Genome Medicine

RESEARCH

Open Access

Advantages of updated WHO mutation catalog combined with existing whole-genome sequencing-based approaches for *Mycobacterium tuberculosis* resistance prediction

Yiwang Chen[1,2†], Xuecong Zhang[1†], Jialei Liang[2], Qi Jiang[3], Mijiti Peierdun[4], Peng Xu[1], Howard E. Takiff[5] and Qian Gao[1,2*]

Abstract

Background The WHO recently released a second edition of the mutation catalog for predicting drug resistance in *Mycobacterium tuberculosis* (MTB). This study evaluated its effectiveness compared to existing whole-genome sequencing (WGS)-based prediction methods and proposes a novel approach for its optimization.

Methods We tested the accuracy of five tools—the WHO catalog, TB Profiler, SAM-TB, GenTB, and MD-CNN—for predicting drug susceptibility on a global dataset of 36,385 MTB isolates with high-quality phenotypic drug susceptibility testing (DST) and WGS data. By integrating the genotypic DST predictions of these five tools in an ensemble machine learning framework, we developed an improved computational model for MTB drug susceptibility prediction. We then validated the ensemble model on 860 MTB isolates with phenotypic and WGS data collected in Shenzhen, China (2013–2019) and Valencia, Spain (2014–2016).

Results Among the five genotypic DST tools for predicting susceptibility to ten drugs, MD-CNN exhibited the highest overall performance (AUC 92.1%; 95% CI 89.8–94.4%). The WHO catalog demonstrated the highest specificity of 97.3% (95% CI 95.8–98.4%), while TB Profiler had the best sensitivity at 79.5% (95% CI 71.8–86.2%). The ensemble machine learning model (AUC 93.4%; 95% CI 91.4–95.4%) outperformed all of the five individual tools, with a specificity of 95.4% (95% CI 93.0–97.6%) and a sensitivity of 84.1% (95% CI 78.8–88.8%), principally due to considerable improvements in second-line drug resistance predictions (AUC 91.8%; 95% CI 89.6–94.0%).

Conclusions The second edition of the WHO MTB mutation catalog does not, by itself, perform better than existing tools for predicting MTB drug resistance. An integrative approach combining the WHO catalog with other genotypic DST methods significantly enhances prediction accuracy.

Keywords *Mycobacterium tuberculosis*, Whole-genome sequencing, Drug resistance prediction, WHO mutation catalog, Ensemble machine learning model

[†] Yiwang Chen and Xuecong Zhang contributed equally to this work.

*Correspondence:
Qian Gao
qiangao@fudan.edu.cn
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Drug-resistant tuberculosis (DR-TB) represents a challenge to global tuberculosis (TB) treatment and control. The Global TB Report estimated that in 2023 at least 410,000 individuals developed multidrug-resistant or rifampicin-resistant TB (MDR/RR-TB) [1]. However, only 149,511 cases were diagnosed and reported [1], highlighting the considerable deficiencies in DR-TB diagnostic capabilities. Although the global treatment success rate for MDR/RR-TB has recently improved, globally it remains alarmingly low at 63% [1]. MDR/RR-TB generally evolves due to inappropriate TB treatment, but recent molecular epidemiological studies have shown that person-to-person transmission is now the primary contributor to the global DR-TB burden [2, 3]. Delays in drug susceptibility testing (DST) to detect drug-resistant *Mycobacterium tuberculosis* (MTB) strains will prolong their transmission and exacerbate the epidemic [4, 5].

Phenotypic DST methods remain the gold standard for detecting DR-TB, though WHO endorses a composite reference standard for rifampicin resistance [6]. These methods are labor-intensive and require biosafe laboratory infrastructure [7]. Molecular methods, such as Xpert MTB/RIF and GenoType MTBDRplus, offer quicker results than phenotypic DST but will only detect a subset of all drug-resistant mutations (DRMs) for only a few anti-TB medications [8]. The advantage of whole-genome sequencing (WGS) is that it can detect all putative DRMs across the entire MTB genome and can be quicker than phenotypic methods requiring MTB cultures. The current WGS-based tools to predict MTB drug resistance fall into two categories: (1) tools that search WGS data for variants from a list of known DRMs, such as TB Profiler [9], PhyResSE [10], and SAM-TB [11], and (2) machine learning-based tools that use phenotypic DST data and WGS data to establish a prediction model for drug resistance, such as GenTB [12], HANN [13], and MD-CNN [14]. Despite significant advances in the WGS-based methods for detecting drug resistance in MTB, there are, at present, no universally accepted global standards for genotypic DST.

In 2021, the WHO released its inaugural catalog of MTB mutations associated with drug resistance, derived from the analysis of WGS data and phenotypic DST results of 38,215 global isolates [15]. This catalog serves as a reference standard for interpreting genotypic DST results. To incorporate the findings from the vast number of data sets now available, the WHO mutation catalog was updated in 2023 [16]. In this study, we compared the performance of the updated WHO catalog against existing WGS-based tools for predicting drug resistance in MTB. We then integrated the WHO catalog with existing WGS-based tools to develop a more accurate method for predicting drug resistance in MTB.

Methods

Datasets

To evaluate the performance of the 2023 WHO catalog and other WGS-based genotypic DST methods for predicting drug resistance, as well as the development of ensemble machine learning models, we used a globally sourced dataset comprised of high-quality WGS data and phenotypic DST results from 36,385 MTB isolates collected from 45 countries across six continents [5]. To assess the performance of the ensemble models, we used an external testing cohort that included 860 MTB isolates from two retrospectively collected sources: 154 MDR-TB isolates with phenotypic DST results collected from 2013 to 2019 in the Shenzhen Center for Chronic Disease Control, China [17], and 706 MTB isolates with phenotypic DST results collected from 2014 to 2016, in the 25 clinical laboratories in Valencia, Spain [18]. All isolates had undergone WGS previously and their WGS data (Additional files 1, 2) were downloaded from the NCBI-SRA database (https://www.ncbi.nlm.nih.gov/sra) [19].

