Visualization of tuberculosis patient and *Mycobacterium tuberculosis* complex genotype data via host-pathogen maps

Kristin P. Bennett[1,2], Cagri Ozcaglar[1], Janani Ranganathan[3], Srivatsan Raghavan[3],
Jacob Katz[2], Dan Croft[2], Bülent Yener[1], Amina Shabbeer[1]
(1) Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY
(2) Mathematical Sciences Department, Rensselaer Polytechnic Institute, Troy, NY
(3) Information Technology, Rensselaer Polytechnic Institute, Troy, NY
bennek@rpi.edu, ozcagc2@cs.rpi.edu, rangaj@rpi.edu, sraghavan@judlau.com,
katzj2@rpi.edu, croftd2@rpi.edu, shabba@cs.rpi.edu, yener@cs.rpi.edu

***Abstract*—DNA fingerprints of *Mycobacterium tuberculosis* complex bacteria (MTBC) are routinely gathered from tuberculosis (TB) patient isolates for all tuberculosis patients in the United States to support TB tracking and control efforts but few tools are available for visualizing and discovering host-pathogen relationships. We present a new visualization approach, host-pathogen maps, for simultaneously examining MTBC genotyped by multiple DNA fingerprinting methods such as spoligotypes and restriction fragment length polymorphisms (RFLP) along with associated patient surveillance data. The host-pathogen maps are dynamically coupled with spoligo-forests or other phylogenetic tree approaches to allow easy navigation within the pathogen genotyping space. Visualization of New York State and New York City (NYC) TB patient data from 2001-2007 is used to illustrate how host-pathogen maps can be used to rapidly identify potential instances of uncontrolled spread of tuberculosis versus disease resulting from latent reactivation of prior infection, a critical component of tuberculosis control. Host-pathogen maps also reveal trends and anomalies in the relationships between patient groups and MTBC genetic lineages which can provide critical clues in epidemiology and contact investigations of TB.**

***Index Terms*—tuberculosis, visualization, host-pathogen maps, tree-maps, spoligotypes.**

I. Introduction

Tuberculosis (TB) stubbornly persists as a leading cause of death worldwide. According to the World Health Organization, one third of the human population is infected, either latently or actively, with TB. *Mycobacterium tuberculosis* complex (MTBC) is the causative agent of tuberculosis. DNA fingerprinting of MTBC has proven useful for tracking and control of TB. Isolates from TB patients are routinely genotyped using multiple biomarkers, which include spacer oligonucleotide types (spoligotypes), Mycobacterial Interspersed Repetitive Units - Variable Number Tandem Repeats (MIRU-VNTR), and IS6110 Restriction Fragment Length Polymorphism (RFLP). TB fingerprints of a single type can be visualized using phylogenetic trees or specialized methods for spoligotypes such as spoligoforests [1], [2], [3]. The spoligoforest illustrating the spoligotypes found in 206 TB patients in New York State is shown in Figure 1. Each node in the spoligoforest represents a spoligotype, and each edge represents a possible mutation event from parent spoligotype to child spoligotype. However, spoligoforests have limitations in that only the node size is used to convey patient information (here the number of patients infected with that spoligotype on a log scale). The available information from other biomarkers and patient information does not appear. Current visualization tools from molecular epidemiology of tuberculosis focus on examining the data from either the pathogen or host perspective, but not both. No visualization approach deals with multiple biomarkers. This study introduces a new visualization method for host-pathogen relationships with multiple biomarkers.

Host-pathogen maps provide a graphical representation of strain and patient associations. Patients are represented as nodes within nested rectangles depicting MTBC strains as typed by various biomarkers. The visualization displays each strain by telescopic boxes depending on the number of biomarkers uploaded. The host-pathogen maps are based on the design of tree-maps [4]. In an analogy to how tree-maps are applied to visualize the activity of a stock exchange, genetically identical strains of MTBC can be thought of as a company and patients who acquire a strain of tuberculosis can be thought of as buyers purchasing that company's stock. The tree-map hierarchy captures the genetic hierarchy imposed by the biomarkers much like a tree-map represents sectors of the stock market. The size of the boxes reflect the number of patients with that strain. The patient nodes within the box can reflect patient properties critical for molecular epidemiology investigations such as country of birth, risk factors, site of infection, and drug resistance.

