Detecting mixed *Mycobacterium tuberculosis* infection and differences in drug susceptibility with WGS data

Arlin Keo* & Thomas Abeel1,2,*

1) Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands  
2) Broad Institute of MIT and Harvard, Cambridge, MA, USA  
*t.abeel@tudelft.nl

*Mycobacterium tuberculosis* is a clonal, bacterial pathogen that causes the pulmonary disease tuberculosis (TB) and infects and kills millions of people [1]. The study of genetic diversity within the *M. tuberculosis* complex (MTBC) is complicated by mixed TB infections, which happen when a person is infected with more than one distinct strain type of MTBC. This often results in poor diagnosis and treatment of patients as the bacterial subpopulation may have undetected differences in drug susceptibility [2].

Whole genome sequencing (WGS) yields a great number of single nucleotide polymorphisms (SNPs) and offers an increased resolution to distinguish distinct strains of MTBC [3]. Here, I present a tool that maps sample reads against 21 bp cluster-specific SNP markers to detect a possible mixed infection and estimate the frequencies of present subpopulations.

MTBC SNP-based phylogenetic tree with hierarchical clusters

A SNP-based phylogenetic tree was constructed from a global dataset of 5992 MTBC samples. In this bifurcating tree, each branch that represents a cluster of strains splits into two new monophyletic subclusters of strains that are genetically more closely related. These “spits” were used to define clusters and subclusters of ≥10 strains, indicated by the vertical dark and light gray bars, colored bars indicate lineage 1-6 and *M. bovis*. 308 Clusters (MTBC root excluded) were defined.

Global SNP association

Global SNP association was done for each hierarchical cluster to get cluster-specific SNPs. Based on the number of samples that harbor a specific SNP, a SNP was associated to a cluster when the true positive rate, true negative rate, positive predictive value, and negative predictive value were >0.95.

Detect mixed TB infection and estimate frequencies

By mapping sample reads against these 21 bp cluster-specific SNP markers the tool determines the presence of paths (with present clusters) in the phylogenetic tree that start at the MTBC root node. Paths that split indicate the presence of multiple strains and thus a mixed infection.

Drug resistance detection

Drug resistance SNP

Mixed infection samples with ambiguous drug resistance mutations indicate drug susceptibility differences of the present subpopulations.

Prevalence of mixed TB infections in global dataset

7661 TB samples were tested, present strain(s) and frequencies could be predicted for 7495 samples of which 914 (~12%) are mixed infections.

<table>
<thead>
<tr>
<th>Number of subpopulations</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6581</td>
</tr>
<tr>
<td>2</td>
<td>798</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td>&gt;3</td>
<td>21</td>
</tr>
</tbody>
</table>

References