Infection, Genetics and Evolution 12 (2012) 767–781

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

Web tools for molecular epidemiology of tuberculosis

Amina Shabbeer, Cagri Ozcaglar, Bülent Yener, Kristin P Bennett *

Departments of Mathematical Science and Computer Science, Rensselaer Polytechnic Institute, Troy, NY-12180, United States

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form 14 August 2011
Accepted 19 August 2011
Available online 28 August 2011

Keywords:
Molecular epidemiology
Tuberculosis
Spoligotype
MIRU-VNTR
Genomic databases
MTBC classification

ABSTRACT

In this study we explore publicly available web tools designed to use molecular epidemiological data to extract information that can be employed for the effective tracking and control of tuberculosis (TB). The application of molecular methods for the epidemiology of TB complement traditional approaches used in public health. DNA fingerprinting methods are now routinely employed in TB surveillance programs and are primarily used to detect recent transmissions and in outbreak investigations. Here we present web tools that facilitate systematic analysis of *Mycobacterium tuberculosis* complex (MTBC) genotype information and provide a view of the genetic diversity in the MTBC population. These tools help answer questions about the characteristics of MTBC strains, such as their pathogenicity, virulence, immunogenicity, transmissibility, drug-resistance profiles and host-pathogen associativity. They provide an integrated platform for researchers to use molecular epidemiological data to address current challenges in the understanding of TB dynamics and the characteristics of MTBC.

1. Introduction

Over the past two decades, the development of methods for the molecular epidemiology of tuberculosis (TB) have helped create a better understanding of this disease and its causative agent, *Mycobacterium tuberculosis* complex (MTBC). DNA fingerprinting methods such as spoligotyping, Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR) typing and IS6110 restriction fragment length polymorphism (RFLP) typing have provided insights into the genetic diversity of the population structure of the MTBC (Mathema et al., 2006). Primarily, these typing methods aid traditional epidemiological approaches to detect unsuspected transmission links, thus addressing the shortcomings of standard contact tracing methods in identifying transmission events. Since epidemiologically-linked patients have MTBC isolates with identical fingerprints, the fingerprint can serve as a basic tool to distinguish between reactivation of latent infections and recent transmissions and in identifying chains of transmissions (CDC, 2011). Additionally, DNA fingerprint data have been useful in population-based studies and have helped develop a deeper understanding of the disease dynamics. There is great potential in further insights that can be created using routinely collected genotype information.

In this study, we explore available web-based tools that may be applied to existing molecular epidemiologic data to address current challenges in TB research. A summary of tools surveyed in this paper are presented in Table 1 and in the companion website at http://tbinsight.cs.rpi.edu/molepisurvey.html. Throughout this paper, we utilize the surveillance data obtained from the New York State Department of Health (henceforth referenced as NYS), comprised of spoligotype and MIRU type information of MTBC strains from patients diagnosed during the period 2004–07. The NYS dataset is comprised of 674 isolates: 268 distinct spoligotypes, 361 distinct MIRU types and 500 distinct RFLP patterns. This genotype information augmented with expert-assigned major lineage labels is used to explore and test the various tools presented.

In Section 2, we provide some background of the molecular methods utilized in the epidemiology of TB. In subsequent sections, we present tools that can be categorized as follows: databases, transmission and mutation models, classification tools, and visualization tools.

In Section 3, we explore available DNA fingerprint databases that help explore the genetic diversity and bio-geographic distribution of MTBC strains worldwide, and explore potential applications of these data. We also list some databases that investigate MTBC at the detailed genomic level. These databases provide a platform for researchers to share their data, and analyze their results in conjunction with data from other studies.

In Section 4, we look at mathematical models of the transmission and mutation of MTBC strains that use DNA fingerprint information to characterize TB dynamics. We explore the application of these models in detecting potential outbreaks.

In Section 5, we analyze various classification models. Phylogenetic analyses have shown that MTBC strains may be classified into

* Corresponding author. Tel.: +1 518 276 6899; fax: +1 518 276 4824.
E-mail addresses: shabba@cs.rpi.edu (A. Shabbeer), ozcagc2@cs.rpi.edu (C. Ozcaglar), yener@cs.rpi.edu (B. Yener), bennek@rpi.edu (K.P. Bennett).

1567-1348/$ - see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.meegid.2011.08.019

Table 1
Summary of the functionality and features of the web tools surveyed. The tools help analyze the genetic diversity of MTBC strains and characterize TB dynamics using molecular epidemiological data.

Category	Tool[a]	Features
Databases	MIRU-VNTR*plus*	*Highly curated, multi-marker database* ∗Provides similarity search or tree based analysis tools for strain identification ∗Provides tools for comparison with reference strains ∗Provides standardized nomenclature for MIRU types
	SITVIT	*Largest international collection of spoligotype and MIRU (>40k strains)* ∗Represents global distribution of spoligotypes ∗Facilitates search for occurrences of strains of interest in database ∗Provides standardized nomenclature for spoligotypes (SIT) and MIRU types(VIT)
	TBDB	*Central repository for annotated genome sequence data, gene expression data* ∗Provides access to analysis and visualization tools ∗Facilitates discovery of new drug targets, vaccine antigens, diagnostics
Transmission & Mutation models	spolTools: Spoligoforests	*Visualize evolutionary relationships between strains* ∗Produces visualization that takes into account domain knowledge about mutation process e.g. Probability of length of deletion follows Zipf distribution
	spolTools: DESTUS	*Detect emerging strains* ∗Identifies emerging strains as those with number of observed mutations significantly less than predicted (for a given cluster size and estimated ratio of transmission to mutation rate in a dataset) ∗Makes inferences based on spoligotype mutation model ∗Helps predict outbreaks, focus control efforts
Classification	TB-Insight:Rules	*Classify MTBC strains into major lineages based on expert-defined rules augmented with bioinformatic approach* ∗Automated rule based system for classification of MTBC strains using spoligotype and MIRU24 locus ∗Uses Bayes Net to predict modern/ancestral when MIRU24 locus unavailable ∗Visualize distribution of lineages identified in dataset using spoligoforests
	TB-Insight: CBN	*Classify MTBC strains into major lineages using Bayesian Network* ∗Uses different blends of biomarkers for lineage-prediction based on availability ∗Provides confidence in prediction (probability of strain belonging to class) ∗Exploits known properties of biomarkers
	TB-Insight: SPOTCLUST	*Classify MTBC strains into sublineages based on mixture model* ∗Takes into account domain knowledge about spoligotype mutation process ∗Provides confidence in prediction (probability of strain belonging to class) ∗Mixture model verifies clusters used in SpolDB3
Visualization	TB-Insight: Spoligoforests	*Visualization of genetic diversity in MTBC population and distribution by lineage* ∗Depicts genetic relatedness of strains based on spoligotype and MIRU patterns ∗Colors nodes by lineage
	TB-Insight: Host-pathogen treemaps	*Visual representation of associations between patient and strain groups.* ∗Helps identify trends and spot anomalies in associations between patient characteristics and genotype of strains ∗Helps focus control efforts

[a] URLs listed under the Websites section.

related genetic groups using various biomarkers. We look at some tools that can classify strains efficiently using only the DNA fingerprint, and will help in the investigation of phenotypic characteristics shared by strains within each lineage.

In Section 6, we cover visualization methods that represent surveillance data in ways that help study the diversity in strain and host populations. These can reveal unobserved epi-links and help identify typical as well as anomalous associations between strain and host groups.

2. DNA fingerprinting methods

In this section, we present a brief description of current methods used for MTBC genotyping that are referenced in this survey. Although, earlier studies found negligible genetic diversity between MTBC strains (Frothingham et al., 1994; Sreevatsan et al., 1997; Musser et al., 2000), the advent of molecular epidemiology has revealed considerable inter-strain diversity. We discuss potential applications of such methods to answer some of the questions facing TB researchers today. A more detailed discussion of the methods can be found at (Van Soolingen, 2001; Barnes and Cave, 2003; Mathema et al., 2006).

2.1. IS6110 restriction fragment length polymorphism (RFLP)

This method is a Southern blot hybridization technique and is the gold standard for molecular epidemiology of MTBC strains (van Embden et al., 1993). Strains are typed based on the copy number of IS6110 insertion sequences and the variability in the positions of these sequences. Strains isolated from epidemiologically linked patients are believed to have identical RFLP patterns and hence can be used to identify/verify clustered cases. The method has high discriminative power for strains with copy number greater than six. It was shown that this method can be used to distinguish closely related strains, but is not suitable to study the evolutionary history of strains, since the order of events cannot be inferred (Fang et al., 1998). The data format makes it difficult to compare results between labs (Mathema et al., 2006). Moreover, this method involves a complicated and time-consuming process requiring sub-culturing of isolates to obtain sufficient quantities of DNA. Other PCR-based

techniques have been developed to overcome these disadvantages.

2.2. Spoligotyping

Spoligotyping is a PCR-based reverse hybridization technique for MTBC genotyping (Kamerbeek et al., 1997). It is a frequently used genotyping tool, and performed on all newly identified culture positive case of tuberculosis in the United States (US). The method exploits the polymorphisms in the direct repeat (DR) locus to distinguish strains. The DR locus contains 36 bp repeats interspersed with non-repetitive short sequences called spacers. Spoligotypes are represented as a 43-bit long binary string, constructed on the basis of presence or absence of these spacers. The portable data format facilitates easy inter-laboratory comparisons. The loss of spacers can occur due to homologous recombination, or due to the transposition of IS6110 insertion sequences (Fang et al., 1998; Legrand et al., 2001). Since such mutations by the loss of spacers in the DR region are irreversible; spoligotypes can be used in the construction of evolutionary history. However, some caution must be used when using spoligotyping to study the evolution of strains, as convergent evolution of phylogenetically unrelated strains to a common spoligotype has been observed (Warren et al., 2002).

2.3. MIRU-VNTR analysis

Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats based analysis exploits the polymorphisms observed in a selected number of 41 identified mini-satellite like regions distributed on the chromosome of MTBC. MIRU typing is based on the number of repeats observed at each of the 12, 15 or 24 selected MIRU loci determined using a PCR-based method. Each locus was found to differ in its discriminative power and in its variability in alleles (Cowan et al., 2002; Aminian et al., 2010). Twelve or ideally 15 selected loci can be used for genotyping for epidemiological discrimination of strains (Supply et al., 2006). It was determined that optimally 24 loci should be used to capture the genetic diversity of strains for phylogenetic studies(Supply et al., 2006). A comparative analyses of typing methods has shown that the discrimination power of MIRU typing is higher than spoligotyping, and only slightly lower than IS6110-RFLP (Supply et al., 2001). The present CDC standard requires spoligotyping and MIRU typing to be performed on isolates from every TB case identified in the US.