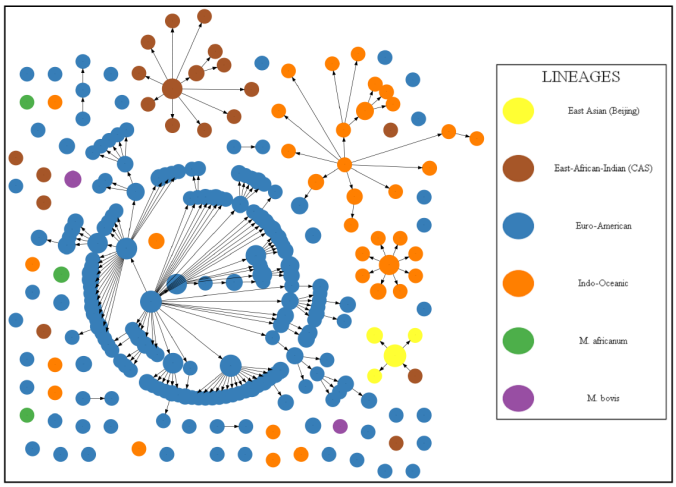

Fig. 1: The spoligoforest of the NYS dataset. Each node in the spoligoforest represents a spoligotype, and each edge represents a potential mutation event from the parent spoligotype into the child. The nodes are colored by the lineage of MTBC strain associated with the spoligotype, and size of the node represents the number of isolates with the specified spoligotype in log scale.

Figure 2 shows the host-pathogen map of NYC patients from 2001-2007 infected with strains belonging to East Asian (Beijing). In the figure, the outer bounding box represents the spoligotype and the inner bounding box represents the RFLP pattern. For example "S00069" in the upper right corner is a spoligotype that contains 9 distinct RFLP's. Other biomarkers such as MIRU-VNTR and single nucleotide polymorphisms may also be used. Different patient char-

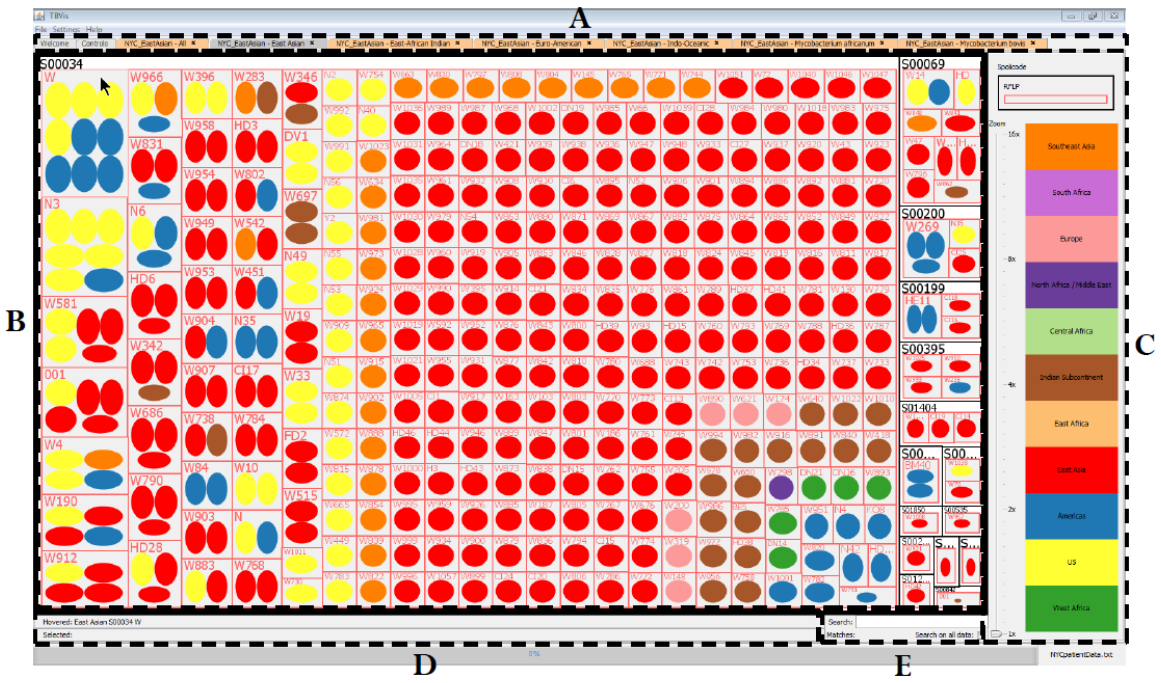

Fig. 2: Host-pathogen tree-map of patients infected with East-Asian (Beijing) strains in the NYC dataset. The sections of the host-pathogen tree-map are: A) Tabbed pane, B) Display panel, C) Legend panel, D) Query panel, E) Search panel. The outer box represents spoligotype, and the inner box represents the RFLP of MTBC strains. The host-pathogen map shows the predominance of East-Asian (Beijing) strains in patients from East Asia and the US. Possible outbreaks are indicated by boxes with many patients.

