



Consistency of *Mycobacterium tuberculosis* Complex Spoligotyping between the Membrane-Based Method and *In Silico* Approach

Charlotte Genestet,^{a,b}  Elisabeth Hodille,^{a,b} Albin Bernard,^a Maxime Vallée,^b  Gérard Lina,^{a,b} Adrien Le Meur,^c Guislaine Refrégier,^c Oana Dumitrescu,^a on behalf of the Lyon TB study group

^aCentre International de Recherche en Infectiologie (CIRI), Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon-1, Inserm, Lyon, France

^bHospices Civils de Lyon, Institut des Agents Infectieux, Laboratoire de Bactériologie, Lyon, France

^cCentre national de la recherche scientifique (CNRS), AgroParisTech, Ecologie Systématique et Evolution, Université Paris-Saclay, Gif-sur-Yvette, France

Charlotte Genestet and Elisabeth Hodille contributed equally to this article. Author order was determined in alphabetical order of surnames.

ABSTRACT To tackle the spread of tuberculosis (TB), epidemiological studies are undertaken worldwide to investigate TB transmission chains. Clustered regulatory interspaced short palindromic repeats (CRISPR) locus diversity, also called spoligotyping, is a widely used genotyping assay for the characterization of *Mycobacterium tuberculosis* complex (MTBC). We compared herein the spoligotyping of MTBC clinical isolates using a membrane-based method (following an initial PCR step) and whole-genome sequencing (WGS)-based method (i.e., *in silico* spoligotyping). All MTBC strains isolated at the Lyon University Hospital, France, between November 2016 and December 2020 were included ($n = 597$). Spoligotyping profiles were also used for species identification among the MTBC. Outputs of both methods were analyzed, and discrepant results were investigated thanks to CRISPRbuilder-TB. The overall agreement was 85.7%. Spacer discrepancies observed between the methods were due to the insertion of IS6110 within the direct repeat (DR) sequence upstream or downstream of spacers, mutated DR sequences, or truncated spacers. Discrepancies did not impact species identification. Although spoligotyping-based species identification was inconclusive for 29 isolates, SNP-based phylogeny conducted after WGS allowed the identification of 23 *M. tuberculosis* (Mtb), 2 *M. canettii*, and 4 mixed MTBC infections. WGS yielded very few discrepancies compared to membrane-based spoligotyping. Overall agreement was significantly improved (92.4%) by the CRISPR locus reconstruction using CRISPRbuilder-TB for the MTBC isolates with the shared international type 53 *in silico* spoligotyping. A smooth transition from the membrane-based to the *in silico*-based genotyping of *M. tuberculosis* isolates is, therefore, possible for TB diagnosis and epidemiologic survey.

IMPORTANCE Whole-genome sequencing (WGS) has profoundly transformed the perspectives of tuberculosis (TB) diagnosis, providing a better discriminatory power to determine relatedness between *Mycobacterium tuberculosis* complex (MTBC) isolates. Previous genotyping approaches, such as spoligotyping consisting of an initial PCR step followed by reverse dot hybridization, are currently being replaced by WGS. Several pipelines have been developed to extract a spoligotype from WGS data (*in silico* spoligotyping) allowing for the continuity of MTBC molecular surveys before and after WGS implementation. The present study found very good overall agreement between hybridization to membrane-based spoligotyping and *in silico* spoligotyping, indicating the possibility of a smooth transition from the traditional to the *in silico*-based genotyping of MTBC isolates for TB diagnosis and epidemiological survey.

Editor Florence Claude Doucet-Populaire, University Paris-Saclay, AP-HP Hôpital Antoine Bécélère, Service de Microbiologie, Institute for Integrative Biology of the Cell (I2BC), Commissariat à l'énergie atomique et aux énergies (CEA), CNRS

Copyright © 2022 Genestet et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Elisabeth Hodille, elisabeth.hodille@chu-lyon.fr.

The authors declare no conflict of interest.

Received 19 January 2022

Accepted 7 April 2022

Published 25 April 2022

KEYWORDS tuberculosis, *Mycobacterium tuberculosis* complex, spoligotyping, membrane-based spoligotyping, *in silico* spoligotyping, whole-genome sequencing, CRISPR-builder-TB

Control of *Mycobacterium tuberculosis* complex (MTBC) transmission in high-income and low tuberculosis (TB) prevalence countries remains a public health priority given the constant changes in MTBC epidemiology worldwide. Key measures for TB control rely on the linkage of cases and identification of transmission chains, through a population-based systematic molecular TB survey, to uncover outbreaks, even between unrelated cases (1, 2).