Analysis of WGS data

We employed a previously established pipeline for identifying single nucleotide polymorphisms (SNPs) [20]. Initially, the Sickle tool [21] trimmed WGS data to preserve reads with a Phred base quality above 20 and lengths exceeding 30 nucleotides. The reads were then aligned to the MTB H37Rv reference genome (GenBank AL123456) using bowtie2 (v2.2.9) [22]. SNP-calling was conducted with SAMtools (v1.3.1) [23], requiring a mapping quality above 30. Varscan (v2.3.9) [24] identified fixed SNPs (frequency ≥ 75%) supported by at least five reads, applying a strand bias filter. We excluded SNPs located in noisy or repetitive genomic regions, such as PPE/PE-PGRS family genes, phage sequences, and mobile genetic elements [25]. MTB lineages were classified using lineage-specific SNPs [26].

Baseline methods

The updated WHO mutation catalog, two lists from known DRMs-based tools (TB Profiler and SAM-TB), and two machine learning-based tools (GenTB and MD-CNN) were considered as the baseline methods in this study. All baseline methods were used as is, without modification or retraining. The newly updated 2023 WHO mutation catalog served as a reference standard for interpreting MTB DRMs [16]. Only variants classified as “Category 1: Associated with resistance” or “Category 2: Associated with resistance-interim” were considered as known DRMs. TB Profiler (v5.0.0) [9] and SAM-TB (v1)

[11], both widely used tools, predict MTB resistance by incorporating known DRMs into their frameworks. All three catalog-based methods relied on the same systematic WGS data analysis pipeline described in the Analysis of WGS data section, but each predicted resistance based on their respective DRM catalogs. GenTB (v1) employs a conventional random forest (RF) classifier, which has frequently been used as a comparator in earlier TB drug resistance prediction studies [12, 13]. MD-CNN (v1.0) [14] is a multi-drug convolutional neural network that significantly enhances the interpretability of predictions over traditional machine learning-based approaches. Evaluations were conducted using ten-fold cross-validation, with results including the mean area under receiver operator curves (AUC), sensitivity, specificity, and the corresponding 95% confidence intervals (95% CI) for each baseline method.

Identification of potential explanatory mutations for phenotypically resistant but genotypically susceptible strains

To identify potential mutations explaining phenotypically resistant but genotypically susceptible strains, we applied two criteria. First, the mutations must be present in phenotypically resistant but genotypically susceptible strains and absent in phenotypically susceptible strains. Second, these mutations were required to appear at least five times among the phenotypically resistant but genotypically susceptible strains.

Ensemble model development and validation

A stacking ensemble strategy based on the five baseline methods was utilized to improve the performance of the individual methods. For each of the 10 drugs, a single-drug ensemble model was developed using a standardized procedure. The global dataset was randomly split into training and validation sets in a 3:1 ratio using the train_test_split function from Python (v3.9.2) [27] with the Scikit-learn library (v1.3.2) [28]. The stacking framework consisted of two levels: base classifiers and a meta-classifier. The base classifiers included models such as the updated WHO catalog, TB Profiler, and SAM-TB, which produced binary predictions (0 for susceptibility, 1 for resistance), as well as GenTB and MD-CNN, which generated graded numerical predictions (0 to 1, representing the probability of resistance). These outputs were combined to create a new dataset, which served as input for the meta-classifier. The meta-classifier, implemented as a decision tree (DT) using Scikit-learn, was optimized through grid search with tenfold cross-validation, focusing on parameters such as "splitter," "criterion," "max_features," and "max_depth." Model performance was evaluated using the mean area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and their respective 95% confidence intervals (CIs). Furthermore, a merged dataset from the two additional cohorts was used for the external evaluation of the ensemble model.

Statistical analysis

Categorical variables are reported as count (%) and continuous variables as mean with standard deviations (SD) or median with interquartile ranges (IQR). To compare the AUC, sensitivity, and specificity between two methods, we used the Wilcoxon signed-rank test with *p* values adjusted according to the Benjamini–Hochberg procedure. For comparisons among three or more methods, we applied the Friedman test to evaluate AUC, sensitivity, and specificity. All statistical analyses were conducted using R software (version 4.3.3) [29], with a significance threshold set at $p < 0.05$.

Results

Overview of the global dataset and baseline methods

The global dataset comprised 36,385 MTB isolates collected from 45 countries across six continents (Fig. 1A). Ten lineages of the *Mycobacterium tuberculosis* complex (MTBC) were represented in the dataset: seven human-associated (L1 to L7) and three animal-associated (*M. bovis*-La1, *M. caprae*-La2, and *M. orygis*-La3) lineages. The majority of isolates belonged to lineage L4 (17,134), followed by L2 (9695), L3 (5438), and L1 (3434). Lesser numbers of genomes belonged to La1 (426), La3 (121), L5 (63), L6 (59), La2 (12), and L7 (3) (Fig. 1A). Not all isolates had phenotypic DST results for all anti-TB drugs (Fig. 1A) but were most frequently available for isoniazid (91.3% [33,224 of 36,385]) and rifampicin (90.1% [32,785 of 36,385]). There were fewer phenotypic DST results for new drugs such as delamanid (21.5% [7808 of 36,385]) and bedaquiline (22.9% [8335 of 36,385]; Fig. 1B). The prevalence of resistance across 15 drugs varied from a low of 0.08% for bedaquiline resistance to 34.3% for isoniazid resistance (Fig. 1B). Among the five baseline genotypic DST methods employed, the WHO catalog, TB Profiler, and SAM-TB could predict resistance for 15 drugs. In contrast, GenTB and MD-CNN did not predict resistance for 5 and 4 drugs, respectively, as they had not been trained for these drugs. All methods enabled resistance prediction for 10 drugs (Fig. 1B), which were used for subsequent analyses.

MD-CNN outperformed the other baseline methods

When we evaluated performance of each baseline genotypic DST method by comparing their predictions of drug resistance with the phenotypic DST results, the five methods exhibited varying AUCs (Fig. 2A).

