Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

# Spoligotyping of *Mycobacterium tuberculosis* – Comparing *in vitro* and *in silico* approaches

Zofia Bakuła[a], Mikołaj Dziurzyński[b], Przemysław Decewicz[c], Daiva Bakonytė[d], Laima Vasiliauskaitė[e,f,g], Birutė Nakčerienė[f,h], Rafał Krenke[i], Petras Stakėnas[d], Tomasz Jagielski[a,*]

[a] *Department of Medical Microbiology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Poland*
[b] *Department of Biology (DBIO), University of Florence, via Madonna del Piano 10, Sesto Fiorentino 50019, Italy*
[c] *Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Poland*
[d] *Department of Immunology and Cell Biology, Institute of Biotechnology, Life Sciences Center, Vilnius University, Lithuania*
[e] *Department of Physiology, Biochemistry, Microbiology and Laboratory Medicine, Institute of Biomedical Sciences, Vilnius University, Lithuania*
[f] *Institute of Biotechnology, Life Sciences Center, Vilnius University, Lithuania*
[g] *Centre of Laboratory Medicine, Laboratory of Infectious Diseases and Tuberculosis, Vilnius University Hospital Santaros klinikos, Lithuania*
[h] *Department of Programs and State Tuberculosis Information System, Vilnius University Hospital Santaros klinikos, Vilnius, Lithuania*
[i] *Department of Internal Medicine, Pulmonary Diseases & Allergy, Medical University of Warsaw, Warsaw, Poland*

ARTICLE INFO

*Keywords:*
*Mycobacterium tuberculosis*
Molecular typing
Spoligotyping
Spoligoprofiles

ABSTRACT

Spoligotyping is one of the molecular typing methods widely used for exploring the genetic variety of *Mycobacterium tuberculosis*. The aim of this study was to compare the spoligoprofiles of *M. tuberculosis* clinical isolates, obtained using *in vitro* and *in silico* approaches.

The study included 230 *M. tuberculosis* isolates, recovered from Poland and Lithuania between 2018 and 2021. Spoligotyping *in vitro* was performed with a commercially available kit. Whole genome sequencing (WGS) was done with Illumina NovaSeq 6000 sequencer. Spoligotype International Types (SITs) were assigned according to the SITVIT2 database or using three different *in silico* tools, and based on WGS data, namely SpoTyping, SpolPred, and lorikeet.

Upon *in vitro* spoligotyping, the isolates produced 65 different spoligotypes. Spoligotypes inferred from the WGS data were congruent with *in vitro* generated patterns in 81.7% (188/230) for lorikeet and 81.3% (187/230) for SpolPred and SpoTyping. Spacers 18 and 31 produced the highest ratio of discrepant results between *in vitro* and *in silico* approaches, with their signals discordantly assigned for 15 (6.5%) and 9 (3.9%) isolates, respectively.

All three *in silico* approaches used were similarly efficient for *M. tuberculosis* spoligotype prediction. However, only SpoTyping could predict spoligotypes without a need for manual curation. Thus, we consider it as the most accurate tool. Its use is further advocated by the shortest time of analysis. A relatively high (*ca.* 20%) discordance between *in vitro* and *in silico* spoligotyping results was observed. While we discourage comparing conventional spoligotyping with *in silico* equivalents, we advise the use of the latter, as it improves the accuracy of spoligopatterns, and thus depicts the relatedness between the isolates more reliably.

## 1. Introduction

Tuberculosis (TB) is caused by a group of phylogenetically closely related, slowly growing bacteria, collectively known as the *Mycobacterium tuberculosis* complex (MTBC). Of nine species in the MTBC, *M. tuberculosis* is the most common cause of TB in humans worldwide (Kanabalan et al., 2021). Each year, the disease affects 10 million people and kills nearly 1.5 million globally (World Health

[*] Corresponding author.

*E-mail addresses:* zofiabakula@uw.edu.pl (Z. Bakuła), mikolaj.dziurzynski@unifi.it (M. Dziurzyński), p.decewicz@uw.edu.pl (P. Decewicz), daiva.bakonyte@bti.vu.lt (D. Bakonytė), laima.vasiliauskaite@santa.lt (L. Vasiliauskaitė), birute.nakceriene@santa.lt (B. Nakčerienė), rafal.krenke@wum.edu.pl (R. Krenke), petras.stakenas@bti.vu.lt (P. Stakėnas), t.jagielski@uw.edu.pl (T. Jagielski).

https://doi.org/10.1016/j.meegid.2023.105508
Received 9 May 2023; Received in revised form 20 September 2023; Accepted 25 September 2023
Available online 26 September 2023

Organization, 2022). One of the cornerstone priorities of TB control is to break the cycle of community transmission, and thus to curb the spread of the disease. A powerful tool to identify patients involved in the same chain of recent transmission is TB genotyping, a laboratory-based approach aimed at assessing the genetic relatedness of strains, and thus confirming or rejecting their epidemiological linkage. Among a wide array of genotyping methods, which have been developed for TB, spacer oligonucleotide genotyping (spoligotyping) has been one of the earliest and most widely adopted approaches for investigating molecular epidemiology of TB. Spoligotyping has become particularly useful in performing phylogenetic reconstructions and inferring evolutionary scenarios associated with *M. tuberculosis* (Tulu and Ameni, 2018*;* Song et al., *2020; * Ali et al., *2019;* Mokrousov et al., *2002*). Spoligotyping interrogates the genetic diversity of the direct repeat (DR) region, which is found in the genomes of MTBC. Technically, it detects the presence of 43 unique spacer sequences (spacers) through a reverse line blot hybridization assay, with the results (spoligopatterns) expressed in a digital format (Jagielski et al., 2014). Establishing *M. tuberculosis* lineages and sublineages, Spoligotype International Types (SITs) designations, and comparative analyses are easily achievable with a publicly available international SITVIT2 web database, which encompasses spoligotypes of 103,856 *M. tuberculosis* isolates, originating from 131 countries (Couvin et al., 2019). Some important advantages offered by spoligotyping include robustness and reproducibility of the results, along with their amenability to database storage and bioinformatic processing. The method is also highly sensitive, and requires ultra-low inputs of DNA, making it feasible on clinical samples, without the need for prior culture (Jagielski et al., 2016).