Two major genotyping assays for MTBC isolates have been developed and employed across numerous epidemiological studies: the spacer oligonucleotide typing (spoligotyping) and the mycobacterial identification repetitive unit-variable number of tandem repeats 15 (MIRU-VNTR15) typing (3–5). Historically, spoligotyping detected the presence or absence of 43 unique spacers in the direct repeat (DR) region of the clustered regularly interspaced short palindromic repeats (CRISPR) locus of MTBC. This was based on an initial PCR using primers directed to the most frequently occurring DR sequences, called DR0, followed by a reverse line blot hybridization membrane-based revelation method (6, 7). More recently, and using whole-genome sequencing (WGS), 68 spacers for the MTBC were identified (98 including *M. canettii*) (8). Because the diversity of the CRISPR locus has been shown to accurately reflect the phylogeny of MTBC, spoligotyping has not solely been used for epidemiological purposes but also for MTBC species identification using an algorithmic approach in routine TB diagnosis (9).

Since the advent of next-generation sequencing (NGS), WGS has been implemented in high-income countries and profoundly transformed the perspectives of TB diagnosis. WGS provides a better discriminatory power than spoligotyping and MIRU-VNTR15-typing to determine relatedness between MTBC isolates (10). In addition, WGS allows obtaining quick and accurate genotypic antimicrobial susceptibility testing and MTBC species identification using single nucleotide polymorphism (SNP) calling without prior specific PCR amplification (11). Moreover, several pipelines have been developed to extract a spoligotype from WGS data, also called *in silico* spoligotyping (12), enabling continuity of MTBC molecular surveys. Nevertheless, to ensure a smooth transition from the hybridization method to the *in silico*-based assay, the consistency of both outputs needs to be thoroughly investigated. It is well known that the initial amplification step of the former (that is not required for the latter) may introduce bias in the output sequences mainly due to heterogeneous or variant-sensitive primer affinity (13). In a recent study, Bogaerts et al. (14) compared these methods using 166 MTBC from the Belgian National Reference Center, but the discrepancies were not explored. In the present study, we compared WGS-based MTBC identification and spoligotyping (43 spacers) with those obtained by membrane hybridization assays in a French cohort. The discrepancies were further analyzed using CRISPRbuilder-TB, allowing a reconstruction of the MTBC CRISPR locus to understand the underlying causes.

RESULTS

Comparison between MTBC species identification using membrane-based spoligotyping and WGS SNP calling. A total of 597 MTBC isolates were analyzed, among which there were 4 cases of mixed infections (infection with 2 different strains of MTBC), not allowing species or lineage identification using spoligotyping. For the 593 remaining MTBC isolates, identification of MTBC species using membrane-based spoligotyping was feasible for 568 isolates (95.8%). For 25 MTBC isolates, spoligotyping did not allow species identification, including 21 “unknown” spoligotypes, i.e., spoligotyping was not related to an MTBC species in the SITVIT database, and 4 isolates with the shared international type (SIT) 2669 for which none of the 43 spacers was detected.

For these 568 isolates identified by membrane-based spoligotyping, all identifications were concordant with MTBC species found by WGS SNP calling. The majority

