Journal of Clinical Microbiology

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

J. Clin. Microbiol. **2010, 48(9):3403. DOI: 10.1128/JCM.00370-10.**
Published Ahead of Print 30 June 2010.

JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2010, p. 3403–3406 Vol. 48, No. 9
0095-1137/10/$12.00 doi:10.1128/JCM.00370-10

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

Anita C. Schürch,[1,2] Kristin Kremer,[1] Olaf Daviena,[1] Albert Kiers,[3] Martin J. Boeree,[4] Roland J. Siezen,[2] and Dick van Soolingen[1,4]*

Tuberculosis Reference Laboratory, National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control, P.O. Box 1, 3720 BA Bilthoven, Netherlands[1]; Radboud University Nijmegen Medical Centre/NCMLS, Centre for Molecular and Biomolecular Informatics, P.O. Box 9101, 6500 HB Nijmegen, Netherlands[2]; Department of Tuberculosis Control GGD Fryslân, P.O. Box 601, 8901 BK Leeuwarden, Netherlands[3]; and University Centre for Chronic Diseases, Department of Pulmonary Disease, Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, Netherlands[4]

Received 23 February 2010/Returned for modification 19 April 2010/Accepted 22 June 2010

To investigate whether genome sequencing yields more useful markers than those currently used to study the epidemiology of tuberculosis, it was applied to three *Mycobacterium tuberculosis* isolates of the Harlingen outbreak. Our findings suggest that single nucleotide polymorphisms can be used to identify transmission chains in restriction fragment length polymorphism clusters.

Molecular typing contributes significantly to our understanding of the epidemiology of tuberculosis. A variety of genetic markers, such as IS*6110* 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, IS*6110* 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 IS*6110*, 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 IS*6110* 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-

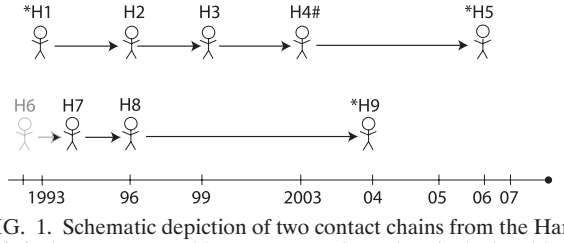

FIG. 1. Schematic depiction of two contact chains from the Harlingen (H) cluster suggested by contact tracing. Time is depicted by the horizontal bar; patients are numbered H1 to H9. Bacterial isolates of patients marked with an asterisk were subjected to genome sequencing. The bacterial isolates are designated with an "S" (for strain) followed by the patient number (e.g., SH1, SH5, and SH9). A bacterial isolate of patient H6 was not available for analysis. #, after being diagnosed with tuberculosis in 2003, patient H4 experienced two relapses of the disease in 2004 and 2005 (referred to as H4.a, H4.b, and H4.c in Fig. 3).

* Corresponding author. Mailing address: National Mycobacteria Reference Laboratory, National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control (CIb/LIS, pb22), P.O. Box 1, 3720 BA Bilthoven, Netherlands. Phone: 31-30-2742363. Fax: 31-30-2744418. E-mail: Dick.van.Soolingen@rivm.nl.

▽ Published ahead of print on 30 June 2010.

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

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

[a] Strains are referred to by an "S" followed by the number of the patient from which the strain was isolated.

gorithm 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.