Although still a relatively expensive solution, whole genome sequencing (WGS), has emerged in recent years as a powerful tool to map, most thoroughly and accurately, genetic diversities of *M. tuberculosis* (Meehan et al., *2019; * Nikolayevskyy et al., *2019*). Several software applications have been developed to predict spoligotype patterns from raw sequence reads. Three of such softwares, namely SpoTyping (Xia et al., 2016), SpolPred (Coll et al., 2012), and lorikeet (Cohen et al., 2015) are most widely used, and have been successfully adopted in several WGS-based studies (Shanmugam et al., 2022; Hijikata et al., 2017; Jiménez-Ruano et al., 2021; Wollenberg et al., 2020; Tarlykov et al., 2020). The principle of all *in silico* spoligotyping methods lies in the detection of the 43 unique spacers, based on the obtained sequence reads. However, different softwares vary in the bioinformatic workflow. Lorikeet utilizes sequenced reads to match known spacer marker sequences, providing a comprehensive assessment of spacer presence or absence for each strain through the analysis of read count totals (Cohen et al., 2015). SpolPred, on the other hand, applies a detection threshold of 4 to address sequencing errors, offering two spoligotype outputs while permitting one 'SNP' per spacer (Coll et al., 2012). Lastly, Spotyping adopts BLAST, allowing for one mismatch in a hit, and its determination of spacer presence relies on the number of hits surpassing a specific threshold, which is directly correlated with the sequence read depth of the locus, providing an alternative perspective for spoligotype analysis (Xia et al., 2016).

The objective of this work was to evaluate the congruence of the spoligotyping patterns of *M. tuberculosis* clinical isolates, produced *in vitro* and using three independent *in silico* analytical tools. This was done to select the software application providing the most accurate prediction of the spoligotypes of *M. tuberculosis* from whole-genome sequence data.

## 2. Materials and methods

### 2.1. Study sample

The study included 230 *M. tuberculosis* isolates, recovered from as many patients from Poland ($n = 86$) and Lithuania ($n = 144$) between 2018 and 2021. Within this number were 130 multidrug-resistant (MDR) and 100 drug-susceptible (DS) isolates. Primary isolation, culturing, species identification, and drug susceptibility testing were performed with standard mycobacteriological methods (CLSI, 2018).

The requirement for informed consent from the study subjects was waived by the Medical University of Warsaw Bioethics Committee (decision no. AKBE/22/2019) since the study sample was collected during routine clinical practices, and all personal data were anonymized prior the study. All experimental protocols and methods were approved by the Medical University of Warsaw Bioethics Committee (decision no. AKBE/22/2019). All methods were carried out in accordance with guidelines regulations of the Medical University of Warsaw.

### 2.2. DNA extraction

Genomic DNA was extracted using PureLink Genomic DNA Mini Kit (ThremoFisher Scientific, USA) or using a modified cetyltrimethylammonium bromide method, as described elsewhere (van Embden et al., 1993). The purified DNA was dissolved in TE buffer and quantified with the NanoDrop OneC Spectrophotometer (ThermoFisher Scientific, USA). The DNA samples were diluted to the required concentration (*ca.* 10 ng/μL) and stored at −20 °C until used.

### 2.3. DNA sequencing and processing of sequencing data

Paired-end libraries were prepared from high-quality genomic DNA with the NovaSeq 6000 Reagent Kits according to the manufacturer's instructions (Illumina, USA). Whole-genome sequencing was done with Illumina NovaSeq 6000 sequencer (Illumina, USA) in 2 × 150 bp paired-end mode. The quality of reads before and after pre-processing was assessed using FastQC v0.11.5 (Leggett et al., 2013) and MultiQC v1.9 (Ewels et al., 2016) tools. The filtering and trimming of raw reads was performed with fastp v0.23.1 tool (Chen et al., 2018) with the following parameters: *--detect_adapter_for_pe --cut_window_size 6 --cut_tail --cut_mean_quality 19 --length_required 50 --n_base_limit 5 --trim_poly_x --poly_x_min_len 10 --correction --overlap_len_require 20 --overlap_diff_limit 5*. In order to evaluate the completeness and contamination level of acquired genomic data and to allow manual verification of discordant spacer we assembled *M. tuberculosis* genomes. Filtered reads were used with SPAdes genome assembler v3.15.3 using *–isolate* and *--kmers* 33,55,77,99,127 flags (Vasilinetc et al., 2015). Quality of assemblies was assessed using QUAST v5.1.0rc1 with *Mycobacterium tuberculosis* H37Rv genome as a reference (Gurevich et al., 2013) and further by remapping 3,000,000 filtered read pairs subsampled with SeqKit v2.1.0 (Shen et al., 2016) with the application of bwa mem v.0.7.17-r1198-dirty (Li, 2013) and samtools v.1.10 (Li et al., 2009). Information about the number of contigs, GC content distribution across contigs in each assembly, as well as the coverage of each contig were used to evaluate assemblies and perform additional filtering of contaminated and low-coverage contigs. This step was conducted manually. Furthermore, the completeness and contamination level of obtained assemblies was assessed with CheckM v1.1.3 (Parks et al., 2015). For *in silico* spoligotyping, complete sets were used to maintain comparable genomes coverage.

The spoligotyping was conducted using only raw reads and assembled genomes were used to evaluate discordant spacers.

The raw reads were deposited under NCBI Bio-Project, accession number PRJNA931475.

### 2.4. Spoligotyping in vitro

Spoligotyping was performed using commercial kits (Ocimum Biosolutions, India) and following the published protocol (Kamerbeek et al., 1997). All profiles were assessed by two independent researchers. SITVIT2 database (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/) was used for classifying Spoligotype International Types (SITs), and spoligotype families for all isolates studied (Couvin et al., 2019).