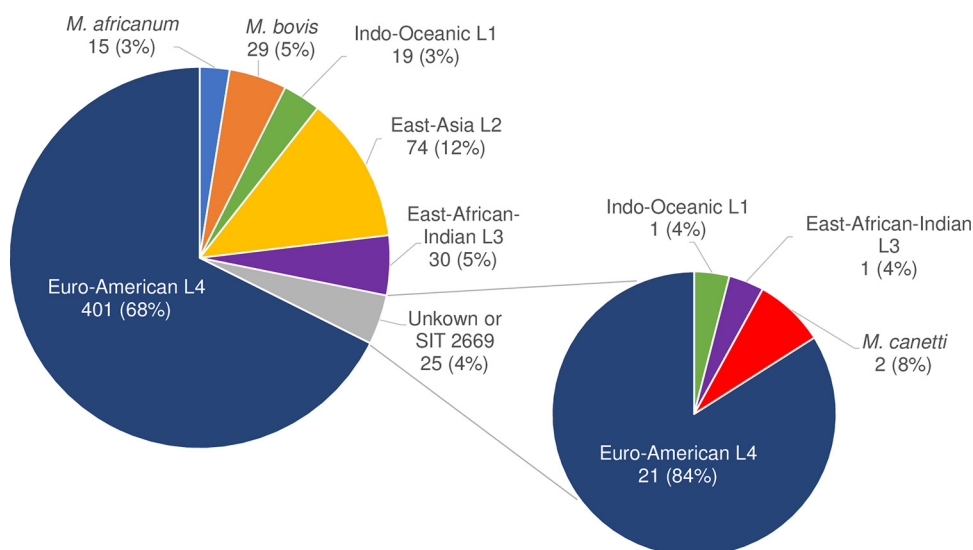


FIG 1 Identification of 593 MTBC isolates. Sector diagram on the left: membrane-based spoligotyping identification. Sector diagram on the right: whole-genome sequencing single nucleotide polymorphism calling identification for MTBC isolates classified as “Unknown” or standard international type (SIT) 2669 by spoligotyping identification.

were *M. tuberculosis* ($n = 524$, 92.3%), including 401 Euro-American lineage 4, 74 East-Asian lineage 2, 30 East African-Indian lineage 3, and 19 Indo-oceanic lineage 1, 29 *M. bovis*, and 15 *M. africanum* (Fig. 1). WGS SNP calling allowed MTBC species identification for the 25 MTBC isolates not identified by spoligotyping. Among these, 2 were *M. canettii* (spoligotyping 2669 for which no spacer was detected) and 23 were *M. tuberculosis*, including 21 Euro-American lineage 4 (including the other two SIT 2669), 1 East African-Indian lineage 3, and 1 Indo-oceanic lineage 1 (Fig. 1).

For the 4 cases of mixed MTBC infection, only WGS SNP calling allowed the accurate identification of the 2 MTBC isolates contained in the samples. In 3 cases, mixed infection was with the *M. tuberculosis* L2-Beijing strain and an *M. tuberculosis* L4-Euro-American strain, and in 1 case the mixed infection was with an *M. tuberculosis* L4-Euro-American strain and an *M. tuberculosis* L3-East African-Indian strain.

Concordance between membrane-based spoligotyping and *in silico* spoligotyping.

Excluding mixed infections, the overall agreement between membrane-based spoligotyping and *in silico* spoligotyping at the sample level was 85.7% (508/593; 95% confidence interval, 95%CI [82.6, 88.4], Table S1). Of the 85 isolates with discordant MTBC spoligotypes, 75 differed by only 1 spacer, and 10 isolates differed by 2 spacers.

At the spacer level, the most discordant was spacer 31, which concerned 61 isolates (71.8%). It was always absent in the membrane-based method but present in *in silico* spoligotyping (Table 1).

Among these discrepancies involving spacer 31, 39/61 (63.9%) were found in MTBC isolates with the SIT 50 on membrane-based spoligotyping and SIT 53 on *in silico* spoligotyping. Of note, in the study data set the SIT 50 was never obtained with *in silico* spoligotyping: all SIT 50 on membrane-based spoligotyping were SIT 53 on *in silico* spoligotyping (Table S1, Fig. 2). Except for spacer 31, Cohen's kappa indicated that the concordance between *in silico* and membrane-based spoligotyping indicated an almost perfect agreement at the spacer level (Cohen's kappa > 0.81; Table 1).

Discrepancy analysis. To better understand the discrepancies observed between membrane-based spoligotyping and *in silico* spoligotyping, the 85 discrepant spoligotypes were analyzed using CRISPRbuilder-TB, allowing a reconstruction of the CRISPR locus in MTBC (8). This found an insertion of the mobile element IS6110 at 5 nucleotides from the end of the DR sequence upstream of the spacers 6, 10, 11, 15, 20, 26, 31, or 32. An insertion of IS6110 at 6 nucleotides from the start of the DR sequence

TABLE 1 Discordant spacers between membrane-based spoligotyping and *in silico* spoligotyping