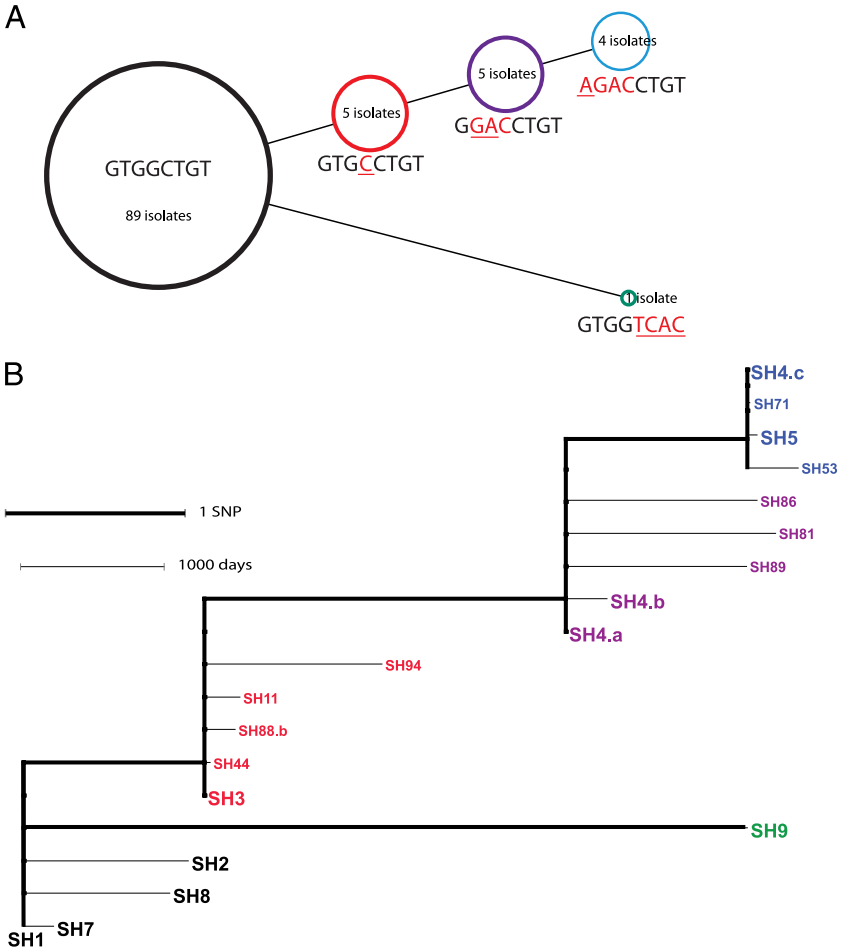

FIG. 2. (A) Clustering of the single nucleotide polymorphism (SNP) types of the isolates of the Harlingen IS*6110* restriction fragment length polymorphism (RFLP) cluster. Application of the SNP markers to the isolates of the Harlingen cluster divided the IS*6110* RFLP cluster into SNP clusters. The 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.

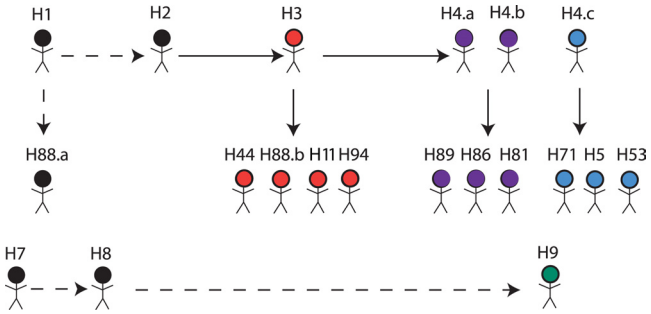

FIG. 3. Most likely transmission scheme suggested by single nucleotide polymorphism (SNP) typing, temporal, and contact tracing data. Black arrows indicate the most likely transmission events based on the SNP type clustering and integration of temporal data and supported by contact tracing information. Arrows with dashed lines indicate transmission events suspected based only on contact tracing information. Stickmen with the same color belong to the same SNP cluster. Black, Harlingen cluster with index case H1; 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.

By applying genome sequencing data, the number of possible infection routes was minimized for all patients. For patients whose isolates were part of the newly identified SNP clusters, a most likely scheme of transmission events was created (Fig. 3).

For one patient who had had two episodes of tuberculosis, in 1996 (isolate SH88.a) and 1999 (isolate SH88.b), SNP typing confirmed a suspected reinfection by the Harlingen strain from a secondary source in the Harlingen outbreak and excluded a possible endogenous reactivation of the disease. The isolate of the first disease episode of this patient was part of the largest cluster (SNP type GTGGCTGT), and the isolate of the second disease episode exhibited SNP type GTGCCTGT (underlining represents a mutation event), which supported a reinfection by patient H3. Patient H88 was married to patient H3; a reinfection by his partner is in agreement with the contact tracing information.