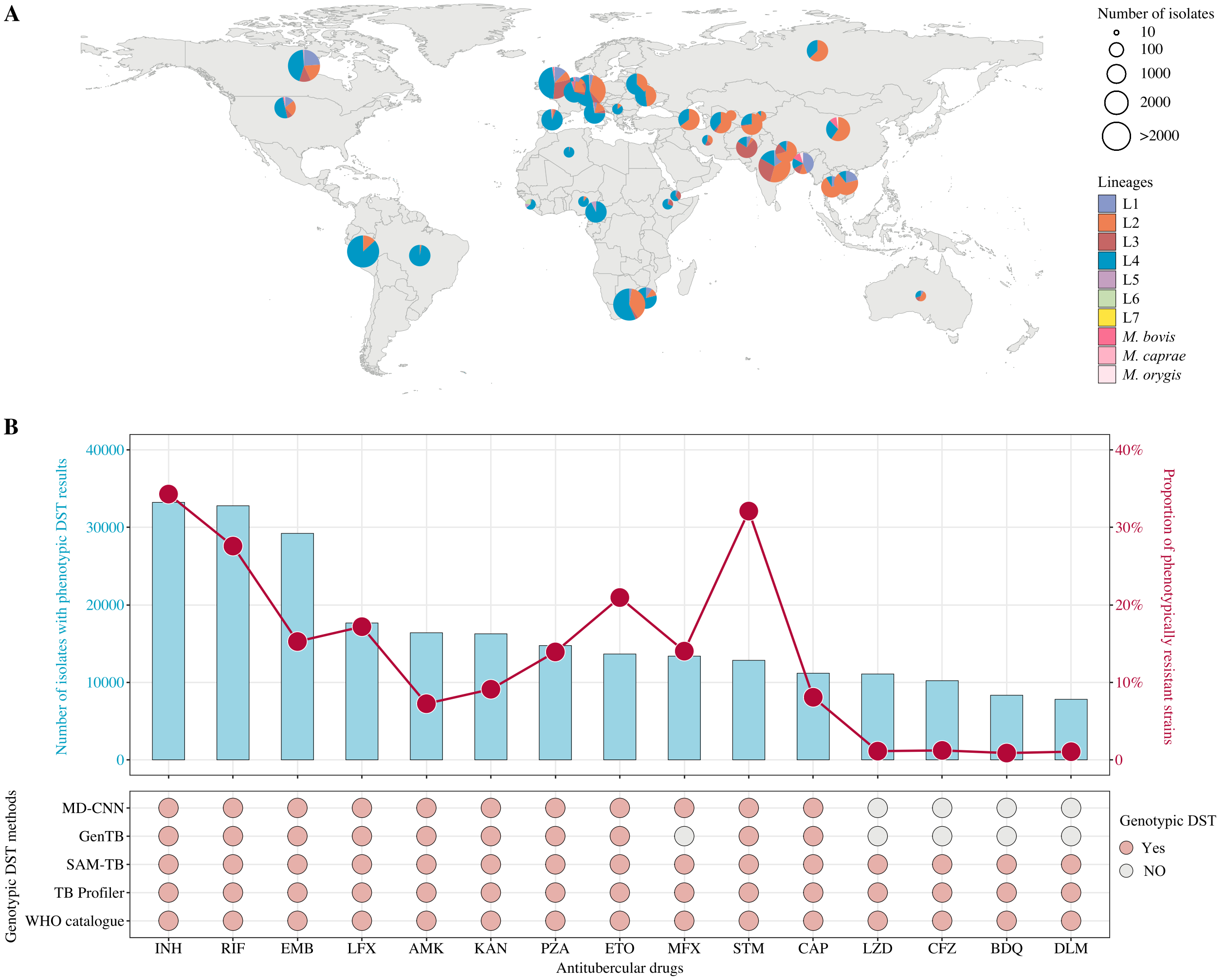

Fig. 1 Overview of the global dataset and baseline methods. **A** Global distribution of *Mycobacterium tuberculosis* in the global dataset. For global lineage distribution, pie charts represent the proportion of each lineage among isolates available from each country. **B** Phenotypic drug susceptibility testing (DST) of strains in the global dataset and the range of resistance prediction by different genotypic DST methods. INH, isoniazid; RIF, rifampicin; EMB, ethambutol; LFX, levofloxacin; AMK, amikacin; KAN, kanamycin; PZA, pyrazinamide; ETO, ethionamide; MXF, moxifloxacin; STM, streptomycin; CAP, capreomycin; LZD, linezolid; CFZ, clofazimine; BDQ, bedaquiline; DLM, delamanid

All methods showed the highest prediction accuracy for rifampicin (RIF; mean AUC 96.4% [95% CI 95.6–97.3%]) and the lowest for ethionamide (ETO; 77.1% [70.4–83.8%]; Fig. 2A). Diagnostic performance was higher for first-line drugs compared to second-line drugs (91.6% [88.6–94.5%] vs 86.4% [84.2–88.5%]; Wilcoxon signed-rank test $p < 0.05$), with the exception of lower accuracy for first-line drug pyrazinamide (PZA; 82.4% [75.9–88.9%]; Fig. 2A).

Using tenfold cross-validation, the mean AUC of MD-CNN (92.1% [89.8–94.4%]) was significantly higher than that of the other four methods: TB Profiler (88.2% [84.4–92.0%]; Wilcoxon signed-rank test with Benjamini–Hochberg FDR $p < 0.05$); the WHO catalog (88.0% [84.4–91.6%]; $p < 0.05$); SAM-TB (87.1% [82.6–91.7%]; $p < 0.05$); and GenTB (86.7% [81.8–91.7%]; $p < 0.05$; Fig. 2B). The WHO catalog demonstrated the highest specificity (97.3% [95.8–98.4%]), while TB Profiler had the highest sensitivity (79.5% [71.8–86.2%]) among the five baseline methods (Fig. 2B). Subsequently, when we compared the prediction performance of the baseline methods for both first-line (Fig. 2C) and second-line drugs (Fig. 2D), MD-CNN achieved a mean AUC of 95.1% [95% CI 92.3–97.9%] and 90.1% [87.9–92.4%], respectively. However, these differences between the five methods were not statistically significant (Friedman test first-line $p = 0.14$;

