Whole-Genome-Based *Mycobacterium tuberculosis* Surveillance: a Standardized, Portable, and Expandable Approach

Thomas A. Kohl, Roland Diehl, Dag Harmsen, Jörg Rothgänger, Karen Meywald Walter, Matthias Merker, Thomas Weniger, Stefan Niemann

Molecular Mycobacteriology, Forschungszentrum Borstel, Borstel, Germany; Institute for Epidemiology, Schleswig-Holstein University Hospital, Kiel, Germany; Department of Periodontology, University Hospital Münster, Münster, Germany; Ridom GmbH, Münster, Germany; Public Health Department Hamburg-Central, Hamburg, Germany; German Center for Infection Research, Borstel Site, Borstel, Germany

Whole-genome sequencing (WGS) allows for effective tracing of *Mycobacterium tuberculosis* complex (MTBC) (tuberculosis pathogens) transmission. However, it is difficult to standardize and, therefore, is not yet employed for interlaboratory prospective surveillance. To allow its widespread application, solutions for data standardization and storage in an easily expandable database are urgently needed. To address this question, we developed a core genome multilocus sequence typing (cgMLST) scheme for clinical MTBC isolates using the Ridom SeqSphere software, which transfers the genome-wide single nucleotide polymorphism (SNP) diversity into an allele numbering system that is standardized, portable, and not computationally intensive. To test its performance, we performed WGS analysis of 26 isolates with identical IS6110 DNA fingerprints and spoligotyping patterns from a longitudinal outbreak in the federal state of Hamburg, Germany (notified between 2001 and 2010). The cgMLST approach (3,041 genes) discriminated the 26 strains with a resolution comparable to that of SNP-based WGS typing (one major cluster of 22 identical or closely related and four outlier isolates with at least 97 distinct SNPs or 63 allelic variants). Resulting tree topologies are highly congruent and grouped the isolates in both cases analogously. Our data show that SNP- and cgMLST-based WGS analyses facilitate high-resolution discrimination of longitudinal MTBC outbreaks. cgMLST allows for a meaningful epidemiological interpretation of the WGS genotyping data. It enables standardized WGS genotyping for epidemiological investigations, e.g., on the regional public health office level, and the creation of web-accessible databases for global TB surveillance with an integrated early warning system.

Tuberculosis (TB) is a global health challenge, with more than one-third of the world’s population infected, around eight million new cases annually, and about 1.5 million deaths every year (1). This global TB epidemic is accelerated by high HIV/TB coinfection rates, e.g., in Sub-Saharan Africa, and the emergence of resistant, multidrug resistant (MDR), and extensively drug resistant (XDR) *Mycobacterium tuberculosis* complex (MTBC) strains, particularly in Eastern Europe, Asia, and some parts of Africa (1, 2). Importantly, recent studies applying molecular strain typing indicated that transmission of MDR strains rather than insufficient treatment is one major driving force for the actual MDR epidemic (3–5).

This illustrates the need to precisely define the factors driving the epidemic locally or on a global level. Of special importance is the accurate tracing of pathogen transmission to develop optimized TB control strategies (3). For clinical MTBC isolates, three classical genotyping techniques have been used during the last few years, IS6110 restriction fragment length polymorphism (RFLP) typing, spoligotyping (clustered regularly interspaced palindromic repeats [CRISPRs]), and mycobacterial interspersed repetitive-unit–variable-number tandem-repeat (MIRU-VNTR) typing of up to 24 loci (7–9). Classical genotyping has been applied to a variety of research questions ranging from local outbreak analyses and longitudinal molecular epidemiological studies to analysis of global population structure, global spread of particular variants, and host pathogen coevolution (10–13).

While classical genotyping methods such as IS6110 DNA fingerprinting have been widely used during previous years, recent studies using whole-genome sequencing (WGS) analysis indicate that these methods lack resolution power to accurately determine transmission chains (3–6). WGS-based genotyping appears to offer an optimal resolution of MTBC isolates in molecular epidemiological studies with the advantage that additional information (e.g., on drug resistance) can be retrieved easily from the sequencing data (3, 6).

