NIH Public Access
Author Manuscript
*Proceedings (IEEE Int Conf Bioinformatics Biomed)*. Author manuscript; available in PMC 2010 October 14.

Published in final edited form as:
*Proceedings (IEEE Int Conf Bioinformatics Biomed)*. 2009 November 1; 2009: 338–343. doi:10.1109/BIBM.2009.86

# Determination of Major Lineages of *Mycobacterium tuberculosis* Complex using Mycobacterial Interspersed Repetitive Units

**Minoo Aminian**, **Amina Shabbeer**, and **Kristin P. Bennett**
Departments of Mathematical Science and Computer Science, Rensselaer Polytechnic Institute

Minoo Aminian: aminim@cs.rpi.edu; Amina Shabbeer: shabba@rpi.edu; Kristin P. Bennett: bennek@rpi.edu

## Abstract

We present a novel Bayesian network (BN) to classify strains of Mycobacterium tuberculosis Complex (MTBC) into six major genetic lineages using mycobacterial interspersed repetitive units (MIRUs), a high-throughput biomarker. MTBC is the causative agent of tuberculosis (TB), which remains one of the leading causes of disease and morbidity world-wide. DNA fingerprinting methods such as MIRU are key components of modern TB control and tracking. The BN achieves high accuracy on four large MTBC genotype collections consisting of over 4700 distinct 12-loci MIRU genotypes. The BN captures distinct MIRU signatures associated with each lineage, explaining the excellent performance of the BN. The errors in the BN support the need for additional biomarkers such as the expanded 24-loci MIRU used in CDC genotyping labs since May 2009. The conditional independence assumption of each locus given the lineage makes the BN easily extensible to additional MIRU loci and other biomarkers.

### Keywords

tuberculosis; MIRU-VNTR; Bayesian network; lineages

## 1. INTRODUCTION

Tuberculosis (TB) is an acute or chronic infection caused by the *Mycobacterium tuberculosis* complex (MTBC). TB is a leading cause of death from infectious diseases world-wide. Increasingly MTBC has developed resistance toward the drugs that cure TB. There were an estimated 0.5 million cases of multi-drug resistant TB in 2007, and by the end of 2008, 55 countries and territories had reported at least one case of extensively drug resistant TB [13].

TB genotyping enriches traditional epidemiological approaches and plays an increasingly important role in TB control strategies. It helps track transmission routes, distinguish reactivation of latent infections from potential recent transmission, identify outbreaks, and quantify their severity. Additionally, laboratory cross-contamination events can be detected. Two types of DNA fingerprinting methods based on PCR are routinely used in the United States for genotyping all culture-positive TB cases: spacer oligonucleotide typing (spoligotyping) and mycobacterial interspersed repetitive units - variable number of tandem repeats (MIRU-VNTR). Spoligotyping is based on the polymorphisms found in the direct repeat locus of the mycobacterial chromosome, while MIRU is the number of repetitive units present in multiple loci [8].

Classification of strains of MTBC may help track transmission routes and enable suitable control measures given recent studies on the existence of stable host-pathogen associations [1] and phylogeographic distributions of strains [9]. Existing work for classifying strains of MTBC predominantly relies on deletion analysis to distinguish lineages [5,7]. Unfortunately

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

deletion analysis results are often not available in large genotyping data collections or for routine public health TB patient investigations. So alternatives such as mathematical models and visual rules for sub-lineage classification based on spoligotyping alone have been developed [6,3]. Traditionally, RFLP-based typing has also been used, however this method is time-consuming and the results are not comparable between labs. MIRU-VNTR*plus* [2] is a multimarker-based curated database that classifies strains by finding their nearest neighbors in the database. High accuracy results were reported using MIRU-VNTR types of strains alone, which were further boosted when augmented with other biomarkers: spoligotypes, large sequence polymorphisms (LSPs), and single nucleotide polymorphisms (SNPs).