2.4. Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNP) that characterize strains of MTBC have been identified. Non-synonymous polymorphisms (nsSNP) are changes that alter the polypeptide sequence and could possibly provide selective advantage to strains. The nsSNP are especially useful in understanding the acquisition and spread of drug-resistance in strains, even in strains that share the same DNA fingerprint (Niemann et al., 2009; Reed et al., 2004; Constant et al., 2002). On the other hand, synonymous SNPs (sSNP), silent mutations that are considered to be functionally neutral, are useful in phylogenetic analyses to study the evolutionary relationships between strains (Sreevatsan et al., 1997; Filliol et al., 2006; Baker et al., 2004; Gutacker et al., 2006; Gutacker et al., 2002).

2.5. Long sequence polymorphisms

Long sequence polymorphisms (LSP) or regions of deletions (RD) play an important role in phylogenetic analyses and studies of inter-strain diversity. Specific sequences occurring in progenitor strains are preserved by all strains that have evolved from it (Gagneux et al., 2006; Flores et al., 2007). Studies have identified RDs that characterize MTBC lineages. e.g. all ancestral strains have the TbD1 sequence conserved, while it is absent in all modern strains (Brosch et al., 2002). This highly clonal structure of the MTBC population is a result of the lack of horizontal gene exchange (Brosch et al., 2002; Hirsh et al., 2004). Studies have also found associations between LSPs and the pathogenicity, drug-resistance and virulence of strains e.g. RD1 (Zhu et al., 2009; Ernst et al., 2007).

3. DNA fingerprint and genomic databases

Centralized repositories of genetic data related to MTBC have been constructed from epidemiological and scientific studies conducted worldwide. DNA fingerprint databases, obtained by the aggregation of surveillance data collected in several countries, provide a view of the bio-geographic diversity in MTBC strains. These databases facilitate the introduction of standardized nomenclature for scientific communication. They also provide a means for performing comparisons between different epidemiological studies and identifying possible anomalies. Strain comparisons based on genetic distances also provide a means of evaluating the genetic relatedness of strains and inferring phylogenies or potential evolutionary relationships between strains. The focus of this survey is mainly on publicly available tools that use methods from molecular epidemiology for tracking and control, and extensions of these methods to understand the structure of the MTBC population and the evolution of strains. However, we also list some detailed genomic databases that provide an integrated platform for the analysis of genomic and experimental data that will help TB researchers in the design and development of new drugs, vaccines and biomarkers. These database tools also provide access to a suite of software for the comparative analysis, interpretation and visualization of the data. In the following section, we look at these database driven tools in further detail.

3.1. MIRU-VNTRplus

MIRU-VNTR*plus* is a highly curated database comprising of the detailed profiles of 186 strains representing all MTBC lineages (Allix-Beguec et al., 2008). The website provides access to the LSP, SNP, IS6110 RFLP fingerprint, spoligotype, 24-locus MIRU profile, and drug resistance profile of these reference strains as well as species, lineage, and epidemiologic information pertaining to these strains. The website provides tools that facilitate analysis of user strains with respect to these reference strains.

- *Input*: User strains comprising of one or more of the following fields in specified format: (i) MIRU-VNTR type (ii) Spoligotype (iii) RD (iv) SNP (v) Susceptibility.
- *Functionality*:
  - Compare user strains with reference strains.
  - Identify strain lineages by similarity search.
  - Construct phylogenetic trees using neighbour-joining (NJ) or Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithms.
  - Helps establish universal nomenclature by facilitating assignment and querying of MtbC15-9 codes.
- URL:http://www.miru-vntrplus.org.

The identification program assigns species labels, lineage labels, SpolDB4 lineage labels (Brudey et al., 2006) and RD Gagneux's lineage labels (Gagneux et al., 2006) to user uploaded strains. Identification by similarity search is a best-match approach which employs a nearest-neighbor or parsimony-based analysis to find

reference strains whose genetic profiles most closely match the user strains. Genetic relatedness between strains is determined based on one of the following genetic distance measures provided.

- Categorical Distance (*default for all markers*): This is the normalized sum of the number of markers that have different alleles.
- Chord distance D_C *(MIRU-VNTR only)*: This is a geometric interpretation based on the "angular distance" between the two strain groups determined as the sum of square root of frequencies of each allele observed in both strain groups (Cavalli-Sforza and Edwards, 1967).
  The following two distance measures are especially applicable for tandem repeat loci following a stepwise mutation model (SMM), such as the MIRU-VNTR. The SMM assumes that at each mutation step, microsatellites only gain or lose one repeat.
- $(\delta_\mu)^2$: This is the average of the difference in repeat numbers of alleles at all loci (Goldstein et al., 1995).
- D_{SW} (stepwise weighted distance): This measures the probability that two alleles are different for two different strain groups, weighted by the absolute value of the difference in the number of repeats for the two strains. (Shriver et al., 1995).

Lineage labels may be assigned to a strain if there exists one or more matching reference strains exist that are within the user-defined distance cut-off. The cut-off specifies the accepted tolerance in finding the closest match. This choice needs to balance a trade-off between choosing a large value to reduce the effect of noise, such as erroneous or irrelevant markers, and a small value, for higher specificity in lineage identification. Additionally, better sensitivity and specificity were reported when multiple markers are used in conjunction as opposed to a single biomarker. These genetic distances (normalized over the number of loci in each biomarker) may be weighted based on user discretion. The tool provides default values based on heuristics to guide the user to getting best results in strain identification.

Users can construct a phylogenetic tree to inspect the genetic relatedness of strains using the neighbor-joining (NJ) algorithm (recommended by (Allix-Beguec et al., 2008)) or the Unweighted Pair Group Method with Arithmetic Means (UPGMA). Single or multiple markers may be used to determine the distance between strains, as with the similarity search. Reference database strains can optionally be included in the tree calculation. The tool guidelines recommend that the tree be rerooted using an*M. cannettii* strain. Strains may be colored based on user input or automatically using fields such as species or lineage. The tree-based identification feature can also be used to refine labels assigned by similarity search. The phylogenetic tree can be viewed and saved in various file formats, and may also be embedded alongside the strain information. The *Calculate-Tree* feature helps view the inferred evolutionary relationships between strains in the dataset based on various distance measures as well as additional information incorporated from the reference dataset. In the following subsection, we illustrate the functionality of this tool by analyzing genotype data in the NYS dataset.

Analysis of NYS dataset using MIRU-VNTRplus: 500 isolates genotyped by spoligotyping and MIRU typing from the NYS dataset were uploaded, and 435 of these were assigned labels using the similarity search tool. Fig. 1 represents a tree in radial format created from this dataset using the NJ algorithm. In most cases, the strains that were not labeled by similarity search, can be easily identified based on the other user and reference strains belonging to the same subtree in the phylogenetic tree. From Fig. 1 it can be seen that the Euro-American lineage comprising of sub-lineages such as LAM (77 strains), Haarlem (101 strains) and X (57 strains) are the most prevalent in the NYS dataset. This is in accordance with the previously observed stable associations between host and MTBC populations (Hirsh et al., 2004), and hypotheses of host-pathogen co-evolution (Hershberg et al., 2008).

3.2. SITVIT

The international spoligotype database SITVIT (Sola et al., 2001; Filliol et al., 2002; Filliol et al., 2003; Brudey et al., 2006) represent the bio-geographic distribution of spoligotypes in the MTBC population worldwide. The most recently published version of the database, SpolDB4, is comprised of 39, 295 isolates collected from more than 120 countries and contains 1939 distinct strains or shared types (STs).

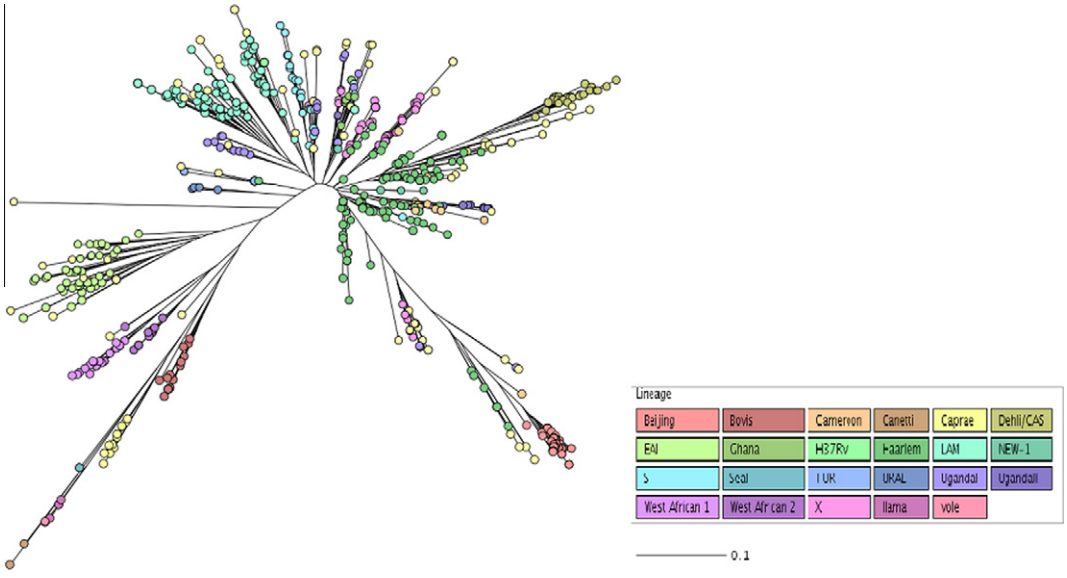

Fig. 1. NJ tree in radial format created from the NYS dataset using the tree-based identification tool on MIRU-VNTR*plus*.org. Lineage labels were assigned using similarity search, followed by tree-based analysis. The strains were assigned colors based on their lineage using the options available on the Calculate Tree tool. The scale of the genetic distance and the colors associated with each lineage are indicated in the legend.

- *Input:*
  - User strains comprising of one or both of the following fields: (i) Spoligotype (ii) MIRU-VNTR type, or,
  - One or more of the following search criteria: (i) Isolation Country (ii) Country of Origin (iii) Investigator (iv) Year of Isolation.
- *Functionality:*
  - Browse database based on search criteria.
  - Find shared-type number (SIT) or MIRU-VNTR type number (VIT) corresponding to user specified spoligotype or MIRU type.
  - View entire database, list of all (i) Spoligotypes (SIT) (ii) MIRU-VNTR types (VIT) (iii) Countries of Isolation (iv) Countries of Origin (v) Investigators.
- URL: http://www.pasteur-guadeloupe.fr:8081/SITVITDemo/.

SITVIT is the largest publicly available, curated database of spoligotypes and MIRU types. The publication of this enormous database of previously observed strains facilitates the systematic analysis of genotype information. These databases provide a view of the frequency distribution of strains worldwide and help identify highly prevalent strains. These databases can also assist in epidemiological studies by determining if strains of interest have been previously observed. Users can query the database by applying various filters to find strains by country or year of isolation, country of origin or the investigator who reported the strains. Alternatively, users may search for a single strain type and find corresponding clinical isolates reported. SITVIT also provides an international forum for scientists to share information. Users may submit their strains to SITVIT. The database assigns shared type (ST) numbers to user strains, thus establishing a standard universal nomenclature and facilitating scientific communication. The aggregation of data from various epidemiological studies allows comparisons to be made between studies and provides a global overview of the MTBC population.