acteristics can be represented by varying the patient node size, shape, and color. In Figure 2, the patients are colored by the continent or world region containing their country of birth. Spoligotype "S00069" contains 9 patients from the regions of East Asia, Southeast Asia, Americas, United States, and Indian Subcontinent. Country of birth is critical for TB control since the majority of TB cases results from latent reactivation of TB in foreign-born patients who acquired the disease abroad. The predominance of red dots in Figure 2 indicates that most patients infected with the East East Asian lineage strains are from Asia. Many TB strains are associated with specific geographical areas. The fact that many boxes contain a single patient indicates that most cases represent latent reactivation of disease acquired abroad. However the box labeled "W" anomalously contains several US patients and other patients from the Americas which is highly indicative of recent transmission within NYC which should be further investigated. Separate host-pathogen maps are drawn for each MTBC genetic lineage allowing the data from 5235 patients in the New York City dataset to be rapidly interrogated.

Host-pathogen maps are available online, as part of the TB-vis tool: http://tbinsight.cs.rpi.edu/run_tb_vis/treemaps.html. Host-pathogen maps are interactive giving the user the flexibility to hover over a component in the map to query more information, search for a component they already know, zoom, pan, select a set of nodes, and label the boxes and nodes with attributes of interest. The tool can be used to address public health questions critical for TB control. Groups of MTBC with identical biomarkers (called clusters) narrow down the search for patients involved in transmission events, leading to possible epidemiological links. The use of nested boxes to depict strains is well suited to capturing the inherent hierarchical relationship between biomarkers used for MTBC genotyping arising from differences in their discriminative abilities. The efficient use of space by tree-maps allows a big-picture view of a large number of strains, and thus enables spotting anomalous behavior of a strain with respect to the others in the population. This could lead to the identification or prediction of recent transmission events (potential outbreaks) versus latent reactivation of disease acquired abroad. The tree-maps are interactive, enabling the user to filter and zoom in on strains of interest. Statistics and details related to patients and strain groups could be shown. Thus, host-pathogen tree-maps provide a compact overview of the patient-strain associations.

The implementation is not specific to TB, thus TB-Vis can be used to visualize any biomarkers and patients for other infectious diseases and agents. A TB-specific version coupling spoligoforests, host-pathogen maps, and patient epidemiology graphs is in progress to provide a comprehensive visualization of TB host-pathogen relationships to support TB public health control efforts.

II. Methods

Host-pathogen maps allow simultaneous overview of patient attribute and strain genotype data. Each patient is represented as a node within nested boxes. Each box surrounding the patient node represents a biomarker of the MTBC strain which infected the patient. Figure 3 shows the generation process of host-pathogen maps. First, the patient dataset is read, and patients are grouped by the associated genotype of the MTBC strain selected by the user (in this case spoligotype and RFLP). The host-pathogen map is generated based on this hierarchical grouping of patients. Patient nodes can be colored and other patient attributes can be depicted by shape or size of the node. Each bounding box represents a biomarker of a MTBC strain, and each level in the nested box adds a new biomarker to the genotype of the strain. Groups of patients with identical biomarkers are called clusters. An individual tuberculosis case may either be the result of recent transmission, or more commonly, latent reactivation

Patient	ID	Biomarker 1	Biomarker 2	TB continent
Patient 1	105	S00669	MY8	Indian Subcontinent
Patient 2	2443	S00210	GD139	Indian Subcontinent
Patient 3	2452	S00210	MY44	Indian Subcontinent
Patient 4	2487	S00247	NO12	East Africa
Patient 5	2492	S00537	GD142	Indian Subcontinent
Patient 6	2493	S00648	GD141	US
...	...	...	...	...

