


SOFTWARE

Open Access

# GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning



Matthias I. Gröschel<sup>1</sup> , Martin Owens<sup>1</sup>, Luca Freschi<sup>1</sup>, Roger Vargas Jr<sup>1,2</sup>, Maximilian G. Marin<sup>1,2</sup>, Jody Phelan<sup>3</sup>, Zamin Iqbal<sup>4</sup>, Avika Dixit<sup>1,5</sup> and Maha R. Farhat<sup>1,6\*</sup>

## Abstract

**Background:** Multidrug-resistant *Mycobacterium tuberculosis* (*Mtb*) is a significant global public health threat. Genotypic resistance prediction from *Mtb* DNA sequences offers an alternative to laboratory-based drug-susceptibility testing. User-friendly and accurate resistance prediction tools are needed to enable public health and clinical practitioners to rapidly diagnose resistance and inform treatment regimens.

**Results:** We present Translational Genomics platform for Tuberculosis (GenTB), a free and open web-based application to predict antibiotic resistance from next-generation sequence data. The user can choose between two potential predictors, a Random Forest (RF) classifier and a Wide and Deep Neural Network (WDNN) to predict phenotypic resistance to 13 and 10 anti-tuberculosis drugs, respectively. We benchmark GenTB's predictive performance along with leading TB resistance prediction tools (Mykrobe and TB-Profiler) using a ground truth dataset of 20,408 isolates with laboratory-based drug susceptibility data. All four tools reliably predicted resistance to first-line tuberculosis drugs but had varying performance for second-line drugs. The mean sensitivities for GenTB-RF and GenTB-WDNN across the nine shared drugs were 77.6% (95% CI 76.6–78.5%) and 75.4% (95% CI 74.5–76.4%), respectively, and marginally higher than the sensitivities of TB-Profiler at 74.4% (95% CI 73.4–75.3%) and Mykrobe at 71.9% (95% CI 70.9–72.9%). The higher sensitivities were at an expense of  $\leq 1.5\%$  lower specificity: Mykrobe 97.6% (95% CI 97.5–97.7%), TB-Profiler 96.9% (95% CI 96.7 to 97.0%), GenTB-WDNN 96.2% (95% CI 96.0 to 96.4%), and GenTB-RF 96.1% (95% CI 96.0 to 96.3%). Averaged across the four tools, genotypic resistance sensitivity was 11% and 9% lower for isoniazid and rifampicin respectively, on isolates sequenced at low depth ( $< 10\times$  across 95% of the genome) emphasizing the need to quality control input sequence data before prediction. We discuss differences between tools in reporting results to the user including variants underlying the resistance calls and any novel or indeterminate variants

**Conclusions:** GenTB is an easy-to-use online tool to rapidly and accurately predict resistance to anti-tuberculosis drugs. GenTB can be accessed online at <https://gentb.hms.harvard.edu>, and the source code is available at <https://github.com/farhat-lab/gentb-site>.

\* Correspondence: [maha\\_farhat@hms.harvard.edu](mailto:maha_farhat@hms.harvard.edu)

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>6</sup>Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Keywords:** Tuberculosis, Drug resistance, Drug-susceptibility testing, Diagnostics, Whole genome sequencing, Machine learning, MDR-TB, XDR-TB

## Background

Human tuberculosis, a chronic infectious disease caused by members of the *Mycobacterium tuberculosis* complex, is a leading cause of death from a bacterial infectious agent [1]. The proliferation of multidrug-resistant tuberculosis (MDR-TB) is threatening TB prevention and control activities worldwide [1]. Timely detection of antimicrobial resistance is vital to guide therapeutic options and contain transmission. Antimicrobial resistance is conventionally determined by in vitro drug susceptibility tests (DST) on solid or liquid antibiotic-containing culture, which uses drug-specific testing breakpoints (“critical concentration”) to classify the infecting strain into drug-susceptible or drug-resistant [2]. Being contingent on mycobacteria’s slow growth rate, these phenotypic tests require days to weeks [3, 4]. In contrast, molecular methods have emerged as rapid resistance prediction alternatives to complement and speed up traditional DST, leveraging known and reliable genotype-phenotype relationships between variants in the *M. tuberculosis* genome and in vitro drug resistance [5].