Although there is no web tool available for classification of strains into lineages, lineage labels for all 1939 strains in SpolDB4 are publicly available in the accompanying publication (Brudey et al., 2006). This study also presents characteristic examples of each lineage which provide a basis for constructing visual rules. Characteristic signatures have been identified in spoligotype sequences using data published in SpolDB4 and other spoligotype-based studies (Streicher et al., 2007). Thus, classification of strains is made possible by matching spoligotype signatures.

In addition to aiding epidemiological studies, such large databases facilitate the application of bioinformatic methods. Statistical approaches from population genetics may be applied in the development of mathematical models of the genetic variation in spoligotypes (Brudey et al., 2006). These models can help detect emerging global trends and identify anomalies in transmission patterns. Phylogenies may be constructed while taking into account the unique evolutionary mechanism of the DR region in order to make inferences about the evolutionary relationships between strains (Sola et al., 2001). Thus, the SITVIT database can be useful in epidemiological studies, population genetics, and in the inference of phylogenies.

Analysis of NYS dataset based on SITVIT: A snapshot of the global distribution of strains in the SITVIT database is represented in Fig. 2.[1] This map was constructed using strains in the NYS dataset that were also reported in SITVIT from various studies conducted worldwide. It indicates the number of clinical isolates observed in SITVIT for various genotype families in each country. The sizes of the dots are representative of the frequency of occurrence of strains. Thus, large dots represent highly prevalent lineages. The dots are colored by the SpolDB4 lineage. It can be seen that strains belonging to the predominant Euro-American lineages in the NYS dataset, such as LAM, Haarlem and X as seen in Fig. 2, are prevalent in the rest of US. As expected, another strain group reported in the NYS dataset that is prevalent in the US and across the world, are strains belonging to the highly virulent East-Asian lineage. This world map showing the global distribution of strains reported in the NYS dataset illustrates a potential application of data from SpolDB4 in providing a global view of the composition of the MTBC population.

3.3. TB Genotyping Information Management System (TB-GIMS)

TB-GIMS is a web-based system provided by the Centers for Disease Control and Prevention (CDC), to store and manage genotype data on TB patients in the US.[2] TB programs and genotyping laboratories affiliated with the CDC have access to this database. This central repository of data facilitates the storage and management of genotyping data of TB patients in the US. It enables the generation of summary statistics about clusters at the county, state and national levels. Users can also perform fine-grained analysis by linking isolates to patient-level data. Users are assigned roles that provide different access privileges. The establishment of the database facilitates efficient transmission of information regarding genotype reports and updates to participating laboratories and TB programs. TB-GIMS plays an increasingly important role in the tracking and control of TB at the national level in the US. More information can be obtained at http://www.cdc.gov/tb/programs/genotyping/tbgims/default.htm.

3.4. TBDB

TBDB is a central repository for annotated genome sequence data, as well as gene expression data (microarray and reverse transcription polymerase chain reaction (RT-PCR) data) of *M. tuberculosis* strains and some related species (Reddy et al., 2009). Such an integrated platform facilitates detailed genomic analysis, visualization, annotation and comparisons of sequences. TBDB also provides a host of analysis tools for micro-array data and enables comparisons with results from up to 1800 assays.

- Functionality:
  - *Genomic Data*:
    - Quick search based on: (i) Gene name (ii) Gene Sequence name (iii) Author name (iv) Title (v) Keyword.
    - BLAST Search based on input query sequence.
    - Feature Search based on: (i) Gene (ii) Epitope (iii) Operon (iv) Polymorphism (v) BlastX (Blast hits) and additional text arguments to filter the search.
    - Graphically view annotated sequence regions using Genome Browser tools: Argo Genome Browser, Feature Map.
    - Perform comparative analysis via Synteny Maps, Dot Plots, Circular Genome Viewer.
    - View expression correlation between genes in *M. tuberculosis* H37Rv while simultaneously viewing orthologous genes in related species using the Operon Browser. (Correlation coefficients computed using 1260 dual-dye micro-array chips).
    - Find polymorphisms in resequenced genomes of 31 selected strains representative of the MTBC population using Diversity Sequencing tools.
    - Download data e.g. sequence, genes, markers in various formats facilitating further analysis.

[1] This map drawing tool is not part of SITVIT. Map generated using Google Maps.

[2] We do not cover this system extensively as access to the system is unavailable to individuals not directly affiliated with public health departments in the US.

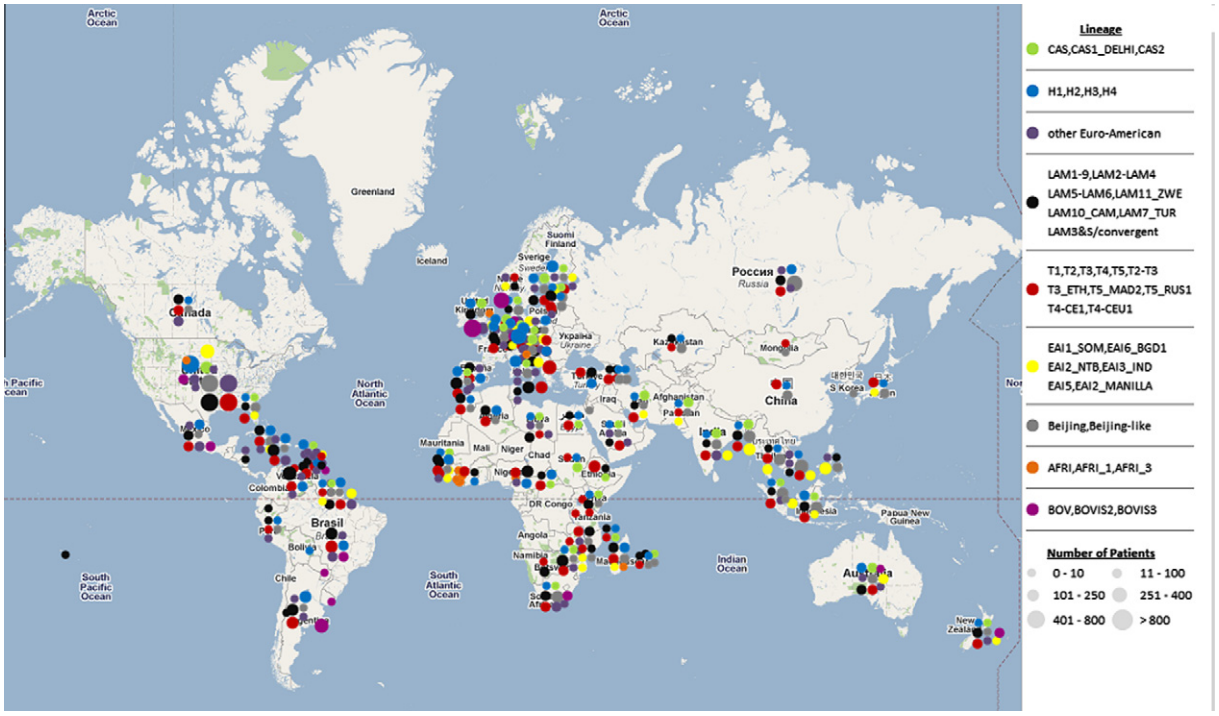

Fig. 2. Map showing the global distribution of strains from the NYS dataset. Each dot represents instances of these strains reported in the SpolDB4 database. The dots are colored by SpolDB4 lineage. The size of the dot is proportional to the number of isolates reported in SITVIT.

– *Expression Data:*
* Search for data from individual microarrays or RT-PCR assays or a publication.
* Cluster gene expression data based on experimental conditions, mutants, strains, publication.
* View gene profile for a dataset from a given publication or for a gene of interest
* Upload and save user expression data for analysis using tools such as hierarchical clustering, imputation of missing values, Gene Set Enrichment Analysis, Singular Value Decomposition and pathway analyses.

– *Protein Information*:
* View protein structure.
* Search for epitopes.

- http://tbdb.org.

As of 2011, TBDB houses thirty-one sequenced strains that represent the phylogenetic diversity of the MTBC. Gene sequence and expression data can be found by a broad level search on publication details or gene of interest, or by specific microarrays or assays. Datasets associated with the experiments conducted for the publication are available for download and analysis.

This website provides access to a host of software for genomic comparison and microarray analysis. Users may search the database for their gene of interest and view details regarding its location within the genome, neighboring genes, protein domain and observed polymorphisms. A BLAST search can be performed on the local database or with sequences in NCBI's database. Users may also analyze genomic data visually using a host of available tools with advanced functionality such as zoom and visualization of associated annotations. Fig. 3 shows the output of the Operon Browser one of the many genomic analysis tools available on TBDB. Microarray analysis tools include Expression Connection that allows users to investigate clustered microarray data for specific genes or genes that are highly correlated to the gene of interest. Users may filter or cluster raw microarray data based on their requirements using a range of methods, such as hierarchical clustering, Singular Value Decomposition, tools from the Stanford Microarray Database and those provided through the Gene Pattern software. Users are provided the option of submitting their pre-publication data to the repository, and gaining access to the tools available, as well as other datasets in the repository for comparison. Pre-publication data can be shared with specific users, or made public, with the permission of the user. All published data are made publicly accessible. Using these analyses tools will help identify new targets for drug design, vaccine antigens, biomarkers and diagnostics.

3.5. Other genomic databases

Several other databases exist that contain genomic data of MTBC strains, as well as of related species of mycobacteria. These genomic databases provide integrated access to sequence and expression data from various studies, as well as tools that facilitate the analysis of these data.

- TB Structural Genomics Consortium (TBSGC) facilitates the determination of the structure of proteins from MTBC allowing inference of protein function and characterization of drug targets (Terwilliger et al., 2003). http://www.webtb.org/.
- TubercuList contains DNA and protein sequences derived from the H37Rv strain (Cole, 1999). http://genolist.pasteur.fr/TubercuList.
- GenoMycDB facilitates large-scale comparative analysis of mycobacterial genomes and the functional classification of mycobacterial proteins. It is based on the comparative analysis of the predicted protein sequences coded by the genomes of 6 sequenced mycobacteria (Catanho et al., 2006). http://157.86.176.108/catanho/genomycdb/.

