OXFORD

Briefings in Bioinformatics, 2022, **23**(2), 1–11
https://doi.org/10.1093/bib/bbac030
Problem Solving Protocol

SAM-TB: a whole genome sequencing data analysis website for detection of *Mycobacterium tuberculosis* drug resistance and transmission

Tingting Yang†, Mingyu Gan†, Qingyun Liu, Wenying Liang, Qiqin Tang, Geyang Luo, Tianyu Zuo, Yongchao Guo, Chuangyue Hong, Qibing Li, Weiguo Tan and Qian Gao

Corresponding authors: Qian Gao, Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Sciences, Shanghai Medical College, Shanghai Institute of Infectious Disease and Biosecurity and Shanghai Public Health Clinical Center, Fudan University, Shanghai, China. Tel.: +86-21-54237195; E-mail: qiangao@fudan.edu.cn; Weiguo Tan, Shenzhen Center for Chronic Disease Control, Shenzhen, China. Tel.: +86-755-25618776; E-mail: twg202@163.com
†These authors contributed equally to this work.

Abstract

Whole genome sequencing (WGS) can provide insight into drug-resistance, transmission chains and the identification of outbreaks, but data analysis remains an obstacle to its routine clinical use. Although several drug-resistance prediction tools have appeared, until now no website integrates drug-resistance prediction with strain genetic relationships and species identification of nontuberculous mycobacteria (NTM). We have established a free, function-rich, user-friendly online platform for MTB WGS data analysis (SAM-TB, http://samtb.szmbzx.com) that integrates drug-resistance prediction for 17 antituberculosis drugs, detection of variants, analysis of genetic relationships and NTM species identification. The accuracy of SAM-TB in predicting drug-resistance was assessed using 3177 sequenced clinical isolates with results of phenotypic drug-susceptibility tests (pDST). Compared to pDST, the sensitivity of SAM-TB for detecting multidrug-resistant tuberculosis was 93.9% [95% confidence interval (CI) 92.6–95.1%] with specificity of 96.2% (95% CI 95.2–97.1%). SAM-TB also analyzes the genetic relationships between multiple strains by reconstructing phylogenetic trees and calculating pairwise single nucleotide polymorphism (SNP) distances to identify genomic clusters. The incorporated mlstverse software identifies NTM species with an accuracy of 98.2% and Kraken2 software can detect mixed MTB and NTM samples. SAM-TB also has the capacity to share both sequence data and analysis between users. SAM-TB is a multifunctional integrated website that uses WGS raw data to accurately predict antituberculosis drug-resistance profiles, analyze genetic relationships between multiple strains and identify NTM species and mixed samples containing both NTM and MTB. SAM-TB is a useful tool for guiding both treatment and epidemiological investigation.

Keywords: whole genome sequencing, drug-susceptibility testing, drug-resistant tuberculosis, transmission, nontuberculous mycobacteria

Introduction

Whole genome sequencing (WGS) of *Mycobacterium tuberculosis* (MTB) has been shown to be clinically useful for predicting drug-resistance, tracing transmission and defining outbreaks [1]. WGS has the potential to determine drug-resistance much faster than traditional phenotypic susceptibility testing (pDST), does not require biological safety infrastructure [2, 3], and can accurately predict resistance to the full range of antituberculosis drugs [4]. Molecular epidemiology using WGS has a higher resolution than strain-typing techniques such as IS6110-RFLP, spoligotyping or MIRU-VNTR because it can identify single nucleotide polymorphism (SNP) differences in strains that are identical with these other

Tingting Yang is a postdoctoral researcher at the Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Sciences, Shanghai Medical College, Fudan University, Shanghai, China.
Mingyu Gan received his PhD degree at Fudan University, Shanghai, China. Presently, he is a research assistant in the Center for Molecular Medicine, Pediatric Research Institute, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China.
Qingyun Liu received his PhD degree at Fudan University, Shanghai, China. He had done his postdoctoral research in Fudan University. Presently, he is a postdoctoral researcher in Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA.
Wenying Liang, MSc, is engaged at Clabee Genomics, Shenzhen, China.
Qiqin Tang, BSc, is engaged at Clabee Genomics, Shenzhen, China.
Geyang Luo is a PhD student at the Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Sciences, Shanghai Medical College, Shanghai Institute of Infectious Disease and Biosecurity and Shanghai Public Health Clinical Center, Fudan University, Shanghai, China.
Tianyu Zuo received his MS degree at Fudan University, Shanghai, China. Presently, he is a data analyst in patsnap, Shanghai, China.
Yongchao Guo, PhD, is the supervisor of Shenzhen Uni-medica Technology Co., Ltd, Shenzhen, China.
Chuangyue Hong, MM, is engaged at Shenzhen Center for Chronic Disease Control, Shenzhen, China.
Qibing Li, PhD, is the managing director of Clabee Genomics, Shenzhen, China
Weiguo Tan is the director of the tuberculosis prevention and treatment department, Shenzhen Center for Chronic Disease Control, Shenzhen, China
Qian Gao is a Professor at the Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Sciences, Shanghai Medical College, Shanghai Institute of Infectious Disease and Biosecurity and Shanghai Public Health Clinical Center, Fudan University, Shanghai, China.

Received: October 27, 2021. **Revised:** December 28, 2021. **Accepted:** January 25, 2022

techniques and can also trace transmission by delineating the order of nucleotide substitutions [1]. Some countries, such as the UK and the Netherlands, have implemented WGS-guided, individualized treatment [5] through WGS-based monitoring of all tuberculosis patients [6].

However, the analysis of genomic sequencing data remains an obstacle to the routine use of WGS technology in clinical tuberculosis because it requires bioinformatics expertise and high-performance computing that are not readily available in most clinical laboratories [3]. In recent years, several tools have been developed for analyzing MTB WGS data, including KvarQ, PhyResSE, TGS-TB, CASTB, Mykrobe, TBProfiler, MTBseq and ReSeqTB-UVP [3, 4, 7–13]. All of these can detect drug-resistance mutations and identify MTB lineages, and some can also perform phylogenetic and clustering/network analysis (TGS-TB and MTBseq) or species identification of nontuberculous mycobacteria (NTM) (Mykrobe), but there is no tool that integrates all of these functions.

Another bottleneck is the lack of an internationally recognized standard pipeline for the analysis of MTB WGS data. The outputs of different pipelines vary, making it difficult or impossible to compare and validate results [14]. Walter *et al.* compared five different pipelines and found that the number and types of single nucleotide variations obtained were inconsistent, leading to differences in genomic clustering and phylogenetic tree reconstruction [15]. This is a serious obstacle for investigators trying to combine results from different groups for further analysis, such as data mining for new resistance mutations, assessing the confidence of mutations for predicting resistance and identifying global outbreaks of specific strains. To compare strains or add strains to analyses reported by other groups, researchers generally must download the raw WGS data and repeat the entire analysis. This results in a waste of computing resources and storage space as well as time and labor costs that could be avoided with a single analysis tool for both private and shareable genomic data that would allow facile subsequent secondary analyses.