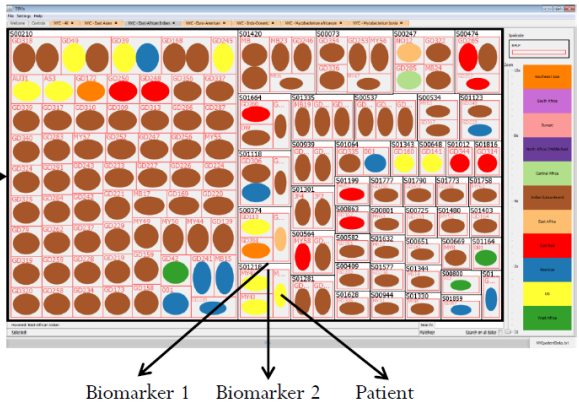

Fig. 3: The generation of host-pathogen maps. First, the patient dataset is read. Patients are grouped by the genotype of the associated MTBC strain and a host-pathogen tree-map is generated. Each bounding box represents a biomarker of the MTBC strain, and each level in the nested box adds a new biomarker to the strain. Within nested boxes, each patient is represented as a node, colored by TB continent of the patient based on country of birth.

of previously acquired infection, the former being more of a public health concern. Most cases of recent transmission occur within a cluster and more rarely across closely related genetic clusters. However, recent transmission may not be occurring in a cluster if the infections are caused by latent reactivation of previously acquired disease.

Host-pathogen tree-maps are dynamic maps based on the design of tree-maps [4]. The user can specify the properties of the treemap including the attributes to be used to determine the boxes, classes and colors of the nodes. In the example here, the classes are the major lineages of MTBC, the boxes are the biomarkers used to determine clusters (genetically distinct groups of MTBC), and each node corresponds to a patient colored by their country or region of birth. Host-pathogen tree-maps allow the user to search for a component, query a component, zoom and pan on the map, or label the boxes and nodes. The user can hover over a box or a node in the tree-map to query the data about a particular component of the map. The user can also search for a particular patient or strain using the search text box, and corresponding components of the map are highlighted on the tree-map. They can also label the patient nodes with attributes of interest.

A. System overview

Host-pathogen maps are implemented as part of the TB-Vis tool, available at http://tbinsight.cs.rpi.edu/about_tb_vis.html. TB-Vis is implemented in Java. The program first reads in the patient data from the dataset submitted by the user. The data is converted to a graph, preserving the hierarchical structure. The graph is converted into a host-pathogen map. The tabs of the TB-Vis screen display are populated with host-pathogen maps for each class selected by the user. In the examples studied here the classes are the major genetic lineages of TB as determined by the TB-Lineage package (http://tbinsight.cs.rpi.edu) [5].

B. The tool

The TB-Vis tool displays host-pathogen maps and gives the user the flexibility to search, query, zoom, and pan over the map. After uploading the patient dataset, the user selects a class, a set of boxes for biomarkers, and a ColorBy attribute to determine the coloring scheme of patients from dropdown boxes. The user has the flexibility to assign colors and choose labels for the patient nodes before drawing the host-pathogen tree-map. The user clicks on the "New graph using Treemap Visualizer" button and tree-maps for the patient dataset and for each class are displayed in each tab of the screen. The user can search for a component on the map using the search panel in the bottom right corner. The user can hover over a node to query more information about the component, which is displayed on the query panel on the bottom left corner. The user can also zoom in and out of the map, as well as pan the map.

III. Case studies

Host-pathogen maps are useful tools to find host-pathogen associations in TB patient datasets. We have used host-pathogen maps to visualize TB patient data as well as associated MTBC genotype data collected for TB surveillance in the United States. We illustrate this new visualization method on two datasets: a New York City (NYC) dataset consisting of 2499 TB patients and a New York State (NYS) dataset consisting of 206 TB patients. In the next section, we analyze patients and strains of these datasets by each major lineage using host-pathogen maps.The major lineages were determined using the TB-lineage tool [3], [5], [6].

A. NYC dataset

The NYC dataset consists of 2499 TB patients. The New York City Department of Health denotes clusters by spoligotype and RFLP. Thus we used 2-level host-pathogen maps to analyze NYC dataset with spoligotype as the first-level biomarker, and RFLP as the second-level biomarker. A different host-pathogen map was drawn for each major lineages using the class feature of TB-Vis.