### 2.5. Spoligotyping in silico

Phylogenetic clades of *M. tuberculosis* were assigned *in silico*, using three different spoligotyping tools available online, *i.e.* (i) SpoTyping (https://github.com/xiaeryu/SpoTyping-v2.0) (Xia et al., 2016); (ii) SpolPred (www.pathogenseq.org/spolpred; available as of July of 2021) (Coll et al., 2012), and (iii) lorikeet (http://genomeview.org/jenkins/lorikeet/) (Cohen et al., 2015). All three applications were run on raw WGS reads after quality control processing. Only raw reads meeting two criteria, *i.e.* (i) high quality of the sequences and (ii) confirmed *M. tuberculosis* origin, were selected for *in silico* spoligotyping. The approach involved analyzing the genome assembly results.

Lorikeet and SpolPred were run separately for each raw read type, forward and reverse, with appropriate value supplemented to *-b* flag in case of SpolPred. Since SpoTyping can compute spoligotype for paired-end reads, each pair was analyzed as a single unit. In case of SpolPred and lorikeet, the programs were run separately on forward and reverse reads, which in 12 cases showed discordant results (for 9 and 3 strains when using SpolPred and lorikeet, respectively). The conflicts were resolved manually by accepting the results with a higher number of detected spacers.

To check, if the presence of spacer 31 was missed due to insertion of IS*6110*, CRISPRbuilder-TB and manual inspection were applied (Guyeux et al., 2021).

Furthermore, a comprehensive examination of CRISPRbuilder-TB results was applied for all isolates with disconcordant *in vitro vs. in*

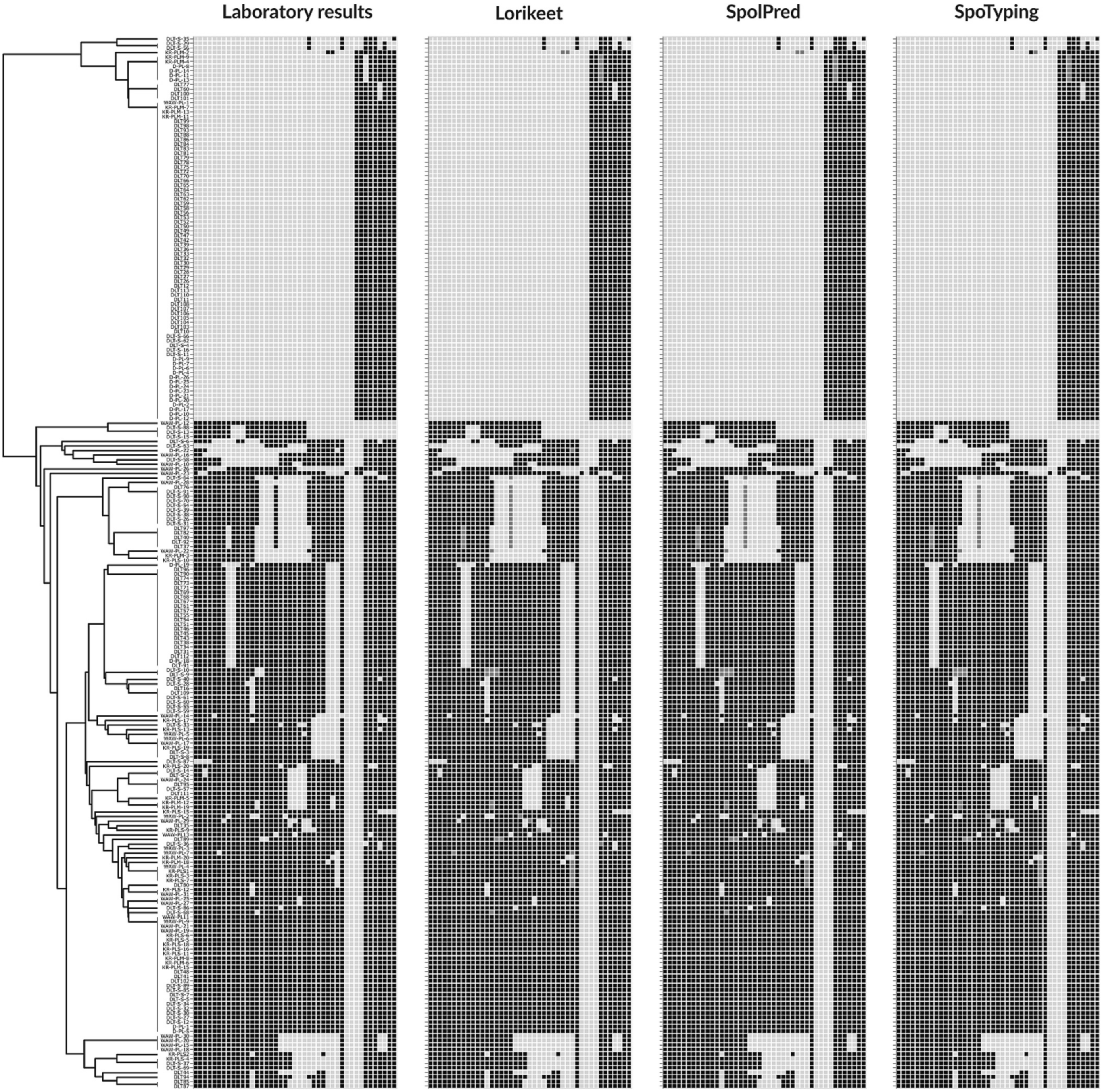

**Fig. 1.** Spoligotype patterns of 230 *M. tuberculosis* isolates determined upon laboratory typing and WGS. Probes differently assigned with *in vitro* and *in silico* methods are marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

*silico* results. Manual inspection was employed to identify any sequence abnormalities within the DR locus. To narrow down this approach, the analysis was conducted only for 17 spacers which gave discrepant results between *in vitro* and *in silico* methods. Furthermore, the 17 spacers were analyzed only in context of 42 isolates which gave discrepant results between *in vitro* and *in silico* methods. This translated into a total of 714 signals manually assessed.

An anomaly was detected when: (i) there was no DR in the upstream/downstream of the spacer, (ii) the DR was truncated downstream of the spacer, (iii) sequence of the DR upstream/downstream of the spacer was qualified as non-canonical, (iv) there was an IS*6110* insertion upstream/downstream of the spacer.