This paper introduces a model to classify isolates into the major genetic lineages using only MIRU types. We accomplish this by creating a fast and accurate probability-based model which is readily extensible to adding further MIRU loci and other biomarkers in the future. The method identifies 6 major lineages of MTBC as identified by LSPs [9] consisting of three ancestral strains: Indo-Oceanic, *M. bovis*, and *M. africanum*, and three modern strains: Euro-American, East-African-Indian (CAS), and East-Asian (Beijing). Note that East-African-Indian lineage (EAI) will be called EAI (CAS) with the East-African-Indian referring to the lineage name in [10] determined by LSPs and CAS referring to the spoligotype family name determined by spoligotype such as in [5]. This convention is also used for East-Asian (Beijing).

A hierarchical Bayesian network (BN) was created using a database maintained by the Centers for Disease Control and Prevention (CDC). The data consists of 4768 distinct 12-loci MIRU-VNTR types pertaining to isolates collected from across the United States. The BN achieves high accuracy on the CDC data and on three other datasets from the New York State Department of Health (NYSDOH), MIRU-VNTR*plus*, and a study in Brussels [14]. We also study the distribution of the MIRU loci and shed light on the strengths and limitations of 12-loci MIRU.

We give background on MIRU in Section 2, and continue with a description of the BN model in Section 3. In Section 4 we describe the data and experiments, followed by discussion and future work.

## 2. MIRU ANALYSIS

MIRU typing is a VNTR analysis bacterial typing scheme for molecular typing of MTBC. It provides a high-throughput reproducible method for molecular typing of MTBC with the degree of discrimination depending on the number of loci used.

In this study, we primarily utilize a database of 4768 distinct MIRU types as determined by 12 MIRU loci amassed from surveillance of tuberculosis in the United States by the CDC. Altogether, there are 41 MIRU loci, of which 12 are used in this study. The names of these 12 loci are 02, 04, 10, 16, 20, 23, 24, 26, 27, 31, 39, and 40. Each MIRU genotype is labeled with one of 6 distinct lineages and weighted with the number of isolates found with that genotype in the United States from approximately 2003–2008, resulting in 31482 records. The labels were determined by Dr. Lauren Cowan of the CDC from spoligotypes and the MIRU locus 24 using a refinement of rules from the spoligotyping literature [9,5,10,7,4,3].

MIRU used in conjunction with spoligotyping has become a standard method for MTBC DNA fingerprinting. MIRU allows high-throughput, discriminatory, and reproducible analysis of clinical isolates. Because of its portable data format, MIRU typing has the potential to be a versatile tool for individual strain identification based on large reference databases or classification models. Beyond studying genetic diversity of MTBC, MIRU has become a major method for epidemiological tracking of MTBC because of its portable data and discriminatory power [11,2].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

## 3. BAYESIAN NETWORK

A hierarchical Bayesian network (BN) is created to predict the 6 lineages. A BN is a graphical representation of a probability distribution. Formally speaking, a BN is a directed acyclic graph *G(N,E)* consisting of a set of nodes $X = \{x_i \mid x_i \in N\}$ to represent the variables and a set of directed links to connect pairs of nodes.

Each node has a conditional probability distribution that quantifies the probabilistic relation between the node and its parents, such that for a network of *k* nodes:

$$P(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} P(x_i \mid parents(x_i))|$$

Therefore, one can compute the full joint probability distribution from the information in the network. In other words, a well-represented Bayesian network can capture the complete nature of the relationship between a set of variables.

### A. Bayesian Network for Efficient MIRU Classification

The 12 MIRU loci are scattered throughout the chromosome of MTBC [15]. Hence, the number of repeats present at each loci are independent of each other. Also, each locus has been shown to exhibit different degrees of allelic diversity [16]. The hierarchical BN given in Figure 1 is designed to exploit this known structure of MIRU. The random variable L represents the lineage and the random variables $M_i$ i∈I= {02, 04, 10, 16, 20, 23, 24, 26, 27, 31, 30 and 40} represent the MIRU loci.