In order to overcome these limitations, we established SAM-TB, a free, function-rich and user-friendly online platform for analysis and sharing of MTB WGS data. The platform integrates functions for variant detection, drug-resistance prediction and genetic relationship analysis of MTB samples, as well as the identification of NTM species and the discovery of mixed samples containing both MTB and NTM.

Materials and methods

Overview

The SAM-TB platform includes three analysis pipelines: 'single sample variants analysis', 'phylogenetic tree reconstruction' and 'pairwise SNP distance' (Figure 1). The 'single sample variants analysis' pipeline consists of the following four modules: (i) reads quality analysis; (ii) NTM species identification; (iii) variants detection and annotation; and (iv) molecular drug susceptibility testing (mDST). The pipeline decides automatically whether to perform NTM species identification, variants detection or both after mapping the sequencing data to the reference genome H37Rv (NC000962.3). If the sequence is considered to be MTB, the variants identified by the 'single sample variants analysis' pipeline can be used to delineate the genetic relationships among strains through analysis of pairwise SNP distances and phylogenetic reconstruction.

Read quality analysis

SAM-TB currently accepts paired-end reads (fastq or fastq.gz format) generated on Illumina sequencing platforms and uses FastQC for quality control [16]. Cutadapt software [17] is used for trimming low-quality bases from 5′ and/or 3′ ends of each read and removing adapter sequences. Sequencing reads with Phred base quality greater than 20 and reads length longer than 35 are kept for analysis. The retained sequencing reads are mapped to the reference genome (H37Rv, NC000962.3) with BWA MEM (v 0.7.15) [18] and then SAMtools (v1.6) [19] is used to calculate the mapping results, including the rates of mapped reads, unique mapped reads, duplicate reads, etc.

NTM species identification

Samples with low rates of mapping to the MTB reference genome may not be MTB or may contain MTB together with sequences from other species. For samples with rates of mapped reads <50% and 10× coverage <50%, SAM-TB performs NTM species identification using mlst-verse software [20]. SAM-TB uses the software's default parameters and reports the species with the highest multilocus sequence typing (MLST) score. For subspecies belonging to the *Mycobacterium abscessus*, *Mycobacterium fortuitum* or *Mycobacterium avium* complexes, SAM-TB reports the corresponding species or complex without identifying the subspecies. For other samples, except those with rates of mapped reads >95% (considered as MTB), SAM-TB performs mixed NTM and MTB samples detection using Kraken2 software [21]. If both NTM and a member of MTB complex (MTBC), such as *M. tuberculosis sensu stricto*, *M. tuberculosis var. bovis*, *M. tuberculosis var. africanum*, *M. canettii*, *etc.*, are detected in a sample, SAM-TB will report both, e.g., '*M. tuberculosis* complex, *M. abscessus*'.

Variant detection and annotation

The SAM-TB analysis platform performs variant detection and annotation for samples containing MTBC, as detected by 100% coverage of reads mapping to MTBC specific sequences (H37Rv region 315947-316534) [22], with an average depth ≥ 5. Duplicate reads are marked and deduplicated using Picard (https://broadinstitute.github.io/picard/), and then SAMtools (v1.6)/VarScan

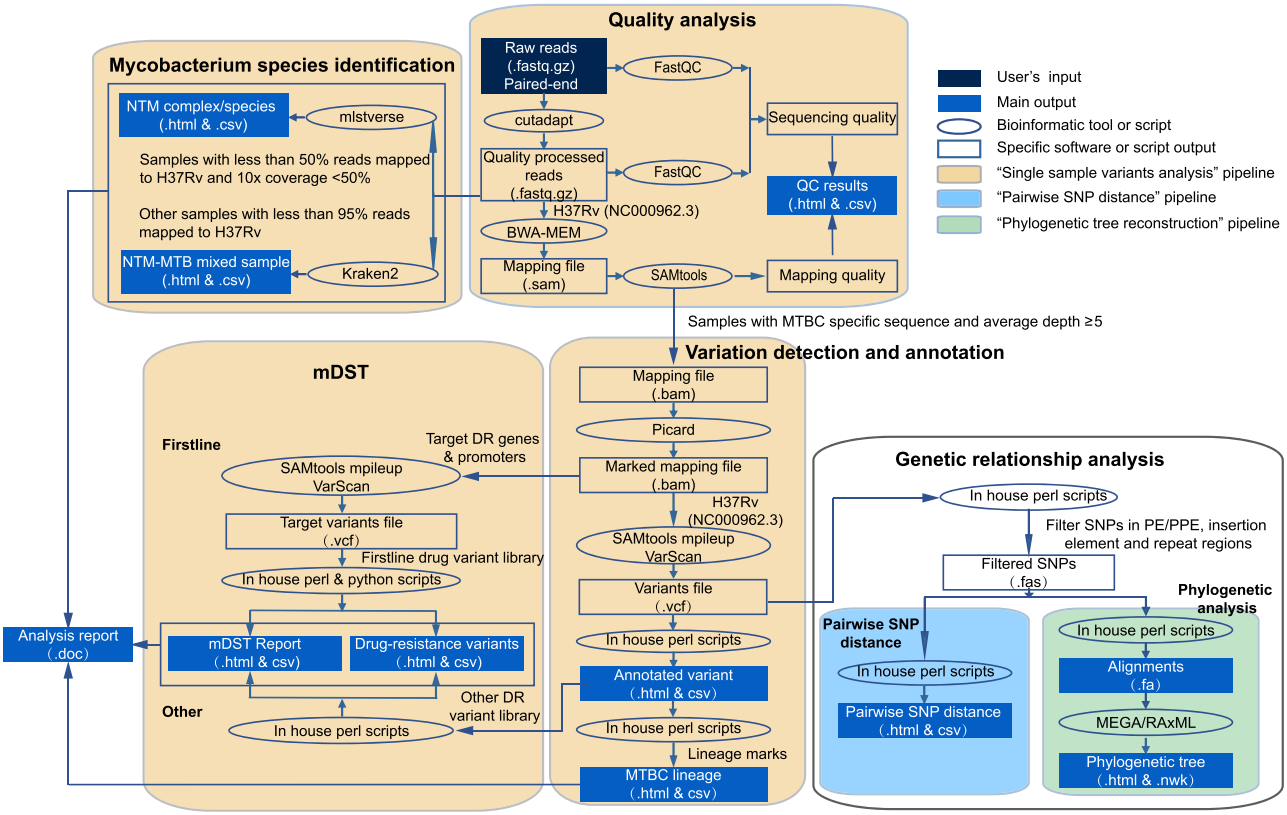

Figure 1. SAM-TB analysis pipelines. The SAM-TB platform includes three analysis pipelines: single sample variant analysis (light brown background), phylogenetic analysis (green background), and pairwise SNP distance (blue background). The single sample variant analysis is composed of four modules: (1) read quality analysis; (2) MTB/NTM species identification; (3) variant detection and annotation; (4) molecular drug-susceptibility test (mDST).