The host-pathogen tree-map of 414 patients infected with East Asian (Beijing) strains in the NYC dataset is shown in Figure 2. As previously discussed above the patients infected with East Asian (Beijing) strains are predominantly from East Asia, as indicated by the patient nodes colored red. Most clusters consist of one or two patients indicating that most cases of East Asian probably arise from recent activation. However, several clusters contain many patients including United States (US) born patients which is indicative of recent transmission. Interactivity of host-pathogen tree-maps enable the user to query these clusters of interest. For example, the user can hover over cluster *W* on the host-pathogen tree-map in Figure 2, and the query panel on the bottom left corner displays "East Asian S00034 W", meaning that the user hovers over the cluster of patients

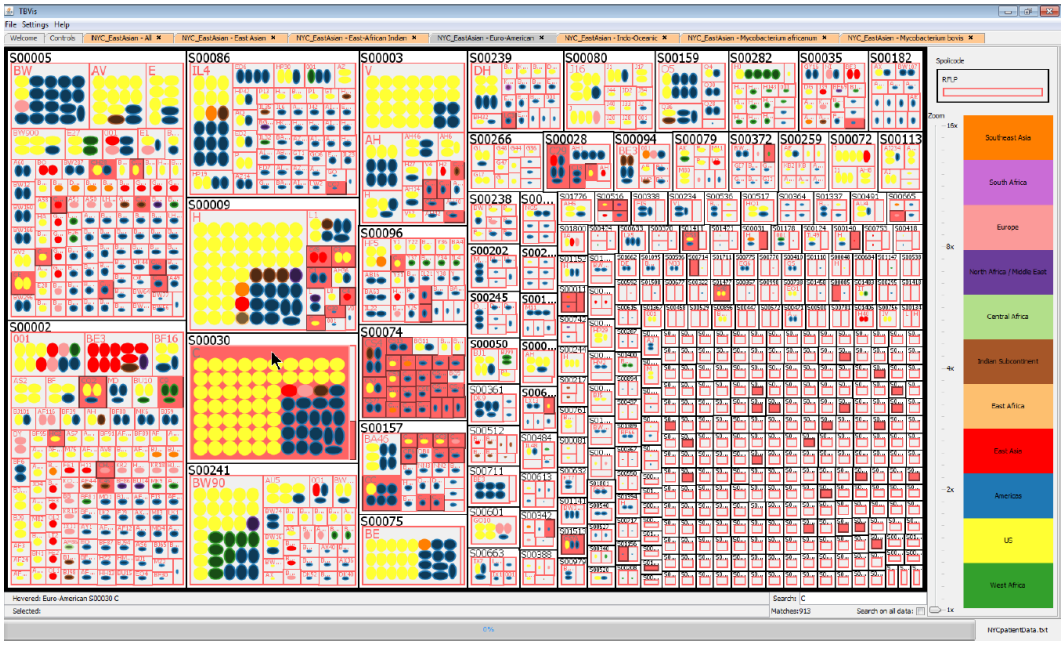

Fig. 4: Host-pathogen tree-map of patients infected with Euro-American strains in the NYC dataset. On the tree-map, the user searches for cluster *C*, using the text box on the search panel on the bottom right corner. Cluster *C* is highlighted in pink, as well as other clusters including the letter C in their RFLP. The user hovers over the largest highlighted cluster to query more information about it, and the query panel on the bottom left corner informs the user that the cluster consists of patients infected with Euro-American strains genotyped by spoligotype S00030 and RFLP C.

infected with MTBC strain genotyped with spoligotype S00034 and RFLP W. These labels are also visible in the figure, given as an option to the user before generating the tree-map. Notice that this cluster *W* is the largest cluster in the tree-map and it is displayed at the top left corner of the tree-map.

Zooming into cluster *W* in Figure 2 gives more insight into patients infected with W-Beijing strains. Cluster *W* is a known initiator of severe outbreaks worldwide and in the US [7], [8]. Strains of cluster *W* have spoligotype S00034, and they infected 4 patients from the US and 5 patients from the Americas. This is unusual since East Asian MTBC typically occurs in Asian patients. This is suggestive of recent transmission and a potential outbreak. There may be possible epidemiological links between patients from the US and patients from Americas. Further cluster and patient investigation is warranted to seek these epidemiological links.

Host-pathogen associations of 1645 patients infected with Euro-American strains in the NYC dataset are displayed in the host-pathogen tree-map in Figure 4. The host-pathogen map is dominated by nodes colored in yellow and blue, which highlight the predominance of patients from the United States and the Americas. Large clusters are readily indicated including the C strain which has been proven to be associated with recent transmission. Host-pathogen tree-maps enable the user to search for patients or strains within the map. In the host-pathogen tree-map in Figure 4, the user types “C” in the text box of the search panel on the bottom right corner, and RFLP boxes including “C” are highlighted in pink. The user then hovers over the largest cluster among the highlighted boxes, and the query panel on the bottom left corner displays “Euro-American S00030 C”, meaning that the cluster belongs to patients infected with Euro-American strains genotyped by spoligotype S00030 and RFLP *C*. As in this scenario, the user can use search and query options of the host-pathogen tree-map to identify the clusters of interest.