Over recent years, whole-genome sequencing (WGS) of *M. tuberculosis* has become an affordable tool to provide genetic information for genotypic resistance prediction and high-resolution outbreak reconstruction [6]. Large scale genotype-phenotype assessments have demonstrated high diagnostic accuracy for clinical use to predict susceptibility to first-line drugs based on WGS [7]. While some evidence suggests that WGS-based mycobacterial diagnostics is feasible with fast turnaround in a clinical research setting further validation studies under routine care conditions are warranted [3]. Following these results, public health authorities have begun to discontinue phenotypic testing when pan susceptibility is predicted from the genotype, a step with considerable cost and time benefits [8]. Start-to-end applications which analyze sequencing data to predict resistance phenotypes and are accessible to non-bioinformatic experts are required as WGS based analyses become part of the standardized diagnostic process in clinical laboratories. A range of published tools available for command-line [9, 10] or web-based/desktop use [11–13] or both [14, 15] exists. These applications vary in quality control and sequence preprocessing steps and rely on detecting pre-defined resistance-conferring mutations such as single nucleotide polymorphisms (SNPs) or small insertions/deletions (indels) in the WGS data to predict the resistance phenotype. They also vary in the type of information fed back to the user including error rates and specific variants detected.

Here, we present GenTB (<https://gentb.hms.harvard.edu>), an open user-friendly start-to-end application to predict drug resistance phenotypes to 13 drugs from WGS data. Resistance prediction is made based on a previously observed set of variant positions spanning 18 resistance-associated genetic loci and a validated random forest (RF) classifier [16] as well as a wide and deep neural network (WDNN) combining a logistic regression model with a multilayer perceptron [17]. GenTB provides access to multivariate statistical models of resistance to non-expert users. These models can consider the simultaneous effect of one or more mutations. GenTB’s online interface allows users to interactively explore the sequencing data, prediction results and geographic distributions. The GenTB analysis pipeline is also available for command-line use wrapped in *Snake-make* [18]. In this study, we benchmark these two classification models implemented in GenTB along with two other tools with a command-line interface, *TB-profiler* [14], and *Mykrobe* [15], on a large dataset of > 20 k clinical *M. tuberculosis* isolates (Additional file 1) starting from raw Illumina sequence data.

## Implementation

### Backend and website build

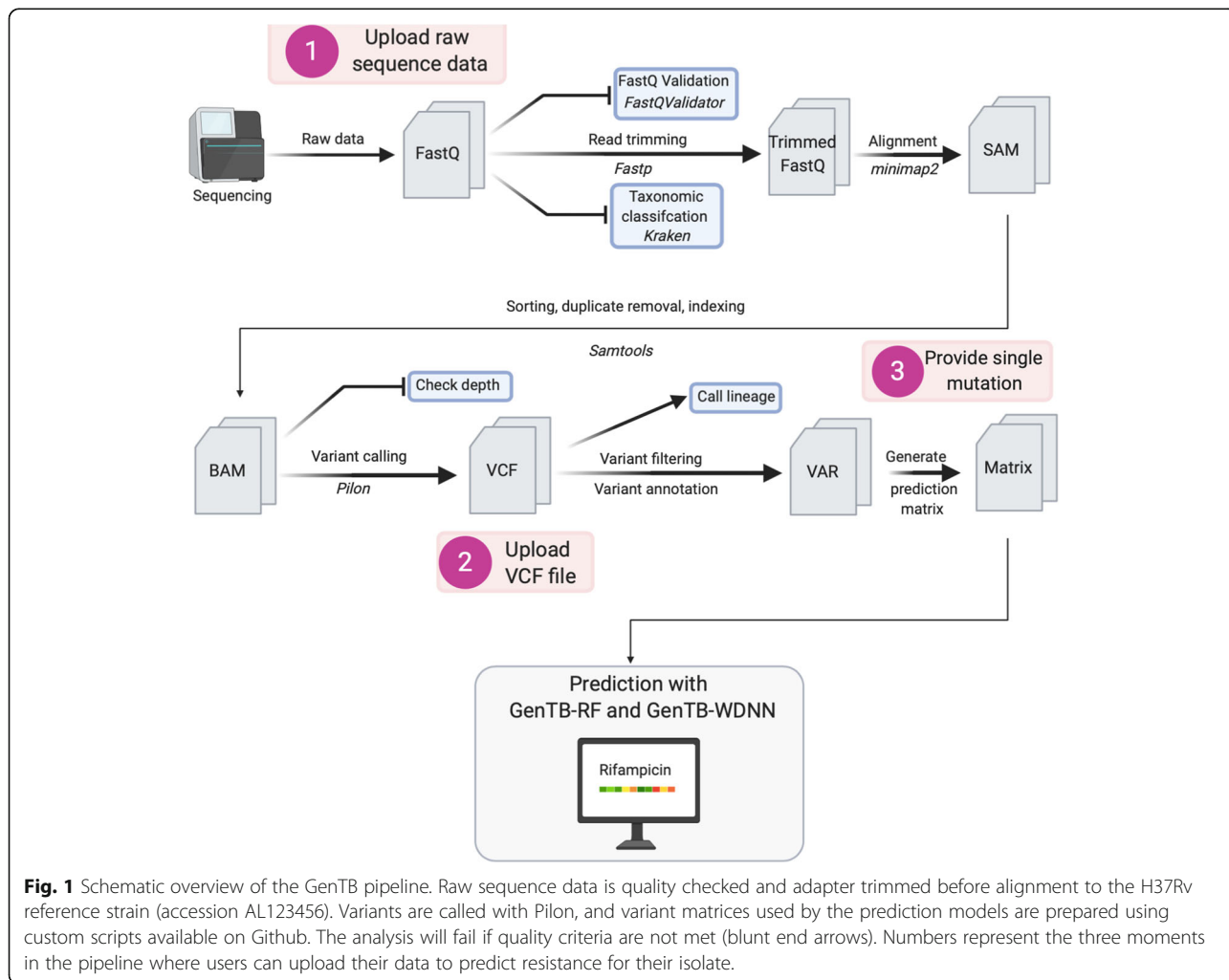
GenTB is a bespoke Django website hosted by the Harvard Medical School O2 high performance computing environment and collaboratively developed on GitHub (<https://github.com/farhat-lab/gentb-site>) [19]. The website uses off-the-shelf frontend components; Bootstrap for styling and mobile-friendly delivery, nvd3 for plots and graphs, resumable.js for robust uploading and supplements these with custom Javascript functionality for integration. The backend is a Python-Django web service using a PostgreSQL database which integrates with Dropbox for file uploading, and python-chore for slurm cluster job submission and management. GenTB predict jobs are run by modular programs organized into pipelines. The modularity allows for easy maintenance and management of dependencies and outputs. Administration screens allow a non-expert developer design new program calls and construct new pipelines and integrate them without redeployment of the website. Further tools provide error tracking. GenTB predict results are integrated into the PostgreSQL database allowing website generated plots to be populated quickly. All generated files for the intermediary pipeline steps are provided for download by the user. GenTB Map uses a PostGIS database to rapidly link strain mutation and lineage