The presence of spacers disconcordantly assigned with *in silico* tools was investigated through the BLAST sequence aligner (to scrutinize the raw data for sequences resembling the spacers). Additionally, lorikeet and SpolPred log files were used, which provide information on the number of hits recovered for each spacer.

## 3. Results

### 3.1. Spoligotyping *in vitro*

For 230 isolates, a total of 65 distinct spoligopatterns were observed. Almost a fifth ($n = 43$; 18.7%) of the isolates were represented by a unique pattern. The remaining isolates ($n = 187$; 81.3%) were split into 22 clusters (2–70 isolates per cluster) (Fig. 1).

Upon comparison with the SITVIT2 database, 29 (12.6%) isolates could not be assigned to any SIT described in the database. Phylogenetically, all isolates of known SITs were classified into 17 families (Supplementary Table 1).

### 3.2. Whole-genome sequencing

Sequencing of 230 *M. tuberculosis* genomes yielded a mean coverage of 531× per genome. The average number of contigs per genome was 76 (± 26) corresponding to an average N50 score of 151 kb (± 54 kb). The genome sizes ranged from 4.330 to 4.535 Mbp (avg. 4.377 Mbp ± 0.158 Mbp). The mean GC content was 65.6% ± 0.02%. The genome completeness ranged from 99.61% to 99.94% (avg. 99.93 ± 0.01) and the contamination level reached up to 0.67% in a single genome.

### 3.3. Concordance between *in vitro* and *in silico* spoligotyping

A total of 188 (188; 81.7%) isolates gave consistent results between *in vitro* spoligotyping and lorikeet. The same isolates, except one (187; 81.3%) produced identical spoligotypes with SpolPred and SpoTyping. Thus, the overall agreement between *in vitro* and *in silico* spoligotyping at the isolate level was 81.7% for lorikeet and 81.3% for SpolPred and SpoTyping (Fig. 1; Table 1).

The isolates with discordant spoligotypes between the two approaches, usually differed by only 1 or 2 spacers (*i.e.* 32 isolates differed in only one spacer, 7 in two spacers, two in three spacers, and one in four spacers).

Among the whole study sample, 26 (26/43; 60.5%) of the spacers consistently gave congruent results between *in vitro* and *in silico* spoligotyping. The remaining (17/43; 39.5%) spacers were either present or absent, compared to *in vitro* assay (Fig. 2; Supplementary Table 1). More specifically, 7 spacers missing with *in silico* tools were detected upon *in vitro* analysis (i); (ii) 9 spacers detected with *in silico* tools were missing upon *in vitro* analysis (ii); and one spacer (no. 37) was either present (1 isolate) or absent (7 isolates), compared to *in vitro* analysis (Fig. 2; Supplementary Table 1).

Among two spacers with the highest ratio of discordantly assigned signals were spacers no. 18 and 31 (Fig. 2; Supplementary Table 1). The spacer no. 18 was not detected upon *in silico* analysis in 15 (6.5%) isolates that produced weak, but visible hybridization signals upon *in vitro* assay (Fig. 3). Of these isolates, all (14; 93.3%) except one (DLT-S-69 of H3 family) were of LAM-RUS family, designated as either SIT254 ($n = 9$) or SIT264 ($n = 5$) according to *in silico* analysis, yet of LAM9 family (SIT766; n = 9) or T1 family (SIT3351; n = 5) according to *in vitro* analysis. Nine (3.9%) isolates had spacer no. 31 present in their *in silico*-generated spoligopatterns, while it was undetected upon hybridization procedure. Half (5/9; 55.5%) of the isolates were either of T1 family (SIT53) or H3 family (SIT50; $n = 4$ and SIT207; $n = 1$), according to *in silico* and *in vitro* spoligotyping, respectively.

### 3.4. CRISPRbuilder analysis of the anomalies within the spacers

In order to explain discrepancies between *in vitro* and *in silico* analyses, the manual inspection of CRISPRbuilder results was employed for 17 spacers of 42 isolates producing such discrepancies (please see Materials and Methods, section 2.5).

Among 23 signals missing with *in silico* tools, yet detected upon *in vitro* analysis (8 spacers; 19 isolates), only one case (1/23; 4.4%) was associated with a sequence anomaly (Supplementary Table 3). On the contrary, for the vast majority (31/35; 88.6%) of signals identified with *in silico* tools, yet missing upon *in vitro* analysis, an anomalous sequence of the corresponding spacer was identified (10 spacers; 30 isolates) (Supplementary Table 3). In case of 14 (14/35; 40%) of those signals, IS*6110* insertion in the upstream (11/14; 78.6%) or downstream (3/14; 21.4%) of the spacer was detected in the corresponding spacer sequence.

Furthermore, as many as 108 anomalies, including (42; 38.9%) insertion of IS*6110,* were found in sequences of signals which gave congruent results between *in silico* and *in vitro* analysis (10 spacers; 32 isolates) (Supplementary Table 4).

### 3.5. Concordance between *in silico* tools

The concordance of the results between SpolPred and lorikeet was 99.6%, since all but one isolates produced identical spoligopatterns (Table 1). The concordance between all three *in silico* tools was 99.1%, as only one isolate had its SpoTyping-derived profile different from that produced by the other two methods (Table 1). The two isolates with discordant *in silico* results are summarized in Fig. 4. The KR-PLM-2 isolate found to be a new orphan type upon *in vitro* analysis, was categorized as either SIT1837 with SpoTyping or SIT1 (Beijing family) with

**Table 1**
Time of analysis and the overall agreement between spoligotyping methods under the study.

| | | Average time of analysis[**] | Method: (n[*]; %) | | | |
|---|---|---|---|---|---|---|
| | | | *in vitro* | SpolPred | SpoTyping | lorikeet |
| Method: (n[*]; %) | *in vitro* | 1 day | 230/230; 100% | 187/230; 81.3% | 187/230; 81.3% | 188/230; 81.7% |
| | SpolPred | 10 min 30 s | 187/230; 81.3% | 230/230; 100% | 228/230; 99.1% | 229/230; 99.6% |
| | SpoTyping | 20 s | 187/230; 81.3% | 228/230; 99.1% | 230/230; 100% | 228/230; 99.1% |
| | lorikeet | 57 s | 188/230; 81.7% | 229/230; 99.6% | 228/230; 99.1% | 230/230; 100% |

[*] n; number of isolates/total number of isolates;

[**] the test was conducted on a single, uncompressed fasta file with 6 million reads, each composed of 150 nucleotides. For each program the test was run on a single core of Intel(R) Xeon(R) CPU E5–2630 v3 @ 2.40GHz processor.