Presence or absence ^a	Discordant spacer	Prevalence in membrane-based spoligotyping, n	No. of concerned isolates	Overall agreement, %	Cohen's kappa	
"0" in membrane, "1" in <i>in silico</i>	6	447	1	99.8	0.99	
	10	419	4	99.3	0.98	
	11	440	2	99.7	0.99	
	14	444	5	99.2	0.98	
	15	441	7	98.8	0.97	
	20	444	3	99.5	0.99	
	26	389	4	99.3	0.99	
	31	320	61	89.7	0.79	
	32	429	2	99.7	0.99	
	38	544	1	99.8	0.99	
	39	515	1	99.8	0.99	
	42	522	2	99.7	0.98	
	"1" in membrane, "0" in <i>in silico</i>	18	447	1	99.8	0.99

^a"0" meaning the absence of the spacer; "1" meaning the presence of the spacer.

downstream of the spacer 31 was also found and was responsible for all conversions of SIT 50 according to membrane-based spoligotyping to SIT 53 according to *in silico* spoligotyping. The CRISPRbuilder-TB analysis found mutated DR sequences leading to DR other than DR0 for some of the discrepancies observed for the spacer 15 (DRb2 downstream of the spacer) and those observed for the spacer 42 (DR6 downstream of the spacer). It also found truncated spacers for the discrepant spacer 38 and some of the discrepancies observed for the spacers 14 and 15. Analysis of the 4 SIT 2669 for which no spacer of the 43 investigated by the conventional spoligotyping methods were detected using CRISPRbuilder-TB found that 2 *M. canettii* isolates had spacers of the 98-spacer spoligotyping format reported by Guyeux et al. (8). For the 2 *M. tuberculosis* isolates SIT 2669, DR/CRISPR regions were completely deleted. CRISPRbuilder-TB did not find any event for the discrepancies observed for the spacers 18 and 39.

DISCUSSION

The MTBC CRISPR locus is the preferential insertion site for the IS6110, possibly disrupting DR or adjacent spacer sequences (15). Both DR variations and IS insertion may hamper primer affinity resulting in incomplete or abortive DNA amplification, thus changing expected spoligotype patterns, despite the presence of spacers in the CRISPR locus (16–18). Thus, as observed in the present study, previous studies not using WGS reported that the insertion of IS6110 around the spacer 31 led to erroneous

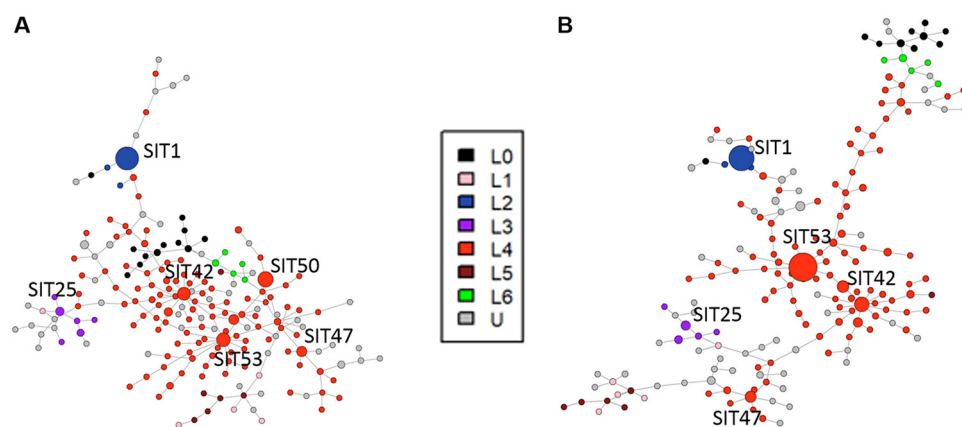


FIG 2 Minimum spanning trees of the test sample featuring spoligotype diversity as studied either using membrane-based spoligotyping or *in silico* WGS-based spoligotyping. (A) Membrane-based spoligotyping. (B) WGS-based spoligotyping. Nodes are colored according to the corresponding Lineage (see legend). L0: animal-adapted lineages, such as *M. bovis*; U: unknown. Node size corresponds to the population size with the corresponding spoligotype. Most prevalent patterns ($n > 7$) are labeled with their standard international type (SIT).

spoligotypes with the conventional PCR first-step methods (16–18), but, unlike herein, they did not report this for spacers 6, 10, 11, 15, 20, 26, and 32. These insertions of IS6110 within the DR sequence upstream or downstream of spacers probably lead to an asymmetrical split of the primer targets and a failure to detect the spacers in conventional spoligotyping methods using an initial PCR step. The presence of other genetic alterations of the CRISPR locus also explained the failure to detect the spacers by the conventional PCR first-step membrane-based spoligotyping methods such as mutations of the DR sequence (that is, likely to reduce the affinity of the primers; spacer 15 and 42), and the truncation of a spacer (that is, likely to prevent the detection by hybridization; spacers 14, 15, and 38).