We turn our focus to known clusters of Euro-American strains which are initiators of possible outbreaks. In Figure 4, we zoom into the host-pathogen tree-map and investigate clusters *C*, *V*, and *BW90*. Cluster *C* consists of strains genotyped by spoligotype S00030, and infected patients from the United States and Americas, as shown with patients colored in yellow and blue. One patient each from East Asia, Europe and Indian subcontinent, shown by circles colored in red, pink and brown respectively, are also infected with the strains of cluster *C*. Cluster *V* consists of strains genotyped by spoligotype S00003, and infected patients are from the United States and the Americas. Cluster *BW90* consists of strains genotyped by spoligotype S00241, which infected patients from a diverse set of TB continents: United States, Americas, West Africa, and North Africa / Middle East. As highlighted by the number of green patient nodes, strains of cluster *BW90* are predominant among patients from West Africa, which might be transmitted to patients in the US and the Americas.

B. *NYS dataset*

NYS dataset consists of 206 patients from 2001 to 2007. The New York State Department of Health determines clusters using three biomarkers, spoligotypes, MIRU, and RFLP, thus a three level host-pathogen map with boxes corresponding to these biomarkers is used. We visualized the patients infected with Euro-American and Indo-Oceanic strains, to visualize trends and potential outbreaks.

The host-pathogen tree-map of 22 patients infected with Indo-Oceanic strains in the NYS dataset is shown in Figure 5. In this tree-map, the user chooses to label the patients with their country of birth. The patients in the same cluster share the same spoligotype, MIRU, and RFLP. The user can immediately observe that patients are primarily orange indicating they are from the Philippines and that most clusters consists of a single patient. This indicates that most cases are likely due to latent activation of disease acquired abroad

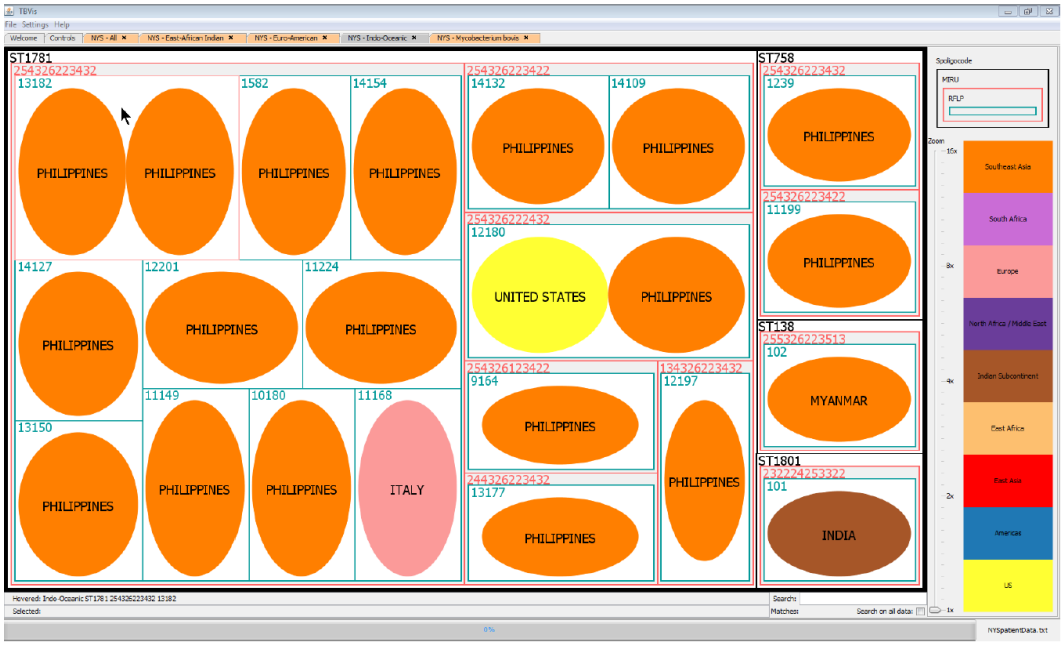

Fig. 5: Host-pathogen tree-map of patients infected with Indo-Oceanic strains in the NYS dataset. The user chooses to label patient nodes with their country of birth. The user hovers over a cluster, and the query panel on the bottom left corner informs the user that the cluster has the set of patients infected with MTBC strains genotyped by spoligotype ST1781, MIRU-VNTR 254326223432, and RFLP 13182.