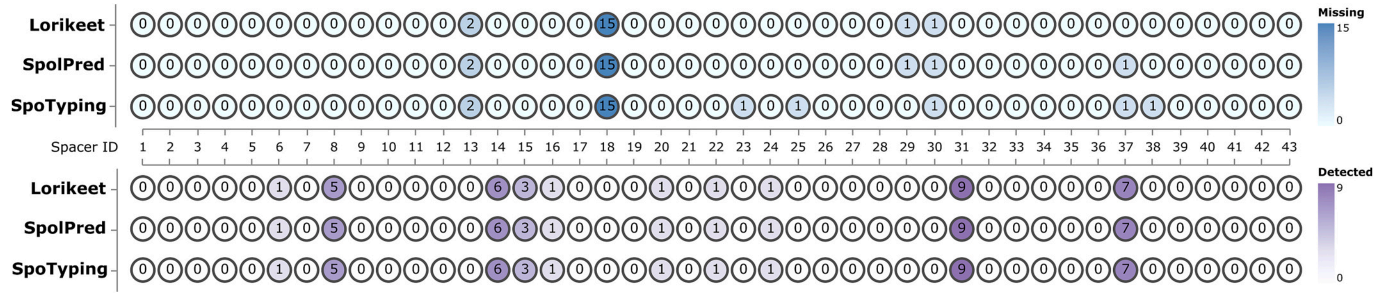

**Fig. 2.** Missing or detected spacer as per *in silico* when compared with *in vitro* analysis. Panel above and below the Spacer ID shows in how many isolates a given spacer was missed (top) or detected (bottom) upon *in silico* analysis, compared hybridization assay.

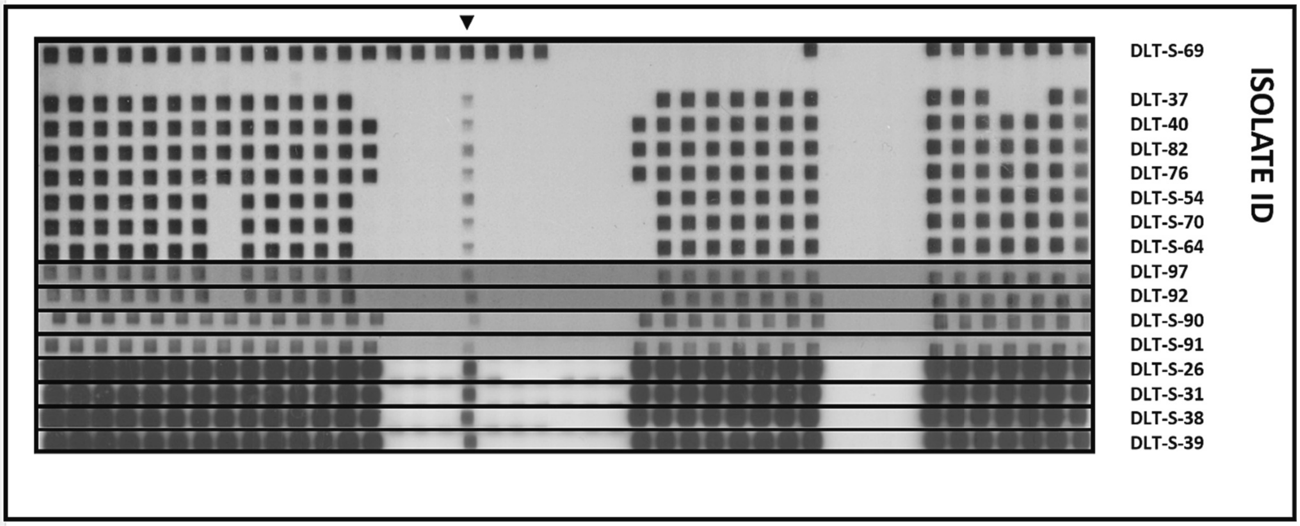

**Fig. 3.** Hybridization patterns using conventional spoligotyping probes. The density of dots for 15 isolates which gave discordant results between *in silico* and *in vitro* analysis (DLT-37 – DLT-S-39), at position 18 (marked with ▼) compared to positive dot (isolate DLT-S-69).

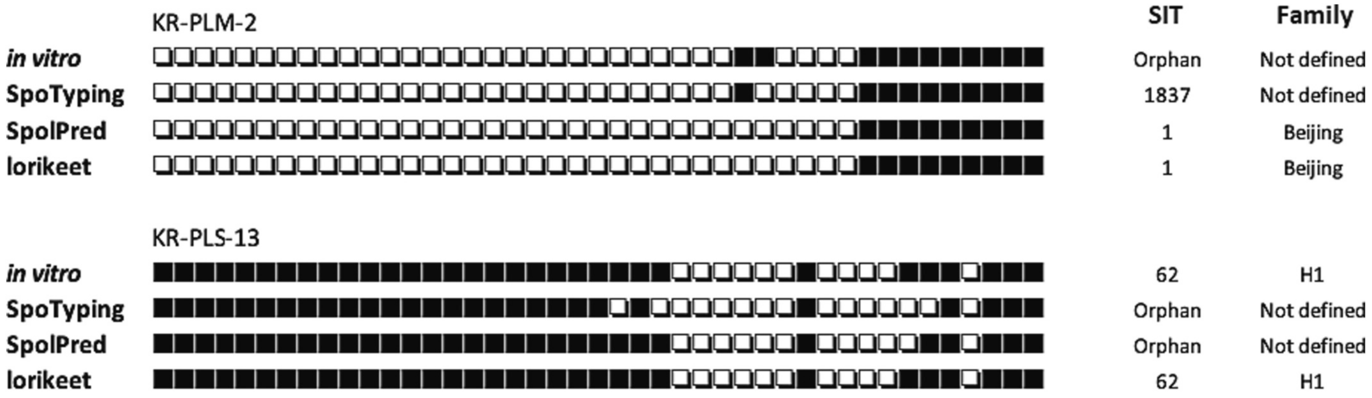

**Fig. 4.** Spoligotyping profiles of two isolates, which produced discordant results upon *in silico* spoligotyping.

SpolPred and lorikeet. Whereas KR-PLS-13 isolate harbored spoligotype characteristic of SIT62 (H1 family), according to lorikeet and *in vitro* assays. This isolate, however, was recognized as representing an orphan type of not defined family, based on SpolPred or SpoTyping analysis.