The overall agreement between membrane-based and *in silico* spoligotyping at the sample level was 85.7% (508/593; 95% CI [82.8, 88.5]) despite the discrepancies due to genetic variations described above. Recently, Bogaerts et al. found a similar overall agreement (89.2%) on a set of 166 MTBC isolates (14). As observed herein, the spacer 31 concerned the most mismatched spacer, and in all cases of mismatch, it was detected using *in silico* spoligotyping and not the traditional spoligotyping method. Although most of these discrepancies were strains identified by membrane-based spoligotyping as SIT 50 that were identified by *in silico* spoligotyping as SIT 53, these followed a distinct evolutionary pathway marked by the insertion of IS6110 within the DR sequence downstream the spacer 31 (6, 19, 20). Due to the phylogenetic relevance of this insertion, the classification of the corresponding strains as SIT 50 instead of SIT 53 should be preferred. Thus, to prevent misclassification, for strains assigned to SIT 53 by *in silico* spoligotyping, analysis of WGS data using CRISPRbuilder-TB may restore the SIT 50 pattern for those harboring the IS6110 insertion within the DR sequence downstream of the spacer 31. This approach would have significantly improved the overall agreement between membrane-based spoligotyping and *in silico* spoligotyping herein from 85.7% (508/593; 95%CI [82.6, 88.4]) to 92.4% (547/593; 95%CI [89.9, 94.4]), $P = 0.0003$; unpublished data).

The present study found that the MTBC strains lacking all 43 spacers investigated by the conventional spoligotyping method and assigned to the SIT 2669 were extremely diverse according to the WGS. Some strains were identified as *M. canetti*, and the others that had a completely deleted DR/CRISPR region belonged to genetically unrelated L4-EuroAmerican *M. tuberculosis* lineage as previously described for the SIT 2669 isolates (21). These observations highlight that the SIT 2669 was not meaningful for both epidemiological studies and rapid species identification. In contrast, for the MTBC isolates tested herein with SITs other than 2669, there was a complete overall agreement for species identification by WGS SNP calling or membrane-based spoligotyping, indicating the relevance of both methods for the species identification. Nevertheless, species identification based on SNP calling from WGS presented two advantages over species assignment by spoligotyping. First, WGS SNP calling allowed the identification of 25 MTBC isolates for which species assignment by spoligotype-based algorithms was inconclusive, including 4 strains of the SIT 2669 for which no spacer was amplified. Second, only WGS allowed accurate identification of the 4 mixed infections whereas a false membrane spoligotype pattern resulted from the superimposed profiles of the mixed strains.

The overall agreement was further improved to 92.4% (547/593; 95%CI [89.9, 94.4]) by supplementary analysis using CRISPRbuilder-TB for certain isolates. In addition to more accurate epidemiological monitoring than that provided by spoligotyping, WGS had an added value in some cases of species identification. These data support a smooth transition from the membrane-based to the *in silico*-based genotyping of *M. tuberculosis* isolates is therefore possible for TB diagnosis and epidemiologic survey.

MATERIALS AND METHODS

MTBC isolates. Between November 2016 and December 2020, MTBC isolated from specimens taken from patients during routine care in the Lyon University Hospital, France were prospectively included.

All data were maintained in an electronic database, in accordance with the ethics committee of the Lyon university hospital, France (Comité d'Éthique du CHU de Lyon, number: 20-216), and the national

data protection commission (Commission nationale de l'informatique et des libertés; reference methodology MR-004 that covered the processing of personal data for purposes of study, evaluation or research that did not involve the individual). In accordance with French legislation, written informed consent from patients was not required.

MTBC conventional spoligotyping. Membrane-based spoligotyping was performed as described elsewhere (6). MTBC spoligotyping-based identification and SIT number determination were provided through the open-access SITVITWEB (22) and SpoLineages software tool (<https://github.com/dcouvin/SpoLineages>) (9).