Genome sequencing of three isolates of the Harlingen cluster allowed us to determine SNPs and distinguish between isolates with higher discriminatory power than IS*6110* RFLP and identified separate transmission chains within the cluster. These SNPs (and their discriminatory power) are specific for the Harlingen cluster and, for example, are not present in strain H37Rv, and determination of SNPs in other clusters requires genome sequencing of patient isolates of those clusters. Although the sequencing of only a few isolates of the Harlingen outbreak led to a phylogenetic discovery bias (17, 20), i.e., other samples of this IS*6110* RFLP cluster most likely have other SNPs in addition to the ones reported here, it did not prohibit the disclosure of secondary and tertiary sources (patients H3 and H4) in the first transmission chain. Moreover, the first and second isolates of patient H4 exhibited the same SNP types as isolates of three other patients (patients H81, H86, and H89) that were infected by this case. In a later year, patient H4 had a reactivation of tuberculosis, and the isolated bacteria had another SNP type, also found in the isolates of another three patients (including patient H5) who were supposedly infected during the reactivation stage of patient H4 (Fig. 3).

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

1. **Ansorge, W. J.** 2009. Next-generation DNA sequencing techniques. N. Biotechnol. **25:**195–203.
2. **Brudey, K., J. R. Driscoll, L. Rigouts, W. M. Prodinger, A. Gori, S. A. Al-Hajoj, C. Allix, L. Aristimuno, J. Arora, V. Baumanis, L. Binder, P. Cafrune, A. Cataldi, S. Cheong, R. Diel, C. Ellermeier, J. T. Evans, M. Fauville-Dufaux, S. Ferdinand, D. Garcia de Viedma, C. Garzelli, L. Gazzola, H. M. Gomes, M. C. Guttierez, P. M. Hawkey, P. D. van Helden, G. V. Kadival, B. N. Kreiswirth, K. Kremer, M. Kubin, S. P. Kulkarni, B. Liens, T. Lillebaek, M. L. Ho, C. Martin, C. Martin, I. Mokrousov, O. Narvskaia, Y. F. Ngeow, L. Naumann, S. Niemann, I. Parwati, Z. Rahim, V. Rasolofo-Razanamparany, T. Rasolonavalona, M. L. Rossetti, S. Rusch-Gerdes, A. Sajduda, S. Samper, I. G. Shemyakin, U. B. Singh, A. Somoskovi, R. A. Skuce, D. van Soolingen, E. M. Streicher, P. N. Suffys, E. Tortoli, T. Tracevska, V. Vincent, T. C. Victor, R. M. Warren, S. F. Yap, K. Zaman, F. Portaels, N. Rastogi, and C. Sola.** 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. **6:**23.
3. **Chong, C. E., B. S. Lim, S. Nathan, and R. Mohamed.** 2006. *In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. In Silico Biol. **6:**341–346.
4. **de Boer, A. S., K. Kremer, M. W. Borgdorff, P. E. de Haas, H. F. Heersma, and D. van Soolingen.** 2000. Genetic heterogeneity in *Mycobacterium tuberculosis* isolates reflected in IS*6110* restriction fragment length polymorphism patterns as low-intensity bands. J. Clin. Microbiol. **38:**4478–4484.
5. **de Vries, G., H. W. Baars, M. M. Sebek, N. A. van Hest, and J. H. Richardus.** 2008. Transmission classification model to determine place and time of infection of tuberculosis cases in an urban area. J. Clin. Microbiol. **46:**3924–3930.
6. **Glynn, J. R., E. Vynnycky, and P. E. Fine.** 1999. Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. Am. J. Epidemiol. **149:**366–371.
7. **Grant, A., C. Arnold, N. Thorne, S. Gharbia, and A. Underwood.** 2008. Mathematical modelling of *Mycobacterium tuberculosis* VNTR loci estimates a very slow mutation rate for the repeats. J. Mol. Evol. **66:**565–574.
8. **Gupta, P. K.** 2008. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. **26:**602–611.