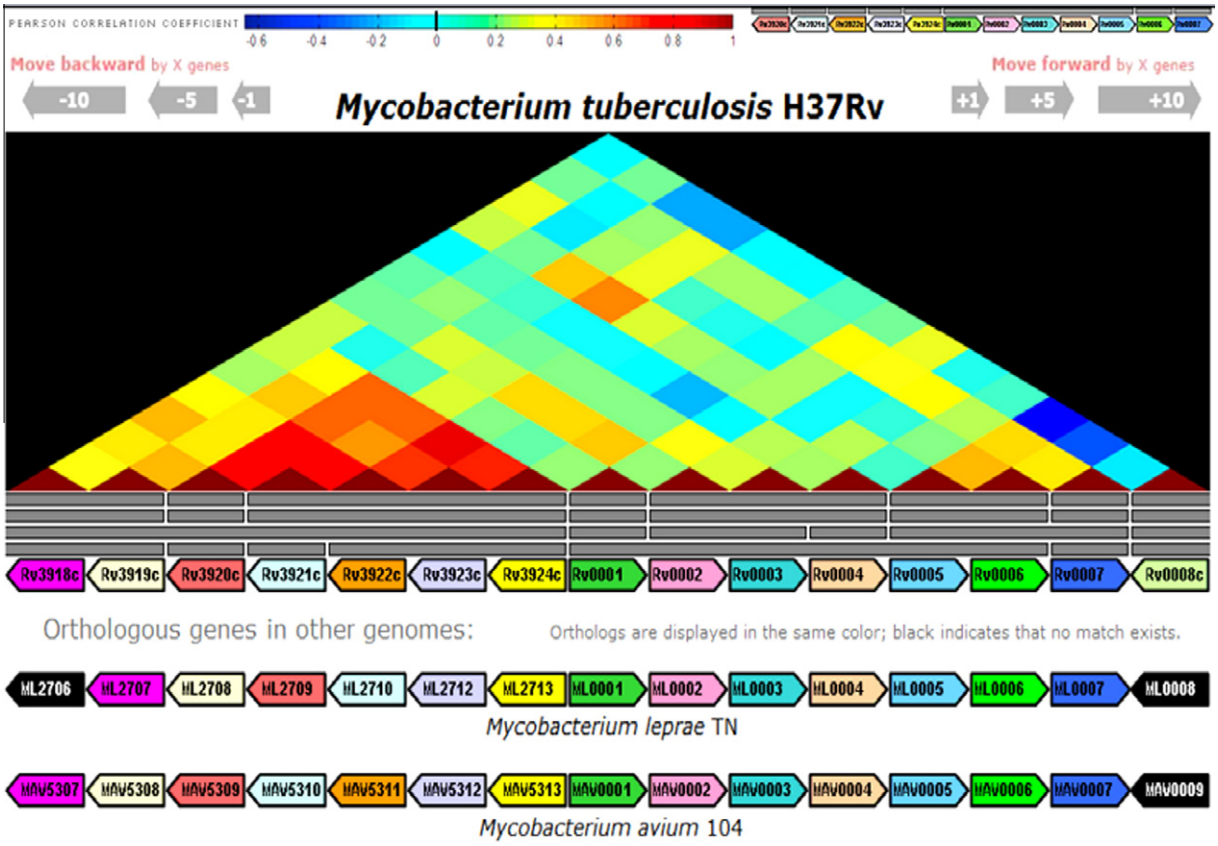

Fig. 3. Visualization of the expression correlation between genes in *Mycobacterium tuberculosis* H37Rv computed using 1260 dual-dye micro-array chips using the Operon Browser. Orthologous genes in genomes of related species are also displayed.

- MGDD: *M. tuberculosis* genome divergence database provides detailed information about genomic variations between 6 fully sequenced MTBC strains and facilitates comparisons based on SNPs (Single Nucleotide Polymorphism), indels, tandem repeats and divergent regions(Vishnoi et al., 2008). http://mirna.jnu.ac.in/mgdd/.
- MyBASE is a database developed to facilitate the study of polymorphisms and gene function for mycobacterial pathogens. It focuses on the influence of LSPs and essential genes on the pathogenicity and virulence of strains (Zhu et al., 2009). http://mybase.psych.ac.cn.
- MTbReglist houses data pertaining to the analysis of transcriptional regulation of *M. tuberculosis* strains (Jacques et al., 2005). http://www.USherbrooke.ca/vers/MtbRegList.
- TBrowse is an integrated database containing over half a million data points pertaining to genomic sequences of MTBC. It also provides interoperability with other databases by means of the Distributed Annotation System, and inbuilt access to sequence analysis tools from NCBI (Bhardwaj et al., 2009). http://tbrowse.osdd.net.

4. Transmission and mutation models

In this section, we explore some web-based tools that make inferences about TB dynamics based on MTBC transmission and mutation models using DNA fingerprint information. Mathematical modeling has provided great insight into the dynamics of tuberculosis and helps guide control efforts. Better models can be designed by exploiting the wealth of information that can be gleaned from DNA fingerprint surveillance data (Murray, 2002). A few such models that incorporate knowledge of the biomarkers and their mutation mechanisms are discussed in this section. First, we look at quantitative measures of transmission and mutation. Next, we present spoligoforests, a visualization based on cluster-graphs, that depicts evolutionary relationships between strains. We then explore some available tools that model TB transmission as a stochastic process, while taking into account various factors such as mutation rate and sampling frequency. These tools employ methods from statistics and population genetics to address the need for methods that detect outbreaks and quantify their severity, determine the rate of active transmission, and model the MTBC population dynamics.

4.1. SpolTools

SpolTools is a suite of software that provides a means of statistical and visual analysis of spoligotype data. The quantitative in-depth analysis facilitates a better understanding of the MTBC population being studied and provides valuable information to guide control efforts.

- *Input:* Spoligotype strains in binary or octal format, which is internally transformed into Rich Spoligotype Format (RSF).
- *Functionality:*
  - Generate *summary statistics* that provide a glimpse of the genetic diversity of strains in the dataset.
  - Generate *spoligoforests* to create a visual representation of possible evolutionary history of strains.

  - Determine emerging strains with corresponding q-value indicating significance of statistical tests performed (DESTUS).
- *URL:* http://www.emi.unsw.edu.au/spolTools/.

4.1.1. Summary statistics

This tool employs methods from population genetics to define summary statistics that measure the degree of *clusteredness* in the population of interest. These statistics provide a basic estimate of the extent of disease transmission. They quantify *genetic homogeneity*, a measure of the extent of recent transmission, or the *genetic diversity* which may be an indication of reactivations of latent infection. The rate of active transmission in the population can be better quantified by such approaches that incorporate information from molecular epidemiological studies. There is also a need to have easily available tools that quantify the role of other host and population specific characteristics and risk factors in determining the extent of recent transmission, such as exogenous sources of infection, characteristics and risk factors associated with the host population, the variability of strains in terms of their virulence, immunogenicity and transmissibility (Murray and Nardell, 2002; Nardell et al., 1986; Bifani et al., 2002; de Jong et al., 2008). Details of the statistics provided by this tool and their significance are presented in Table 2. The values of these indices determined on the NYS dataset listed in Table 2 indicate a large number of clustered cases implying a high rate of transmission. However, it is important to interpret these results in the light of using all available genotype information. The NYS Department of Health considers isolates with identical spoligotype, MIRU type *and* RFLP pattern as belonging to a cluster. Using this definition of a cluster the values of the indices change significantly: the average cluster size $\frac{n}{g} = 1.15$, clustering index $RTI_{n-1} = 0.13$, clustering rate $RTI_n = 0.11$ and virtual heterozygosity $H = 0.99$.

4.1.2. Spoligoforests

SpolTools creates *spoligoforests* to visualize the inferred evolutionary history of the MTBC population under investigation (Reyes et al., 2008). Spoligotype datasets are depicted as a forest of trees, in which each node represents a cluster of isolates of a spoligotype, and each directed edge represents the mutation of the parent spoligotype into the child.This representation builds on the cluster-graph approach developed in (Tanaka and Francis, 2005). It incorporates the Infinite Alleles (IA) assumption that stipulates each node can have only a single parent (since every mutation creates a new, previously unobserved genotype). For every spoligotype, a set of candidate parents are generated by performing a pairwise comparison with all other strains in the dataset. The single most likely parent spoligotype for a strain is chosen based on the edge weight. The edge-weight is a function of the frequency of occurrence of the parent strain and the probability of occurrence of the observed deletion that caused the mutation from the parent to the child spoligotype. Parameters of the model of deletion of spoligotypes were selected using Maximum Likelihood Estimation, such that they maximize the probability of the observed frequencies of deletions. The authors (Reyes et al., 2008) found the Zipf distribution to be the most parsimonious model amongst several evaluated that capture the mechanism of mutation of the DR region. Note that the frequencies of deletions used to construct the model were determined using only unambiguous deletions. The advantage of spoligoforests over traditional tree-based analysis, such as phylogenetic trees, is that they are based on prior knowledge of the mutation process of the genetic markers, rather than on genetic distances alone.

Additionally, spoligoforests, and other visualizations such as IA forests, may be used to analyze diversity in the MTBC population, infer information about the age of the strain and hence deduce transmission and mutation events (Reyes et al., 2008). Future tools can build on the foundations laid by these models. A possible direction is considering the effects of changing one or more of the assumptions made by these methods: (i) The length of the deletion is not affected by the cause of the mutation, or the specific positions in the DR region where the deletion is observed. This may not always hold as hot-spots for deletions and IS6110 insertions have been observed. (ii) The Infinite Alleles (IA) Assumption, each genotype mutation gives rise to new genotype. Convergent evolution has been observed in spoligotype sequences, albeit infrequently (Warren et al., 2002). (iii) Each strain may have descended only from one of the candidate parents observed in the dataset. This does not always hold as many of the cases of TB are due to reactivations of latent infections, possibly acquired elsewhere. Findings in (Borile et al., 2011; Shabbeer et al., 2011; Ozcaglar et al., 2011) that demonstrate the advantages of using contiguous deletions of spacers in establishing the genetic relatedness of strains may be employed in the generation of spoligoforests. Thus, spoligoforests are an efficient model for spoligotype evolution, and provide a foundation to examine MTBC genetic diversity.

Spoligoforest of NYS dataset:

Fig. 4 represents a spoligoforest generated from the NYS dataset using SpolTools. A majority of the edges are solid edges indicating that they were chosen unambiguously. There are a large number of orphan nodes, indicating that no spoligotype sequence within the dataset was a suitable candidate parent. This is quite likely, since a majority of the cases in the NYS dataset are believed to be reactivations of latent infections acquired elsewhere.

4.1.3. DESTUS

DESTUS is a tool designed to use spoligotypes to identify emerging strains by performing statistical tests to determine whether they are spreading faster than the background rate (Tanaka and Francis, 2006). Strains identified as emerging could indicate poten-

Table 2
Summary statistics determined by SpolTools for NYS dataset. Formulae for indices quantifying transmission of TB in a population and their interpretations are provided. nc = number of cases that belong to a cluster of size 2 or greater, c = number of genotypes with cluster of size 2 or greater, n = number of cases, ni = number of cases of the ith genotype, g = number of genotypes. The values of these indices computed on the NYS dataset are also shown in the table.