(v2.3.6) [19, 23] suite is used for calling SNPs and short indels, whereas large deletions (≥50 bp) are identified with Delly software [24]. By default, SAM-TB annotates the variants using in-house scripts with the following criteria: base quality ≥30, mapping quality ≥30, depth ≥5, no less than two reads supporting the mutated allele and frequency of the mutated allele ≥75%, but these parameters can be modified by the users. The MTB lineage is assigned by comparing the variants identified in the reads to lineage specific variants [25].

Molecular drug susceptibility testing

A complete and reliable library of drug-resistance mutations is the foundation for accurate mDST. SAM-TB performs mDST with a library that integrates two mutation sets that have been verified with more than 10 000 strains by both sequencing data and pDST results [4, 26]. The mutation set from the CRyPTIC Consortium and 100 000 Genomes Project (library 1) identifies both resistance and harmless mutations (considered to be consistent with susceptibility) in genes associated with resistance to the first-line drugs—isoniazid, rifampicin, ethambutol and pyrazinamide. The targeted genes include *katG*, *inhA*, *fabG1*, *ahpC*, *rpoB*, *embA*, *embB*, *embC*, *pncA* and their promoter regions. The second mutation set (library 2) was extracted from the TBProfiler tool and contains mutations associated with resistance to 16 antituberculosis drugs. The integrated library was manually corrected by adding ofloxacin/levofloxacin resistance mutation *gyrB* E459K [27], removing phylogenetic SNPs [28] and excluding mutations considered to be resistance-conferring in library 2 but annotated as harmless in the library 1. The combined library is shown in Supplementary Tables 1 and 2.

The integrated mutation library can be used to detect resistance to 17 antituberculosis drugs and susceptibility to the four first-line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide. To increase the accuracy of drug-susceptibility prediction, the SAM-TB remaps sequence reads to the targeted regions. Mutations detected by both genome-wide and target mapping are analyzed for mDST. The mDST for the first-line drugs is implemented as previously described [26]: if a drug-resistance mutation is detected in a sample, the sample is reported as 'Resistant (R)' to the drug; if no mutation is detected in the targeted regions or all detected mutations are annotated as harmless, the sample is reported as 'Susceptible (S)' to the drug; if the targeted regions for a first-line drug fail the quality control criteria or contain mutations not known to be associated with drug-resistance but also not annotated as harmless, the mDST result is reported as 'No resistance mutations detected (N.D.)' for that drug. For each of the other 13 drugs: if a drug-resistance mutation is detected in a sample, the sample is reported as 'R' to that drug; otherwise, the mDST result is 'N.D.' (Supplementary Figure 1). Because the pDST can detect

resistance for samples with 1% resistant strains, and sequencing errors are always a possibility, we set the frequency for calling drug-resistance mutations to 10%.

Many drug-resistance mutations have been identified by the association of phenotype with genotype, but without *in vitro* experimental verification. Therefore, SAM-TB includes the level of confidence, as assessed by Miotto *et al.* [27], for 225 drug-resistance mutations in 17 genes used to predict resistance to 10 drugs (isoniazid, rifampicin, ethambutol, pyrazinamide, moxifloxacin, ofloxacin/levofloxacin, amikacin, capreomycin, kanamycin, streptomycin), If the level of confidence in the association of a detected mutation with drug-resistance is minimal or unknown, SAM-TB will annotate this in the 'analysis report'.

Pairwise SNP distance analysis

SAM-TB can analyze the genetic relationship between multiple samples by determining the pairwise SNP distances. This analysis can be performed on samples that have completed variant detection and have an average depth ≥20 and 10× coverage ≥95%. The pairwise distance is calculated using only fixed SNPs (frequency ≥ 75%) that are not in drug-resistance associated genes or repetitive regions of the genome (e.g. PPE/PE-PGRS family genes, phage sequence, insertion or mobile genetic elements).

Phylogenetic tree reconstruction

SAM-TB can also construct phylogenetic trees to display the genetic relationships between strains. The SNPs detected in MTBC isolates that meet the above criteria for pairwise SNP distance analysis are combined into a single consensus and nonredundant list. Nucleotide positions with gaps (possibly due to indels, low coverage or poor mapping quality) in more than 5% of the strains are excluded. The alignments of the polymorphic positions in the strains are used for maximum parsimony tree reconstruction with MEGA v10.0.4 (https://www.megasoftware.net) or for maximum likelihood tree reconstruction using RAxML-NG v.1.0.2 with '–model GTR + G + ASC_LEWIS' [29]. By default, SAM-TB performs 100 bootstrap replicates. The parameters of site screening, the method used for phylogeny reconstruction and the bootstrap values can all be modified by users.

Test data and statistical analysis

To evaluate the accuracy of SAM-TB in performing mDST and identifying NTM species, we downloaded WGS data for 3569 isolates from NCBI, representing 46 mycobacterial species/subspecies (Supplementary Tables 3 and 4). SAM-TB was used to analyze 3177 sequenced NCBI MTB samples with an average depth no less than 50×, 10× coverage of no less than 97% of the genome, and pDST results for at least one antituberculosis drug. These pDST results were used to evaluate SAM-TB for its sensitivity and specificity in predicting drug susceptibility and resistance. The other 392 samples were used to assess the performance of NTM species identification by comparing the species identified by SAM-TB to the species designation accompanying the NCBI WGS data.

Results

SAM-TB functions

SAM-TB is a free-to-use MTB WGS platform with functions for data analysis and data sharing (Table 1). Once users create an account they can upload sequencing data and conduct analyses. When a sample is introduced, SAM-TB performs variant detection and reports genome-wide variants, mDST, MTB lineage classification, and when detected, NTM species identification. The variant SNPs detected in the sequencing data can be used to analyze the genetic relationships among multiple strains using phylogenetic reconstruction and pairwise SNP distance analysis to look for evidence of potential transmission chains. The data and analysis results in SAM-TB are both private and shareable. Each user can only access data and analyses in their own account, but users can share data and analysis results with other users they select.