information with geo-spatial objects; these are fed into the leaflet.js display to render strain information to the user. Map allows users to display strain data groupings by country, lineage, drug resistance phenotype, or specific genetic mutation through tabs that can nest the groups in any order.

**Raw read processing**

Upon uploading single-end or paired-end FastQ files, GenTB first validates the input using *fastQValidator* (Fig. 1). Low-quality reads and sequencing adapters are then trimmed with *fastp* [20]. *Kraken* is used as a quality control step to assess the percentage of reads that map to the *M. tuberculosis* complex using a custom-built *Kraken* database comprising *M. tuberculosis* complex reference sequences [21]. Reads not classified as *M. tuberculosis* complex are filtered internally using *seqtk* (<https://github.com/lh3/seqtk>). Paired-read matching is performed with *fastq-pair* (<https://github.com/linsalrob/fastq-pair>) followed by *minimap2* alignment (parameters: -ax sr) of reads to the H37Rv reference genome

(AL123456) [22]. In the present benchmarking we removed unclassified reads only for those isolates with > 10% non-*M. tuberculosis* reads. *Samtools* is used for sorting the aligned reads, removing duplicates, and indexing [23]. After read mapping is performed the coverage of drug resistance genes is confirmed (Additional file 2: Table S1). Sequence read datasets with a coverage of < 95% at 10x or less across these resistance genes will not be further processed, and an error message is displayed to the user. Variants are called with *pilon* (parameters: default) [24] to obtain SNPs and indels in the variant calling format (VCF) requiring that they have a PASS or Amb filter tags with read allele frequency > 0.40. The allele frequency threshold of 0.40 was chosen based on our observation that lower thresholds only marginally increased sensitivity of resistance prediction with a larger decrease in specificity (see Farhat et al [16]). *Fast-Lineage-Caller* then detects the *M. tuberculosis* lineage based on five lineage typing schemes as implemented by Freschi et al. [25]. Subsequently, invariant sites in the VCF file are removed, and



**Fig. 1** Schematic overview of the GenTB pipeline. Raw sequence data is quality checked and adapter trimmed before alignment to the H37Rv reference strain (accession AL123456). Variants are called with Pilon, and variant matrices used by the prediction models are prepared using custom scripts available on Github. The analysis will fail if quality criteria are not met (blunt end arrows). Numbers represent the three moments in the pipeline where users can upload their data to predict resistance for their isolate.

a custom Perl script annotates each variant as frameshift, synonymous or non-synonymous, stop codon, indel along with the H37Rv locus tag for each respective gene. A custom python script generates a matrix file with all model features/variables in the columns used as input to the two prediction steps specified below. These scripts are available from Github [19] and are open source (AGPLv3 license). All intermediate sequence files are accessible to the user for download and verification.

For runtime evaluation, we pulled start and end time of all successfully completed pipeline runs submitted between April 12th and May 12th and computed average and median processing times.