Formula	Interpretation	Value
$\frac{n}{g}$	Average cluster size, indicative of number of recent transmissions, as well as reactivations.	2.48
Singleton count	Number of singletons, indicative of number of reactivations of latent infections.	189
$RTI_{n-1} = \frac{n_c - g}{n} = \frac{n-g}{n-1}$	Clustering index, indicative of the number of transmissions and is the difference between the number of cases observed less the number of source cases.	0.60
$RTI_n = \frac{n_c}{n}$	Clustering rate, a measure of genetic homogeneity that indicates degree of transmission.	0.40
$H = 1 - S = 1 - \sum_{n_i} \frac{n_i(n_i-1)}{n(n-1)}$	Virtual heterozygosity, a measure of genetic diversity that represents probability that any 2 individuals chosen from the same dataset have different genotype.	0.02
ζ	Maximum likelihood estimate of the ratio of transmission rate to mutation rate, used to determine if an individual strain is transmitting at a rate higher than other strains in the dataset (computed over the entire dataset).	166.54

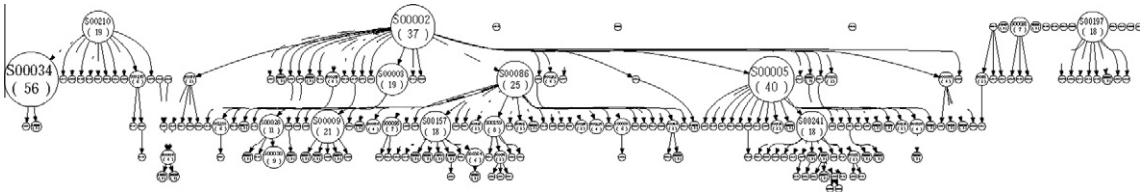

Fig. 4. Snapshot of the spoligoforest of the NYS dataset generated using SpolTools. Each node represents a cluster of strains with the same spoligotype and has size proportional to the number of isolates. Each edge represents a mutation from the *parent* spoligotype sequence into a new *child* spoligotype sequence by the loss of one or more adjacent spacers i.e. a contiguous deletion. There are 3 types of edges, represented by solid, dashed and dotted lines, that represent edges that were selected unambiguously, those with weights between 0.5 and 1, and weights less than 0.5, respectively. By the Infinite Alleles model assumption, each node has only a single incoming edge from one of its candidate parents. The most probable edge is selected based on the hypothesis that lengths of deletions in spoligotypes follow a Zipf distribution.

tial outbreaks. This information can be used to guide control efforts.

The method builds on earlier approaches that measure transmissibility on the basis of cluster size (Luciani et al., 2008). It has been observed that a large cluster size can be attributed to one of two causes: (i) The strain is transmitting rapidly or (ii) The strain has been present in the population for a long time (Tanaka and Francis, 2006). The number of inferred mutations can be used to distinguish between these two causes, since the older the strain the more mutations it has accumulated. The number of mutations a strain can accumulate was shown to be a function of its copy number (number of spacers present), mutation rate per copy and the number of cases in the history of the strain. To quantify the number of mutations observed, the authors use IA forests (Tanaka and Francis, 2005). In an IA forest, every node has a single inbound edge and the outdegree is the number of mutations observed for the corresponding strain. The strain is deemed emerging if the observed number of mutations for a particular strain is significantly less than the number predicted based on the assumption that the number of mutations accumulated follows a Poisson distribution. The probability that the number of mutations predicted (for a given IA forest and strain) by the model is less than those observed, is calculated using ζ (the rate of transmission relative to mutation over all strains in the dataset), the observed cluster size and copy number of the strain. The p-value from testing the null hypothesis that the number of mutations predicted is at most the number observed, is the sum of these probabilities over all possible forests. A low p-value implies the observed number of mutations is less than predicted and not consistent with the model, implying the strain is emerging. Multiple hypothesis tests are performed to test the significance of the result. The False discovery rate (FDR) or q-value is calculated to determine the proportion of significant tests that are falsely rejected. A low q-value hence implies the number of transmissions (cluster size) is greater than expected, given the observed number of mutations. Strains with low p-value and/or q-value that are deemed significant are found to be "emerging".

Detecting emerging strains in the NYS dataset: We tested the DESTUS programs on the NYS dataset. Table 3 shows the shared types in the NYS dataset, i.e. strains that were found to be emerging strains by the Benjamini–Hochberg and Storey tests. No strain was found to be emerging by the Dunn-Sidak test which is the most conservative of the three statistical tests performed. One of the strains that was deemed emerging belongs to the notorious Beijing lineage (strain S00034), and is known to be spreading at a rate higher than the background transmission. By these tests, strains S00197, S00091, S00615 are rapidly transmitting strains and could also be a cause for concern. However, an analysis of additional genotype information such as the MIRU types and RFLP patterns of all the 18 isolates with these spoligotypes indicates otherwise. Based on the NYS Department of Health's definition of a cluster (identical spoligotype, MIRU and RFLP patterns), there are only two clusters of size 2 associated with these spoligotypes. This indicates that the large number of cases associated with these spoligotypes can be attributed to the age of the strain, and are a result of reactivations of latent infections or injections from foreign populations, rather than recent transmissions of emerging strains. Whereas, spoligotypes S01283 and S00173 are associated with clusters of size 3 and 5 respectively, and are indeed rapidly transmitting strains. This suggests the use of additional genotype information in models in order to make more accurate inferences about the transmission rates of strains.

Table 3
Top strains in the NYS dataset identified as emerging strains by SpolTools-DESTUS based on both the Storey and Benjamini–Hochberg tests for FDR control.

Strain name	pFDR control (Storey) with cutoff q-value = 1			Benjamini–Hochberg FDR control value = 1		
	Rank	p-value	q-value	Rank	p-value	$-\log(p)$
S00034	2	.063	.999	1	.063	2.76
S00197	18	.065	.999	2	.065	2.74
S00091	10	.127	.999	5	.127	2.06
S00173	6	505	999	11	.505	.68
S00615	14	.513	.999	12	.513	.67
S01283	15	.532	.999	15	.532	.63

5. Classification tools

Classification of MTBC strains into lineages provides insight into the genetic diversity of the strains being investigated and helps identify the predominant genetic groups in a population. Further, strains associated with different lineages have been found to vary in their immunogenicity, pathogenicity, virulence, transmissibility and drug susceptibility (van der Spuy et al., 2009; Reed et al., 2009; Gagneux and Small, 2007; Gagneux et al., 2006). The observed associations between clades previously identified by phylogenetic analysis and geographical regions indicate the influence of social factors such as migration, and other host-related factors on disease dynamics (Gagneux and Small, 2007). We need methods to classify strains into these groups using only the DNA fingerprint data, collected as part of routine TB surveillance.

Although visual rules and nearest neighbor based approaches performed on highly curated databases can be used to classify strains, these do not constitute scalable solutions. Visual methods may involve considerable human effort and nearest neighbor approaches may involve a large number of pairwise comparisons. To address this problem, computational methods for lineage classification have been developed. These tools complement existing visual rules. Classification results will help recognize variations in phenotypic characteristics of lineages. In this section, we present some web-based automated tools that allow classification to be performed easily and efficiently on large datasets. Such tools can provide perspective on differences in phenotypic characteristics,

and phylogeographic associations of MTBC strains with host populations.

5.1. TB-Insight : SPOTCLUST

SPOTCLUST is a mixture-model-based approach to determining MTBC lineages of spoligotypes (Vitol et al., 2006). The MTBC lineages predicted by the tool correspond to highly cohesive genetic groups identified by the two different probability models (i) the 36-component SpolDB3 based Model (Filliol et al., 2002; Filliol et al., 2003), and (ii) Randomly Initialized Model (48 components).

- *Input:* Spoligotype strains in binary or octal format.
- *Functionality:*
  - Determine MTBC lineages of the spoligotype strains using (i) SpolDB3 model (ii) Randomly Initialized Model.Results also include the probability that the spoligotype belongs to the lineage which indicates the confidence of the model.
- *URL:* http://tbinsight.cs.rpi.edu/run_spotclust.html.

The two models are created differently, using different number of clusters k and initial class probabilities. Randomly Initialized Model (RIM) used Monte Carlo Cross Validation to determine an initial value of the number of clusters, while the SpolDB3 model is based on findings in (Filliol et al., 2002). The *families* identified by the SpolDB3 model closely match the lineages defined in SpolDB3 (Filliol et al., 2002). In both models, spoligotypes are represented as multivariate Bernoulli distributions (Vitol et al., 2006). The probability of the presence of each spacer is a parameter of the model, and each of the 43 parameters are treated as conditionally independent. Expectation Maximization (EM) was used to find maximum likelihood estimates of the model parameters: the class probabilities and the 43-variable Bernoulli distribution (the conditional probabilities of the presence of the 43 spacers given the class). Both SPOTCLUST models take into account the fact that spacers are lost but rarely gained. To do so, the models use hidden variables called 'Hidden Parents', whose 'children', the observed strains, may lose a spacer with a small but non-negligible probability. The probability of a child gaining a spacer however, is extremely small. A Naive Bayes classifier determines the most likely class label c_j for a spoligotype x, as the one with highest $P(c_j|x)$. The $P(c_j|x)$ is determined for each lineage based on the product of probabilities of presence or absence of spacers given the class (as defined by the model) while accounting for the Hidden Parent model.

Analysis of NYS dataset using SPOTCLUST: Based on the correspondence between SpolDB3 lineages and the major lineages, 99.25% of the labels assigned to the NYS dataset by SPOTCLUST were correct. A sample output for 3 strains from the NYS dataset obtained by running through SPOTCLUST is provided in Fig. 5.

5.2. TB-Insight: TB-Lineage

TB-Lineage is a set of web tools that provide automated methods to classify MTBC strains and visualize genotype datasets (Shabbeer et al., 2011; Aminian et al., 2010).

- *Input:* (i) Spoligotype strains in binary or octal format (ii) (locus MIRU24), 12, 15 or 24 loci of MIRU-VNTR (Optional)
- *Functionality:*
  - Classify strains into 6 major lineages using (i) A rule-based system that uses spoligotype and the MIRU24 locus (when available) to assign lineage labels. (ii) A Conformal Bayes Network (CBN) that uses a blend of PCR-based methods.
- *URL:* http://tbinsight.cs.rpi.edu/run_tb_lineage.html.

