High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster

Anita C. Schürch, Kristin Kremer, Olaf Daviena, Albert Kiers, Martin J. Boeree, Roland J. Siezen and Dick van Soolingen

Published Ahead of Print 30 June 2010.

Updated information and services can be found at:
http://jcm.asm.org/content/48/9/3403

REFERENCES

This article cites 25 articles, 8 of which can be accessed free at:
http://jcm.asm.org/content/48/9/3403#ref-list-1

CONTENT ALERTS

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»
Molecular typing contributes significantly to our understanding of the epidemiology of tuberculosis. A variety of genetic markers, such as IS6110 restriction fragment length polymorphism (RFLP) and variable-number tandem repeat (VNTR) typing, are currently used for DNA fingerprinting of Mycobacterium tuberculosis isolates (2, 7, 12–14, 23). Unfortunately, these markers do not distinguish primary and subsequent sources of infection in long-term DNA fingerprinting surveillance, as the turnover of these markers is not in range with the pace of transmission (4–6). Therefore, molecular typing is inaccurate when applied for extended time periods in a given area.

In the Netherlands, IS6110 RFLP typing has been routinely used for molecular epidemiology since the early 1990s. A remarkably large outbreak began in the city of Harlingen in 1992, and this cluster grew to over 100 cases in 2008 and is still expanding (10, 11). Although a small subset of isolates of this cluster exhibited a single transposition or deletion of IS6110, it soon became impossible to distinguish sources of infection and secondary and subsequent cases in the cluster. Some contact chains in the Harlingen cluster were suggested by contact tracing, performed according to the stone-in-the-pond principle (15, 25), but the exact transmission chains could not be validated by fingerprinting of the M. tuberculosis isolates, as most of the isolates revealed the same DNA fingerprints.

For this study, three isolates from two chains of transmission in the Harlingen cluster that could be accurately determined by contact tracing were selected for genome sequencing (Fig. 1). The bacterial isolates exhibited no change in antituberculosis drug resistance or any other observable change in phenotype. Sequencing and analysis of strains SH1 and SH5, as well as the tempo and mode of evolutionary changes between these two isolates, were described in one of our earlier studies (19). The DNA of strain SH9, purified according to the method of van Soolingen et al. (24), was de novo sequenced on a GS FLX Titanium system, and assembly of raw sequencing reads with an average read length of 400 bases was performed by using the Genome Sequencer software, version 2.0.0.22. Sequence reads, contigs, and quality scores were provided by Microsynth AG, Switzerland. The SH9 sequence consisted of 214,283,462 high-quality bases assembled in 401 contigs with 4,207,440 bases (50.9-fold coverage). In total, 95.4% of the theoretical genome size of 4.41 Mb was available for analysis. From the in silico comparison of the three genomes, eight polymorphic single nucleotide polymorphisms (SNPs) were verified by subsequent resequencing on an ABI 3730xl sequencer (Table 1).

All 104 isolates of the IS6110 Harlingen cluster were tested for the presence of these eight SNPs. The bacterial isolates were designated with an “S” (for strain) followed by their patient number (for example, H44 for Harlingen patient 44). Identified polymorphic sites were concatenated, and these aligned sequences were clustered using a neighbor-joining al-
algorithm with ClustalW version 2.0.1 (22), which divided the Harlingen cluster into five SNP clusters (Fig. 2A).

To assign index cases within an SNP cluster, the dates of isolation of the strains were included for isolates of patients in one of the contact chains (Fig. 1) and for isolates in one of the new SNP clusters. The patient with the earliest isolate in each SNP cluster was defined as the index case. The earliest identified case is, however, not necessarily the first source of transmission. Delays in the timely diagnosis of tuberculosis may occur because of differences in health care-seeking behavior and lack of symptoms or expertise of care providers. In our study, the contact tracing information supported the index cases assigned by this model. For all other isolates, the time that had passed since the diagnosis of the index case was calculated and added as separate branches in Fig. 2B.