Overall, 37 (37/43; 86%) of the spacers produced congruent results across all *in silico* spoligotyping methods for the entire study sample, and matched with the *in vitro* results. Thus, 6 spacers (*i.e.* 29 and 30 for KR-PLM-2 and 23, 25, 37, 38 for KR-PLS-13) gave discordant results (Fig. 4), in a way that they were not identified using at least one of the *in silico* spoligotyping tools. Manual inspection, involving use of BLAST software for sequence alignment of raw reads, was employed to check if the data contained any traces of missing spacer sequences. In case of strain KR-PLM-2, manual inspection showed that there were no sequences significantly homologous to spacer 29, however there was a substantial number of raw reads covering spacer 30. The same approach showed that for strain KR-PLS-13, the *in silico* methods should have detected all missing spacers. Further investigation revealed that the disconcordance of *in vitro* and *in silico* results for strain KR-PLS-13 was likely due to the spacers' hit counts falling just below the expected detection threshold. The reason behind discrepancies for KR-PLM-2 isolate is unknown.

Due to the internal construction of SpolPred and lorikeet software, spoligotypes for 9 (3.9%) and 3 (1.3%) isolates, respectively, required additional manual curation. Using SpoTyping all isolates had their spoligotypes assigned without a need for manual curation.

## 4. Discussion

For more than two decades, spoligotyping has been one of the most widely used genotyping method for epidemiological studies of TB. Although being a relatively simple, cost-effective, and high-throughput method, it suffers from low discriminatory power and thus limited use in phylogenetic and transmission studies. Nowadays, the optimal option to fully explore the phylogenetic branching and variation on strain level and to justifiably draw epidemiological conclusions on TB disease is WGS analysis. However, integrating WGS into the routine workflow is too expensive for many clinical laboratories. Therefore, spoligotyping is still in use.

In total, 80% of the *in silico* predictions of spoligopatterns matched the experimental (*in vitro*) data. This value is slightly lower than those published previously for SpolPred (88.6%) (Coll et al., 2012) and SpoTyping (85.7%–90.1%) (Xia et al., 2016; Genestet et al., 2022; Bogaerts et al., 2021). The differences were explained by a number of possible factors including suboptimal hybridization, laboratory cross contamination, PCR contamination, ambiguous hybridization patterns or genetic alterations of the DR locus (Genestet et al., 2022; Bogaerts et al., 2021). Weak hybridization signals due to sequence variations within the spacers may particularly lead to interpretation ambiguities (Genestet et al., 2022; van Embden et al., 2000; Honisch et al., 2010). Furthermore, the lack of a spacer (signal) upon *in vitro* spoligotyping, yet its preservation upon *in silico* spoligotyping might occur with the insertion of transposable elements into the DR locus (Genestet et al., 2022; Filliol et al., 2000; Legrand et al., 2001). DR region is known to be a hotspot for IS*6110* insertion (Roychowdhury et al., 2015). The discrepancies between *in vitro* and *in silico* spoligotyping can also be attributed to WGS-related technical issues, such as low genomic sequence quality or errors during data analysis (Xia et al., 2016; Coll et al., 2012). These difficulties, however, can be overcome due to optimization of DNA isolation and analytical conditions (Xia et al., 2016; Hijikata et al., 2017; Genestet et al., 2022).

In this study, the two most error-prone spacers were nos. 18 and 31 which were discordantly assigned for 6.5% and 3.9% of the isolates, respectively. The spacer no. 18 was absent upon *in silico* approach in 15 isolates, yet presented faint signal with *in vitro* analysis. Low signal intensity for spacer no. 18 upon *in vitro* spoligotyping has been described previously and linked to inaccurate design of the oligonucleotides. The use of redesigned probe, more specific to spacer no. 18 clearly resolved those weak signals as false-positives (van der Zanden et al., 2002).

As for spacer no. 31, it was detected with *in silico* yet missed upon *in vitro* analysis in 9 isolates. As observed in previous studies, insertion of IS*6110* into DR adjacent to spacer no. 31 might disrupt its target. This in turn makes signal undetected in conventional spoligotyping method, as exemplified by conversion of SIT50 (*in vitro*) to SIT53 (*in silico*) (Genestet et al., 2022; Filliol et al., 2000; Legrand et al., 2001). Here, only 2 (2/9; 22.2%) isolates with discordant results for spacer no. 31 had an insertion of IS*6110*, in DR downstream of the spacer no. 31 (Supplementary Table 2). In our sample, 4 out of 9 isolates were of SIT50 according to *in vitro* spoligotyping and of SIT53 upon *in silico* analysis.