One major caveat for using WGS-based genotyping is the inherent difficulty of data standardization and integration into a readily accessible and expandable classification scheme. One way to overcome these problems is a genome-wide gene-by-gene analysis extending multilocus sequence typing to the genome level (core genome MLST [7]). By transferring genome-wide single nucleotide polymorphism (SNP) diversity into an allele-numbering system, the cgMLST (or MLST approach) allows for standardized WGS-based genotyping, the creation of web-accessible databases such as BIGSdb (7), and nomenclature servers (8). cgMLST has been successfully applied for a few pathogens such as *Strepto*
**coccus pneumoniae, Escherichia coli, and Neisseria meningitidis (7, 9, 10).** However, no data are available for highly monomorphic bacteria such as MTBC.

Here, we used a newly available software (SeqSphere+ version 1.0; Ridom GmbH, Münster, Germany) (11) to develop an MTBC cgMLST typing scheme and evaluated its performance in comparison with a genome-wide SNP-based approach for discrimination of a longitudinal MTBC outbreak. The outbreak comprised 26 patients (notified between 2001 and 2010) and has been defined by classical genotyping in a molecular epidemiological study in the federal state of Hamburg, Germany (3, 12). Next-generation sequencing (NGS) of all 26 patient isolates exhibiting identical IS6110 DNA fingerprint and spoligotyping patterns (Fig. 1) was performed with the SeqSphere and analyzed both by an SNP-based pipeline and cgMLST typing system. Reads were mapped to the H37Rv reference sequence (11) to develop an MTBC genotyping system. Reads were mapped to the H37Rv reference sequence and analyzed both by an SNP-based pipeline and cgMLST typing with the SeqSphere+ software. For data comparison, we calculated minimum spanning trees (MST) and performed a cluster analysis. In addition, the degree of correlation with contact tracing data was evaluated.

**MATERIALS AND METHODS**

**Study population.** Longitudinal prospective molecular epidemiological surveillance has been performed in Hamburg since 1 January 1997 (3, 12). All culture-confirmed TB cases obligatorily reported on the basis of the German Infection Protection Act to the Hamburg Public Health Department were prospectively enrolled in the study. Up to 31 December 2012, isolates from 2,150 patients have been analyzed by classical strain typing. Of these, 26 Haarlem strains showed identical IS6110 DNA fingerprint and spoligotype patterns and were chosen for further investigation by WGS in this study.

The molecular epidemiological investigation is embedded in mandatory routine surveillance and contact investigation work performed by the Public Health Offices according to the legal mandate of the German Infection Protection Act. It was approved by the Hamburg Commissioner for Data Protection.

**Experimental data sets.** Case data were collected prospectively by trained public health staff using a standardized questionnaire as reported previously (3).

**Classical genotyping.** Extraction of genomic DNA from mycobacterial strains, DNA fingerprinting using IS6110 as a probe, and spoligotyping were performed by use of standardized protocols, as described previously (13, 14).

**Whole-genome sequencing.** Isolated genomic DNA of individual strains was sequenced using the Illumina MiSeq sequencer, Nextera XT library preparation kits, and MiSeq reagent kits as instructed by the manufacturer (Illumina, San Diego, CA, USA). Resulting reads were mapped to the *M. tuberculosis* H37Rv genome (GenBank accession number NC_000962.3) using the exact alignment program SARUMAN (15). All isolates were sequenced with a minimum coverage of 50-fold.

**SNP-based analysis pipeline.** Single nucleotide polymorphisms were extracted from mapped reads by customized Perl scripts using a minimum coverage of 10 reads and a minimum allele frequency of 75% as thresholds for detection as reported previously (3). Variants were excluded if another SNP was detected within a window of 12 bases, if they had been reported as resistance conferring, or if they were located in repetitive regions in the genome (16).