### Operation

GenTB is a free tool and registration is open to everyone. User registration is needed for security and to allow users to run predictions, track intermediary files and results. Users with low internet bandwidth can use the *Dropbox* integration to upload files. Both raw sequence reads and variants in variant call format (VCF) can be uploaded for resistance prediction. The required minimum genomic input, i.e., in case of targeted sequencing data, is specified on the input page and derived from Farhat et al. [16]. The user can select an option to delete uploaded source data after prediction or otherwise to save it for their future access through GenTB. Files are user-specific and not shared or accessible by others. Users can submit their genomic data for prediction and log off and will be sent an email with a link to the results when they are completed. Their prediction result will be stored indefinitely unless the user deletes it. Raw sequence data and intermediary files will be stored for three days.

The upload and processing stability of the GenTB online interface has been tested with up to 300 isolates uploaded in one batch. For batch processing of larger numbers of raw sequence data, we provide a command-line GenTB workflow based on *Snakemake* v5.20.1 [18] where dependent software will be sourced via *conda* [26]. The *Snakemake* workflow can be accessed via Github (<https://github.com/farhat-lab/gentb-snakemake>) [27]. This repository contains a README file detailing the installation process and a description of the output files.

### Validation sequencing and phenotype data

We collated a database of 20,408 Illumina raw sequence read datasets for which laboratory-based phenotypic DST data was available from public sources (Additional file 1). Sequence data was downloaded from NCBI nucleotide databases. Custom scripts were used to pool the phenotype data from Patric [28], ReseqTB [29], Zignol et al. [30], Wollenberg et al. [31], Phelan et al. [32], Hicks et al. [33], Coll et al. [34], and Dheda et al. [35]

(scripts available at <https://github.com/farhat-lab/resdata-ng>). Phenotypic testing was performed using WHO endorsed methods (Additional file 2: Table S2). Sequence data was merged in case of multiple sequencing runs per isolate for downstream processing and resistance prediction. The 20,408 sequence read datasets are not completely independent of the original training datasets with a small overlap (< 5%) and we thus do not expect this to affect the diagnostic accuracy.

### Genotypic resistance prediction using two statistical models

Two multivariate models are used to predict the resistance phenotype, an RF model (GenTB-RF) and a WDNN (GenTB-WDNN). GenTB-RF was trained on isolates with available resistance phenotype data and was validated as described in Farhat et al. [16]. Briefly, 1397 clinical isolates sampled as detailed in Farhat et al. [16] underwent targeted sequencing at 18 drug resistance loci using molecular inversion probes and in parallel underwent binary drug culture-based DST to 13 drugs (Additional file 2: Table S1). One RF was built for each drug using the *randomForest* R package (v. 4.6.7) with a subset of the total 992 SNPs/indels observed. Variants of highest importance for resistance prediction to each drug were selected by iteratively paring down the model and measuring loss of performance. Important variants are shown in Additional file 2: Fig. S1 for isoniazid and rifampicin.

Pyrazinamide resistance is known to rely on a large number of individually rare variants. Given the large increase in published *M. tuberculosis* WGS and linked DST data as well as the recent implication of novel resistance loci we retrained the pyrazinamide RF here using a newer version of *randomForest* R package (v. 4.6.-14) on variants in the genes *pncA*, *panD*, *clpC1*, and *clpP* [36]. We used 75% (15,267 isolates) of the dataset to train the model and 25% (5098 isolates) to validate its performance. During retraining, we excluded silent variants, those that occurred only in phenotypically susceptible isolates, and the final model was trained on 393 variants occurring in 3,262 phenotypically pyrazinamide resistant isolates [25]. We chose the *randomForest* *mtry* variable that yielded the smallest out-of-bag error and varied the *classwt* variable to maximize the sum of sensitivity and specificity.

GenTB-WDNN is a multitask logistic regression model combined with a multilayer perceptron. It has been previously shown to have equal or higher performance than the RF architecture when both are trained on the same data [17]. GenTB-WDNN was trained on 3,601 isolates (sampled as detailed in Chen et al. [17]) for 11 drugs using the *Keras* 2.2.4 library in Python 3.6 with a *TensorFlow* 1.8.0 backend. The model uses 222

features (i.e., SNPs or small insertions/deletions) along with derived variables (i.e., the number of non-synonymous SNPs across all resistance-conferring genes) to predict the resistance phenotype. GenTB-WDNN was trained on the same genetic loci like GenTB-RF plus the resistance genes *rpsA* (plus its promoter region) and *eis* (plus its promoter region) (Additional file 2: Table S1).