The BN is a generative model. The value of MIRU24 generates the lineage which in turn determines the number of repeats in the remaining loci. This hierarchical structure exploits the fact that MIRU24 is known to correspond to the TbD1 deletion, a known marker for ancestral versus modern strains [2,12]. Modern strains have MIRU24 with less than 2 repeats. With rare exceptions, ancestral strains have 2 or more repeats at MIRU24. Thus the top-level variable, $M_{24}$, indicates whether MIRU24 is less than two (indicating modern lineages with high probability) or at least two (indicating ancestral lineages with high probability). The joint probability function represented by the Hierarchical BN is:

$$P(L, M_{24}, M_I) = \prod_{i \in I \setminus 24} P(M_i \mid L) P(L \mid M_{24}) P(M_{24})$$

where L ∈ {Indo-Oceanic, *M. africanum*, *M. bovis*, East-Asian(Beijing), East-African-Indian (CAS), Euro-American}, $M_{24}$∈ {< 2, ≥ 2}, M i∈ {0,1,2,3,4,5,6,7,8, ≥ 9}.

We also assume that the MIRU loci are conditionally independent given the lineage - a reasonable assumption since the MIRU loci are distributed throughout the genome. Each MIRU locus is modeled as a multinomial distribution with possible values 0, 1…8, and ≥9. Note all values greater than 9 are binned together since they are very rare. Standard Laplacian smoothing with parameter 1 was used for parameter estimation in the BN.

Bayes' rule can predict the lineage for new data by determining the lineage with maximum probability:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$P(L=c \mid M) \propto \prod_{i \in I \setminus 24} P(M_i \mid L=c) P(L=c \mid M_{24})$$

### B. MIRU Datasets

Four datasets were used in this study. The primary CDC dataset was used to construct the classification models. Three additional datasets collected and labeled in independent studies were used to test the models.

The CDC dataset consists of 4768 distinct MIRU types that correspond to genotypes of 31482 occurrences of TB cases across the United Sates. The dataset was pre-processed to exclude MIRU types with unknown numbers of repeats. The records were assigned labels based on expert-defined rules [5,9,7,4,3] that use spoligotype and MIRU24. The distribution of number of cases (number of distinct genotypes weighted by the number of occurrences of the genotype) by lineage was as follows: Indo-Oceanic (n=4409), *M. africanum* (n=123), *M. bovis*(n=583), Euro-American (n=20965), East-Asian (Beijing) (n=4188), and EAI (CAS) (n=1214).

Three additional datasets comprising MIRU data collected as part of independent studies were used to test the model constructed using the CDC dataset. There exists some overlap of genotypes between the CDC dataset and the three other datasets. The first dataset contains 163 MIRU types from MIRU-VNTR*plus* with 16 Indo-Oceanic, 29 *M. africanum*, 11 *M. bovis*, 87 Euro-American, 10 East-Asian (Beijing), and 10 EAI (CAS) MIRU types. The dataset from the NYSDOH consists of 668 MIRU isolates with 68 Indo-Oceanic, 5 *M. africanum*, 7 *M. bovis*, 435 Euro-American, 57 East-Asian (Beijing), and 43 EAI (CAS) records. Finally, the Brussels dataset as described in [14] contains 27 Indo-Oceanic, 13 *M. africanum*, 17 *M. bovis*, 333 Euro-American, 15 East-Asian (Beijing) and 30 EAI (CAS) MIRU types.

## 4. COMPUTATIONAL RESULTS

Two types of computational studies were performed. The first examined the overall accuracy of the model on the CDC dataset. The second study examined the predictive accuracy of the BN modeling method in out-of-sample testing.

### A. Overall Accuracy

The model was trained using the entire CDC dataset. Table 4.1 shows the confusion matrix for the model trained on the CDC dataset. The diagonal elements represent the number of strains predicted correctly for each class. Note that the total number of strains is reported (i.e. each distinct genotype is weighted by the number of occurrences). The sensitivity for each lineage as a metric for accuracy of the model is given in the first column of Table 1. A sensitivity rate of over 98% is obtained for all classes except EAI (CAS) which obtains 93%. The specificity for all lineages is quite high with that of the relatively rare *M. africanum* being slightly lower due. The errors in East-African-Indian (CAS) are primarily due to confusion with East-Asian (Beijing) resulting from the close similarity of the MIRU associated with these strains. We discuss, in Section 4.3, the source of this confusion using MIRU signatures. Spoligotypes can help more clearly differentiate these two modern strains.