9. **Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F. X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman, and G. Dougan.** 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. Nat. Genet. **40:**987–993.
10. **Kiers, A., A. P. Drost, D. van Soolingen, and J. Veen.** 1996. Border-crossing source tracing in tuberculosis via DNA fingerprint technique. Ned. Tijdschr. Geneeskd. **140:**2290–2293. (In Dutch.)
11. **Kiers, A., A. P. Drost, D. van Soolingen, and J. Veen.** 1997. Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in the Netherlands. Int. J. Tuberc. Lung Dis. **1:**239–245.
12. **Kik, S. V., S. Verver, D. van Soolingen, P. E. de Haas, F. G. Cobelens, K. Kremer, H. van Deutekom, and M. W. Borgdorff.** 2008. Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. Am. J. Respir. Crit. Care Med. **178:**96–104.
13. **Kremer, K., B. K. Y. Au, P. C. W. Yip, R. Skuce, P. Supply, K. M. Kam, and D. van Soolingen.** 2005. Use of variable-number tandem-repeat typing to differentiate *Mycobacterium tuberculosis* Beijing family isolates from Hong Kong and comparison with IS*6110* restriction fragment length polymorphism typing and spoligotyping. J. Clin. Microbiol. **43:**314–320.
14. **Kremer, K., D. van Soolingen, R. Frothingham, W. H. Haas, P. W. M. Hermans, C. Martin, P. Palittapongarnpim, B. B. Plikaytis, L. W. Riley, M. A. Yakrus, J. M. Musser, and J. D. A. van Embden.** 1999. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. **37:**2607–2618.
15. **Lambregts-van Weezenbeek, C. S., M. M. Sebek, P. J. van Gerven, G. de Vries, S. Verver, N. A. Kalisvaart, and D. van Soolingen.** 2003. Tuberculosis contact investigation and DNA fingerprint surveillance in the Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. Int. J. Tuber. Lung Dis. **7:**S463–S470.
16. **Niemann, S., C. U. Koser, S. Gagneux, C. Plinke, S. Homolka, H. Bignell, R. J. Carter, R. K. Cheetham, A. Cox, N. A. Gormley, P. Kokko-Gonzales, L. J. Murray, R. Rigatti, V. P. Smith, F. P. Arends, H. S. Cox, G. Smith, and J. A. Archer.** 2009. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS One **4:**e7407.
17. **Pearson, T., R. T. Okinaka, J. T. Foster, and P. Keim.** 2009. Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect. Genet. Evol. **9:**1010–1019.
18. **Pettersson, E., J. Lundeberg, and A. Ahmadian.** 2009. Generations of sequencing technologies. Genomics **93:**105–111.
19. **Schürch, A. C., K. Kremer, A. Kiers, O. Daviena, M. J. Boeree, R. J. Siezen, N. H. Smith, and D. van Soolingen.** 2010. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infect. Genet. Evol. **10:**108–114.
20. **Smith, N. H., R. G. Hewinson, K. Kremer, R. Brosch, and S. V. Gordon.** 2009. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat. Rev. Microbiol. **7:**537–544.
21. **Tettelin, H., and T. Feldblyum.** 2009. Bacterial genome sequencing. Methods Mol. Biol. **551:**231–247.
22. **Thompson, J. D., T. J. Gibson, and D. G. Higgins.** 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics **2:**Unit 2.3.
23. **van Deutekom, H., S. P. Hoijng, P. E. de Haas, M. W. Langendam, A. Horsman, D. van Soolingen, and R. A. Coutinho.** 2004. Clustered tuberculosis cases: do they represent recent transmission and can they be detected earlier? Am. J. Respir. Crit. Care Med. **169:**806–810.
24. **van Soolingen, D., P. De Haas, and K. Kremer.** 2001. Restriction fragment length polymorphism typing of mycobacteria, p. 165–203. *In* T. Parish and N. G. Stoker (ed.), *Mycobacterium tuberculosis* protocols. Humana Press Inc., Totowa, NJ.
25. **Veen, J.** 1992. Microepidemics of tuberculosis: the stone-in-the-pond principle. Tuber. Lung Dis. **73:**73–76.
26. **Voelkerding, K. V., S. A. Dames, and J. D. Durtschi.** 2009. Next-generation sequencing: from basic research to diagnostics. Clin. Chem. **55:**641–658.
27. **Willcox, P. A.** 2000. Drug-resistant tuberculosis. Curr. Opin. Pulm. Med. **6:**198–202.