#### Performance of GenTB and comparison with other tools

To assess the performance of GenTB for predicting resistance, all isolates were processed through the GenTB pipeline. We compared the diagnostic accuracy with two leading resistance prediction tools, *TB-profiler* 2.8.12 [14] and *Mykrobe* v0.9.0 [15], that were run with default parameters. These two tools rely on a curated list of mutations and make a resistance call once they identify one of these mutations in a sample. The two tools and two GenTB prediction models' predictive ability was obtained by comparing the genotypic prediction to the phenotype data that was considered the ground truth. We calculated the true positive rate (sensitivity), the true negative rate (specificity), and area under the receiver operating curve (AUC for short) to measure test accuracy for each drug and tool. We evaluated 1,000 probability thresholds per drug to call resistance or susceptibility for GenTB-RF while using the GenTB-WDNN thresholds described in Chen et al. [17] (Additional file 2: Fig. S2 and Additional file 2: Fig. S3).

#### Statistical analyses and data visualization

Prediction files from all tools were parsed and analyzed in Jupyter Notebooks running Python 3.7 using the Pandas [37] and JSON libraries. Receiver operating characteristic curves were plotted using the Seaborn library [38]. The Vioplot package was used for violin plots [39]. We used the randomforestExplainer v0.10.1 package to visualize important variants in random forest models. Summary tables were created in R version 3.6.3 [40] using the packages from the tidyverse [41] and kable (<https://cran.r-project.org/web/packages/kableExtra/index.html>). Sequencing depth in resistance loci was calculated and plotted using *Mosdepth* version 0.2.9 [42]. Confidence intervals were obtained by bootstrapping, comparing 5000 predictions per tool and drug on a resampled dataset.

#### Comparison of output between tools

We collated the output files and information produced by the GenTB online application, the webserver of TB-Profiler (<https://tldr.lsh.ac.uk>, version 3.0.0), and the Desktop version of Mykrobe (MacOS app v0.90) using one example raw sequence dataset (accession ERR1664619). The tools' output was compared based on the following criteria: (1) type and accessibility of output

data formats; (2) communication of genotypic prediction results, i.e., binary classification versus probability; (3) disclosure of the prediction model's error rate; (4) description of known resistance-conferring variants identified; (5) reporting any novel mutation not listed in the resistance variant database; (6) detailed account of detected lineage variants and what lineage typing scheme was used; and (7) report quality metrics on the input sequence data.