5.2.1. TB-Lineage: TB-Rules

TB-Rules is a rule-based classification system available as part of the TB-Insight suite of tools that classifies strains of MTBC into major lineages: the *modern* lineages *Euro-American, East-African Indian (CAS), East-Asian (Beijing)* and the *ancestral* lineages *M. africanum, M. bovis, Indo-Oceanic*. The rules were culled from prior literature on spoligotype signatures and various epidemiological studies conducted worldwide (Brudey et al., 2006; Sebban et al., 2002; Ferdinand et al., 2004; Streicher et al., 2007; Filliol et al., 2002) and refined based on expert guidelines from the Centers for Disease Control and Prevention (CDC) (Shabbeer et al., 2011). Prediction is done using the spoligotype sequence and, if available, the MIRU24. The locus MIRU24 is known to correlate with the TbD1 deletion, and therefore, can be used to characterize strains as modern or ancestral (Sun et al., 2004). However, some strains do not conform to the observed patterns in the number of repeats at MIRU24. In some cases, as with historical datasets, MIRU data are not available. For such cases, the tool employs a Naive Bayes classifier that uses spoligotype data alone to predict whether a strain is *modern* or *ancestral*. The prediction is then used as a surrogate for the MIRU24 locus in the rules. This system relies on certain signature deletions which are assumed to correspond to speciation events. Although the classification process is automated, the labels assigned by the rules are easily verifiable.

5.2.2. TB-Insight: Conformal Bayes Net

This tool employs a hierarchical Bayesian network to classify MTBC strains into the major genetic lineages using different blends of PCR-based biomarkers (Aminian et al., 2010). The design of the probability-based model exploits known properties about the structure, position and mutation mechanisms of spoligotypes and MIRUs. The MIRU loci are distributed across the genome of the MTBC but mostly away from the DR locus. The assumption of independence between the MIRU loci and between the MIRU loci and the spacers is made on the basis of this fact. The number of repeats at each MIRU locus is frequently between 0 and 9, and in some cases greater than 9. Therefore, each MIRU locus is modeled as multinomial distributions with possible values $0, 1, \ldots, 8$, and $\geqslant 9$ (Aminian et al., 2009). Spoligotypes are modeled as a multivariate Bernoulli distribution with 'Hidden Parents', as in the SPOTCLUST model (Vitol et al., 2006). The CBN model exploits the correlation between the number of repeats at locus MIRU24 and the conservation of the TbD1 deletion (Sun et al., 2004) and represents this as a causal relationship. Therefore, a top-level classification into *modern* and *ancestral* strains, characterized by the absence or presence of the TbD1 deletion respectively, is done on the basis of locus MIRU24. Thus, the CBN models the distributions of the MIRU loci and the spoligotypes of various lineages, and provides a method

			SpolDB3-based Model		Randomly initialized Model	
1						
2	ID	Spoligotype	Most probable family	Probability	Most probable family	Probability
3	12	000000000000000000000000000000000000000111111	M. tuberculosis Beijing	0.927945388	Family N3	0.935536155
4	14	111011111111111111111111111111111100001111111	M. tuberculosis T1	0.999991199	Family N24	0.999893855
5	15	111111111111111111111111110100001011111111	M. tuberculosis EAI5	0.97869962	Family N14	0.999997082

Fig. 5. Most probable family using RIM and SpolDB3 model of SPOTCLUST. The probability alongside the lineage prediction is a measure of the confidence of the prediction.

of assigning labels for strains based on the spoligotype and/or any number of MIRU loci.

The independence assumption between all loci allows for predictions to be made using any number of biomarkers depending on availability. The design of the model allows it to be trained using all available data, even if the data is incomplete, i.e. not all biomarkers are available. This model is flexible and extensible to including different types of biomarkers as they become available. Host-related characteristics and risk factors may also be included into such a model, as suggested in (Getoor et al., 2004). Hybrid approaches that incorporate a rule base into a Bayesian Network model have been developed recently that benefit from the advantages of both rule-based and probabilistic approaches (Aminian et al., 2011).

Analysis of NYS dataset using TB-Lineage: The CBN and TB-Rules were tested on the NYS dataset. All the lineage labels assigned by TB-Rules matched the labels assigned by human experts at the CDC exactly. The CBN classified 673 isolates correctly, and misclassified a single strain, ST 2, as East-Asian instead of Euro-American. This misclassification occurs since this spoligotype sequence, 000000000000000000000000100000010000111111, closely matches the signature of Beijing strains. The error can possibly be explained by the fact that the Bayesian Network uses individual spacers as features, and this does not account for the fact that one or more adjacent spacers may be deleted in a single mutation event. Such errors can potentially be avoided by using contiguous deletions as features as in (Shabbeer et al., 2011; Borile et al., 2011; Ozcaglar et al., 2011), instead of the individual spacers.

Other computational approaches have also been applied to accomplish classification of strains of MTBC. Data mining approaches have been specified in (Sebban et al., 2002; Ferdinand et al., 2004) that apply decision tree based approaches to classify strains. Feature selection was performed on the basis of information gain to identify the smallest subset of loci that need to be investigated to classify strains. In addition to web tools for classification of strains into major lineages, there is a need for tools that will perform automatic classification of strains into sub-lineages, i.e. genetic groups of finer granularity. Existing methods can be extended for this purpose by incorporating expert knowledge based on previously defined visual rules for sub-lineage classification (Brudey et al., 2006; Streicher et al., 2007).

6. Visualization tools

Visualization of public health data is emerging as a popular aid to traditional methods of epidemiology. Modeling and visualizing genetic relatedness and patterns of mutation over relatively short periods of time are crucial for epidemiological studies as they help analyze recent transmission trends. Identifying previously unrecognized epi-links and associations between patient and strain groups helps focus public health efforts in an effective manner. In this section, we look at web-based tools that address the need for such visualizations and discuss other possibilities that could help further the application of visual analytics for TB epidemiology.

6.1. Spoligoforests

- *Input:* (i) Spoligotype strains in binary or octal format (ii) 12, 15 or 24 loci of MIRU-VNTR
- *Functionality:*
  - Draw spoligoforest colored by lineage (obtained from TB-Lineage) depicting genetic diversity and relatedness of strains in dataset.
- *URL:* http://tbinsight.cs.rpi.edu/run_tb_vis/spoligoforests.html.

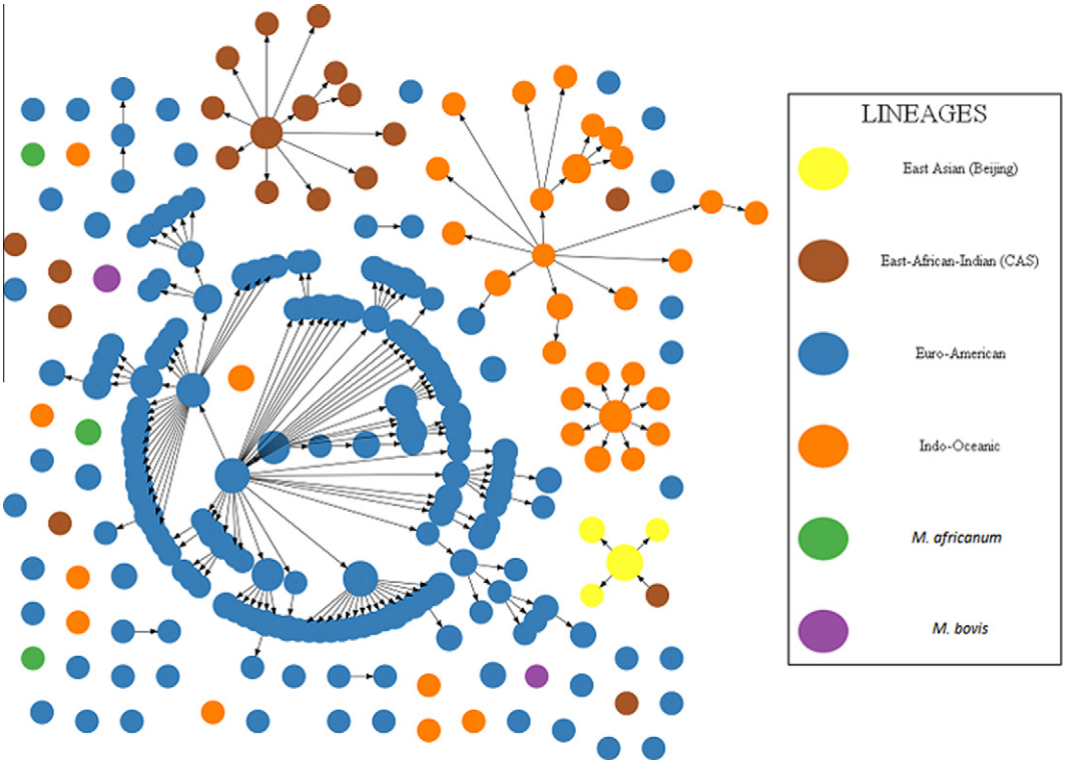

Fig. 6. Spoligoforests of 268 distinct spoligotype strains from the NYS dataset of 674 isolates generated using the visualization tool of TB-Lineage. Each lineage corresponds to a unique color as shown in the legend. Each node represents a cluster of strains of the same spoligotype and the node size represents the number of isolates on a log scale. Each edge represents a mutation from a parent spoligotype sequence to the child sequence by the loss of one or more adjacent spacers i.e. a contiguous deletion. Note lineages are highly cohesive with few edges between lineages. This indicates the high degree of genetic relatedness between strains within a lineage.

TB-Vis provides a visualization tool based on spoligoforests designed by (Reyes et al., 2008) that depicts the genetic diversity in the MTBC strain population by lineage and the possible evolutionary relationships between strains. Fig. 6 represents a spoligoforest constructed from the NYS dataset using TB-Vis. The MTBC strain population is depicted in the form of a forest of radial trees in which each node represents a distinct spoligotype that may be associated with one or more MIRU types. The sizes of the nodes represent the number of distinct MIRU types associated with the spoligotypes and are an indication of the inter-strain genetic diversity. An edge represents a possible mutation. The children of each node *i*, thus represent the strains that *i* can mutate into. Spoligotype mutation is modeled by deletion of one or more adjacent spacers, whereas a change in the number of repeats at a MIRU locus is regarded as a single mutation event for the MIRU type. The evolutionary relationships between strains is modeled based on genetic distances between MIRU patterns and spoligotypes of strains. Each strain may have multiple candidate parents. A single parent is chosen for each strain based on a comparison of the genetic distances between each parent-child pair. The following distance measures are used to select the most likely parent from amongst the candidates generated: (i) Hamming distance between MIRUs (the number of loci in which the two MIRU types differ) (ii) Hamming distance between spoligotypes (the number of spacers in which the two spoligotype sequences differ) (iii) Euclidean distance between MIRUs (root sum of squared differences in the numbers of repeats at each MIRU locus of the two strains). The nodes are colored based on the lineage identified by TB-Insight (TB-Rules). Thus, the visualization provides insight into the relatedness of strains based on the spoligotype and MIRU type of strains, as well as a view of the distribution of strains by lineage.

6.2. Host-Pathogen Treemaps

- *Input:*
  - Genotype information which may include (i) Spoligotype strains in binary or octal format (ii) 12, 15 or 24 loci of MIRU-VNTR (iii) RFLP (iv) SNP.
  - Patient's continent of birth.
- *Functionality:*
  - Draw host-pathogen treemap to visualize trends in associations between strain and patient groups, and identify anomalies.
- *URL:* http://tbinsight.cs.rpi.edu/run_tb_vis/treemaps.html.