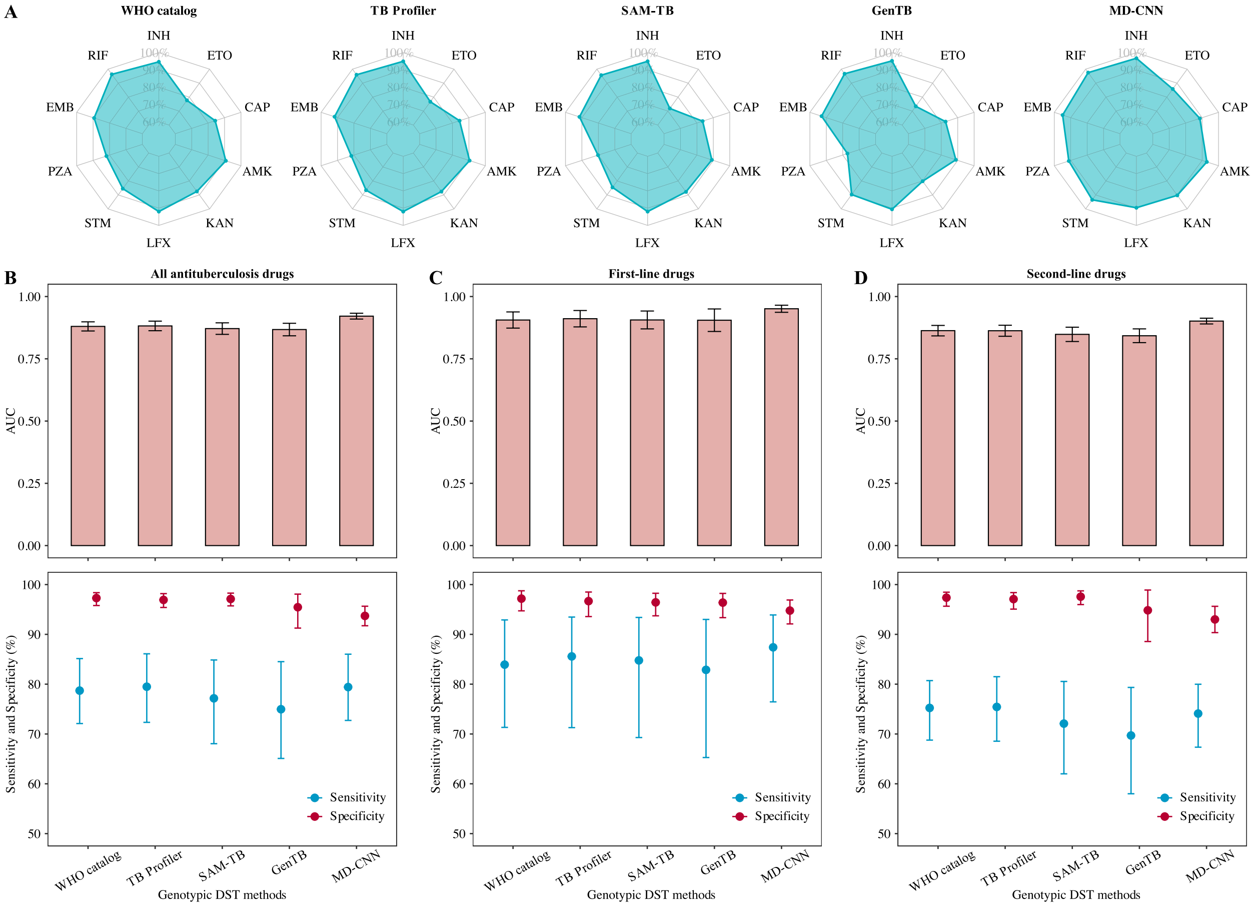

Fig. 2 Diagnostic performance of the five prediction tools across antituberculosis drugs. **A** Radar plot of mean AUC of the baseline method for each drug. Data are presented as pooled mean AUC, sensitivity, specificity, and the corresponding 95% CI of each baseline method for all (**B**), first-line (**C**), and second-line (**D**) drugs. First-line drugs: INH, RIF, EMB, and PZA. Second-line drugs: STM, LFX, KAN, AMK, CAP, and ETO. AUC, area under the receiver operating characteristic curve; INH, isoniazid; RIF, rifampicin; EMB, ethambutol; PZA, pyrazinamide; STM, streptomycin; LFX, levofloxacin; KAN, kanamycin; AMK, amikacin; CAP, capreomycin; ETO, ethionamide

second-line $p=0.08$), indicating only minor variability in the ability of the five baseline methods to detect drug resistance.

Discrepancies between phenotypic and genotypic DST

To identify specific challenges in genotypic DST methods, we detailed the discrepancies between genotypic and phenotypic DST results. The discrepant results were placed into two categories: (1) phenotypic DST detected drug resistance where genotypic DST indicated susceptibility (Fig. 3A), or false negatives, and (2) genotypic DST predicted resistance whereas phenotypic DST indicated susceptibility, or false positives (Fig. 3B). The first type of discrepancy is attributable to an incomplete catalog of DRMs, especially for ETO (43.4% [95% CI 35.5–51.4%]). This suggests that additional, novel mutations that can confer ETO resistance are not present in the DRM catalogs employed by the five tools (Fig. 3C). To identify novel potential DRMs that may confer ETO resistance, we searched for unique mutations in phenotypically resistant but genotypically susceptible strains that were not present in phenotypically susceptible strains. We identified 7 novel mutations that appeared repeatedly in these phenotypically ETO-resistant strains (Fig. 3D).

The second scenario of genotypically resistant strains that were phenotypically susceptible was most frequently observed for ethambutol (EMB; 32.2% [29.8–34.5%]), particularly when using machine learning-based genotypic DST methods (Fig. 3E). This phenomenon was predominantly due to DRMs associated with resistance to EMB (Fig. 3F), and the phenotypic incongruency (40.8% [605 of 1483]) most commonly involved the *embB* M306I DRM (Fig. 3F).