**cgMLST-based analysis pipeline.** First, a cgMLST scheme was defined using the MLST+ Target Definer tool of the Ridom SeqSphere+ software (Ridom GmbH, Münster, Germany) with default settings. The finished genome of the *M. tuberculosis* strain H37Rv (GenBank accession number NC_000962.3) served as the reference genome (4,018 genes). Subsequently, query genomes were compared with the reference genome to establish a list of core genome genes. The following six query genomes were used: *M. tuberculosis* H67Rv (strain CDC1551 [NC_002755.2], strain F11 [NC_009565.1], and strain KZN 4207 [NC_016768.1]), *M. africanum* (strain GM041182 [NC_015758.1]), and *M. bovis* (strain BCG str. Pasteur 1173P2 [NC_008769.1] and strain AF2122/97 [NC_002945.3]). Here, default settings include the removal of the shorter in genome-based clusters. SNP- and cgMLST-based minimum
spanning trees were both calculated and drawn with the SeqSphere\textsuperscript{+} software.

**Nucleotide sequence accession number.** For all isolates, next-generation sequencing data have been submitted to ENA’s Sequence Read Archive (accession number PRJEB6276).

**RESULTS**

**Development of a cgMLST scheme.** To allow for standardized genome-based genotyping of clinical MTBC strains, we developed a cgMLST/MLST-based analysis pipeline. Using *M. tuberculosis* strain H37Rv (GenBank accession number NC_00962.3) as the reference genome (4,018 genes) and the genome of a further six MTBC strains as query genomes, we defined a standard set of 3,257 genes (76.8% of the whole reference genome) for the cgMLST scheme (see Materials and Methods).

We then evaluated the performance of the developed cgMLST scheme by testing its capacity to discriminate strains from a longitudinal outbreak caused by a strain of the Haarlem lineage affecting 26 patients during the time period from 2001 to 2010 in the city of Hamburg, Germany, in comparison with a classical genome-wide SNP approach. All strains had identical IS6110 DNA fingerprinting and spoligotyping patterns (Fig. 1). The outbreak has a mixed epidemiological composition, with, on the one hand, patients having confirmed transmission links and, on the other hand, a high proportion of patients for whom no link was established (see below). Therefore, this cluster is well suited for evaluation of the cgMLST approach and the added value of WGS-based genotyping for longitudinal molecular epidemiology of TB. The cluster comprised 17 men and 9 women with fully susceptible TB (Table 1). While the majority of patients were German born (14 out of 26 [54%]), initial contact tracing indicated the transmission of the strain to be favored by close contacts in a neighborhood setting involving persons with foreign nationality of mainly Turkish migration background (Table 1).

Contact tracing also indicated the majority of patients to have been infected by one super spreader (Fig. 2), primarily via personal contacts involving three different families and other social contacts over a period of 11 years. Overall, likely transmission links were established for 14 patients, 10 of whom were probably directly infected by the proposed super-spreader index patient (7679-03) (Fig. 2). Subsequent transmission events supposedly occurred in the family setting and via contacts during “joint smoking.”

However, despite in-depth contact tracing investigations, no epidemiological links could be established for the remaining 11 patients of mostly German nationality (8 out of 11 [73%]), including the first patient notified to be diseased with the outbreak strain in 2001.