### B. Predictive Accuracy

As reported in Table 2, we performed out-of-sample testing to examine the predictive accuracy of the model. First we applied the BN created on the CDC data as reported in section 4.1 to the MIRU-VNTR*plus*, NYSDOH, and Brussels datasets. In addition, the table presents predictive accuracy as on the CDC data as determined by 10-fold cross validation of the distinct

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

MIRU genotypes. Thus the 10-fold training and testing sets are disjoint. The BN performed well on all datasets with results similar to those observed in Section 4.1. The relatively few errors correspond to confusion of *M. africanum* with Indo-Oceanic and EAI (CAS).

### C. MIRU Lineage Signatures

We also studied the distribution of each MIRU locus in each lineage and provided the probability distribution map of each locus per lineage as shown in Figure 2. We observe that the numbers of repeats at a given loci for a lineage tend to take values that lie in a close range. Distinct patterns of MIRU loci distributions were found for each lineage but it is difficult to capture these patterns in simple rules or decision trees [5]. Probability-based models such as the proposed BN can do a better job of capturing the lineages than rules.

The existence of a broad pattern within a lineage and the significant difference in patterns across lineages might explain the success behind nearest-neighbor approaches [2] as well as this model. However, the nearest neighbor approach used in MIRU-VNTRplus involves selecting a suitable distance measure and cut-off. Changing the distance cut-off value yields varying results – a large value reduces the effect of erroneous or irrelevant values of markers, but results in multiple matches, thus making boundaries between classes less distinct. In contrast, the BN determines the probability of lineage of the strain without tuning or parameter choices.

Note that East-Asian (Beijing) and East-African-Indian (CAS) have similar MIRU probability signatures differing primarily only in two loci (10 and 16). This helps explain the misclassifications of East-African-Indian (CAS) as East-Asian (Beijing). The greater discriminatory power of 24-loci MIRU will help further resolve the difference between these two lineages. This contrasts with spoligotyping, where the patterns for East-Asian (Beijing) and East-African-Indian (CAS) are quite distinct. One can clearly see how MIRU24 discriminates between the ancestral and modern strains with high probability. But there are rare exceptions where MIRU24 does not discriminate between ancestral and modern strains. Thus the hierarchical BN has problem differentiating these strains at times.

## 5. CONCLUSIONS AND FUTURE WORK

We have created a model using BN to accurately predict the major lineages of strains of MTBC using MIRU alone. The assumption that the loci are conditionally independent given the class makes the model immediately applicable to 24-loci MIRU which became the new United States standard in 2009. Spoligotypes can easily be added to the model by employing the strategies used in SPOTCLUST, a BN for lineage identification using spoligotypes [6].

The hierarchical BN developed here can be further improved. The current model assumes that MIRU24 corresponds to ancestral and modern strains and the remaining loci are independent. Other BN structures could be explored to further improve performance. Performance on other types of classifiers may also be tested and results compared to determine the optimum method for lineage identification given MIRU data.

The BN model is accurate, fast, and easy to train and use. An additional advantage of classifying strains using MIRU is that results may be compared with lineages determined using models that employ other biomarkers. Any discrepancies will help flag quality control problems in the genotyping laboratories.

Future work will involve using additional MIRU loci and other biomarkers such as spoligotype in the model. This should further improve accuracy. Present methods that use spoligotype data involve matching patterns with expert-defined signatures. However, strains belonging to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

different lineages have been shown to exhibit similar patterns involving certain spacers indicating the occurrence of convergent evolution. Thus spoligotype signatures alone are not entirely reliable to classify strains into lineages.

In addition, the model may be applied to classifying strains into sub-lineages and understanding variability of MIRU genotypes indicative of recent transmission.