The SAM-TB platform is easy-to-use, with both English and Chinese interfaces, and requires only a computer connected to the Internet. Users can create an analysis for a single sample or conduct batch analyses on multiple samples (Supplementary Figure 2). The results of the analyses can be examined on the SAM-TB webpage or downloaded to the user's computer (Supplementary Figures 3–8). The results of the analysis of multiple samples for quality control, detection of genome-wide variants and mDST by detection of drug-resistance mutations can be exported in batch (Supplementary Figure 3 and Supplementary Tables 5–8). The variants of multiple samples are integrated to show their distribution in each of the samples. A detailed description of the SAM-TB outputs is presented in Supplementary Table 9.

Drug-resistance mutation library

SAM-TB integrates two highly reliable drug-resistance mutation libraries to conduct mDST (Supplementary Table 10) [4, 26]. The integrated mutation library comprises 9169 polymorphisms at 1382 nucleotide positions in 34 loci (six promoters and 28 coding regions), covering mutations associated with resistance to 17 drugs: isoniazid, rifampicin, ethambutol, pyrazinamide, streptomycin, ethioniamide, amikacin, kanamycin, capreomycin, ofloxacin/levofloxacin, moxifloxacin, para-aminosalicylic acid, cycloserine, linezolid, clofazimine, bedaquiline and delamanid (Supplementary Tables 1, 2, 11, and 12). Each mutation detected is annotated with the level of confidence that it truly confers resistance, as determined by a previous study [27]. SAM-TB also identifies 6617 harmless variants in nine of the genes and promoters associated with resistance to the four first-line antituberculosis drugs (Supplementary Tables 1 and 11) [26].

Table 1. The functions of SAM-TB platform

Function category	Function	Detail
Data analysis	Detect genome-wide variants	SAM-TB performs variants detection by mapping the sequencing reads to the reference genome H37Rv (NC000962.3) and provides annotation for all identified variants.
	Predict drug-resistance	SAM-TB performs mDST analysis for 17 anti-tuberculosis drugs and provides the confidence level of the mutations for predicting resistance. It can also predict susceptibility to the four first-line drugs.
	Analyze the genetic relationship between strains	The integration of the phylogenetic tree and the pairwise SNP distances can be used to analyze the genetic relationship between strains, providing the basis for identifying clusters and inferring recent transmission.
	Identify NTM species/complex	SAM-TB integrates mlstverse software, which is able to identify 175 NTM species. SAM-TB will also identify samples containing both NTM and MTB.
Data sharing	Share sequencing data	SAM-TB users can share sequencing data and analysis results with each other.
	Share analysis results	The shared data and analysis can be viewed and used for further analysis.

The accuracy of SAM-TB in performing mDST

SAM-TB automatically performs quality control, genome mapping, variant detection and mDST on each submitted sample, using either the default or user-set parameters. The running time for a sample with a sequencing depth of 20–500-fold is about 30–150 minutes.

To evaluate the accuracy of mDST predictions using SAM-TB, we downloaded 3177 MTB samples with high quality WGS data and results of pDST. Among these, 1390 (43.8%) were pan-susceptible, 1470 (46.3%) were MDR-TB (resistant to at least rifampicin and isoniazid), and the remaining 317 (10.0%) were resistant to at least one drug but were not MDR-TB (Supplementary Table 3). The most complete pDST results were available for the first-line drugs rifampicin ($N = 3119$; 98.2%) and isoniazid ($N = 3071$; 96.7%), but fewer for the second-line agents (e.g. moxifloxacin, $N = 375$, 11.8%) (Supplementary Table 13). The evaluation excluded linezolid, clofazimine, bedaquiline and delamanid, because there were too few strains with pDST for these drugs to allow a valid analysis.

When compared to the pDST results with only sensitive and resistant predictions included, the sensitivities and specificities of SAM-TB drug-resistance prediction for the first-line drugs were: 97.2% and 98.5% for isoniazid; 97.6% and 96.9% for rifampicin; 95.1% and 88.2% for ethambutol; and 94.7%and 94.8% for pyrazinamide. When the variants reported as 'N.D.' were included, the sensitivities and specificities of SAM-TB drug-resistance prediction were; 94.9% and 94.0% for isoniazid; 96.7% and 92.8% for rifampicin; 92.5% and 78.6% for ethambutol; and 89.3% and 92.3% for pyrazinamide (Table 2). The sensitivity and specificity for predicting MDR-TB were 93.9% (95% CI 92.6-95.1%) and 96.2% (95% CI 95.2-97.1%), respectively.

Sensitivity for predicting resistance to other antituberculosis agents ranged from 31.3% for para-aminosalisylic acid to 89.2% for streptomycin (Table 3). The low sensitivity for para-aminosalicylic acid could result from either difficulties with pDST or the incomplete identification of mutations conferring resistance to it [4], which also likely explains the sensitivity below 90% for streptomycin [30]. The sensitivity for fluoroquinolones was 85.8% for moxifloxacin and 84.2% for ofloxacin. The specificities for predicting resistance to the drugs were all higher than 90%, except for streptomycin (73.3%), moxifloxacin (72.2%) and ethionamide (59.0%). Seventy strains that were moxifloxacin-sensitive by pDST were predicted as resistant by SAM-TB, of which 59 (84.3%) had high-confidence mutations for predicting phenotypic resistance. Of these 59 strains, 34 (57.6%) had resistant mutations associated with low minimum inhibitory concentrations: 27 had the GyrA A90V substitution and seven had D94A [31]. The low MICs associated with these mutations could make it difficult to accurately determine the pDST, which would reduce the apparent accuracy of mDST drug-resistance prediction.

For comparison with an existing method, we also ran the TBProfiler software. For most drugs, the two methods (SAM-TB and TBProfiler) had similar power to detect resistance (Tables 2 and 3 and Supplementary Table 14). For the second-line injectable drugs, amikacin, capreomycin and kanamycin, SAM-TB (81.8%, 79.3% and 85.7%) had notably higher sensitivities than TBProfiler (75.4%, 73.3% and 81.4%). This was due to the presence of resistance mutation *rrs* A1401C with frequencies of 15.21%–85.39% in some resistant samples that was detected by SAM-TB but not by TBProfiler.