**Evaluation of the cgMLST scheme in comparison with whole-genome SNP approach.** All 26 outbreak isolates were sequenced on the rapid Illumina MiSeq benchtop sequencer with a minimum coverage of 50-fold (average 84-fold). Using a genome-wide SNP-based analysis approach, we identified 322 SNPs that are variable between at least two outbreak isolates (see Table S1 in the supplemental material). Of these, 31 were detected outside coding sequences. The remaining 291 polymorphisms were subdivided into 176 nonsynonymous and 115 synonymous SNPs.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Country of birth</th>
<th>Age (yr)</th>
<th>Gender</th>
<th>Prior known contact to TB patient</th>
<th>Previous TB disease</th>
<th>Onset of TB disease (yr)</th>
<th>AFB positivity\textsuperscript{a}</th>
<th>Drug resistance</th>
<th>Presumed infection by or index case of cluster patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024-01</td>
<td>Germany</td>
<td>49</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2001</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4155-03</td>
<td>Germany</td>
<td>19</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2003</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6821-03</td>
<td>Germany</td>
<td>54</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8073-03</td>
<td>Iran</td>
<td>20</td>
<td>Female</td>
<td>Yes</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1114-03</td>
<td>Turkey</td>
<td>20</td>
<td>Female</td>
<td>Yes</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>164-04</td>
<td>Germany</td>
<td>6</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7679-03</td>
<td>Germany</td>
<td>23</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>9956-03</td>
<td>Nigeria</td>
<td>2</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2387-04</td>
<td>Germany</td>
<td>27</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2004</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>1608-04</td>
<td>Nigeria</td>
<td>15</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>2004</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6631-04</td>
<td>Germany</td>
<td>25</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>2004</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>7684-04</td>
<td>Germany</td>
<td>20</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2004</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8506-04</td>
<td>Greece</td>
<td>27</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2004</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3176-05</td>
<td>Germany</td>
<td>1</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2005</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3501-05</td>
<td>Turkey</td>
<td>39</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2005</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3543-05</td>
<td>Yugoslavia</td>
<td>23</td>
<td>Female</td>
<td>Yes</td>
<td>No</td>
<td>2005</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>10118-05</td>
<td>Germany</td>
<td>28</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>2005</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6753-05</td>
<td>Germany</td>
<td>26</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>2005</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>8779-06</td>
<td>Germany</td>
<td>26</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>2005</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6577-07</td>
<td>Germany</td>
<td>42</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2007</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>9952-07</td>
<td>Nigeria</td>
<td>24</td>
<td>Female</td>
<td>Yes</td>
<td>No</td>
<td>2007</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8370-04</td>
<td>Turkey</td>
<td>28</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2008</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>5829-05</td>
<td>Germany</td>
<td>24</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>2008</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>629-09</td>
<td>Iran</td>
<td>28</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2009</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3929-10</td>
<td>Turkey</td>
<td>74</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>2010</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2651-10</td>
<td>Turkey</td>
<td>26</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>2010</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Smear microscopy positive.
The total number of SNPs determined among the outbreak strains investigated appeared to be surprisingly high compared to the 85 SNPs we revealed by WGS genotyping of strains of another longitudinal outbreak involving 86 patients in Hamburg in a recent study (3).

To visualize the precise population structure and spreading of the cluster isolates, we calculated a minimum spanning tree (MST) based on the concatenated SNP sequences (Fig. 3). The SNP-based cluster analysis distinguished the 26 outbreak isolates into one major cluster with 22 isolates in which the maximum SNP distance is not larger than 10 SNPs (within the range proposed for isolates with an epidemiological link [6]) and four outlier isolates which are separated from the cluster by a minimum number of 97 distinct SNPs. Within the cluster, 13 isolates are grouped around the two central nodes (SNP-N1 and SNP-N2) comprising three and six isolates, respectively, mainly having single-, double-, or triple-SNP differences. SNP-N1 includes the main index patient (7679-03).

Applying default quality criteria (e.g., at least 10-fold coverage and no frameshift in a core gene) when analyzing exactly the same reference-mapped 26 WGS data sets with the SeqSphere+ software, 3,041 genes of the cgMLST scheme were present in all isolates with high quality. Therefore, all further cgMLST analysis was done with these 3,041 genes. Using the cgMLST approach based on 3,041 core genome genes (see Table S2 in the supplemental material), we identified 218 allele variants that discriminated the isolates in a comparable manner to the SNP-based approach. The overall topology of the MST based on the allele profiles (Fig. 4) was highly concordant with the results obtained for concatenated SNP sequences, with a major cluster of 22 isolates having a maximum difference of $<10$ allele variants and the separation of the four outlier isolates showing a distance of at least 63 allelic variants to the central cluster. Overall, distances between tree nodes are smaller when derived from the cgMLST scheme, and in three cases isolates separated from the main nodes by one or two SNPs, respectively, are now within the main node cgMLST-N1 (9956-03, 11114-03, and 4155-03). This indicates a slightly lower resolution power of the cgMLST scheme, also seen in a lower overall number of differences between isolates (218 allele variants compared to 322 SNPs).