## Acknowledgments

This work was made possible by and with the great assistance of Dr. Lauren Cowan and Dr. Jeff Driscoll of the Centers for Disease Control and Prevention and Dr. Philip Supply of the Institute Pasteur de Lille. This work was supported by NIH R01LM009731.

## REFERENCES

1. Hirsh AE, Tsolaki A, DeReimer K, Feldman M, Small P. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc. Natl. Acad. Sci 2004;101:4871–4876. [PubMed: 15041743]
2. Allix-Beguec C, Harmsen D, Weniger T, Supply P. Evaluation and strategy for use of MIRU-VNTR*plus*, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J. of Clinical Microbiology 2008:2692–2699.
3. Sola C, Filliol I, Mokrousov I, Vincent V, Rastogi N. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. Emerging Infectious Diseases 2001;3:390–396. [PubMed: 11384514]
4. Streicher EM, et al. Spoligotype signatures in the *Mycobacterium tuberculosis* Complex. J. of Clinical Microbiology 2007;45:237–240.
5. Filliol I, et al. Snapshots of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J. of Clinical Microbiology 2003;41:1963–1970.
6. Vitol I, Driscoll J, Kreiswirth B, Kurepina N, Bennett K. Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. Emerging Infectectious Diseases 2006;6:491–504.
7. Brudey K, Driscoll JR, Rigouts L. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiology 2006;6:1–23. [PubMed: 16401340]
8. Supply P, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology 2006;44:4498–4510. [PubMed: 17005759]
9. Ferdinand S, Valetudi G, Sola C, Rastogi N. Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. Research in Microbiology 2004;8:647–654. [PubMed: 15380552]
10. Gagneux S, Small P. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infectious Disease 2007;7:328–337.
11. Sun Y, et al. Use of my cobacterial interspersed repetitive unit variable number tandem repeat typing to examine genetic diversity of *Mycobacterium tuberculosis* in Singapore. J. of Clinical Microbiology 2004:1986–1993.
12. Sun Y, et al. Characterization of ancestral *Mycobacterium tuberculosis* by multiple genetic markers and proposal of genotyping strategy. J. of Clinical Microbiology 2004:5058–5064.
13. WHO Report 2009. Global tuberculosis control epidemiology, strategy, financing.
14. Allix-Beguec C, Fauville-Dufaux M, Supply P. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. J. of Clinical Microbiology 2008:1398–1406.
15. Supply P, et al. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Molecular Microbiology 2000:762–771. [PubMed: 10844663]

16. Mazars E, et al. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. P. Proc. Natl. Acad. Sci 2001:1901–1906. [PubMed: 11172048]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