Analysis report

SAM-TB provides an easy-to-understand report for each sample in the format recommended in the WHO 'technical guide' [31], including sample details, assay details, final results, drug susceptibility and other information (Supplementary Table 15). The 'Final Results' tells whether the sample is MTB, NTM, mixed NTM and MTB, or an unknown species, and includes the results of mDST. The 'Drug Susceptibility' shows whether the sample is MDR, pre-XDR (MDR with additional resistance to fluoroquinolones) or XDR (pre-XDR with additional resistance to linezolid or bedaquiline), as

Table 2. The accuracy of SAM-TB for predicting resistance or susceptibility to the four first-line drugs

Drug	Resistant phenotype					Susceptible phenotype					Sensitivity	Specificity	PPV	NPV	Sensitivity all†	Specificity all†	NGP	RP
	R	S	N.D.	Total	NGP	R	S	N.D.	Total	NGP								
Isoniazid	1536	44	38	1618	2.3	21	1367	66	1454	4.5	97.2	98.5	98.7	96.9	94.9	94.0	3.4	52.7
Rifampin	1508	37	14	1559	0.9	46	1448	66	1560	4.2	97.6	96.9	97.0	97.5	96.7	92.8	2.6	50.0
Ethambutol	907	47	27	981	2.8	206	1546	216	1968	11.0	95.1	88.2	81.5	97.0	92.5	78.6	8.2	33.3
Pyrazinamide	612	34	39	685	5.7	95	1719	49	1863	2.6	94.7	94.8	86.6	98.1	89.3	92.3	3.5	26.9

Note: NGP, no genotypic prediction; NPV, negative predictive value; PPV, positive predictive value; and RP, resistance prevalence. Unless otherwise indicated, percentages are based on genotypic predictions of resistant (R) or susceptible (S) only (i.e. excluding isolates with mutations of unknown resistance association and genotypic predictions that failed because of missing data around a genomic resistance locus [N.D.]). †Percentages were calculated with the total number of isolates (R, S and N.D.) as the denominator.

well as listing the resistance mutations identified and their frequencies. If no drug-resistance mutations were detected, this is reported as 'N.D.'. Mutations with low confidence are annotated as such in the report comments to indicate the need for further verification. For the first-line drugs, if the prediction result is 'N.D.', the comments include the reason for this designation, such as the presence of mutations not known to confer drug-resistance, harmless variants in the target regions or the number of sites that failed quality control.

Genetic relationship analysis

The genetic relationship between MTBC strains can be measured by the number of genome-wide SNP differences. Strains with 12 or less SNP differences may belong to the same transmission chain, whereas strains with fewer than five SNP differences may be the result of recent transmission [32, 33]. To provide clues for inferring recent transmission clusters, SAM-TB can determine the pairwise SNP distances between isolates (Supplementary Figures 9 and 10). Strains with fewer SNP differences are usually located closer to each other on the phylogenetic tree (e.g. SRR2024940 and SRR2024948) than strains with more SNP differences (e.g. SRR2024940 and SRR2024957) (Figure 2A), and therefore the phylogenetic tree can also help to identify genomic clusters (Supplementary Figures 11 and 12).

SAM-TB results are shown in a flexible display that allows the users to alter the SNP distance threshold to identify strains that differ by no more than a specific number of SNPs (Supplementary Figure 10 and Supplementary Table 16). In addition, the aligned SNP sequences (fasta format) used for phylogenetic reconstruction can be downloaded and used in other visualization programs that reconstruct phylogenetic trees (Supplementary Figure 12). The downloaded phylogenetic tree (newick format) can be annotated with visualization tools such as Figtree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL [34]. The users can also mark potentially clustered strains on the phylogenetic tree to identify possible recent transmissions (Figure 2A) or annotate resistance mutations on the tree to study the evolution of the drug-resistance (Figure 2B).

To test whether SAM-TB can correctly infer transmission clusters, we analyzed a data set containing both WGS and epidemiological data [35]. The genomic clusters identified by SAM-TB were identical with those reported in the literature (Supplementary Table 17 and Supplementary Figure 13). Investigation had identified 18 epidemiologic clusters with 46 patients. Of the 46 patients, 30 had their strains sequenced, involving 16 epidemiological clusters. In five epidemiological clusters only one strain was sequenced, but the isolates from 9 of the remaining 11 epidemiologic clusters were also in genomic clusters (Supplementary Table 17). The results indicate that SAM-TB have the capacity to correctly infer transmission clusters.

Table 3. Accuracy of SAM-TB for predicting resistance to other drugs

Drug	Resistant phenotype			Susceptible phenotype			Sensitivity	Specificity	PPV	NPV	RP
	R	N.D.	Total	R	N.D.	Total					
Streptomycin	514	62	576	43	118	161	89.2	73.3	92.3	65.6	78.2
Ethionamide	297	44	341	153	162	315	87.1	51.4	66.0	78.6	52.0
Amikacin	216	48	264	5	358	363	81.8	98.6	97.7	88.2	42.1
Capreomycin	264	69	333	20	350	370	79.3	94.6	93.0	83.5	47.4
Kanamycin	300	50	350	11	309	320	85.7	96.6	96.5	86.1	52.2
Moxifloxacin	101	19	120	71	184	255	84.2	72.2	58.7	90.6	32.0
Ofloxacin	393	65	458	24	284	308	85.8	92.2	94.2	81.4	59.8
Para-aminosalisylic acid	26	57	83	19	395	414	31.3	95.4	57.8	87.4	16.7
Cycloserine	43	108	151	16	317	333	28.5	95.2	72.9	74.6	31.2

Note: NPV, negative predictive value; PPV, positive predictive value; RP, resistance prevalence. If drug-resistance mutation is detected in a sample, the sample is designated as resistant (R) to the drug; otherwise, the prediction result is 'No Resistance Mutations Detected (N.D.)'.