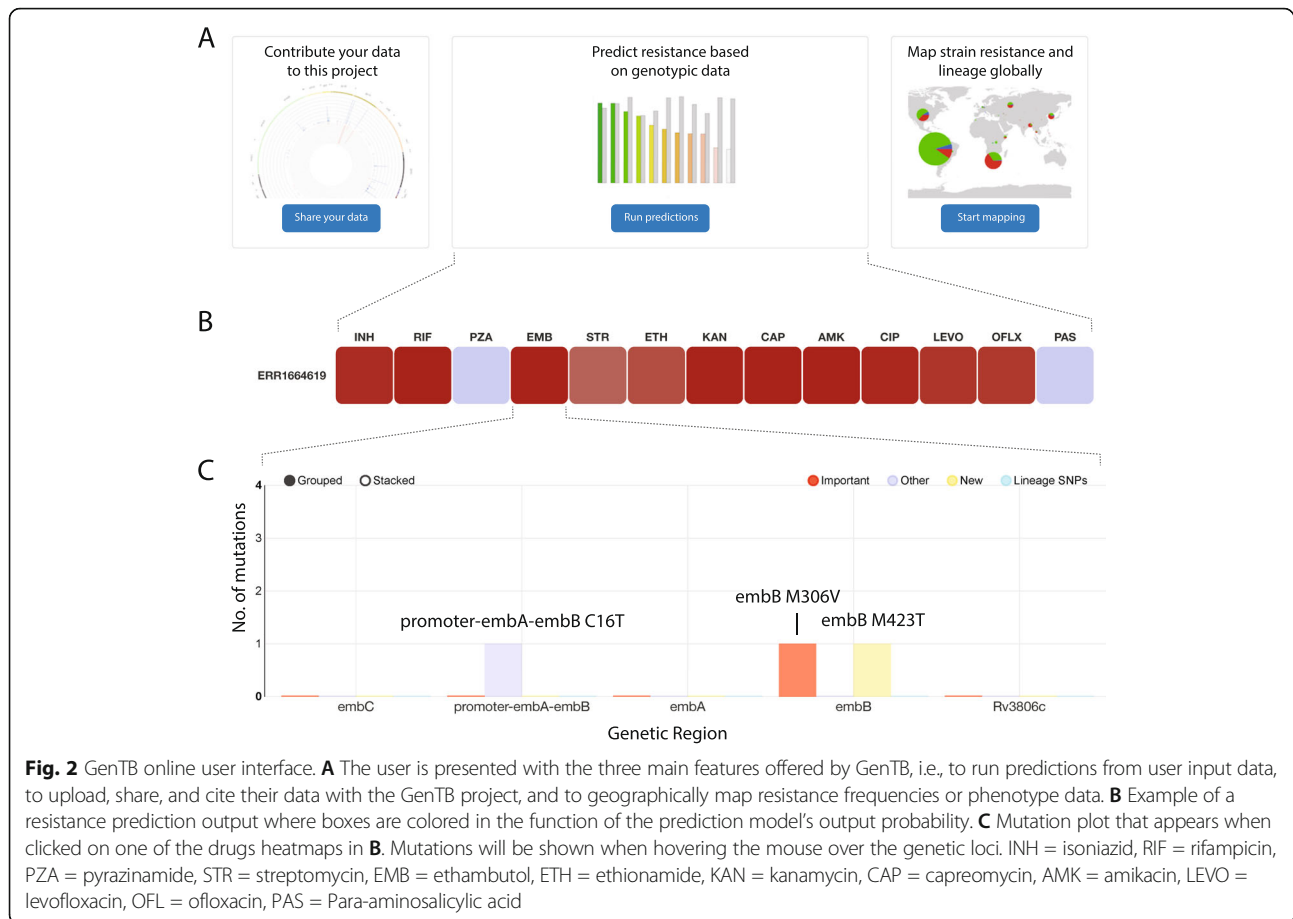
#### Questionnaire

We conducted a survey among past GenTB users in May 2021 to explore how easy the GenTB website is to use (Additional file 3). The survey was developed using Google Forms in English and administered on May 3rd to 166 registered users. The five questionnaire items assessed how (1) pleasing, (2) how clear, (3) how easy to use, (4) how stable, and (5) how usable the GenTB tool is. Responses were recorded on a Likert scale where 0 means "worst" and 10 "best."

#### Results

##### A user-friendly application to analyze *M. tuberculosis* sequencing data

GenTB was developed as a free and benchmarked online application to help public health and clinical practitioners deconvolute the complexity of *M. tuberculosis* WGS data. *GenTB Predict* allows users to predict resistance to 13 anti-TB drugs from a clinical isolate's raw Illumina sequence data (FASTQ) obtained from either WGS or targeted sequencing. Two validated machine learning models are used to make predictions: GenTB-RF and GenTB-WDNN ("Implementation" and [16, 17]). GenTB-RF is the default prediction model. The average computing time based on 23 runs for the prediction pipeline was 35 min (SD 4 min, median 35 min, IQR 33 to 38 min). In addition to the *GenTB Predict* function that we focus on here, the web-application has additional features for sharing, mapping, and exploring *M. tuberculosis* genetic and phenotypic data (Fig. 2). *GenTB Data* enables researchers to store, version, and share *M. tuberculosis* sequence and phenotype data and is powered by the Dataverse research data repository [43]. Users can select an option to delete source files upon processing the prediction. *GenTB Map* enables users to geographically visualize genetic and phenotype data. Users can explore the subset of 20,408 isolates with geographic tags ( $n = 12,547$  isolates) used for GenTB predict validation ("Implementation") or can upload and explore their own data in enriched-VCF format ([https://gitlab.com/doctormo/evcf/-/blob/master/docs/Enriched\\_VCF\\_Format.md](https://gitlab.com/doctormo/evcf/-/blob/master/docs/Enriched_VCF_Format.md)). Raw data and results can be exported to a tabular data format.



A questionnaire-based evaluation of GenTB's user-friendliness among previous users (12 respondents) showed that GenTB is a clear, pleasing, stable, easy to use, and usable application (Additional file 2: Fig. S4).

#### Dataset description

We curated a dataset of 20,408 *M. tuberculosis* isolates with known phenotypic resistance status to benchmark *GenTB Predict* performance ("Implementation" and Additional file 1). We excluded 29 isolates as they failed FastQ validation. Of the remaining, 1339 isolates did not pass our taxonomy filter criterion, and their non-*M. tuberculosis* reads were removed. The GenTB pipeline identified an additional 499 isolates where more than 5% of the genome was covered at depth <10x and these isolates were excluded from further analysis. These isolates had a median depth of 21x (IQR 17 to 26). The remaining 19,880 isolates with high quality sequencing data were majority lineage 4 (52%), with lesser representation of lineage 2 (21%), lineage 3 (15%), lineage 1 (10%), *M. bovis* (0.6%), lineage 6 (0.3%), and lineage 5 (0.2%). Completeness of phenotypic DST data varied by drug and was highest for the first-line drugs rifampicin (98.3%), isoniazid (96.4%), ethambutol (77.5%), and

pyrazinamide (71.5%) (Additional file 2: Table S3). The second and third-line drug phenotype data ranged from 35.1% completeness for streptomycin to 7.8% for ethionamide. Of the 20,408 isolates, 13,817 were phenotypically susceptible to first-line drugs, 4743 (23.3%) were phenotypically MDR (i.e., resistant to isoniazid and rifampicin) and 396 (1.9%) were phenotypically XDR (MDR and resistant to fluoroquinolones and the second-line injectables—amikacin, kanamycin, or capreomycin). We ran GenTB-RF and GenTB-WDNN to predict resistance on 19,880 isolates and compared the predictions to phenotypic data.