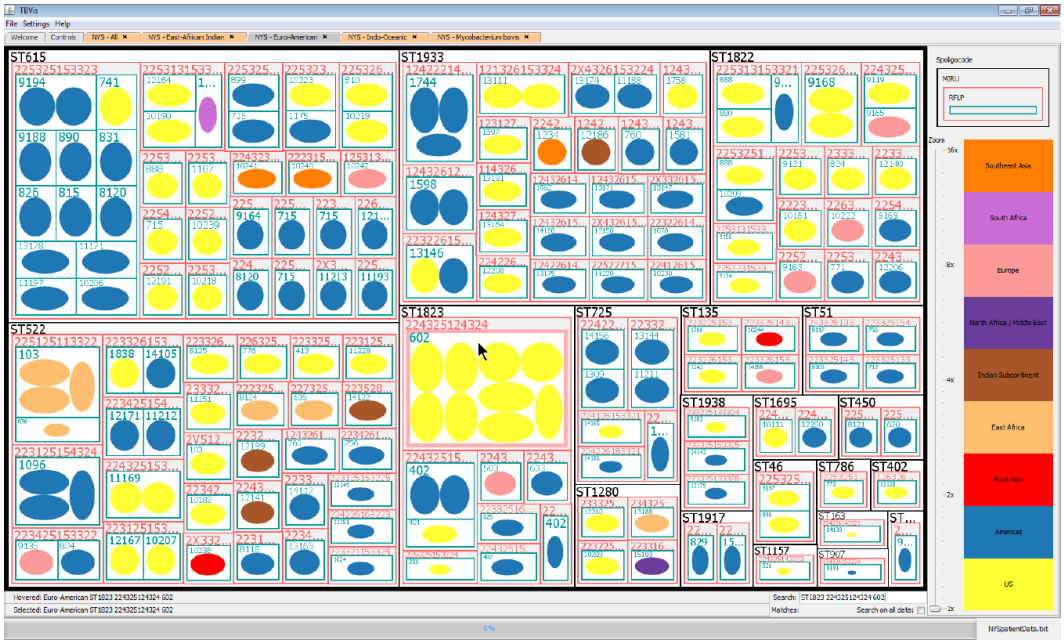

Fig. 6: Host-pathogen tree-map of patients infected with Euro-American strains in the NYS dataset. The user types in the text box of the search panel, and the patients infected with MTBC strain genotyped by spoligotype ST1823, MIRU-VNTR 224325124324, and RFLP 602 are highlighted in pink. Then the user hovers over the highlighted cluster, and the information about the cluster is displayed in the query panel on the bottom left corner.

by the well known Manilla strains of MTBC. Note that only one US patient has been infected by Indo-Oceanic. The small cluster sizes and consistency in the country of birth indicate there is little evidence of recent transmission in the United States.

The host-pathogen tree-map of 180 patients infected with Euro-American strains in the NYS dataset is shown in Figure 6. The patients in the same cluster share the same spoligotype, MIRU, and RFLP which validates the outbreak investigation with more biomarkers. As in the NYC Euro-American example, most patients are from the US or the Americas. In the figure, the user is interested in the largest cluster which consists of 9 patients from United States, infected with MTBC strains genotyped by spoligotype ST1823, MIRU-VNTR 224325124324, and RFLP 602. This cluster is known

to be an outbreak involving recent transmission. The user can see the largest cluster on the tree-map and can hover over the cluster, which will display the lineage, spoligotype, MIRU-VNTR, and RFLP information on the query panel on the bottom left corner. The user can also search for a known cluster: the user types "ST1823 602 224325124324", and 9 matches are found by the tree-map, as shown with the cluster highlighted in pink in Figure 6.