Overall, for all but one of the signals missing with *in silico* tools, yet detected upon *in vitro* analysis, no anomaly was detected in their corresponding spacer sequences (Supplementary Table 3). This confirms that most probably technical issues are responsible for those discrepancies (Genestet et al., 2022; Bogaerts et al., 2021). On the contrary, for the signals detected with *in silico* tools, yet missing upon *in vitro* analysis, nearly 90% were due to anomalies in their corresponding spacer sequences. In 15 instances, the absence of a DR region (either upstream or downstream) adjacent to the spacer sequence was identified as the underlying cause for the absence of a signal *in vitro*. Additionally, in 11 cases, a combination of a non-canonical DR region and an IS*6110* insertion accounted for the misidentifications. Notably, IS*6110* insertion alone was found to be responsible for 14 instances of misidentification. Interestingly, as many as 109 anomalies, including 41 IS*6110* insertions were found among sequences corresponding to 425 signals which gave congruent results between *in silico* and *in vitro* analysis (Supplementary Table 4). This demonstrates that not all within-spacer aberrations might be responsible for *in vitro vs. in silico* discrepancies. Consequently, not all anomalies within the spacers should be included for correction of WGS-based detection of the spacers (Mokrousov et al., 2016). The existence of subtle, not necessarily IS*6110*-insertion linked, variations within the spacer and DR region sequences might potentially lead to false negative results when using *in vitro* spoligotyping method. However, the analysis of those particular sequences was beyond the scope of this article, since it would necessitate a more extensive research effort. Ideally, such an endeavor would involve a comprehensive dataset encompassing all available *M. tuberculosis* raw sequencing reads, coupled with corresponding results from standard laboratory practice spoligotyping.

As for the concordance between *in silico* tools, the implemented analysis showed, that the observed differences might be largely explained by the adopted criteria. It is important to note that the threshold for detection varies among different methods. Lorikeet utilizes a sophisticated statistical model to determine whether the number of recovered hits is sufficient for a spacer to be considered present or absent. On the other hand, SpolPred and SpoTyping have more stringent thresholds: SpolPred uses a strict criterion of more than 4 hits, while SpoTyping defines spacer presence based on a numerical parameter multiplied by the average sequencing depth. Fortunately, all programs allow for threshold parameterization. Our findings indicate that there is potential for further optimization of the threshold values to improve the accuracy of spacer detection. Fine-tuning these thresholds may lead to more reliable results and enhance the overall efficacy of these methods in spacer analysis.

All three *in silico* approaches tested were easy-to-use and fast, yet SpoTyping was the most time efficient (Table 1). Furthermore, only with SpoTyping, spoligopatterns could be assigned without a need for manual curation. This is due to the fact that SpoTyping accepts paired-end reads as an input, while lorikeet and SpolPred cannot and must be run on single reads separately (Xia et al., 2016; Coll et al., 2012; Cohen et al., 2015). Therefore, SpoTyping is considered as the most accurate easily operated tool for the prediction of the spoligotypes of *M. tuberculosis* from whole-genome sequence data.

It is important to acknowledge that the DNA sequencing approach used in our study played a crucial role in determining the range of applicable *in silico* spoligotyping tools. While long-read sequencing technologies, such as PacBio SMRT or Oxford Nanopore, would render the use of the same tools impossible, alternative tools are already available for spoligotyping based on long-read data, such as Galru (Page et al., 2020) or LAMBDR (James et al., 2019). However, their accuracy has not yet been confirmed on datasets as big as ours.

*In silico* spoligotyping was developed as a replacement for conventional *in vitro* techniques, which can sometimes yield ambiguous results with weak hybridization signals, leading to potential misclassification of strains. Additionally, spacer and DR sequence modifications, such as insertion of transposable element into the DR locus or hypothesized simple nucleotide variations can result in false-negative outcomes in conventional spoligotyping. These limitations might be overcome by using *in silico* spoligotyping. This approach, however, relies heavily on the quality of sequence reads and the criteria adopted for bioinformatic analysis, including detection thresholds. Since historically established SITs were based on *in vitro* spoligotyping, the more accurate *in silico* spoligotyping might occasionally “misassign” the *M. tuberculosis* genotype, as it does not include the detection of abnormalities within the DR locus. The reference databases used in *in silico* spoligotyping were based on *in vitro* analyses. In conclusion, *in silico* spoligotyping improves the accuracy of spoligopatterns, providing a more reliable depiction of relatedness between strains. However, for a meaningful comparison with the currently existing SIT databases, the analysis should incorporate anomalies within the DR locus. Optimally, current SIT databases should integrate both *in vitro* and *in silico* spoligotyping results. This combined approach would shed light on the SITs that are most frequently affected by discrepancies between these two methodologies. This will be critical to establish the compatibility of *in silico* spoligotyping results with historical records and databases.

Finally, as our analysis showed a profound impact of CRISPR locus anomalies on both *in vitro* and *in silico* spoligotyping, we hypothesize that further, detailed analysis of intrinsic differences between spacer and DR sequence variants may help indicate whether certain SITs are more prone to misclassification (*e.g.* SIT50).

It is of note that during manuscript preparation a new tool, *i.e.* Spolpred2 was developed (Napier et al., 2023). However, since its major advances are faster data processing and higher data input flexibility, it was not included in our analysis during the revision state of the article.

## 5. Conclusions

All three *in silico* approaches were similarly efficient for the prediction of *M. tuberculosis* spoligotypes. Given a relatively high (*ca.* 20%) discordance between the *in vitro* and *in silico* results, we discourage from comparing conventional spoligotyping with *in silico* equivalents. Since SIT databases were based on *in vitro* spoligotyping, and *in silico* spoligotyping does not include the detection of abnormalities within the DR locus, *M. tuberculosis* genotypes can be wrongly predicted upon bioinformatic analysis. However, as it improves the accuracy of spoligopatterns, the use of *in silico* spoligotyping is recommended whenever WGS data are available.

## CRediT authorship contribution statement

**Zofia Bakuła:** Investigation, Methodology, Writing – original draft. **Mikołaj Dziurzyński:** Investigation, Methodology. **Przemysław Decewicz:** Investigation, Methodology. **Daiva Bakonytė:** Investigation, Methodology. **Laima Vasiliauskaitė:** Methodology. **Birutė Nakčerienė:** Methodology. **Rafał Krenke:** Supervision. **Petras Stakėnas:** Investigation, Funding acquisition. **Tomasz Jagielski:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

## Funding

This work was supported by the National Science Center (2017/27/L/NZ6/03279) and Research Council of Lithuania (S-LL-18-4) «DAINA» Programme.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The raw reads were deposited in the GenBank database, under NCBI Bio-Project, accession number PRJNA931475 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA931475/).

Other data supporting this study are included within the article and supporting materials.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2023.105508.

