, i.e. close to 1, were observed across all lineages. Detailed results of the classification of strains from the CDC dataset and CDC2011 using both spoligotype and MIRU are provided in confusion matrices in Supplementary Tables 5 and 6. The rules work equally well when only spoligotypes are available as shown by *f*-measure values across all lineages and all datasets in Supplementary Table 7. Detailed results of the classification of strains from the CDC dataset using spoligotypes alone are shown in Supplementary Table 8.

Reasons for discordant lineage assignment were investigated. There are several instances of strains that satisfy the clauses of rules for more than one lineage. Strains are assigned the label corresponding to the first rule satisfied. Although the precedence was defined to maximize precision and recall, some strains that satisfy multiple rules may get assigned discordant labels. The less than 100% congruence between the number of repeats at MIRU24 and TbD1 may also result in a discordant lineage assignment. An

**Table 3**
Confusion matrix showing 99.8% correspondence between expert-assigned labels and those determined by the tool at www.MIRU-VNTRplus.org using the distinct pairs of spoligotype and 12 loci of MIRU-VNTR type of strains in the CDC dataset.

| | MIRU-VNTR*plus* lineage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | East Asian | East-African Indian | Euro-American | Indo-Oceanic | West African 1 | West African 2 | *M. bovis* | Unlabeled |
| *CDC labeled lineage* | | | | | | | | |
| East Asian | 653 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| East-African Indian | 0 | 531 | 0 | 0 | 0 | 1 | 0 | 164 |
| Euro-American | 0 | 0 | 6710 | 0 | 17 | 1 | 0 | 613 |
| Indo-Oceanic | 0 | 0 | 0 | 1707 | 4 | 1 | 0 | 189 |
| West African 1 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 |
| West African 2 | 0 | 0 | 0 | 0 | 0 | 48 | 0 | 3 |
| *M. bovis* | 0 | 0 | 0 | 0 | 0 | 0 | 127 | 5 |

**Table 4**
Comparison of ***f***-measure of classification in the five datasets using both spoligotype and MIRU data.

| | East Asian | East-African Indian | Euro-American | Indo-Oceanic | West African 1 | West African 2 | *M. bovis* |
|---|---|---|---|---|---|---|---|
| CDC | 1.0 | 0.9935 | 0.9993 | 0.9985 | 0.9424 | 0.9873 | 0.9985 |
| CDC2011 | 0.9926 | 0.9926 | 0.9969 | 0.9916 | 0.95 | 1.0 | 0.9907 |
| MIRU-VNTR*plus* | 1.0 | 0.9524 | 0.9942 | 1.0 | 1.0 | 1.0 | 1.0 |
| Brussels | 0.9696 | 1.0 | 0.9941 | 0.9474 | 0.8000 | 1.0 | 1.0 |
| SpolDB4 | 0.9926 | 0.9926 | 0.9969 | 0.9905 | 0.9268 | 1.0 | 0.9907 |

analysis of records from the CDC dataset that were misclassified is provided in the Supplementary Section in the Supplementary Table 9.

In addition to assigning lineage, TB-Insight provides the TB-Vis tool to analyze genotype datasets by visualizing the number of occurrences of strains, their distribution by lineage, and potential evolutionary relationships between strains. The visualization of spoligotypes provided by TB-Vis builds on the design of spoligoforests (Reyes et al., 2008), wherein each node represents a unique spoligotype and each edge between the 'parent' spoligotype and the 'child' represents a putative mutation event. The spoligoforest created based on the genotypes in the CDC dataset is shown in Fig. 2. Each lineage corresponds to a unique color. A node represents a cluster of strains of the same spoligotype but different MIRU types, and is assigned a color based on the lineage to which the spoligotype belongs. Node sizes are representative of the number of occurrences of strains of that spoligotype in the dataset on a log scale (base 2). Edges indicate potential evolutionary relationships resulting from a contiguous deletion of spacers and changes in the number of repeats at loci of MIRU-VNTR types associated with the spoligotype, and are thus an indication of the relatedness of strains.

The structure of the MTBC population and the genetic relatedness of strains in the CDC dataset are illustrated in the spoligoforest in Fig. 2. The lineage assigned by rules may be further verified by making a visual examination of the generated spoligoforest, wherein the relatedness of strains is established based on their genetic similarity. Connected components, i.e. spoligotypes linked by edges within the graph may be viewed as belonging to a single group that share evolutionary history. The seven lineages appear in distinct connected components in the resulting spoligoforest, reflecting the fact that strains belonging to the same lineage have