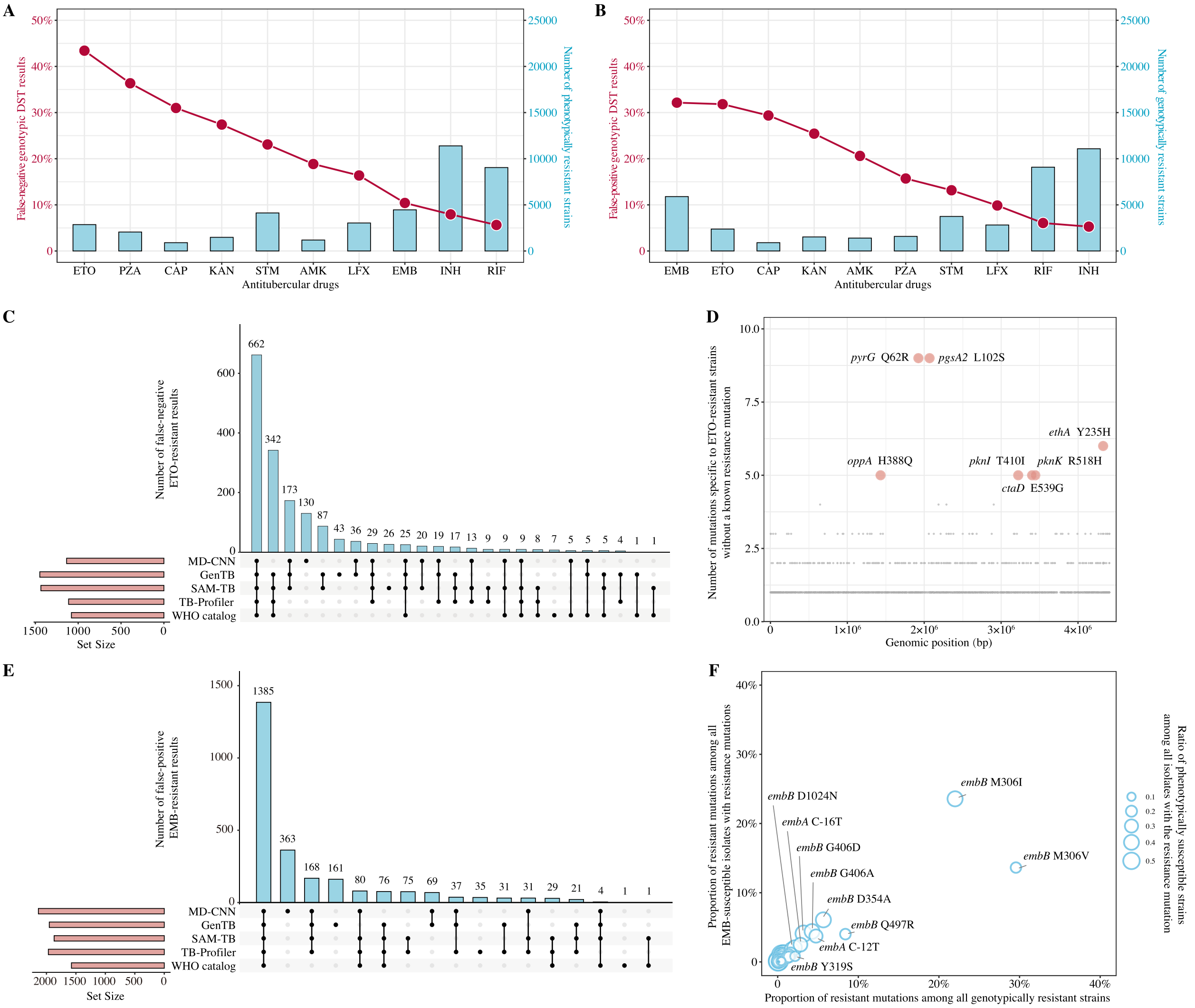

Fig. 3 The discrepancies between genotypic and phenotypic DST approaches. **A** The number of phenotypically resistant MTB strains and the proportion of genotypically susceptible strains among them. **B** The number of genotypically resistant strains and proportion of phenotypically susceptible strains among them. **C** Comparison of false-negative ETO susceptibility predictions (phenotypically resistant but genotypically susceptible) with different baseline methods. **D** The number of mutations specific to ETO-resistant strains without a known resistance mutation. **E** Comparison of different baseline methods for false-positive EMB susceptibility predictions (genotypically resistant but phenotypically susceptible). **F** The proportion of strains phenotypically susceptible to EMB among genotypically resistant strains carrying different EMB DRMs. INH, isoniazid; RIF, rifampicin; EMB, ethambutol; PZA, pyrazinamide; STM, streptomycin; LFX, levofloxacin; KAN, kanamycin; AMK, amikacin; CAP, capreomycin; ETO, ethionamide

Enhanced performance of drug resistance prediction by stacking ensemble model

To enhance the predictive accuracy of genotypic DST methods, we implemented an ensemble learning strategy (Fig. 4A). The stacking ensemble model showed robust predictive performance across all drugs, particularly improving accuracy for PZA (mean AUC 92.9% [95% CI: 90.3–95.6%]) and ETO (88.2% [85.8–92.0%]; Fig. 4B). The average AUC for the stacking model was 93.4% [91.4–95.4%], significantly surpassing the 92.2% [89.9–94.6%] of MD-CNN for all 10 drugs studied (Wilcoxon signed-rank test with Benjamini–Hochberg FDR $p < 0.05$). It also exhibited a higher sensitivity (84.1% [78.8–88.8%]) than all baseline methods (Fig. 4C). For first-line drugs, the stacking model's mean AUC of 95.8% [93.5–98.1%] was comparable to other methods (Friedman test $p = 0.27$). For second-line drugs, the ensemble model (91.8% [89.6–94.0%]) outperformed MD-CNN (90.3% [87.9–92.7%]), although this result was only marginally significant (Wilcoxon signed-rank test with Benjamini–Hochberg FDR