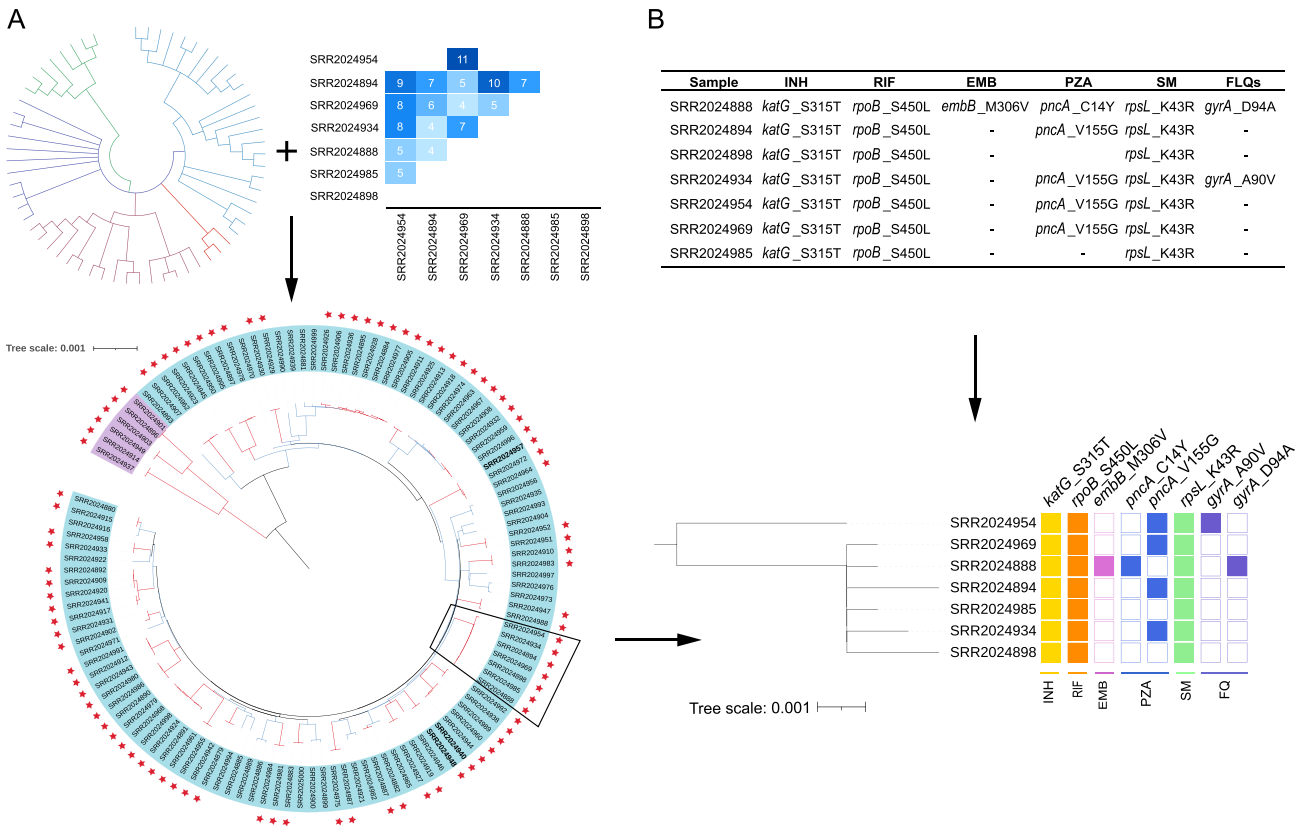

Figure 2. Inference of transmission clusters and annotation of drug-resistance mutations acquired during transmission. (A) The schematic diagram shows the inference of recent transmission clusters based on the results of pairwise SNP distance and phylogenetic analysis. The upper left is a schematic diagram of the phylogenetic tree, with different colors indicating different lineages. To the right of this is a schematic diagram of the SNP distance between strain pairs whose distance is less than a given threshold. On the lower tree the red branches indicate genomic clusters and the red stars indicate clustered strains (SNP distance threshold ≤12). (B) The diagram shows the evolution of drug-resistance during transmission by annotating the resistance mutations on the phylogenetic tree. The colors indicate mutations conferring resistance to different drugs. INH, isoniazid; RIF, rifampicin; EMB, ethambutol; PZA, pyrazinamide; SM, streptomycin; FQ, fluoroquinolone.

NTM species identification

Because some clinical isolates thought to be MTB may actually be NTM, the SAM-TB pipeline integrates the mlstverse software [20] for NTM species identification. The software can distinguish 175 NTM subspecies based on an MLST database of 184 genes. To access the accuracy of NTM species identification, we downloaded the WGS data of 392 mycobacterial strains from NCBI, including 13 strains covering three MTB members (*M. tuberculosis sensu stricto*, *M. tuberculosis var. bovis*, *M. tuberculosis var. africanum*) and 379 strains covering 43 NTM species/subspecies. The mlstverse software infallibly distinguished between MTBC and NTM and also identified the species for some strains that were not identified in the NCBI documentation. For example, the strains labeled by NCBI as *M. sp. CF00131-00135* and *M. bacterium 1482268.1* were identified by mlstverse as *Mycobacterium arosiense* and *Mycobacterium agri*, respectively (Fig. 3). However,