Host-pathogen maps available at TB-Vis, provide a graphical representation of strain and patient associations. Patients are represented as nodes within the nested boxes depicting strains. The visualization depicts each strain by telescopic boxes depending on the number of biomarkers uploaded. In Fig. 7, the nested boxes represent the spoligotype, MIRU type and RFLP pattern, respectively. Other biomarkers such as SNPs may also be used. Patient characteristics such as birth-place are represented by color coding the nodes by continent of birth. This visualization provides a means of tracking trends in transmissions between patients infected with the strains of interest. It can help reveal previously unrecognized epidemiological links between patients. Anomalous behavior of strain groups can also be identified. Epidemiological investigations require the investment of significant time and resources. Therefore, identifying suspicious clusters using such visual tools will help towards the efficient allocation of efforts for case investigations.

The host-pathogen maps are based on the design of treemaps (Johnson, and Shneiderman, 1991). Hence, the use of nested boxes to depict strains is well-suited to capture the inherent hierarchical relationship between biomarkers used for MTBC genotyping arising out of differences in their discriminative abilities (Kremer et al., 2005; Kremer et al., 1999). The efficient use of space by treemaps allows a big-picture view of a large number of strains, and thus enables identifying typical and anomalous behavior of a strain with respect to the others in the study. This could lead to the identification or prediction of outbreaks. The treemaps are interactive enabling the user to search for and zoom in on strains of interest. Thus, treemap-based host pathogen maps provide a compact overview of the patient-strain associations, represent transmission trends and help in the identification of exceptions in these trends.

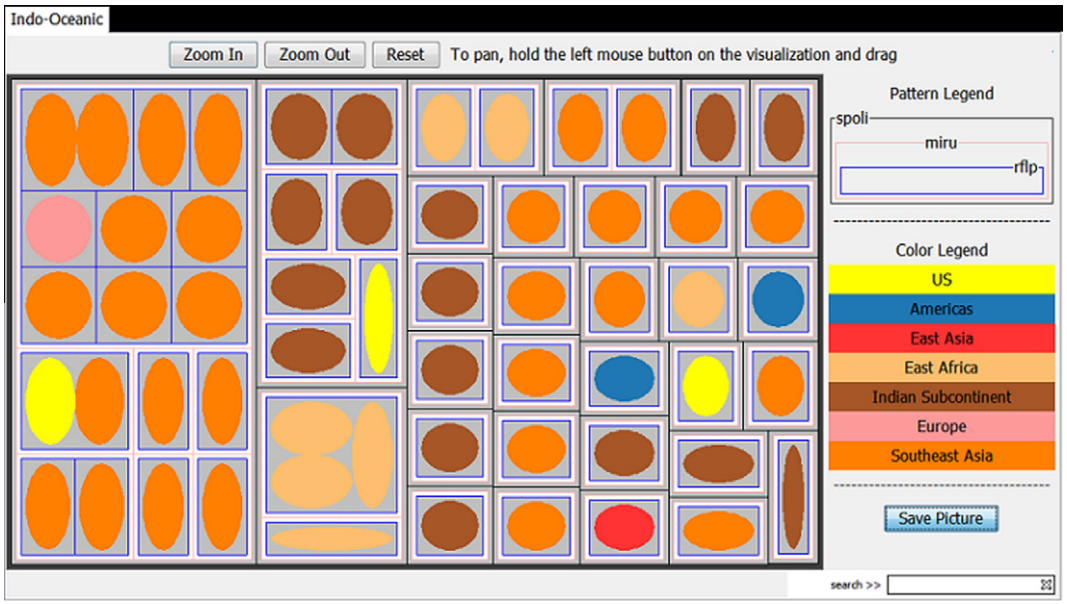

Fig. 7. Host-pathogen maps of patients from the NYS dataset infected with strains of the Indo-Oceanic lineage that visualize associations between the genotype and host characteristics. Strains are represented by triples of spoligotype, MIRU and RFLP patterns and are depicted by nested boxes. Patients are depicted as nodes colored by region of birth. The visualization shows the predominance of strains of the Indo-Oceanic lineage in patients from South-East Asia and the Indian subcontinent. Clusters of cases with identical associated genotype appear in bigger boxes, thus bringing attention to possible outbreaks.

While these existing programs can help epidemiologists make more informed decisions, there is much scope for the application of information visualization to develop tools for molecular epidemiology. There is the need to incorporate time in order to visualize TB dynamics. Interactivity in visual representations can offer the ability to filter and zoom in on individual or groups of strains and/or patients and view relevant statistics. Graph visualizations can be used to represent social networks inferred from epidemiological investigations. The application of visual analytics is a promising new direction for TB epidemiology.

7. Concluding remarks

In this survey, we explored computational tools that utilize molecular epidemiological data to address current challenges in the understanding of the genetic diversity of MTBC, and the disease dynamics and the pathogenesis of TB. Molecular epidemiology integrates molecular biology with traditional epidemiological approaches to study the influence of factors identified at the molecular level on the characteristics of MTBC, and the distribution and control of TB. TB surveillance and control programs now routinely perform DNA fingerprinting for almost all culture positive cases identified in the US. Although DNA fingerprints of strains are used primarily to identify clustered cases and help detect recent transmissions and outbreaks, they capture information that could potentially be used to develop more advanced tracking and control measures. Some web tools that utilize this information were covered in this study, and can be categorized as follows (i) Databases (ii) Transmission and mutation models (iii) Classification tools (iv) Visualization tools. Findings and inferences obtained from these tools can be used to guide control measures.

Large databases of DNA fingerprint information representing the bio-geographic diversity of strains in the MTBC population have been amassed by TB researchers worldwide. This wealth of information can be utilized to study the influence of factors identified at the molecular level on the distribution and control of disease. We need tools that provide access to these data and resources for the systematic analysis and interpretation of these data. The sequencing of the genome of several strains belonging to the MTBC has changed many previously held beliefs about the homogeneity of MTBC strains. There is greater genetic diversity within strains of the MTBC than previously believed and these variations have several important phenotypic consequences (Hershberg et al., 2008). Several studies have shown that such variations in the virulence, immunogenicity, transmissibility, drug-resistance, and host-pathogen associativity of strains are due to genomic polymorphisms (Gagneux and Small, 2007; Gagneux et al., 2006). The discovery of potential drug targets and development of vaccines can be made possible by the development of tools that use and integrate findings from these studies. There is a need for a shared platform to facilitate the detailed comparative analysis of genomic data obtained from various strains to establish these differences and their potential causes. We presented some such databases and potential applications of the data in this review. These databases allow comparisons to be made between epidemiological studies that help reveal previously unrecognized trends. The collection of DNA fingerprint data as part of routine TB surveillance has also helped identify associations between host populations and strain groups.

Classification of strains into lineages provides perspective on the phenotypic consequences of the genetic variations of MTBC strains. Phylogenetic analyses conducted using LSP and SNP of MTBC strains have helped investigate the history of evolution of MTBC and established the existence of distinct clades. Further investigations into the diversity of MTBC strains reveal differences in phenotype of strains belonging to the various clades. Computational methods that classify strains into these clades using DNA fingerprint data were discussed in this survey.

Genetic information will also help determine precise quantitative measures for transmission dynamics and augment classical epidemiological models. The incorporation of strain information into mathematical models of TB will result in richer analysis tools. DNA fingerprinting methods have served mainly to distinguish recent transmissions from reactivations of latent infections. However, methods are needed that exploit collections of DNA fingerprint data to do more complex tasks e.g. quantifying the extent of active transmission, identifying potential outbreaks and quantifying the severity of outbreaks. We look at transmission and mutation models for MTBC strains that have been developed with the application of methods from statistics and population genetics on DNA fingerprint data. Analysis of the mutation mechanisms involved in the biomarkers used for molecular epidemiological studies has helped create models of evolution of MTBC strains. These models have helped make inferences about the incidence of TB in a population and rates of active transmission. We also explore the use of these evolutionary models in visualization tools and in detecting emerging strains. This information can be used to set up appropriate control mechanisms.

In this survey, we discussed the need for tools that address several open questions about the pathogenesis and disease dynamics of TB, and characteristics of MTBC. Mathematical modeling and the application of methods from statistics and population genetics to data obtained from molecular epidemiological studies will play an important role in answering these questions facing TB researchers today.

8. Websites

A list of the URLs of websites surveyed in this paper are provided here for ready reference.

Tool	URL
MIRU-VNTR*plus* (Allix-Beguec et al., 2008)	www.MIRU-VNTRplus.org
SITVIT (Brudey et al., 2006)	www.pasteur-guadeloupe.fr&8081/SITVITDemo/
TB-GIMS	www.cdc.gov/tb/programs/genotyping/tbgims
TB-DB (Reddy et al., 2009)	www.tbdb.org/
MyBASE (Zhu et al., 2009)	mybase.psych.ac.cn
TBrowse (Bhardwaj et al., 2009)	tbrowse.osdd.net
MTbReglist (Jacques et al., 2005)	www.USherbrooke.ca/vers/MtbRegList
TBSGC (Terwilliger et al., 2003)	www.webtb.org/
TubercuList (Cole, 1999)	genolist.pasteur.fr/TubercuList
GenoMycDB (Catanho et al., 2006)	157.86.176.108/catanho/genomycdb/
MGDD (Vishnoi et al., 2008)	mirna.jnu.ac.in/mgdd/
SpolTools (Tanaka and Francis, 2006; Reyes et al., 2008)	www.emi.unsw.edu.au/spolTools/
SPOTCLUST (Vitol et al., 2006)	www.tbinsight.cs.rpi.edu/run_spotclust.html
TB-lineage (Shabbeer et al., 2011; Aminian et al., 2010)	www.tbinsight.cs.rpi.edu/run_tb_lineage.html
TB-Vis	www.tbinsight.cs.rpi.edu/run_tb_vis.html
Summary of Tools Surveyed	www.tbinsight.cs.rpi.edu/molepisurvey.html

Acknowledgments

This work was made possible by and with the assistance of Dr. Vincent Escuyer of the Wadsworth Center, New York State Department of Health, Dr. Jeffrey R. Driscoll and Dr. Lauren Cowan of the CDC, and Dr. Natalia Kurepina (PHRI). We would like to thank Dr. Nalin Rastogi (Institute Pasteur de Guadeloupe) and Dr. Andrew Francis (University of Western Sydney) for their valuable suggestions. This work is supported by NIH R01LM009731.

References

Allix-Beguec, 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. Journal of Clinical Microbiology 46 (8), 2692–2699.

Aminian, M., Shabbeer, A., Bennett., K., 2009. Determination of major lineages of *Mycobacterium tuberculosis* using mycobacterial interspersed repetitive units. In: Proceedings IEEE Int Conf Bioinformatics Biomed, pp. 338–343.