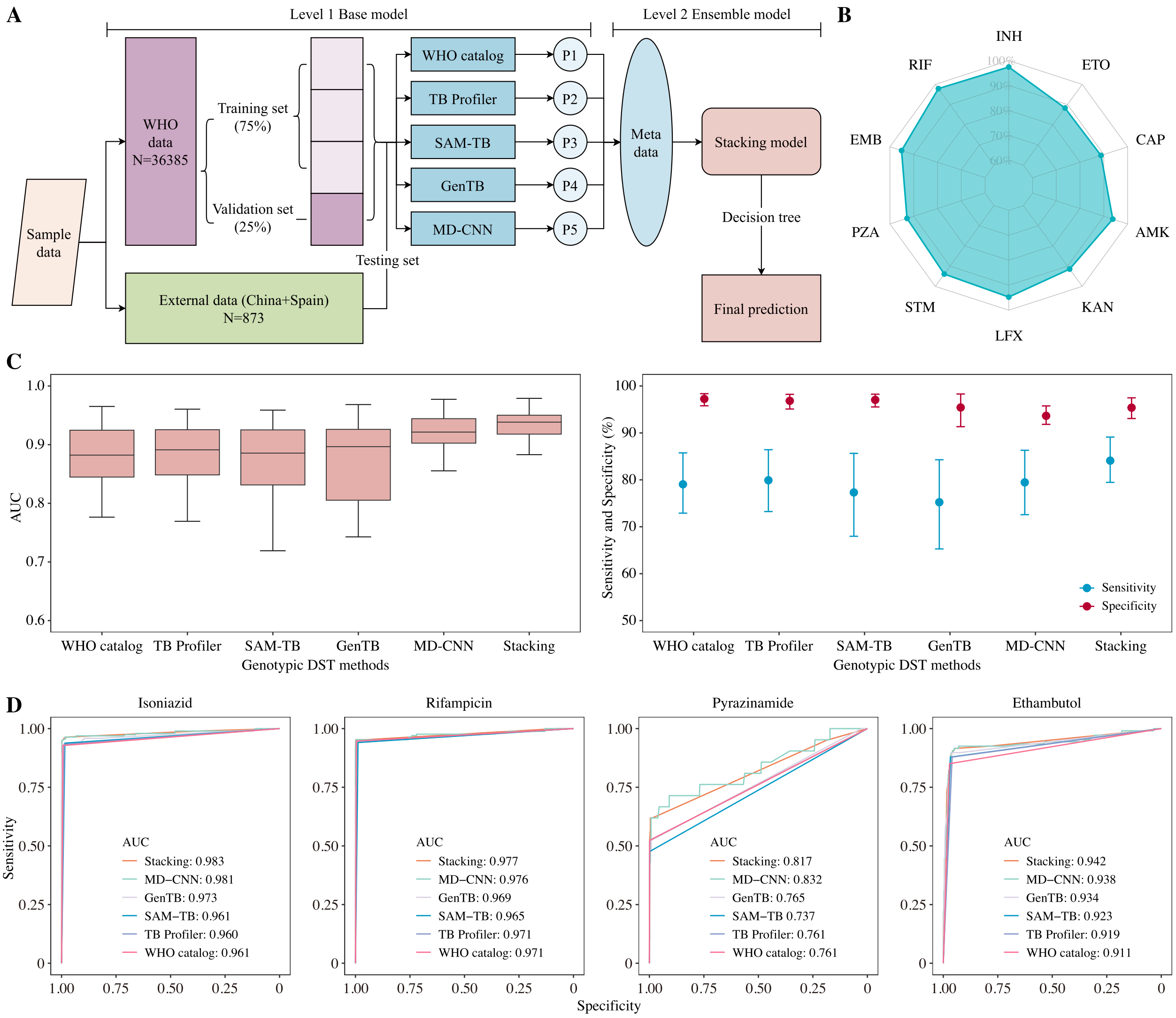

Fig. 4 The stacking ensemble machine learning model development and validation. **A** Schematic diagram of the stacking ensemble model. **B** Radar plot of mean AUC of the stacking ensemble model for each drug. **C** Comparison of mean AUC, sensitivity, and specificity, obtained under the stacking ensemble model and other baseline genotypic DST methods. **D** AUCs of the stacking ensemble model for isoniazid, rifampicin, pyrazinamide, and ethambutol with the external validation datasets. AUC, area under the receiver operating characteristic curve; INH, isoniazid; RIF, rifampicin; EMB, ethambutol; PZA, pyrazinamide; STM, streptomycin; LFX, levofloxacin; KAN, kanamycin; AMK, amikacin; CAP, capreomycin; ETO, ethionamide

$p = 0.09$). Validation using an external test dataset further confirmed that the stacking model overall performs better than the other baseline models (Fig. 4D).

Discussion

This study conducted a systematic comparison of the WHO catalog and other genotypic DST methods using a large collection of MTB WGS and phenotypic DST data from around the world. Our findings reveal that MD-CNN, with a mean AUC of 92.1% [95% CI 89.7–94.4%], surpasses the other four methods in prediction performance. Building on this, we used a stacking ensemble strategy to develop a new machine learning model (93.4% [91.5–95.3%]) that outperformed the five existing methods, particularly for predicting resistance to second-line drugs (91.8% [89.7–93.8%]).

Although we found minor variations in the accuracy of the five baseline genotypic DST methods for predicting drug resistance, the differences were not significant. In our initial evaluation of the five methods for predicting resistance to ten anti-TB drugs, MD-CNN surpassed the other four in overall prediction performance (Fig. 2B).