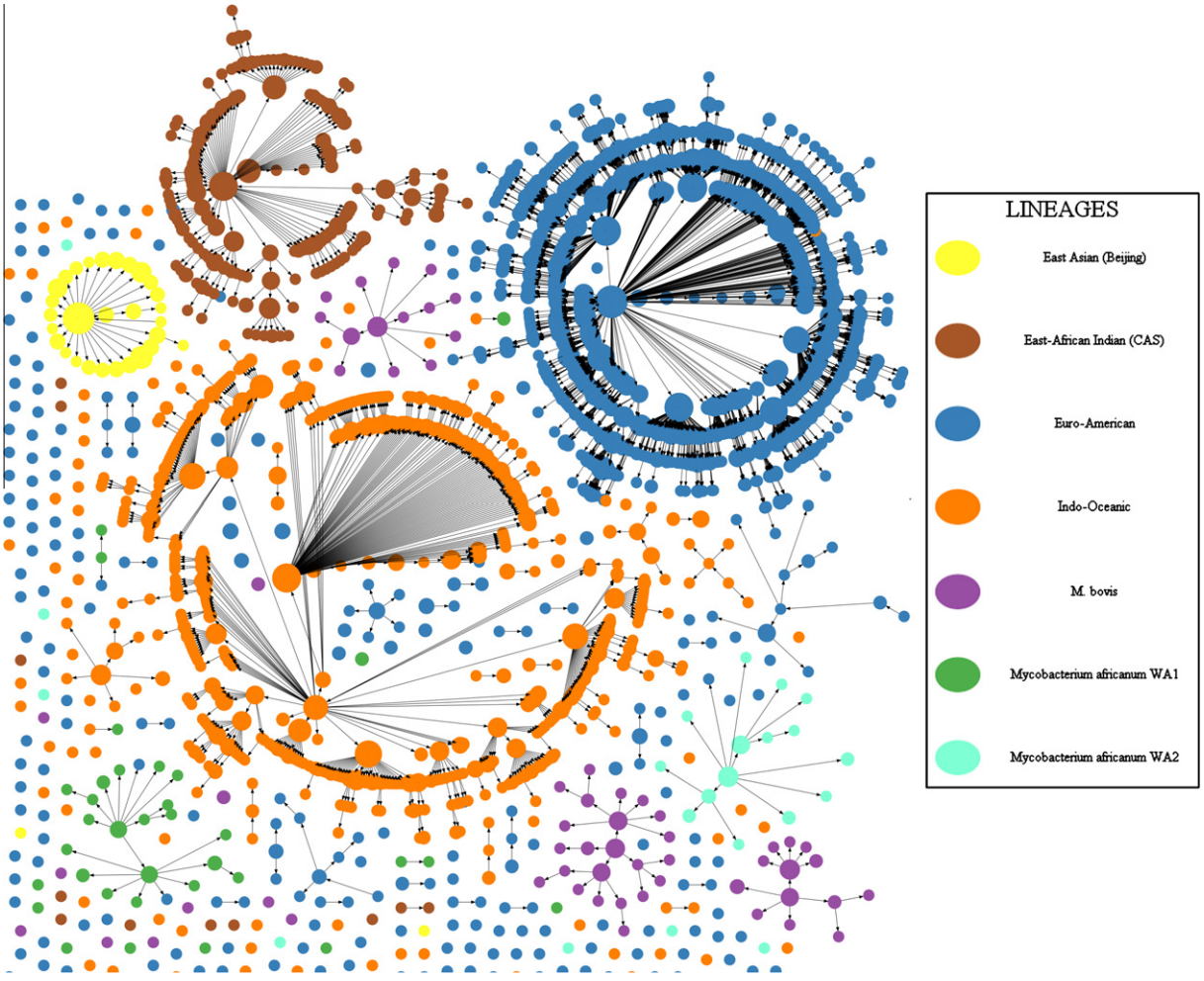


**Fig. 2.** Spoligoforest representation of genetic diversity of MTBC strains in 37,066 isolates collected from TB patients in the United States from 2004 to 2008. Each lineage corresponds to a unique color as shown in the legend. Each node represents a cluster of strains of the same spoligotype but different MIRU types, and the size represents the number of isolates on a log scale. The lineages are highly cohesive with few edges between lineages. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

evolved from a common progenitor strain. Since the edges on the spoligoforests represent possible mutations, spoligotypes that belong to the same connected component potentially share an evolutionary relationship. Therefore, it can be inferred that the rules group related strains into lineages that capture this relationship. Visually one can see that the lineages are well separated since there are few edges that connect isolates of different lineages. Note that the lineages are not used to determine the edges of the spoligoforests or placement of nodes. Thus, the visualization provides a means of verifying and validating the labels assigned by the rules. The segregation accuracy of the groups based on the percentage of edges that occur within a lineage in the spoligoforests for the CDC dataset is greater than 99%. This is represented in the Supplementary Table 10.