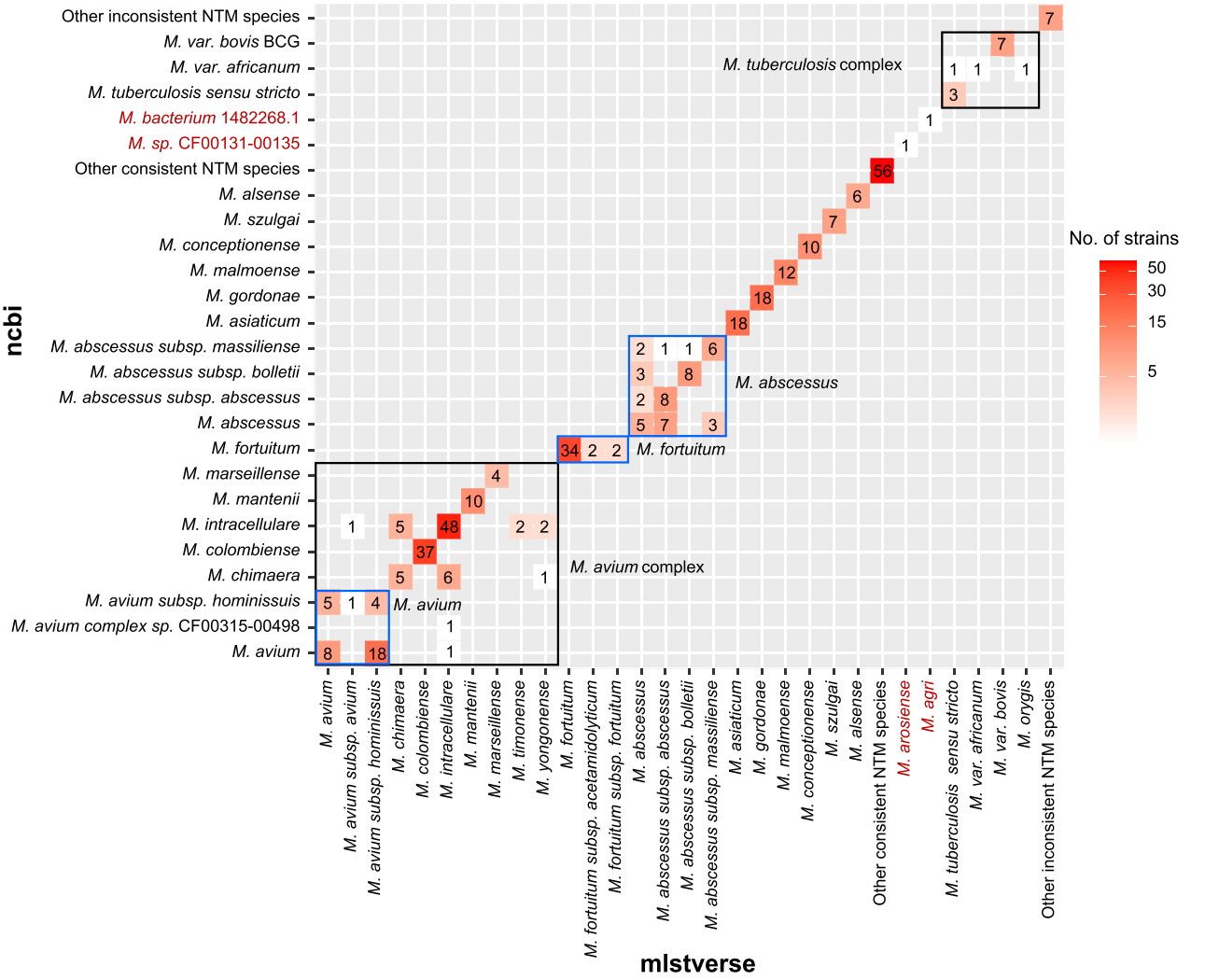

Figure 3. Prediction of mycobacterial species with SAM-TB. Rows represents the species/subspecies reported in NCBI and columns represents the species/subspecies identified by the mlstverse software in SAM-TB. The black boxes indicate the different species in the *Mycobacterium avium* or *Mycobacterium tuberculosis* complexes and the blue boxes indicate the different subspecies of *Mycobacterium avium*, *Mycobacterium fortuitum* and *Mycobacterium abscessus*. The red font indicates two strains whose species were unspecified by NCBI but identified with mlstverse.

although mlstverse software reliably recognized strains belonging to the *M. abscessus*, *M. fortuitum* and *M. avium* complex, it was not completely accurate at identifying the subspecies within these species or complexes. Out of 159 strains belonging to *M. avium* complex, 18 were identified as belonging to a species different from that designated by NCBI and one was a different subspecies. Two of 46 *M. abscessus* strains were identified as subspecies other than those designated by NCBI. In addition, the species identification of another seven strains was different from the species designation by NCBI. In total, there were 28 strains for which the species identification was inconsistent with the designation by NCBI. Therefore, the accuracy of mlstverse in NTM species identification was 92.6% (351/379). However, because SAM-TB only reports that strains belong to the *M. fortuitum*, *M. abscessus* and *M. avium* complex, without identifying subspecies, its estimated accuracy for identifying NTM species was 98.2%.

MTB and NTM mixed sample identification

When not all of a sample's sequencing reads can be mapped to the H37Rv genome, the sample may contain other species in addition to MTB, most commonly NTM. SAM-TB integrates the Kraken2 software [21] for MTB and NTM mixed samples detection. Three downloaded samples (ERR2513283, ERR2517475 and ERR2514390) with MTB mapping rates of 39.2%, 70.9% and 81.2%, and 10× coverage 94.60%, 92.56% and 98.51% were determined by SAM-TB to consist of mixtures of MTB and NTM sequences (Supplementary Table 18). The first two contained MTB and *M. abscessus*, while the third was a mixture of MTB and an *M. avium* complex bacteria.

The similarities between NTM and MTB in certain genomic regions, such as the *rpoB* gene, will affect the calling of variants and drug-resistance mutations in mixed samples, especially for rifampicin resistance. In the three rifampicin-sensitive mixed samples, SAM-TB detected 20, 17 and 9 rifampicin resistance mutations

(Supplementary Table 18) with frequencies between 30.3–71.5%, 31.6–71.3% and 10.1–15.83%, respectively. In order to understand the impact of mixed samples on variants detection, we mixed sequencing reads of MTB and several common NTM species in varying proportions (Supplementary Table 19). Compared to unmixed MTB sample, more variants were detected in mixed samples when <95% of reads mapped to MTB.

Discussion

The SAM-TB website is a free, easy-to-use, function-rich platform for MTB WGS data analysis and sharing. It performs genome-wide variant detection, mDST for 17 antituberculosis drugs and analyzes the genetic relationship between multiple MTB samples for inferring transmission clusters. It can also detect when samples contain MTB mixed with NTM and can accurately identify NTM species.

As WGS has been proven useful for predicting MTB drug-resistance and identifying transmission clusters, several WGS analysis tools for these tasks have been developed for MTB. SAM-TB combines the functions of these prior tools and also integrates and modifies them (Supplementary Table 20). Although other existing tools can detect mDST for up to 14 drugs, SAM-TB and TBProfiler are the only ones that can predict mDST for nearly all antituberculosis drugs [4, 9]. In addition, SAM-TB indicates when there is susceptibility to first-line drugs and annotates any low-confidence resistance mutations that are detected. SAM-TB is also the only website that can performs both MTBC genetic relationship analysis and NTM species identification. SAM-TB allows users to set the SNP distance threshold for identifying related strains. It is thus a flexible tool for identifying outbreaks and inferring recent transmission that can help guide epidemiological investigations. SAM-TB can identify 175 subspecies of NTM, 135 more than that can be identified with Mykrobe [11]. This will facilitate the selection of treatment options, which vary depending upon the specific NTM identified.

Determining resistance to antituberculosis drugs by WGS has been shown to be feasible and cost-effective [4], but the performance depends upon the integrity and accuracy of the drug-resistant mutation library employed. SAM-TB mDST integrates two highly reliable drug-resistant mutation libraries that can predict MDR-TB with a sensitivity 93.9% and specificity 96.2%. Knowledge of mutations conferring resistance is incomplete for some drugs, including recently introduced agents bedaquiline and delamanid/pretomanid. The library will be updated annually for both resistance and harmless mutations, along with their confidence grade, based on the recent publications and consensus. This may not be as flexible as TBProfiler, where users can input their own mutation library to predict resistance [4], but will guarantee the reliability of the mutation library.

Although there is a large amount of published WGS data on MTB and NTM, there are few reports of mixed samples containing both. When a sample contains both NTM and MTB the detection of variants is altered, which in turn affects the drug-susceptibility predictions, clustered strain identification and transmission tracing. In addition, the identification of mixed NTM and MTB samples can be critically important for devising treatments, but prior to SAM-TB, Mykrobe was the only tool that integrates drug-resistance prediction and mixed sample identification.

Compared with other MTB WGS analysis tools, SAM-TB has a variety of user-friendly functions (Supplementary Table 20), including online batch analysis for multiple samples. As all analysis is performed online, users do not have to install any software on their own computers. SAM-TB is both safe and shareable: the data and analysis results are absolutely private, but users have the option of sharing data and analysis with each other, thus saving time, storage space and computing resources. The website is managed by specialized personnel who perform maintenance and storage space management to ensure that the website functions smoothly. However, as a webserver, SAM-TB cannot overcome potential disadvantages such as the time required to upload the WGS files and processing cues on the webserver side.