In subsequent comparisons for first-line and second-line drugs, however, MD-CNN did not show superior accuracy and the performance of the five methods was similar in predicting resistance across both drug categories (Fig. 2C, D). Aside from MD-CNN, only TB Profiler showed a higher overall performance than SAM-TB (mean AUC 88.2% [84.4–92.0%] vs 87.1% [82.6–91.7%]; Wilcoxon signed-rank test with Benjamini–Hochberg FDR $p = 0.04$), and there were no significant performance differences among the other tools. Most current genotypic DST methods, despite diverse principles, show similar accuracy and share similar challenges concerning false negatives and false positives. Our findings indicate that these methods still fall short of the WHO-defined standards for next-generation DST tools [30], with only a few achieving the minimal sensitivity and specificity thresholds for a limited number of drugs (Additional file 3: Fig. S1). Future efforts should prioritize addressing these limitations by expanding knowledge about mutations conferring resistance rather than developing comparable alternative methods to detect mutations, as only through such advancements can genotypic DST methods move closer to meeting WHO benchmarks and gaining broader clinical acceptance.

The better performance of the stacking ensemble model derives from its improved accuracy for predicting resistance to second-line drugs. There are strong correlations of specific genetic mutations with resistance to first-line drugs [5, 31], but the correlation is often ambiguous and complex for second-line drugs. Resistance to several second-line drugs is associated with mutations in several different genes, the functions and interactions of which are not fully understood [32]. This explains why there is less accuracy for predicting susceptibility to second-line drugs [11, 12]. In this study, the stacking ensemble model markedly improved both prediction performance and sensitivity relative to the other methods, but with decreased specificity, highlighting the critical need for a better understanding of the mutations that confer resistance to second-line drugs.

Errors in genotypic DST, such as false positives (phenotypically susceptible but genotypically resistant) and false negatives (phenotypically resistant but genotypically susceptible), carry distinct implications for tuberculosis treatment. False negatives are particularly dangerous because they can lead to inappropriate treatment of resistant infections, thus prolonging transmission and increasing morbidity and mortality [33]. Reducing false negatives by increasing sensitivity is therefore often preferable, even if it increases false positives as a result of a reduced specificity [34]. False positives, however, can lead to inappropriate antibiotic use, potentially harming the patient and heightening the risk of developing resistance to critical last-line antibiotics [35].

False negatives highlight the need for a better understanding of resistance mechanisms. The lists of DRM include most mutations conferring resistance to both first- and second-line drugs, although there are less common variants, especially for second-line agents, which may not be included in the lists of DRM and thus will not be queried [36]. Classical single-variant-based association tests have low statistical power for detecting rare variants unless the sample sizes or effect sizes are exceptionally large [37]. Therefore, improvements in prediction accuracy will require analyzing additional clinical MTB genomic and phenotypic data, especially strains from patients who were treated with second-line agents. Our analysis of WGS data from strains that are phenotypically resistant but genotypically susceptible to ETO identified several potential new DRMs. In addition to the Y235H mutation in the *ethA* gene, which is known to be associated ETO resistance gene, we identified six additional mutations associated with ETO resistance. These included variants in *pyrG*, which has shown epistatic interactions with *ethA* [38]. However, additional mutant library screening and in vitro experimental studies are needed to confirm that these mutations truly influence MTB susceptibility to ETO.

Epistatic interactions may account for the false-positive predictions by genotypic DST, particularly in strain backgrounds not represented in DRM discovery sets. Previous research has demonstrated that the genetic background of MTB strains can influence the level of resistance conferred by DRMs in vitro [39], but the precise mechanisms behind these epistatic interactions are still unclear. Comprehensive empirical testing is necessary to elucidate the variability of specific mutations on drug susceptibility in different strain backgrounds [40]. Innovative statistical methods, combined with large-scale pathogen genomic datasets, could facilitate the identification of new epistatic interactions [41].

One of the main limitations of this study is the potential bias due to the overlap between our dataset and the WHO mutation catalog, which may favor the performance of WHO catalog-based tools. Future evaluations on independent datasets are crucial to validate the generalizability and robustness of these tools across diverse populations. Another limitation is that we did not perform genotypic DST for new and repurposed drugs such as bedaquiline, delamanid, and clofazimine because knowledge of the DRMs that confer resistance to these drugs is known to be incomplete. Finally, the external dataset was only used to validate the stacking ensemble model's prediction accuracy for first-line drugs, because we could not identify a suitable external dataset to validate its performance for second-line drugs.

Conclusions

In conclusion, our study compared the accuracy of the WHO catalog and four other genotypic DST methods for predicting drug susceptibility in a large, globally representative MTB dataset with high-quality WGS and phenotypic data. Based on this evaluation, we developed a new stacking ensemble model that further enhances the performance of genotypic DST methods. We showed that a stacking ensemble strategy is feasible and has the potential to optimize exiting genotypic DST methods.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01458-0.

Additional file 1. This file includes accession number, dataset of origin, resistance phenotypes, and relevant mutations utilized by different approaches to infer resistance for isolates in the WHO dataset

Additional file 2. This file includes accession number, dataset of origin, resistance phenotypes, and relevant mutations utilized by different approaches to infer resistance for isolates in the external validation dataset

Additional file 3. Fig. S1 Sensitivity and specificity of five prediction tools across antituberculosis drugs

Additional file 4. This file includes detailed statistical analysis data, performance values, 95% confidence intervals, and computed *p* values for each model comparison.

Acknowledgements

Not applicable.

Authors' contributions

All authors designed the paper by discussing the key concerns to be include. Y. Chen, X. Zhang, P. Xu and Q. Gao designed research. Y. Chen, X. Zhang and J. Liang collected the data or processed samples. Y. Chen, X. Zhang, Q. Jiang and M. Peierdun analysed the data. Y. Chen, X. Zhang, H.E. Takiff and Q. Gao wrote and edited the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by Shenzhen High-level Hospital Construction Fund (grant number G2022157), Shenzhen Clinical Research Center for Tuberculosis (grant number 20210617141509001), the National Natural Science Foundation of China (grant number 82272376 and 32470186 to QG), Shanghai Municipal Science and Technology Major Project (grant number ZD2021CY001), and the Natural Science Foundation of Hubei Province of China (grant number 2023AFB063 to QJ).