<table>
<thead>
<tr>
<th>SNP no.</th>
<th>Coordinate in strain H37Rv</th>
<th>Single nucleotide polymorphism</th>
<th>Polymorphisms first identified between the genome sequences of strains</th>
<th>Polymorphisms first identified between strains SH1 and SH5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>332,437</td>
<td>G→A</td>
<td>SH1 and SH5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2,187,212</td>
<td>T→G</td>
<td>SH1 and SH5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3,611,558</td>
<td>G→A</td>
<td>SH1 and SH5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4,388,976</td>
<td>G→C</td>
<td>SH1 and SH5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1,483,748</td>
<td>C→T</td>
<td>SH1 and SH9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2,976,989</td>
<td>T→C</td>
<td>SH1 and SH9</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3,165,490</td>
<td>G→A</td>
<td>SH1 and SH9</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3,904,206</td>
<td>T→C</td>
<td>SH1 and SH9</td>
<td></td>
</tr>
</tbody>
</table>

* Strains are referred to by an “S” followed by the number of the patient from which the strain was isolated.

---

**TABLE 1.** Single nucleotide polymorphisms identified by genome sequencing of three *Mycobacterium tuberculosis* Harlingen isolates and used as markers in this study

---

**FIG. 2.** (A) Clustering of the single nucleotide polymorphism (SNP) types of the isolates of the Harlingen IS6110 restriction fragment length polymorphism (RFLP) cluster. Application of the SNP markers to the isolates of the Harlingen cluster divided the IS6110 RFLP cluster into SNP clusters. The SNP types of the clusters are shown as an eight-position SNP code. Positions one to four indicate SNPs that were identified between strains SH1 and SH5 and positions five to eight represent SNPs that were identified between strains SH1 and SH9 (Table 1). Positions in red represent acquired SNPs, and underlined positions represent mutation events. (B) Dendrogram of clustered SNP types with integration of a time scale. All bacterial strains are designated with an “S” followed by the respective patient number. Only isolates of patients of one of the contact chains or strains with an SNP type different from the SNP type of SH1 were used for clustering. Branches in bold indicate the neighbor-joining-based clustering of the SNP types. Other horizontal branches indicate the number of days that have passed since the isolation date of the first patient isolate in each transmission cluster. Samples with the same color belong to one SNP cluster. Black, Harlingen cluster; red, SNP cluster with index case H3; purple, SNP cluster with index case H4.a; blue, SNP cluster with index case H4.c; green, SNP cluster with index case H9. A bacterial isolate of patient H6 (SH6) is missing from the Dutch tuberculosis database, because this patient was not diagnosed in the Netherlands.
The isolates of patients H7 and H8 exhibited no polymorphisms compared to SH1. The suggested contact chain leading to patient H9 was therefore not confirmed by this study. Moreover, no other isolates with the SNPs 5 to 8 were found in the Harlingen cluster. As all diagnosed active tuberculosis cases in the Netherlands are part of the Dutch database, it can be assumed that either patient H9 did not infect other persons in the Netherlands or infected cases did not (yet) progress to active disease.

After careful considerations of the costs and the technological limitations, genome sequencing will probably become the new standard method for typing of *M. tuberculosis* in the future and will replace existing typing methods because of its higher discriminatory power (16) and the decreasing prices for sequencing (8, 18, 26). Genome sequencing will elucidate transmission chains among patients that are clustered by currently used DNA fingerprinting methods. Future sequencing techniques will probably identify more polymorphisms between isolates because of technical progress, such as longer sequencing reads (1, 9, 21) and improved read accuracies. When genome sequencing of *M. tuberculosis* isolates becomes routinely available, identification, prediction of drug resistance, and epidemiological typing can be included in a single rapid analysis (3, 26, 27). In summary, we expect that genome sequencing will become a useful diagnostic tool with unprecedented possibilities.

We thank the staff at the Tuberculosis Reference Laboratory, RIVM, for technical assistance. This work was supported by the RIVM SOR project S/230136/01/GA and by the EU-supported TBadapt project 37919.

**REFERENCES**