When considered in more detail, it became obvious that only a very small subset of 25 SNPs differentiates the 22 isolates within the main cluster. Of these, 17 are retained in the cgMLST scheme causing allele variants of core genome genes. From the remaining eight SNPs, three are located in intergenic regions (at positions 78930, 1613084, and 2049377), two in genes not classified by the cgMLST target definer into the core genome consisting of 3,257 genes (Rv2124c and Rv2407), and three SNPs are located in core genome genes for which the sequencing data did not meet the SeqSphere+ quality criteria (Rv0174, Rv1108c, Rv2339).

Concordance of WGS genotyping with contact tracing data. When concordance of WGS genotyping with epidemiological data is considered, one important point is that WGS analysis following either approach grouped the majority of patients linked to the main index patient isolate (7679-03) either directly in the central node (SNP-N1 or cgMLST-N1, respectively), without any difference detected in the genomes, or with few differences around

---

The diagram in Fig. 2 shows established epidemiological links among the 26 cluster patients. The 26 patients are presented in boxes with the specific strain number. Established epidemiological links are visualized by color coding and position in the figure. All patients without a defined epidemiological link are in white boxes.
this central node, forming a star-like structure (Fig. 3 and 4) in support of the presence of a super spreader.

Both SNP- and cgMLST-based WGS analyses support 10 out of 14 established epidemiological links such as family environment or joint smoking, with a maximum of three SNPs or two allelic differences (Fig. 5). In four cases, SNP data suggest a different transmission chain with either an intermediate carrier (8073-03, 3543-05, or 9952-07), or in the case of 1608-04, direct transmission from the index patient. For cgMLST typing, only the previously assumed direct link between the index patient and isolate 9952-07 is refuted by its position in the MST.

WGS-based genotyping by both methods also suggests that 7 out of the 11 patients with no established epidemiological link from classical contact tracing are indeed part of the outbreak and are closely related to other outbreak isolates, with a maximum distance of two SNPs or one allele variant to the closest outbreak isolate. This indicates an epidemiological relationship with the other patients in the outbreak and involvement in a recent transmission chain that has not been detected by conventional contact tracing.

The four remaining isolates with no epidemiological links to the index patient had >60 SNPs/allele variants compared to the main cluster, thus clearly excluding recent transmission. Interestingly, two of these outlier isolates (1024-01 and 3929-10) had just five SNP differences from each other, pointing to a yet undefined recent transmission event not detected by contact tracing.

**DISCUSSION**

Our study demonstrates that a cgMLST approach based on approximately 3,000 core genome genes is suitable for high-resolution discrimination of MTBC isolates, thus opening the door for the widespread application of WGS-based strain typing for molecular epidemiology and local as well as global disease surveillance. cgMLST nearly fully keeps the discriminatory power and tree topology obtained by WGS genotyping based on SNPs. Due to its easy, nondemanding data format, it also allows for the development of web-based nomenclature servers that can facilitate global strain tracking and universal strain classification.

We and others have recently shown that WGS-based genotyping is superior to classical genotyping by offering a higher resolution for MTBC outbreaks and better spatiotemporal correlation with the spread of the pathogen (3–6). The data obtained in this study confirm a low variability of the MTBC genome under human-to-human transmission without any sign of a mutation burst as described previously (17). In line with previous studies (3, 6), the tree topology obtained could be used to better interpret the mode and timing of the transmission events. The star-like structure (SNP-N1 and cgMLST-N1) confirms the existence of a super spreader who infected at least 10 patients. Importantly, later transmission events are separated in the tree. Here, the WGS tree indicates the presence of a second super spreader linked to 8 patients without previously established epidemiological links in the “joint smoking” subbranch (Fig. 3 and 4). Likely these patients have not...
reported contacts because of the illegal drug use, pointing to the limited efficiency of conventional contact tracing as already observed in our previous work (12).