In conclusion, SAM-TB is a multifunctional platform for analysis and sharing of MTB WGS data. Its analysis functions include variant detection, mDST determination, genomic cluster inference, detection of mixed NTM and MTB samples and identification of NTM species. It can analyze single samples to guide patient management or perform batch analysis on multiple samples to provide information useful for both epidemiology and more basic tuberculosis research.

Key Points

- SAM-TB integrates functions for variant detection, mDST determination, genomic cluster inference, detection of mixed NTM and MTB samples and identification of NTM species.
- SAM-TB performs mDST for 17 antituberculosis drugs and predicts susceptibility to first-line tuberculosis drugs.
- SAM-TB provides the confidence level of the mutations for predicting resistance.
- The results of the analysis of multiple samples can be exported in batch.
- The data and analysis results in SAM-TB are both private and shareable.

Authors' contributions

All authors idealized and planned the platform; T.Y., M.G., Q.L., W.L., Q.T., G.L., T.Z. and Q.L. developed and tested the platform; T.Y., Y.G. and C.H. collected test data; T.Y., W.T. and Q.G. wrote the first draft of the manuscript; and the finalized manuscript contained contributions from all authors.

Data availability

The Illumina project accession numbers are presented in Supplementary Tables 3 and 4. The data used to validate the ability of SAM-TB to infer transmission was derived from EBI (accession number PRJEB5162).

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank Howard E Takiff for helping to edit the manuscript and suggestions for improving the SAM-TB website.

Funding

The Sanming project of Medicine in Shenzhen (SZSM201611030), National Natural Science Foundation of China (81661128043, 81871625), National Science and Technology Major Project of China (2017ZX10201302 and 2018ZX10715012), Natural Science Foundation of Guangdong Province of China (2020A1515011086) and China Postdoctoral Science Foundation (2019M661365).

References

1. Takiff HE, Feo O. Clinical value of whole-genome sequencing of *Mycobacterium tuberculosis*. *Lancet Infect Dis* 2015;**15**(9):1077–90.
2. Shea J, Halse TA, Lapierre P, *et al.* Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of *Mycobacterium tuberculosis* in New York State. *J Clin Microbiol* 2017;**55**(6):1871–82.
3. Coll F, McNerney R, Preston MD, *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* 2015;**7**(1):51.
4. Phelan JE, O'Sullivan DM, Machado D, *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 2019;**11**(1):41.
5. Groschel MI, Walker TM, van der Werf TS, *et al.* Pathogen-based precision medicine for drug-resistant tuberculosis. *PLoS Pathog* 2018;**14**(10):e1007297.
6. Tagliani E, Cirillo DM, Kodmon C, *et al.* EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. *Lancet Infect Dis* 2018;**18**(4):377.
7. Steiner A, Stucki D, Coscolla M, *et al.* KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* 2014;**15**:881.
8. Feuerriegel S, Schleusener V, Beckert P, *et al.* PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 2015;**53**(6):1908–14.
9. Sekizuka T, Yamashita A, Murase Y, *et al.* TGS-TB: total genotyping solution for *Mycobacterium tuberculosis* using short-read whole-genome sequencing. *PLoS One* 2015;**10**(11).
10. Iwai H, Kato-Miyazawa M, Kirikae T, *et al.* CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis (Edinb)* 2015;**95**(6):843–4.
11. Bradley P, Gordon NC, Walker TM, *et al.* Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* 2015;**6**:10063.
12. Kohl TA, Utpatel C, Schleusener V, *et al.* MTBseq: a comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ* 2018;**6**:e5895.
13. Ezewudo M, Borens A, Chiner-Oms A, *et al.* Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci Rep* 2018;**8**(1):15382.
14. Meehan CJ, Goig GA, Kohl TA, *et al.* Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol* 2019;**17**(9):533–45.
15. Walter KS, Colijn C, Cohen T, *et al.* Genomic variant-identification methods may alter *Mycobacterium tuberculosis* transmission inferences. *Microb Genom* 2020;**6**(8):mgen000418.
16. Andrews S. *FastQC: a quality control tool for high throughput sequence data*. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10–2.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013; https://arxiv.org/abs/1303.3997v2.
19. Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and Samtools. *Bioinformatics* 2009;**25**(16):2078–9.
20. Matsumoto Y, Kinjo T, Motooka D, *et al.* Comprehensive subspecies identification of 175 nontuberculous mycobacteria species based on 7547 genomic profiles. *Emerg Microbes Infect* 2019;**8**(1):1043–53.
21. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**(1):257.
22. Zong Z. Bio-marker screening for tuberculosis and non-tuberculosis mycobacterial disease differentiation (in Chinese). 2019. http://cdmd.cnki.com.cn/Article/CDMD-87112-1019251652.htm. 15 Jan 2021, date last accessed.
23. Koboldt DC, Zhang Q, Larson DE, *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**(3):568–76.
24. Rausch T, Zichner T, Schlattl A, *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**(18):i333–9.
25. Napier G, Campino S, Merid Y, *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med* 2020;**12**(1):114.
26. The CRyPTIC Consortium and the 100 000 Genomes Project, Allix-Béguec C, Arandjelovic I, *et al.* Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N Engl J Med* 2018;**379**(15):1403–15.
27. Miotto P, Tessema B, Tagliani E, *et al.* A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J* 2017;**50**(6):1701354.
28. Merker M, Kohl TA, Barilar I, *et al.* Phylogenetically informative mutations in genes implicated in antibiotic resistance in *Mycobacterium tuberculosis* complex. *Genome Med* 2020;**12**(1):27.
29. Kozlov AM, Darriba D, Flouri T, *et al.* RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 2019;**35**(21):4453–5.

30. Rocha DM, Magalhães C, Cá B, *et al.* Heterogeneous streptomycin resistance level among *Mycobacterium tuberculosis* strains from the same transmission cluster. *Front Microbiol* 2021;**12**:659545.
31. World Health Organisation. The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in *mycobacterium tuberculosis* complex: technical guide. 2018. https://apps.who.int/iris/bitstream/handle/10665/274443/WHO-CDS-TB-2018.19-eng.pdf?sequence=1&isAllowed=y. 15 Nov 2021, date last accessed.
32. Yang C, Luo T, Shen X, *et al.* Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis* 2017;**17**(3):275–84.
33. Walker TM, Ip CL, Harrell RH, *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013;**13**(2):137–46.
34. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;**44**(W1):W242–5.
35. Walker TM, Lalor MK, Broda A, *et al.* Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2014;**2**(4):285–92.