#### Predictive performance of the GenTB-Random Forest

We assessed each tools' predictive performance by comparison with phenotypic culture-based DST results. Overall, the four tools had comparable performance characterized by varying sensitivities and high specificities (Tables 1 and 2, Fig. 3A, Additional file 2: Fig. S5). Diagnostic performance was better for first-line than second-line drugs. As sensitivity varied most widely, we discuss it by drug class below. Specificities varied less by tool or by drug. GenTB-RF's diagnostic specificity was > 92% for all drugs including the second-line injectables

**Table 1** Diagnostic accuracy of GenTB RandomForest and GenTB wide and deep neural network compared with two other leading prediction tools on a depth filtered dataset

Drug name	Phenotype		GenTB - RF		GenTB - WDNN		Mykrobe		TB-Profiler	
	R (n)	S (n)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Isoniazid	6,043	13,112	91% (91 to 92)	98% (97 to 98)	90% (89 to 91)	99% (99 to 99)	87% (86 to 88)	98% (98 to 98)	91% (90 to 92)	98% (97 to 98)
Rifampicin	5,068	14,474	93% (93 to 94)	98% (98 to 98)	88% (88 to 89)	99% (99 to 99)	90% (89 to 91)	98% (98 to 99)	92% (91 to 93)	98% (98 to 99)
Ethambutol	2,936	12,362	86% (85 to 87)	92% (92 to 93)	82% (80 to 83)	93% (93 to 94)	79% (77 to 80)	93% (93 to 94)	86% (85 to 88)	92% (92 to 93)
Pyrazinamide	508	1,544	79% (76 to 83)	94% (93 to 95)	80% (79 to 82)	95% (94 to 95)	72% (71 to 74)	98% (97 to 98)	83% (80 to 86)	96% (96 to 97)
Amikacin	618	3,458	67% (63 to 71)	99% (99 to 100)	66% (62 to 70)	99% (99 to 100)	63% (60 to 67)	99% (99 to 100)	55% (51 to 59)	99% (99 to 100)
Capreomycin	648	3,733	63% (59 to 67)	97% (97 to 98)	57% (53 to 61)	98% (98 to 99)	60% (56 to 64)	98% (98 to 99)	56% (52 to 60)	96% (95 to 96)
Ethionamide	502	1,094	67% (63 to 72)	78% (75 to 80)	-	-	-	-	70% (66 to 74)	73% (70 to 76)
Kanamycin	576	3,707	68% (64 to 72)	99% (98 to 99)	66% (62 to 70)	100% (99 to 100)	66% (63 to 70)	99 (99 to 100)	68% (64 to 71)	98% (98 to 99)
Streptomycin	2,126	4,968	82% (80 to 83)	89% (88 to 90)	87% (85 to 88)	87% (86 to 88)	68% (66 to 70)	95% (95 to 96)	71% (70 to 73)	95% (95 to 96)
Ofloxacin	743	4,038	68% (65 to 72)	99% (98 to 99)	62% (58 to 66)	96% (95 to 96)	62% (58 to 65)	99% (98 to 99)	67% (63 to 70)	98 (98 to 99)

and fluoroquinolones with the exception of ethionamide (specificity = 78% [95% CI 75–80]) and streptomycin (specificity = 89% [95% CI 88–90]). GenTB-RF's specificities were similar or higher than the other three tools with the exception of pyrazinamide (94% [95% CI 93–95]) and streptomycin (89% [95% CI = 88–90]) compared to TB-Profiler (96% and 95%, respectively) as well as Mykrobe (98% and 95%, respectively).

#### First-line drugs

Rifampicin resistance prediction by GenTB-RF was most accurate compared to other tools: AUC 0.96 (95% CI = 0.95–0.96), sensitivity 93% (95% CI = 93–94), and specificity 98% (95% CI 98–98), second highest sensitivity was for TB-Profiler at 92% (95% CI = 91–93) (Tables 1 and 2, Fig. 4). The accuracy of isoniazid resistance prediction was high and comparable across three of the four tools including GenTB-RF (sensitivity 91% [95% CI = 91–92], specificity 98% [95% CI 97–98]). For ethambutol,

