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.

© 2011 Published by Elsevier B.V.

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 epidemiological 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...
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 regions distributed on the chromosome of MTBC. MIRU typing is served in a selected number of 41 identified mini-satellite like repeats.

Tandem Repeats based analysis exploits the polymorphisms observed in a selected number of 41 identified mini-satellite like repeats 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.

2.3. 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 repeats 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 analysis 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; Fililiol 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 TdD1 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-VNTRplus 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 type (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.

- $(d_{ii})^2$: 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.](image-url) NJ tree in radial format created from the NYS dataset using the tree-based identification tool on MIRU-VNTRplus.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.
make inferences about the evolutionary relationships between the unique evolutionary mechanism of the DR region in order to emerging global trends and identify anomalies in transmission pat-
spoligotypes (Brudey et al., 2006). These models can help detect development of mathematical models of the genetic variation in cal approaches from population genetics may be applied in the abases facilitate the application of bioinformatic methods. Statisti-
characteristic examples of each lineage which provide a basis for constructing visual rules. Characteristic signatures have been identified in spoligotype se-
sequences using data published in SpolDB4 and other spoligotype-
functions 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 se-
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 data-
bases 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 inher-
ence of phylogenies.

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 labora-
tories 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 transcrip-
tion polymerase chain reaction (RT-PCR) data) of M. tuberculo-
sis 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 compari-
sions 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 coeffi-
cients computed using 1260 dual-dye micro-array chips).
    - Find polymorphisms in resequenced genomes of 31 selected strains representative of the MTBC population using Diverse Sequencing tools.
    - Download data e.g. sequence, genes, markers in various for-
mats facilitating further analysis.

---

* This map drawing tool is not part of SITVIT. Map generated using Google Maps.

* 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.
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.


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.

- 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/.
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.


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 the statistics provided by this tool and their significance 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 \( \frac{g}{n} = 1.15 \), clustering index \( R_{\text{cl}} = 0.13 \), clustering rate \( R_{\text{cl}} = 0.11 \) 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 frequently (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; Ozcanlar 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-

<p>| 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. n = 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. |</p>
<table>
<thead>
<tr>
<th>Formula</th>
<th>Interpretation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{g}{n} )</td>
<td>Average cluster size, indicative of number of recent transmissions, as well as reactivations. Number of singletons, indicative of number of reactivations of latent infections.</td>
<td>2.48</td>
</tr>
<tr>
<td>( R_{\text{cl}} = \frac{n}{n-1} - \frac{c}{n} )</td>
<td>Clustering index, indicative of the number of transmissions and is the difference between the number of cases observed less the number of source cases.</td>
<td>0.60</td>
</tr>
<tr>
<td>( H = 1 - \sum_{i=1}^{n} \frac{n_{i}}{n} \left( \frac{n_{i}}{n} \right)^{2} )</td>
<td>Virtual heterozygosity, a measure of genetic diversity that represents probability that any 2 individuals chosen from the same dataset have different genotype.</td>
<td>0.40</td>
</tr>
<tr>
<td>( \frac{\text{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).}}{\text{Cluster}} )</td>
<td>166.54</td>
<td></td>
</tr>
</tbody>
</table>
In the NYS dataset, i.e. strains that were found to be emerging,
strains with low significance of the result. The False discovery rate (FDR) or
falsely rejected. A low calculated to determine the proportion of significant tests that are
emerging. Multiple hypothesis tests are performed to test the sig-
ificant strains associated with different lineages have been found to vary
in their immunogenicity, pathogenicity, virulence, transmissibility
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
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.

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 Francis, 2005). In an IA forest, every node has a single inbound
edge and the outdegree is the number of mutations observed for the
extending the work of other investigators (Gagneux and Small, 2007; Gagneux et al., 2006). The observed
clusters of size 3 and 5 respectively, and are indeed rapidly trans-
migrants, rather than recent transmissions of emerging strains.
This indicates that the large number of cases associated with these
spoligotypes can be attributed to the age of the strain, and are a re-
sult 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 trans-
mittting strains. This suggests the use of additional genotype infor-
mation in models in order to make more accurate inferences about the
transmission rates of strains.
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](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 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. 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; Sebhan 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 markers (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 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

- **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](http://tbinsight.cs.rpi.edu/run_tb_lineage.html).