Data availability

Data is provided within the supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]National Clinical Research Center for Infectious Diseases, Shenzhen Clinical Research Center for Tuberculosis, Shenzhen Third People's Hospital, Shenzhen, Guangdong, China. [2]Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Sciences, Shanghai Medical College, Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Shanghai, China. [3]School of Public Health, Public Health Research Institute of Renmin Hospital, Wuhan University, Wuhan, China. [4]Department of Epidemiology and Biostatistics, School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang, China. [5]Instituto Venezolano de Investigaciones Cientificas (IVIC), Caracas, Venezuela.

Received: 23 September 2024 Accepted: 13 March 2025
Published online: 26 March 2025

References

1. WHO. Global tuberculosis report 2023. World Health Organization; 2023. Available from: https://www.who.int/publications/i/item/9789240083851.
2. Kendall EA, Fofana MO, Dowdy DW. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. Lancet Respir Med. 2015;3(12):963–72.
3. Yang C, Luo T, Shen X, Wu J, Gan M, Xu P, 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.
4. Finci I, Albertini A, Merker M, Andres S, Bablishvili N, Barilar I, et al. Investigating resistance in clinical *Mycobacterium tuberculosis* complex isolates with genomic and phenotypic antimicrobial susceptibility testing: a multicentre observational study. Lancet Microbe. 2022;3(9):e672–82.
5. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations associated with drug resistance: a genotypic analysis. Lancet Microbe. 2022;3(4):e265–73.
6. WHO. Technical report on critical concentrations for drug susceptibility testing of isoniazid and the rifamycins (rifampicin, rifabutin and rifapentine). World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240017283.
7. Danchuk SN, Solomon OE, Kohl TA, Dreyer V, Barilar I, Utpatel C, et al. Challenging the gold standard: the limitations of molecular assays for detection of *Mycobacterium tuberculosis* heteroresistance. Thorax. 2024;79(7):670–5.
8. MacLean E, Kohli M, Weber SF, Suresh A, Schumacher SG, Denkinger CM, Pai M. Advances in molecular diagnosis of tuberculosis. J Clin Microbiol. 2020;58(10):e01582–619.
9. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11(1):41.
10. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, 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.
11. Yang T, Gan M, Liu Q, Liang W, Tang Q, Luo G, et al. SAM-TB: a whole genome sequencing data analysis website for detection of Mycobacterium tuberculosis drug resistance and transmission. Brief Bioinform. 2022;23(2):bbac030.
12. Gröschel MI, Owens M, Freschi L, Vargas R, Marin MG, Phelan J, et al. GenTB: a user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. Genome Med. 2021;13(1):138.
13. Jiang Z, Lu Y, Liu Z, Wu W, Xu X, Dinnyés A, et al. Drug resistance prediction and resistance genes identification in Mycobacterium tuberculosis based on a hierarchical attentive neural network utilizing genome-wide variants. Brief Bioinform. 2022;23(3):bbac041.

14. Green AG, Yoon CH, Chen ML, Ektefaie Y, Fina M, Freschi L, et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*. Nat Commun. 2022;13(1):3817.
15. WHO. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240028173.
16. WHO. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd ed. World Health Organization; 2023. Available from: https://www.who.int/publications/i/item/9789240082410.
17. Li J, Yang T, Hong C, Yang Z, Wu L, Gao Q, et al. Whole-genome sequencing for resistance level prediction in multidrug-resistant tuberculosis. Microbiol Spectr. 2022;10(3): e0271421.
18. Garcia-Marin AM, Cancino-Munoz I, Torres-Puente M, Villamayor LM, Borras R, Borras-Manez M, et al. Role of the first WHO mutation catalogue in the diagnosis of antibiotic resistance in *Mycobacterium tuberculosis* in the Valencia Region, Spain: a retrospective genomic analysis. Lancet Microbe. 2024;5(1):e43–51.
19. Sequence Read Archive (SRA). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2009. Available from: https://www.ncbi.nlm.nih.gov/sra/.
20. Chen Y, Jiang Q, Peierdun M, Takiff HE, Gao Q. The mutational signatures of poor treatment outcomes on the drug-susceptible Mycobacterium tuberculosis genome. eLife. 2023;12:e84815.
21. Joshi NA, Fass JN. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33) [Software]. 2011. Available from: https://github.com/najoshi/sickle2011.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
24. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
25. Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L, et al. The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. Genome Biol. 2017;18(1):71.
26. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. Genome Med. 2020;12(1):114.
27. Python Software Foundation. Python programming language (version 3.9.2) [Software]. 2021. Available from: https://www.python.org/downloads/release/python-392/.
28. Fabian P. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
29. R Core Team. The R project for statistical computing (version 4.3.3) [Software]. 2024. Available from: https://cran.r-project.org/bin/windows/base/old/4.3.3/.
30. WHO. Target product profile for next-generation drug-susceptibility testing at peripheral centres. World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240032361.
31. Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K. Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. J Antimicrob Chemother. 2018;73(5):1138–51.
32. The CRyPTIC Consortium. Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. PLoS Biol. 2022;20(8): e3001755.
33. Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. J Clin Microbiol. 2019;57(3):e01405–18.
34. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. J Clin Microbiol. 2014;52(4):1182–91.
35. Lange C, Dheda K, Chesov D, Mandalakas AM, Udwadia Z, Horsburgh CR. Management of drug-resistant tuberculosis. Lancet. 2019;394(10202):953–66.
36. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. EBioMedicine. 2019;43:356–69.
37. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
38. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2018;50(2):307–16.
39. Castro RAD, Ross A, Kamwela L, Reinhard M, Loiseau C, Feldmann J, et al. The genetic background modulates the evolution of fluoroquinolone-resistance in *Mycobacterium tuberculosis*. Mol Biol Evol. 2020;37(1):195–207.
40. Chen Y, Takiff HE, Gao Q. Phenotypic instability of *Mycobacterium tuberculosis* strains harbouring clinically prevalent drug-resistant mutations. Lancet Microbe. 2023;4(5): e292.
41. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet. 2017;13(2): e1006508.
42. Chen Y, Zhang X. Available from: https://github.com/zxcwuzheng/stopTB2024.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.