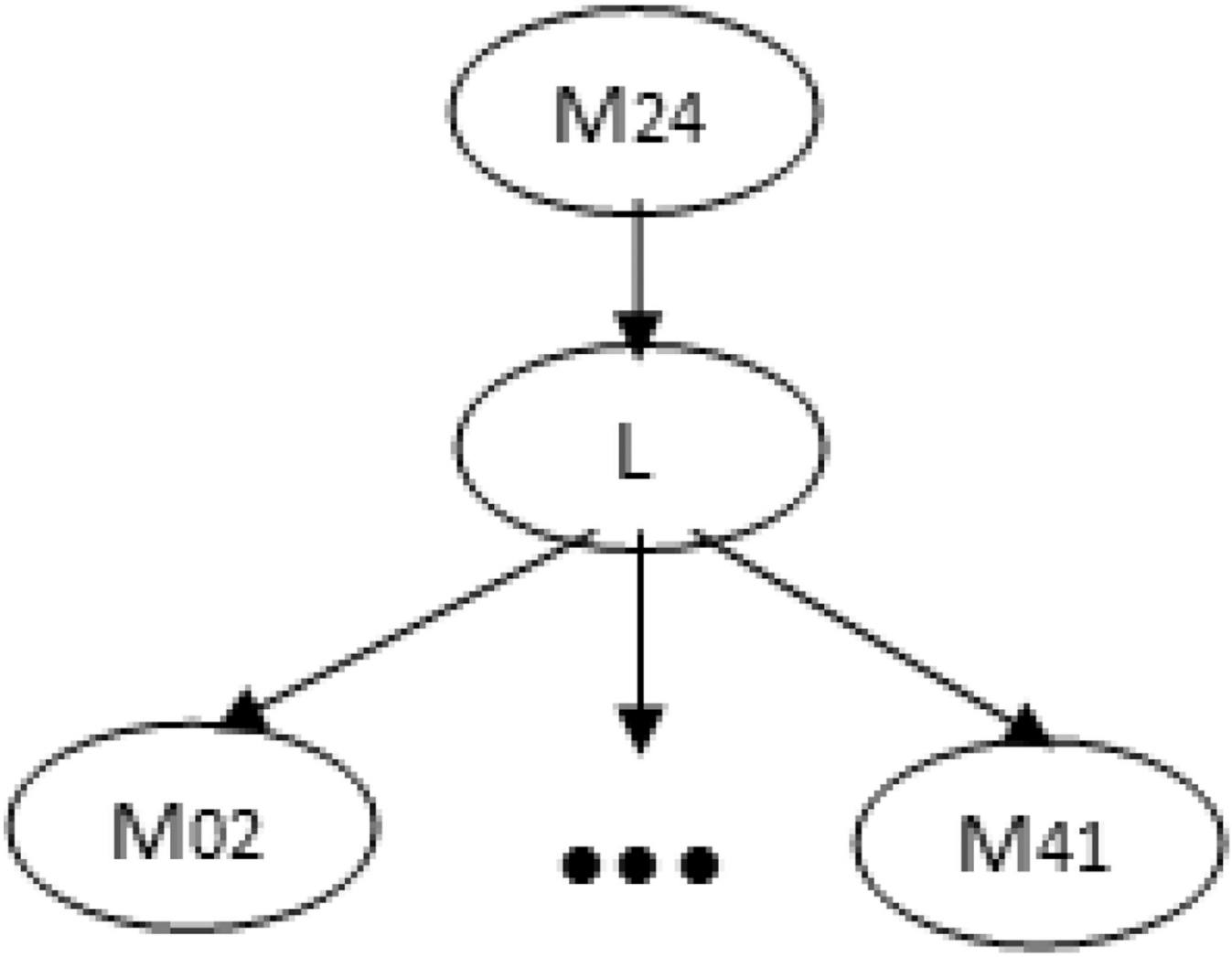


**Figure 1.**
Bayesian network used for lineage classification via MIRU

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

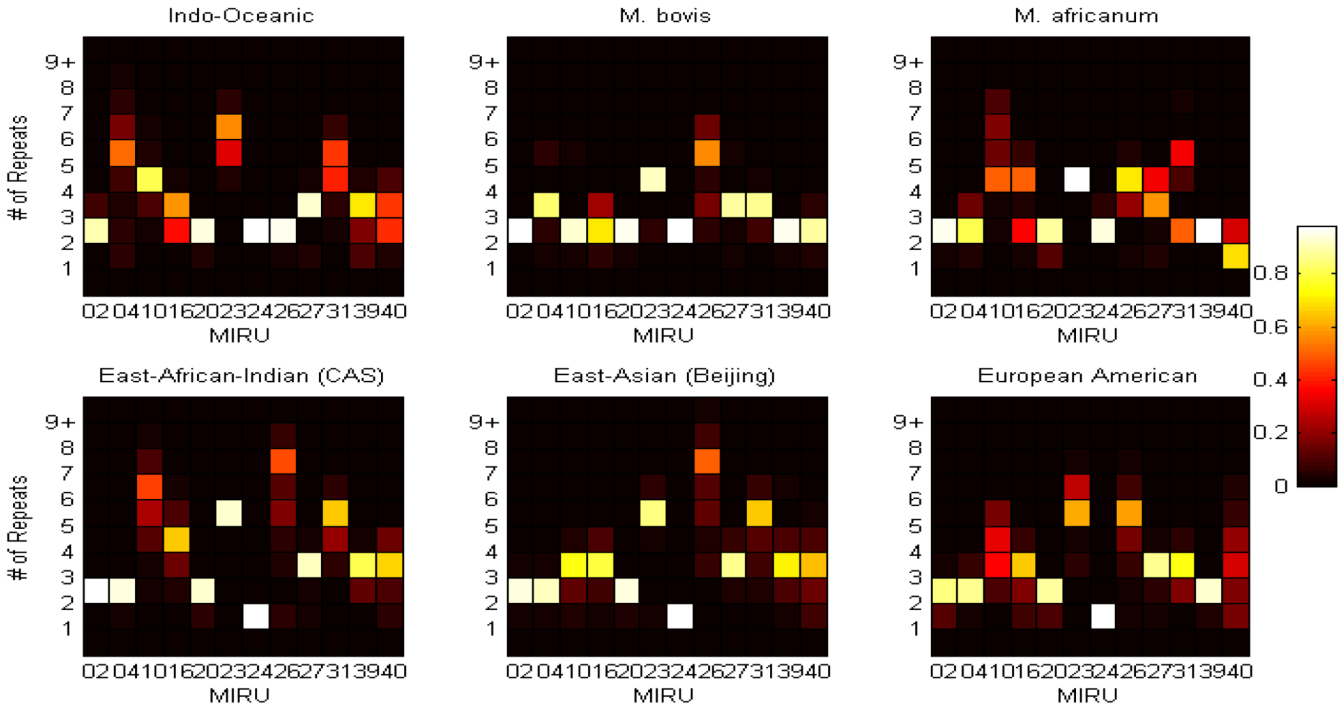


**Figure 2.**

Heat Map indicating probability distribution of the number of tandem repeats at the 12 loci of MIRU for each lineage. X axis: MIRU loci, Y axis: Number of tandem repeats. The bar alongside shows the probability values ranging from 0 (black) to 1 (white)

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript



**TABLE 1**

Overall Accuracy of BN on CDC Data

| Sensitivity % | Predicted Lineage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Indo-Oceanic** | ***M. africanum*** | ***M. Bovis*** | **Euro-American** | **East-Asian** | **EAI (CAS)** | | |
| 99.6 | 4393 | 8 | 1 | 3 | 2 | 2 | **Indo-Oceanic** | **Actual Lineage** |
| 98.4 | 0 | 121 | 0 | 2 | 0 | 0 | ***M. africanum*** | |
| 99.9 | 0 | 2 | 577 | 4 | 0 | 0 | ***M. bovis*** | |
| 99.7 | 8 | 6 | 5 | 20894 | 20 | 30 | **Euro-American** | |