MTBC WGS. For MTBC WGS, genomic DNA was purified from cleared lysates using the Maxwell RSC Instrument (Promega, Madison, WI, USA) automated DNA extraction system and the Maxwell RSC Blood DNA kit (Promega). Libraries were generated using a bead-based tagmentation system (DNAprep; Illumina, San Diego, CA, USA). A nanoliter liquid handler (mosquito HV; SPTLabtech, Hertfordshire, UK) was used to reduce the by 10 times the reaction volumes. Miniaturized libraries were sequenced on the Nextseq or Miseq system (Illumina) to produce 150 or 300 base-pair paired-end reads, respectively. Reference genome coverage was at least 96% and depth of coverage at least 30×.

MTBC *in silico* spoligotyping. MTBC *in silico* spoligotyping was determined through the open-access tool SpoTyping (<https://github.com/xiaeryu/SpoTyping-v2.0>) (12). MTBC spoligotyping-based identification and SIT number determination were provided as described above. Moreover, discrepancies between membrane-based spoligotyping and *in silico* spoligotyping were determined using CRISPRbuilder-TB (<https://github.com/cguyeux/CRISPRbuilder-TB>) (8) to identify events responsible for a spacer found absent "0" in membrane-based but present "1" in *in silico* spoligotyping or vice versa.

MTBC WGS SNP calling identification. The metrics relating to the quality of raw WGS reads were measured using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and potential cross-species contaminations were monitored by FastQ Screen (23). Mapping quality controls were performed using Samtools stats (24). All reports in a sequencing run were compiled using MultiQC (25). Quality control samples then underwent mapping on the *Mycobacterium tuberculosis* reference genome (NC_000962.3) using Burrow-Wheeler Aligner (BWA) (<https://github.com/lh3/bwa>) (26). Following the 2021 guidelines from the Genome Analysis Toolkit (GATK) (<https://github.com/broadinstitute/GATK-for-Microbes>), samples were aligned both on reference and on shifted reference. Duplicated reads were identified using Picard MarkDuplicates (<http://broadinstitute.github.io/picard/>). Variant calling with MuTect2 in microbial mode was performed (27).

Finally, variant calls were processed using the open-access SNP-IT tool (<https://github.com/samlipworth/snpit>) (28) to identify MTBC lineage based on WGS SNP calling (29).

Data analysis. Cohen's Kappa values were calculated using XLSTAT 2020.5.1 (Addinsoft, Paris, France), and interpreted according to Landis and Koch criteria (30). Minimum spanning trees were built using Rstudio: pairwise distances were computed using Manhattan metrics, and the graph was built using the igraph tool in R (31).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank Philip Robinson (DRS, Hospices Civils de Lyon, Lyon, France) for help with manuscript preparation, the GENEPII sequencing platform (Institut des agents infectieux, Hospices Civils de Lyon, Lyon, France) for the MTBC strain sequencing, and the Institute for Integrative Biology of the Cell (I2BC, Université Paris-Saclay, Gif-sur-Yvette, France) for the use of their sharing platform.

This work was supported by the LABEX ECOFECT (ANR-11-LABX-0048) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French national research agency (Agence nationale de la recherche, ANR).

The Lyon TB study group: F. Ader, A. Becker, A. Boibieux, E. Braun, P. Chabert, P. Chauvelot, C. Chidiac, W. Chumbi-Flores, B. Coppere, S. Couraud, G. Devouassoux, O. Dumitrescu, T. Ferry, I. Fredenucci, N. Freymond, S. Gardes, C. Genestet, S. Gerbier-Colomban, Y. Gillet, S. Goutelle, R. Grima, E. Hodille, A. Hot, C. Javaux, L. Kiakouama-Maleka, G. Lina, J. M. Maury, P. Nesme, L. Perard, T. Perpoint, E. Perrot, F. Philit, C. Poudroux, J. P. Rasigade, R. Reix, A. S. Ronnaux-Baron, P. Sève, M. Simon, P. J. Souquet, C. Triffaut-Filit, F. Tronc, F. Valour, and P. Vanhems.

We declare no conflict of interest.