5.2. TB-Lineage: 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).

**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, 000000000000000000010000100000001111111111, 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; Ozçaglar 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.

![Spoligoforests](http://tbinsight.cs.rpi.edu/run_tb_vis/spoligoforests.html)
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, thus represent the strains that 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.

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.](image-url)
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.

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRU-VNTRplus (Allix-Béguec et al., 2008)</td>
<td><a href="http://www.MIRU-VNTRplus.org">www.MIRU-VNTRplus.org</a></td>
</tr>
<tr>
<td>SITVIT (Brudey et al., 2006)</td>
<td><a href="http://www.pasteur-guadeloupe.fr&amp;8081/SITVITDemo/">www.pasteur-guadeloupe.fr&amp;8081/SITVITDemo/</a></td>
</tr>
<tr>
<td>TB-GIMS</td>
<td><a href="http://www.cdc.gov/tb/programs/genotyping/tbgims">www.cdc.gov/tb/programs/genotyping/tbgims</a></td>
</tr>
<tr>
<td>TB-DB (Reddy et al., 2009)</td>
<td><a href="http://www.tbdb.org/">www.tbdb.org/</a></td>
</tr>
<tr>
<td>MyBASE (Zhu et al., 2009)</td>
<td>mybase.psych.ac.cn</td>
</tr>
<tr>
<td>TBrowse (Bhardwaj et al., 2009)</td>
<td>tbrowse.osdd.net</td>
</tr>
<tr>
<td>MTbRegList (Jacques et al., 2005)</td>
<td><a href="http://www.USherbrooke.ca/vers/MtbRegList">www.USherbrooke.ca/vers/MtbRegList</a></td>
</tr>
<tr>
<td>TBSCC (Terwilliger et al., 2003)</td>
<td><a href="http://www.webtb.org/genolist.pasteur.fr/Tuberculist">www.webtb.org/genolist.pasteur.fr/Tuberculist</a></td>
</tr>
<tr>
<td>Tuberculist (Cole, 1999)</td>
<td><a href="http://www.emi.unsw.edu.au/Tuberculist">www.emi.unsw.edu.au/Tuberculist</a></td>
</tr>
<tr>
<td>GenoMycDB (Catanho et al., 2006)</td>
<td>157.86.176.108/catanho/genomycdb/</td>
</tr>
<tr>
<td>MGDD (Vishnoi et al., 2008)</td>
<td>mirma.jnu.ac.in/mgdd/</td>
</tr>
<tr>
<td>SpoTTools (Tanaka and Francis, 2006; Reyes et al., 2008)</td>
<td><a href="http://www.emi.unsw.edu.au/spotTools/">www.emi.unsw.edu.au/spotTools/</a></td>
</tr>
<tr>
<td>SPOTCLUST (Vitol et al., 2006)</td>
<td><a href="http://www.tbinsight.cs.rpi.edu/run_spotclust.html">www.tbinsight.cs.rpi.edu/run_spotclust.html</a></td>
</tr>
<tr>
<td>TB-lineage (Shabbeer et al., 2011; Aminian et al., 2010)</td>
<td><a href="http://www.tbinsight.cs.rpi.edu/run_tb_lineage.html">www.tbinsight.cs.rpi.edu/run_tb_lineage.html</a></td>
</tr>
<tr>
<td>TB-Vis</td>
<td><a href="http://www.tbinsight.cs.rpi.edu/run_tb_vis.html">www.tbinsight.cs.rpi.edu/run_tb_vis.html</a></td>
</tr>
<tr>
<td>Summary of Tools Surveyed</td>
<td><a href="http://www.tbinsight.cs.rpi.edu/molepisurvey.html">www.tbinsight.cs.rpi.edu/molepisurvey.html</a></td>
</tr>
</tbody>
</table>
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 (Institute Pasteur de Guadeloupe). 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 R01LM099731.

References