## References

Ali, S., Khan, M.T., Anwar Sheed, K., Khan, M.M., Hasan, F., 2019. Spoligotyping analysis of *Mycobacterium tuberculosis* in Khyber Pakhtunkhwa area, Pakistan. Infect. Drug Resist. 12, 1363–1369.

Bogaerts, B., et al., 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.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

CLSI, 2018. Laboratory Detection and Identification of Mycobacteria, 2nd edition.

Cohen, K.A., et al., 2015. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. PLoS Med. 12, e1001880.

Coll, F., et al., 2012. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. Bioinformatics 28, 2991–2993.

Couvin, D., David, A., Zozio, T., Rastogi, N., 2019. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the *Mycobacterium tuberculosis* genotyping database. Infect. Genet. Evol. 72, 31–43.

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.

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.

Genestet, C., et al., 2022. Consistency of *Mycobacterium tuberculosis* complex spoligotyping between the membrane-based method and *in silico* approach. Microbiol. Spectr. 10 e00223–22.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075.

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).

Hijikata, M., et al., 2017. Spoligotyping and whole-genome sequencing analysis of lineage 1 strains of *Mycobacterium tuberculosis* in Da Nang, Vietnam. PloS One 12, e0186800.

Honisch, C., et al., 2010. Replacing reverse line blot hybridization spoligotyping of the *Mycobacterium tuberculosis* complex. J. Clin. Microbiol. 48, 1520–1526.

Jagielski, T., et al., 2014. Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. Biomed. Res. Int. 2014, 1–21.

Jagielski, T., et al., 2016. Methodological and clinical aspects of the molecular epidemiology of *Mycobacterium tuberculosis* and other mycobacteria. Clin. Microbiol. Rev. 29, 239–290.

James, R.S., et al., 2019. *LAMBDR:* Long-range amplification and Nanopore sequencing of the *Mycobacterium bovis* direct-repeat region. A novel method for *in-silico* spoligotyping of *M. bovis* directly from badger faeces. Molecular Biology. https://doi.org/10.1101/791129.

Jiménez-Ruano, A.C., et al., 2021. Whole genomic sequencing based genotyping reveals a specific X3 sublineage restricted to Mexico and related with multidrug resistance. Sci. Rep. 11, 1870.

Kamerbeek, J., et al., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Kanabalan, R.D., et al., 2021. Human tuberculosis and *Mycobacterium tuberculosis* complex: a review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. Microbiol. Res. 246, 126674.

Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D., Davey, R.P., 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front. Genet. 4.

Legrand, E., Filliol, I., Sola, C., Rastogi, N., 2001. Use of spoligotyping to study the evolution of the direct repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. J. Clin. Microbiol. 39, 1595–1599.

Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Preprint at. http://arxiv.org/abs/1303.3997.

Li, H., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Meehan, C.J., et al., 2019. Whole genome sequencing of *Mycobacterium tuberculosis:* current standards and open issues. Nat. Rev. Microbiol. 17, 533–545.

Mokrousov, I., et al., 2002. Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. Res. Microbiol. 153, 629–637.

Mokrousov, I., et al., 2016. Next-generation sequencing of *Mycobacterium tuberculosis*. Emerg. Infect. Dis. 22, 1127–1129.

Napier, G., et al., 2023. Comparison of in silico predicted *Mycobacterium tuberculosis* spoligotypes and lineages from whole genome sequencing data. Sci. Rep. 13, 11368.

Nikolayevskyy, V., et al., 2019. Role and value of whole genome sequencing in studying tuberculosis transmission. Clin. Microbiol. Infect. 25, 1377–1382.

Page, A.J., Alikhan, N.-F., Strinden, M., Le Viet, T., Skvortsov, T., 2020. Rapid *Mycobacterium tuberculosis* spoligotyping from uncorrected long reads using Galru. Bioinformatics. https://doi.org/10.1101/2020.05.31.126490.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.

Roychowdhury, T., Mandal, S., Bhattacharya, A., 2015. Analysis of IS6110 insertion sites provide a glimpse into genome evolution of *Mycobacterium tuberculosis*. Sci. Rep. 5, 12567.

Shanmugam, S.K., et al., 2022. *Mycobacterium tuberculosis* lineages associated with mutations and drug resistance in isolates from India. Microbiol. Spectr. 10 e01594–21.

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS One 11, e0163962.

Song, S.-E., et al., 2020. Spoligotype variation of *Mycobacterium tuberculosis* strains prevailing in Korea. Can. J. Infect. Dis. Med. Microbiol. 2020, 1–5.

Tarlykov, P., Atavliyeva, S., Alenova, A., Ramankulov, Y., 2020. Genomic analysis of Latin American-Mediterranean family of *Mycobacterium tuberculosis* clinical strains from Kazakhstan. Mem. Inst. Oswaldo Cruz 115, e200215.

Tulu, B., Ameni, G., 2018. Spoligotyping based genetic diversity of *Mycobacterium tuberculosis* in Ethiopia: a systematic review. BMC Infect. Dis. 18, 140.

van der Zanden, A.G.M., et al., 2002. Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. J. Clin. Microbiol. 40, 4628–4639.

van Embden, J.D., et al., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

van Embden, J.D.A., et al., 2000. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. 182, 2393–2401.

Vasilinetc, I., Prjibelski, A.D., Gurevich, A., Korobeynikov, A., Pevzner, P.A., 2015. Assembling short reads from jumping libraries with large insert sizes. Bioinformatics 31, 3262–3268.

Wollenberg, K., et al., 2020. A retrospective genomic analysis of drug-resistant strains of *M. tuberculosis* in a high-burden setting, with an emphasis on comparative diagnostics and reactivation and reinfection status. BMC Infect. Dis. 20 (17).

World Health Organization, 2022. Global Tuberculosis Report 2022. World Health Organization at. https://apps.who.int/iris/handle/10665/346387>.

Xia, E., Teo, Y.-Y., Ong, R.T.-H., 2016. SpoTyping: fast and accurate *in silico Mycobacterium* spoligotyping from sequence reads. Genome Med. 8, 19.