REFERENCES

- Genestet C, Paret R, Pichat C, Berland J-L, Jacomo V, Carret G, Fredenucci I, Hodille E, Rasigade J-P, Boisset S, Carricajo A, Lina G, Ronnaux-Baron A-S, Mornex J-F, Grando J, Sénéchal A, Ader F, Dumitrescu O, Lyon TB study group. 2020. Routine survey of *Mycobacterium tuberculosis* isolates

- reveals nosocomial transmission. *Eur Respir J* 55:1901888. <https://doi.org/10.1183/13993003.01888-2019>.
2. Genestet C, Tatai C, Berland J-L, Claude J-B, Westeel E, Hodille E, Fredenucci I, Rasigade J-P, Ponsoda M, Jacomo V, Vachée A, Gaudart A, Gaillard J-L, Roux A-L, Ader F, Tararbit K, Terpent G, Bryant JE, Lina G, Dumitrescu O, Lyon TB Study Group. 2019. Prospective whole-genome sequencing in tuberculosis outbreak investigation, France, 2017–2018. *Emerg Infect Dis* 25: 589–592. <https://doi.org/10.3201/eid2503.181124>.
 3. Pichat C, Couvin D, Carret G, Frédenucci I, Jacomo V, Carricajo A, Boisset S, Dumitrescu O, Flandrois J-P, Lina G, Rastogi N. 2016. Combined genotypic, phylogenetic, and epidemiologic analyses of mycobacterium tuberculosis genetic diversity in the rhône Alpes region, France. *PLoS One* 11: e0153580. <https://doi.org/10.1371/journal.pone.0153580>.
 4. Alonso-Rodríguez N, Martínez-Lirola M, Herráiz M, Sanchez-Benitez M, Barroso P, Bouza E, García de Viedma D, INDAL-TB group. 2008. Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in Mycobacterium tuberculosis molecular epidemiology studies. *BMC Microbiol* 8:34. <https://doi.org/10.1186/1471-2180-8-34>.
 5. Rasoahantirisoa R, Rakotosamimanana N, Stucki D, Sola C, Gagneux S, Rasolofo Razanamparany V. 2017. Evaluation of spoligotyping, SNPs and customised MIRU-VNTR combination for genotyping Mycobacterium tuberculosis clinical isolates in Madagascar. *PLoS One* 12:e0186088. <https://doi.org/10.1371/journal.pone.0186088>.
 6. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol* 35:907–914. <https://doi.org/10.1128/jcm.35.4.907-914.1997>.
 7. Refrégier G, Sola C, Guyeux C. 2020. Unexpected diversity of CRISPR unveils some evolutionary patterns of repeated sequences in Mycobacterium tuberculosis. *BMC Genomics* 21:841. <https://doi.org/10.1186/s12864-020-07178-6>.
 8. Guyeux C, Sola C, Noûs C, Refrégier G. 2021. CRISPRbuilder-TB: “CRISPR-builder for tuberculosis”. Exhaustive reconstruction of the CRISPR locus in mycobacterium tuberculosis complex using SRA. *PLoS Comput Biol* 17: e1008500. <https://doi.org/10.1371/journal.pcbi.1008500>.
 9. Couvin D, Segretier W, Stattner E, Rastogi N. 2020. Novel methods included in SpolLineages tool for fast and precise prediction of Mycobacterium tuberculosis complex spoligotype families. *Database (Oxford)* 2020:baaa108. <https://doi.org/10.1093/database/baaa108>.
 10. Wyllie DH, Davidson JA, Grace Smith E, Rathod P, Crook DW, Peto TEA, Robinson E, Walker T, Campbell C. 2018. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying mycobacterium tuberculosis transmission: a prospective observational cohort study. *EBioMedicine* 34:122–130. <https://doi.org/10.1016/j.ebiom.2018.07.019>.
 11. Genestet C, Hodille E, Berland J-L, Ginevra C, Bryant JE, Ader F, Lina G, Dumitrescu O, Lyon TB study group. 2020. Whole genome sequencing in drug susceptibility testing of Mycobacterium tuberculosis in routine practice in Lyon, France. *Int J Antimicrob Agents* 55:105912. <https://doi.org/10.1016/j.ijantimicag.2020.105912>.
 12. Xia E, Teo Y-Y, Ong RT-H. 2016. SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. *Genome Med* 8:19. <https://doi.org/10.1186/s13073-016-0270-7>.
 13. Stadhouders R, Pas SD, Anber J, Voermans J, Mes THM, Schutten M. 2010. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *J Mol Diagn* 12: 109–117. <https://doi.org/10.2353/jmoldx.2010.090035>.