This also exemplifies the difficulties in interpreting results from classical molecular epidemiological studies in which an epidemiological link can usually be established only for a fraction of the clustered cases. Especially in crowded urban settings such as in the city of Hamburg, transmission via short, but intensive, contacts might play a significant role (3). Furthermore, definite transmission links are difficult to establish in mixed family/social milieu settings involving a high number of short contacts and complex interactions. As a consequence, different epidemiological scenarios are possible. Therefore, patient interview-independent information gained directly from the topology of the WGS tree is of the outmost importance for more informed contact tracing investigation.

This is in line with the clear separation of four isolates (15% of the 26 patients grouped into one cluster by classical genotyping) with a >60-SNP/allele variant difference from the index patient isolate clearly excluding an involvement in one recent transmission chain. Such large SNP differences among isolates with identical IS6110 patterns in a low-incidence setting is rather unexpected, as this has been considered a reliable marker for isolates belonging to one transmission chain (18). In this regard, we extend similar findings we made previously for two Beijing lineage isolates from an Eastern European high-incidence setting (19) to Haarlem lineage isolates from a Western European low-incidence setting, suggesting that classical genotyping is overestimating rates of recent transmission in general.

Although WGS-based genotyping analysis appears to be the optimal approach to trace pathogen transmissions, its widespread use is hampered by bioinformatic challenges in basic data analysis and standardization (6). At present, WGS data analysis mostly relies on SNP detection from reads mapped to a reference sequence, followed by calculation of phylogenetic trees from concatenated SNP sequences (3, 6). As this procedure depends on the chosen parameters of oftentimes highly customized in-house pipelines, comparisons across laboratories are nearly impossible. Furthermore, the creation of large databases necessary for longitudinal molecular epidemiological investigations is difficult, as the addition of new sets of isolates would require reanalysis of the total data set.

A possible solution to this problem has recently been suggested (7, 20) by extending the MLST concept to the genome level (cgMLST), meaning that genomic sequence data are analyzed by comparison to a set of loci (e.g., the genes of the core genome) and allele variants indexed. So far, cgMLST has been successfully applied for few pathogens such as S. pneumoniae and N. meningitidis (7, 10, 20).

However, cgMLST has not been explored for highly monomorphic pathogens such as MTBC with a very restricted level of genome variation that might still require a full-genome SNP approach for optimal resolution of clinical isolates (21). To investigate this question, we used the newly available SeqSphere" software.
ware for an MTBC cgMLST approach based on 3,041 genes. The resulting cluster analysis nearly fully resembled the data obtained by the genome-wide SNP approach, with only slightly lower resolution power (Fig. 3 and 4). The overall topologies of the SNP-based and allele-based MSTs are highly similar, completely retaining all information needed for epidemiologically informative reading of the tree, e.g., the star-like structure indicating the presence of a super spreader. Similarly, subsequent transmission events (family environment, joint smoking), as well as the clear separation of the four outlier isolates, are also displayed in the cgMLST minimum spanning tree.

We conclude that a cgMLST approach based on NGS data presents an ideal option for a more standardized way to analyze NGS sequence data for molecular epidemiological investigations of community-transmitted MTBC and other pathogens. General implementation of a cgMLST scheme will allow for meaningful data exchange between laboratories and the establishment of consistent online databases, e.g., using the BIGSdb system (7, 20). Further studies are needed to define parameters for cgMLST-based molecular epidemiological studies and the comparability of data sets generated with different platforms, sequencing chemistries, and laboratories. Another important issue is the agreement of the TB community on a standardized WGS typing scheme, which has been the main reason for the success of classical MTBC genotyping.

ACKNOWLEDGMENTS

We thank T. Ubben, I. Radzio, T. Struwe-Sonnenschein, and J. Zallet, Research Center Borstel, for their excellent technical assistance.

Parts of this work have been supported by grants from the European Community’s Seventh Framework Program (FP7/2007-2013) under grant agreement 278864 in the framework of the European Union PathoNGenTrace project and grant agreement 223681 in the framework of the TB-PAN-NET project.

The following authors have competing interests as defined by the Nature Publishing Group, or other interests that might be perceived to influence the results and/or discussion reported in this article. D.H., T.W., and J.R. are shareholders and T.W. and J.R. are employees of Ridom GmbH (Münster, Germany). The other authors have no competing interests.

REFERENCES