IV. Discussion

Visualization via host-pathogen maps enables molecular epidemiologists and TB controllers to interpret patient and strain data simultaneously to infer host-pathogen associations. Patients infected with genetically identical MTBC strains may potentially be infected by the same chain of transmission, and are named clustered cases [9]. Host-pathogen maps make use of both genotype data and clinical data for genotype cluster investigation. Trends in genetically related strains such as in country of birth of foreign born patients can be readily identified. Anomalies indicative of recent transmission and possible outbreak such as large clusters and transmission to US-born patients can be rapidly identified. The case studies here focused on country of birth of patients and three biomarkers. But in TB-Vis, any biomarker or attribute can be used as boxes and colors can be used to represent other attributes of interest such as drug-resistance, time in the US, age, risk factors, and site of infection. In the future, Host-pathogen maps can be further enhanced by incorporating various patient and genotype attributes. The shape of patient nodes can be used to depict the time spent in the US by the patient, which will help identify disease transmission involving foreign-born patients. Having more clinical data available, patient epi-links can be added between patient nodes to include the details of TB transmission within a host-pathogen map. Statistical details for genotype clusters can be incorporated to the host-pathogen map. For example, within a genotype cluster, the illustration of the time spent in the US versus age of the patient can explain the distribution of infection by the selected MTBC isolate to immigrants and locals. With the available clinical data including the date of infection, epi-curves can help molecular epidemiologists view the infection with a specific MTBC isolate the months or years.

Genetic similarity among genotype clusters in a host-pathogen map can also be represented effectively. Motivated by the design of spoligoforests, genetically similar spoligotype clusters can be linked to preserve genetic proximity among MTBC isolates [2], [5]. Current implementation of spoligoforests in TB-Vis tool can be placed at the top level, and the user can zoom into a spoligotype cluster to view host-pathogen associations within the cluster.

Visual analytics can also be incorporated into host-pathogen maps. The genotype clusters can be highlighted by default if the transmission rate within the cluster is higher than the background rate, which can be derived from the transmission rate of clinical data displayed. This automatic anomaly identification helps epidemiologists spot possible outbreaks.

V. Conclusion

We developed host-pathogen maps to simultaneously visualize genetically similar tuberculosis strains and attributes of patients infected by them. The view from perspectives of both the host and the pathogen helps molecular epidemiologists and public health workers understand trends and anomalies in host pathogen data. These can help distinguish latent reactivation of disease from recent transmission and identify possible outbreaks which can then be verified by more traditional epidemiology investigations helping to make more effective use of short public health resources. The TB-Vis web tool is flexible allowing any biomarkers or host properties to be used as the user desires. Using host-pathogen maps, the users can search for a known genotype cluster or identify a genotype cluster they spotted on the tree-map. This interactive nature of the host-pathogen map lets the users find what they might have missed in data analysis by visual recognition. The general approach and the TB-Vis tool (http:\\tbinsight.cs.rpi.edu) are generic and thus applicable to other diseases and biomarkers. Host-pathogen maps can be further extended to include more patient attributes such as age range or risk factors by varying the patient node shape.

References

[1] M. M. Tanaka and A. R. Francis, "Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information," *Infection, Genetics and Evolution*, vol. 5, no. 1, pp. 35 – 43, 2005.

[2] J. Reyes, A. Francis, and M. Tanaka, "Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes," *BMC Bioinformatics*, vol. 9, no. 1, p. 496, 2008.

[3] A. Shabbeer, C. Ozcaglar, B. Yener, K. P. Bennett, "Web tools for molecular epidemiology of tuberculosis," *Infection, Genetics and Evolution*, in press, 2011.

[4] B. Shneiderman, "Tree visualization with tree-maps: A 2-d space-filling approach," *ACM Transactions on Graphics*, vol. 11, pp. 92–99, 1991.

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

[6] M. Aminian, A. Shabbeer, and K. P. Bennett, "A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages," *BMC Bioinformatics*, vol. 11, no. Suppl 3, p. S4, 2010.

[7] J. R. Glynn, J. Whiteley, P. J. Bifani, K. Kremer, and D. van Soolingen, "Worldwide occurrence of beijing/w strains of mycobacterium tuberculosis: a systematic review," *Emerging Infectious Diseases*, vol. 8, no. 8, pp. 843 – 849, 2002.

[8] P. J. Bifani, B. Mathema, N. E. Kurepina, and B. N. Kreiswirth, "Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains," *Trends in Microbiology*, vol. 10, no. 1, pp. 45 – 52, 2002.

[9] M. A. Behr and P. M. Small, "Molecular fingerprinting of mycobacterium tuberculosis: How can it help the clinician?" *Clinical Infectious Diseases*, vol. 25, no. 4, pp. 806–810, 1997.