| 98.2 | 1 | 0 | 0 | 52 | 4113 | 22 | **East-Asian** | |
| 93.0 | 1 | 0 | 0 | 16 | 71 | 1125 | **EAI (CAS)** | |
| | 99.6 | 88.3 | 99.0 | 99.6 | 97.8 | 95.4 | **Specificity%** | |

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript



**TABLE 2**

Predictive Accuracy of BN. Testing accuracy of CDC model on NYState, MIRU-VNTR*plus*, and Brussels datasets, and 10-fold cross validation accuracy on CDC data. Sensitivity and (specificity) are given.

| Lineage Accuracy (Specificity) (%) | | | | | | |
|---|---|---|---|---|---|---|
| | **Indo-Oceanic** | ***M. africanum*** | ***M. bovis*** | **Euro-American** | **East-Asian (Beijing)** | **EAI (CAS)** |
| **NYSDOH** | 100 (100) | 100 (100) | 100 (100) | 99.8 (100) | 100 (97) | 95.4(100) |
| **MIRU-VNTR*plus*** | 81.3 (100) | 100 (93.5) | 90.9 (100) | 100 (95.6) | 100 (100) | 80 (100) |
| **Brussels** | 100 (100) | 92.8 (100) | 94.4 (100) | 98.2 (99.4) | 93.8 (79) | 96.7(90.6) |
| **CDC (10-fold)** | 99.3 (100) | 94.1 (96.1) | 100 (99.6) | 99.6 (99.6) | 97.4 (97) | 93.1 (96) |