Aminian, M., Shabbeer, A., Bennett, K., 2010. A conformal bayesian network for identification of Mycobacterium tuberculosis complex lineages. BMC Bioinformatics 11 (Suppl 3), S4.

Aminian, M., Shabbeer, A., Hadley, K., Ozcaglar, C., Vandenberg, S., Bennett, K., 2011. Knowledge-based bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages. In: Proceedings ACM Conference on Bioinformatics, Computational Biology and Biomedicine.

Baker, L., Brown, T., Maiden, M.C., Drobniewski, F., 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerging Infectious Diseases 10 (9), 1568–1577.

Barnes, P., Cave, M., 2003. Molecular epidemiology of tuberculosis. New England Journal of Medicine 349 (12), 1149–1156.

Bhardwaj, A., Bhartiya, D., Kumar, N., Scaria, V., 2009. Tbrowse: an integrative genomics map of *Mycobacterium tuberculosis*. Tuberculosis (Edinb) 89 (5), 386–387.

Bifani, P.J., Mathema, B., Kurepina, N.E., Kreiswirth, B.N., 2002. Global dissemination of the *Mycobacterium tuberculosis* w-beijing family strains. Trends in Microbiology 10 (1), 45–52.

Borile, C., Labarre, M., Franz, S., Sola, C., Refrégier, G., 2011. Using affinity propagation for identifying subspecies among clonal organisms: lessons from M. tuberculosis. BMC Bioinformatics 12 (1), 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. Proceedings of the National Academy of Sciences of the United States of America 99 (6), 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 (spoldb4) for classification, population genetics and epidemiology. BMC Microbiology.

Catanho, M., Mascarenhas, D., Degrave, W., Miranda, A.B., 2006. Genomycdb: a database for comparative analysis of mycobacterial genes and genomes. Genetics and Molecular Research 5 (1), 115–126.

Cavalli-Sforza, L., Edwards, A., 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19 (3 Pt 1), 233.

CDC, 2011. Guide to the Application of Genotyping to Tuberculosis Prevention and Control. <http://www.cdc.gov/tb/programs/genotyping/manual.htm>.

Cole, S., 1999. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. FEBS letters 452 (7), 1–2.

Constant, P., Perez, E., Malaga, W., Laneelle, M.A., Saurel, O., Daffe, M., Guilhot, C., 2002. Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. Journal of Biological Chemistry 277 (41), 38148–38158.

Cowan, L.S., Mosher, L., Diem, L., Massey, J.P., Crawford, J.T., 2002. Variable-number-tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of Is6110 by using mycobacterial interspersed repetitive units. Journal of Clinical Microbiology 40 (5), 1592–1602.

de Jong, B.C., Hill, P.C., Aiken, A., Awine, T., Antonio, M., Adetifa, I.M., Jackson-Sillah, D.J., Fox, A., Deriemer, K., Gagneux, S., Borgdorff, M.W., McAdam, K.P., Corrah, T., Small, P.M., Adegbola, R.A., 2008. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in the gambia. Journal of Infectious Diseases 198 (7), 1037–1043.

Ernst, J.D., Trevejo-Nunez, G., Banaiee, N., 2007. Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis. Journal of Clinical Investigation 117 (7), 1738–1745.

Fang, Z., Morrison, N., Watt, B., Doig, C., Forbes, K.J., 1998. Is6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. Journal of Bacteriology 180 (8), 2102–2109.

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. Research in Microbiology 155 (8), 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. Emerging Infectious Diseases 8 (11), 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. Journal of Clinical Microbiology 41 (5), 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. Journal of Bacteriology 188 (2), 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. Journal of Clinical Microbiology 45 (10), 3393–3395.

Frothingham, R., Hills, H.G., Wilson, K.H., 1994. Extensive DNA-sequence conservation throughout the *Mycobacterium-tuberculosis* complex. Journal of Clinical Microbiology 32 (7), 1639–1643.

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*. Proceedings of the National Academy of Sciences of the United States America 103 (8), 2869–2873.

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infectious Diseases 7 (5), 328–337.

Getoor, L., Rhee, J.T., Koller, D., Small, P., 2004. Understanding tuberculosis epidemiology using structured statistical models. Artificial Intelligence in Medicine 30 (3), 233–256.

Goldstein, D., Ruiz Linares, A., Cavalli-Sforza, L., Feldman, M., 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proceedings of the National Academy of Sciences 92 (15), 6723.

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. Journal of Infectious Diseases 193 (1), 121–128.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N., Musser, J.M., 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics 162 (4), 1533–1543.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., Gagneux, S., 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biology 6 (12), e311.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proceedings of the National Academy of Sciences of the United States of America 101 (14), 4871–4876.

Jacques, P., Gervais, A., Cantin, M., Lucier, J., et al., 2005. Mtbreglist, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*. Bioinformatics 21 (10), 2563.

Johnson, B., Shneiderman, B., 1991. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In: Proceedings of the 2nd conference on Visualization'91, IEEE Computer Society Press, pp. 284–291.

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. Journal of Clinical Microbiology 35 (4), 907–914.

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. Journal of Clinical Microbiology 43 (11), 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. Journal of Clinical Microbiology 37 (8), 2607–2618.

Legrand, E., Filliol, I., Sola, C., Rastogi, N., 2001. Use of spoligotyping to study the evolution of the direct repeat locus by IS6110 Transposition in *Mycobacterium tuberculosis*. Journal of Clinical Microbiology 39 (4), 1595–1599. <http://jcm.asm.org/cgi/content/abstract/39/4/1595>.

Luciani, F., Francis, A.R., Tanaka, M.M., 2008. Interpreting genotype cluster sizes of Mycobacterium tuberculosis isolates typed with is6110 and spoligotyping. Infection Genetics and Evolution 8 (2), 182–190.

Mathema, B., Kurepina, N.E., Bifani, P.J., Kreiswirth, B.N., 2006. Molecular epidemiology of tuberculosis: current insights. Clinical Microbiology Reviews 19 (4), 658–685.

Murray, M., 2002. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proceedings of the National Academy of Sciences of the United States of America 99 (3), 1538–1543.

Murray, M., Nardell, E., 2002. Molecular epidemiology of tuberculosis: achievements and challenges to current knowledge. Bulletin World Health Organization 80 (6), 477–482.

Musser, J.M., Amin, A., Ramaswamy, S., 2000. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics 155 (1), 7–16.

Nardell, E., McInnis, B., Thomas, B., Weidhaas, S., 1986. Exogenous reinfection with tuberculosis in a shelter for the homeless. New England Journal of Medicine 315 (25), 1570–1575.

Niemann, S., Koser, C.U., Gagneux, S., Plinke, C., Homolka, S., Bignell, H., Carter, R.J., Cheetham, R.K., Cox, A., Gormley, N.A., Kokko-Gonzales, P., Murray, L.J., Rigatti, R., Smith, V.P., Arends, F.P., Cox, H.S., Smith, G., Archer, J.A., 2009. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical dna fingerprints. PLoS One 4 (10), e7407.

Ozcaglar, C., Shabbeer, A., Vandenberg, S., Yener, B., Bennett, K., 2011. Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors. BMC Genomics 12 (Suppl 2).

Reddy, T., Riley, R., Wymore, F., Montgomery, P., DeCaprio, D., Engels, R., Gellesch, M., Hubble, J., Jen, D., Jin, H., et al., 2009. Tb database: an integrated platform for tuberculosis research. Nucleic Acids Research 37 (suppl 1), D499.

Reed, M.B., Domenech, P., Manca, C., Su, H., Barczak, A.K., Kreiswirth, B.N., Kaplan, G., Barry, C.E.r., 2004. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature 431 (7004), 84–87.

Reed, M.B., Pichler, V.K., McIntosh, F., Mattia, A., Fallow, A., Masala, S., Domenech, P., Zwerling, A., Thibert, L., Menzies, D., Schwartzman, K., Behr, M.A., 2009. Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. Journal of Clinical Microbiology 47 (4), 1119–1128.

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 (2), 235–243.

Shabbeer, A., Cowan, L., Driscoll, J., Ozcaglar, C., Rastogi, N., Vandenberg, S., Yener, B., Bennett, K.P., 2011. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Unpublished Manuscript.

Shriver, M.D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R.E., Chakraborty, R., 1995. A novel measure of genetic-distance for highly polymorphic tandem repeat loci. Molecular Biology and Evolution 12 (5), 914–920.

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. Emerging Infectious Diseases 7 (3), 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. Proceedings of the National Academy of Sciences of the United States of America 94 (18), 9869–9874.

Streicher, E.M., Victor, T.C., van der Spuy, G.D., Sola, C., Rastogi, N., van Helden, P.D., Warren, R.M., 2007. Spoligotype signatures in the *Mycobacterium tuberculosis* complex. Journal of Clinical Microbiology 45 (1), 237–240.

Sun, Y.J., Bellamy, R., Lee, A.S., Ng, S.T., Ravindran, S., Wong, S.Y., Locht, C., Supply, P., Paton, N.I., 2004. Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of *Mycobacterium tuberculosis* in singapore. Journal of Clinical Microbiology 42 (5), 1986–1993.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology 44 (12), 4498–4510.

Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., Locht, C., 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. Journal of Clinical Microbiology 39 (10), 3563–3571.

Tanaka, M.M., Francis, A.R., 2005. Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. Infection, Genetics and Evolution 5 (1), 35–43.

Tanaka, M.M., Francis, A.R., 2006. Detecting emerging strains of tuberculosis by using spoligotypes. Proceedings of the National Academy of Sciences 103 (41), 15266–15271.

Terwilliger, T.C., Park, M.S., Waldo, G.S., Berendzen, J., Hung, L., Kim, C., Smith, C., Sacchettini, J., Bellinzoni, M., Bossi, R., et al., 2003. The tb structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. Tuberculosis 83 (4), 223–249.

van der Spuy, G.D., Kremer, K., Ndabambi, S.L., Beyers, N., Dunbar, R., Marais, B.J., van Helden, P.D., Warren, R.M., 2009. Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. Tuberculosis (Edinb) 89 (2), 120–125.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., et al., 1993. Strain identification of *Mycobacterium tuberculosis* by dna fingerprinting: recommendations for a standardized methodology. Journal of Clinical Microbiology 31 (2), 406–409.

Van Soolingen, D., 2001. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. Journal of internal medicine 249 (1), 1–26.

Vishnoi, A., Srivastava, A., Roy, R., Bhattacharya, A., 2008. Mgdd: *Mycobacterium tuberculosis* genome divergence database. BMC Genomics 9, 373.

Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N., Bennett, K.P., 2006. Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. Infection Genetics and Evolution 6 (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. Journal of Clinical Microbiology 40 (12), 4457–4465.

Zhu, X., Chang, S., Fang, K., Cui, S., Liu, J., Wu, Z., Yu, X., Gao, G.F., Yang, H., Zhu, B., Wang, J., 2009. Mybase: a database for genome polymorphism and gene function studies of mycobacterium. BMC Microbiology 9, 40.