## 4. Conclusions

We present a web tool, TB-Lineage, which implements a rule base enabling automatic classification of MTBC strains into lineages or genetic groups based on previous knowledge of spoligotype and MIRU-VNTR type signatures for lineages. Classification of strains of the MTBC can be accomplished easily and efficiently using spoligotype signatures gathered for TB surveillance without the use of expensive and time-consuming additional genomic analysis. Visualization of lineages in the form of spoligoforests provides an alternative perspective on the genotypes in the dataset: the distribution of the strains by lineage, evolutionary relationships between strains, and frequently occurring strains. Such a software-based implementation reduces the effort involved in visual inspection of the spoligotype and MIRU-VNTR types. It enables the fast and efficient classification of MTBC strains using DNA fingerprint data readily available in large-scale databases and a study of the variance in strain characteristics by lineage.

## 5. Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

## Acknowledgments

This work was made possible by and with the assistance of Dr. Philip Supply of the Institute Pasteur de Lille. We thank Veronique Hill, David Couvin and Thierry Zozio at Institut Pasteur de la Guadeloupe and Minoo Aminian and Veronica Ahiati at Rensselaer Polytechnic Institute for their helpful suggestions and help with data analysis. This work was supported by NIH R01LM009731.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2012.02.010.

## References

Allix-Beguec, C., Fauville-Dufaux, M., Supply, P., 2008a. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 46, 1398–1406.

Allix-Beguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008b. 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.

Aminian, M., Shabbeer, A., Bennett, K., 2010. A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages. BMC Bioinformatics 11, S4.

Borile, C., Labarre, M., Franz, S., Sola, C., Refregier, G., 2011. Using affinity propagation for identifying subspecies among clonal organisms: lessons from *M. tuberculosis*. BMC Bioinformatics 12, 224.

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. USA 99, 3684–3689.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, 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., Ly, H.M., Martin, C., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-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 (SpoIDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6.

Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS ONE 4, e7815.

Coscolla, M., Gagneux, S., 2010. Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov. Today: Dis. Mech. 7, e47–e59.

de Jong, B.C., Antonio, M., Gagneux, S., 2010. Mycobacterium africanum—review of an important cause of human tuberculosis in West Africa. PLoS Neglected Trop. Dis. 4, e744.

Ferdinand, S., Valetudie, G., Sola, C., Rastogi, N., 2004. Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. Res. Microbiol. 155, 647–654.

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., Binzi, D.G., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Kallenius, G., Kassa-Kelembho, E., Koivula, T., 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., 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg. Infect. Dis. 8, 1347–1349.

Filliol, I., Driscoll, J.R., van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valetudie, G., Dang, D.A., 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., Kassa-Kelembho, E., Ho, M.L., 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.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, 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.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de 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.

Gagneux, S., Small, P., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect. Dis. 7, 328–337.

Gansner, E.R., North, S.C., 2000. An open graph visualization system and its applications to software engineering. Softw. Pract. Exper. 30, 1203–1233.

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.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 6, e311.

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

Kato-Maeda, M., Gagneux, S., Flores, L., Kim, E., Small, P., Desmond, E., Hopewell, P., 2011. Strain classification of *Mycobacterium tuberculosis*: congruence between

large sequence polymorphisms and spoligotypes. Int. J. Tuberc. Lung Dis. 15, 131–133 (Short communication).

Kremer, K., Arnold, C., Cataldi, A., Gutierrez, M.C., Haas, W.H., Panaiotov, S., Skuce, R.A., Supply, P., van der Zanden, A.G.M., van Soolingen, D., 2005. Discriminatory power and reproducibility of novel DNA typing methods for *Mycobacterium tuberculosis* complex strains. J. Clin. Microbiol. 43, 5628–5638.

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 epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. 37, 2607–2618.

Reyes, J., Chan, C., Tanaka, M., 2011. Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in *Mycobacterium tuberculosis*. Infect. Genet. Evol. http://dx.doi.org/10.1016/j.meegid.2011.05.018.

Reyes, J.F., Francis, A.R., Tanaka, M.M., 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9, 496.

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

Shabbeer, A., Ozcaglar, C., Yener, B., Bennett, K.P., 2011. Web tools for molecular epidemiology of tuberculosis. Infect. Genet. Evol.. http://dx.doi.org/10.1016/j.meegid.2011.08.019.

Sola, C., Filliol, I., Gutierrez, M.C., Mokrousov, I., Vincent, V., Rastogi, N., 2001. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. Emerg. Infect. Dis. 7, 390–396.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M., 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. USA 94, 9869–9874.

Streicher, E.M., Victor, T.C., van der Spuy, G., Sola, C., Rastogi, N., van Helden, P.D., Warren, R.M., 2007. Spoligotype signatures in the *Mycobacterium tuberculosis* complex. J. Clin. Microbiol. 45, 237–240.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., Locht, C., 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. 36, 762–771.

Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N., Bennett, K., 2006. Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. Infect. Genet. Evol. 6, 491–504.

Warren, R.M., Streicher, E.M., Sampson, S.L., van der Spuy, G.D., Richardson, M., Nguyen, D., Behr, A.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.