 14. Bogaerts B, Delcourt T, Soetaert K, Boarbi S, Ceysens P-J, Winand R, Van Braekel J, De Keersmaecker SCJ, Roosens NHC, Marchal K, Mathys V, Vanneste K. 2021. A bioinformatics whole-genome sequencing workflow for clinical mycobacterium tuberculosis complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches. *J Clin Microbiol* 59:e00202-21. <https://doi.org/10.1128/JCM.00202-21>.
 15. Fang Z, Morrison N, Watt B, Doig C, Forbes KJ. 1998. IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related Mycobacterium tuberculosis strains. *J Bacteriol* 180:2102–2109. <https://doi.org/10.1128/JB.180.8.2102-2109.1998>.
 16. Filliol I, Sola C, Rastogi N. 2000. Detection of a previously unamplified spacer within the DR locus of Mycobacterium tuberculosis: epidemiological implications. *J Clin Microbiol* 38:1231–1234. <https://doi.org/10.1128/JCM.38.3.1231-1234.2000>.
 17. Legrand E, Filliol I, Sola C, Rastogi N. 2001. Use of spoligotyping to study the evolution of the direct repeat locus by IS6110 transposition in Mycobacterium tuberculosis. *J Clin Microbiol* 39:1595–1599. <https://doi.org/10.1128/JCM.39.4.1595-1599.2001>.
 18. Zeng X, Li H, Zheng R, Kurepina N, Kreiswirth BN, Zhao X, Xu Y, Li Q. 2016. Spoligotyping of mycobacterium tuberculosis complex isolates by use of ligation-based amplification and melting curve analysis. *J Clin Microbiol* 54:2384–2387. <https://doi.org/10.1128/JCM.00857-16>.
 19. Marin AV, Rastogi N, Couvin D, Mape V, Murcia MI. 2021. First approach to the population structure of Mycobacterium tuberculosis complex in the indigenous population in Puerto Nariño-Amazonas, Colombia. *PLoS One* 16:e0245084. <https://doi.org/10.1371/journal.pone.0245084>.
 20. Streit E, Millet J, Rastogi N. 2015. Mycobacterium tuberculosis polyclonal infections and microevolution identified by MIRU-VNTRs in an epidemiological study. *Int J Mycobacteriol* 4:222–227. <https://doi.org/10.1016/j.ijmyco.2015.05.005>.
 21. Freidlin PJ, Nissan I, Luria A, Goldblatt D, Schaffer L, Kaidar-Shwartz H, Chemtob D, Dveyrin Z, Head SR, Rorman E. 2017. Structure and variation of CRISPR and CRISPR-flanking regions in deleted-direct repeat region Mycobacterium tuberculosis complex strains. *BMC Genomics* 18:168. <https://doi.org/10.1186/s12864-017-3560-6>.
 22. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N. 2012. SITVITWEB—a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infect Genet Evol* 12:755–766. <https://doi.org/10.1016/j.meegid.2012.02.004>.
 23. Wingett SW, Andrews S. 2018. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res* 7:1338. <https://doi.org/10.12688/f1000research.15931.2>.
 24. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10:giab008. <https://doi.org/10.1093/gigascience/giab008>.
 25. Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>.
 26. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* <https://doi.org/10.48550/arXiv.1303.3997>.
 27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. <https://doi.org/10.1101/gr.107524.110>.
 28. Lipworth S, Jajou R, de Neeling A, Bradley P, van der Hoek W, Maphalala G, Bonnet M, Sanchez-Padilla E, Diel R, Niemann S, Iqbal Z, Smith G, Peto T, Crook D, Walker T, van Soolingen D. 2019. SNP-IT tool for identifying subspecies and associated lineages of Mycobacterium tuberculosis complex. *Emerg Infect Dis* 25:482–488. <https://doi.org/10.3201/eid2503.180894>.
 29. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S, Fellenberg K. 2015. PhyResSE: a web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 53:1908–1914. <https://doi.org/10.1128/JCM.00025-15>.
 30. Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–174. <https://doi.org/10.2307/2529310>.
 31. R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.