**Table 2** Area under the receiver operating characteristic curve for GenTB-RF and GenTB-WDNN

Drug	GenTB-RF	GenTB-WDNN
	Area under the ROC curve (95% CI)	
Isoniazid	0.94 (0.94 to 0.95)	0.94 (0.94 to 0.95)
Rifampicin	0.96 (0.95 to 0.96)	0.94 (0.93 to 0.94)
Ethambutol	0.89 (0.88 to 0.9)	0.87 (0.87 to 0.87)
Pyrazinamide	0.90 (0.88 to 0.91)	0.88 (0.87 to 0.88)
Amikacin	0.83 (0.81 to 0.85)	0.83 (0.81 to 0.84)
Capreomycin	0.80 (0.78 to 0.82)	0.78 (0.76 to 0.80)
Ethionamide	0.73 (0.7 to 0.75)	-
Kanamycin	0.83 (0.81 to 0.85)	0.83 (0.81 to 0.85)
Streptomycin	0.85 (0.84 to 0.86)	0.87 (0.86 to 0.88)
Ofloxacin	0.83 (0.82 to 0.85)	0.79 (0.77 to 0.81)

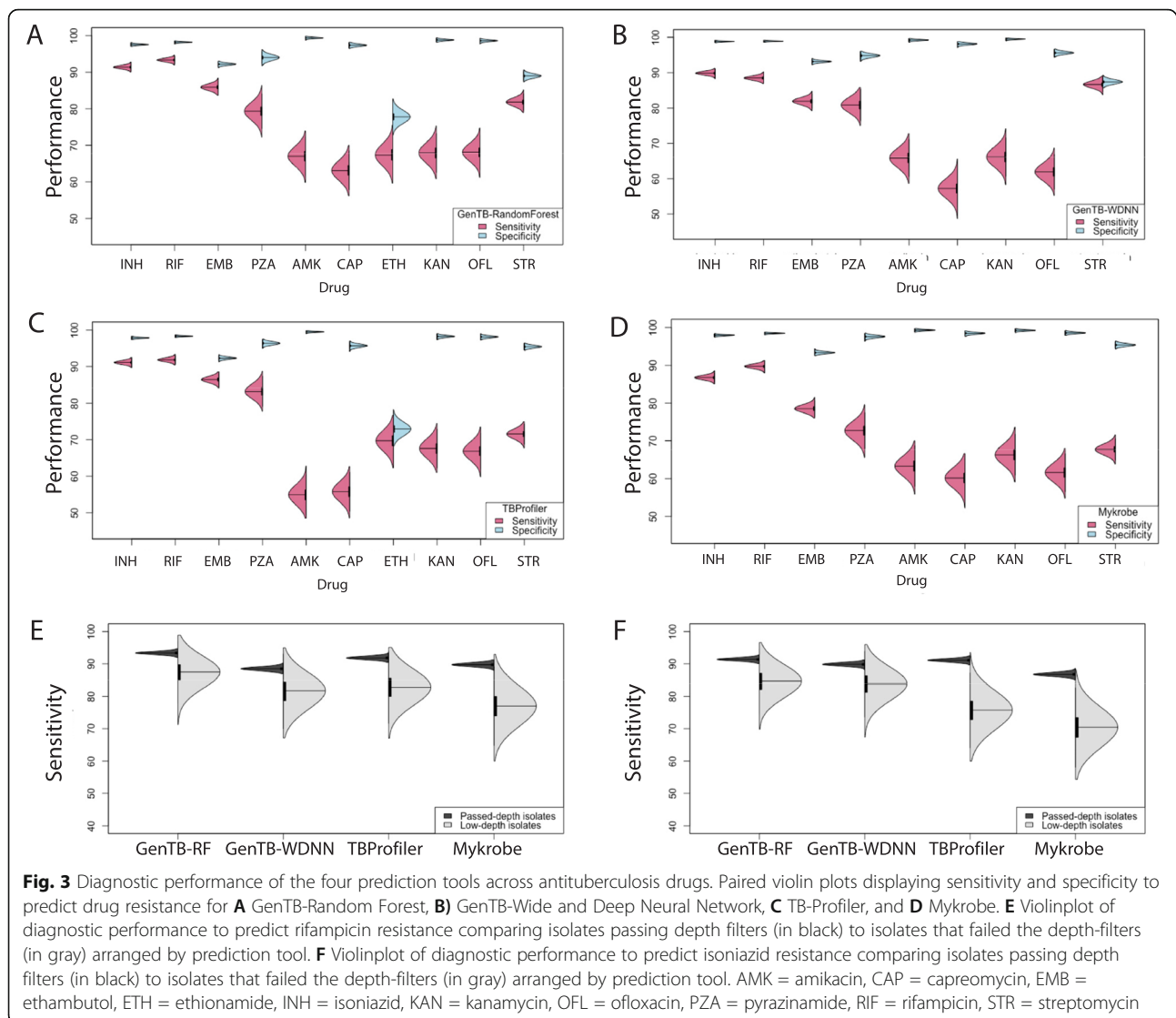
RF = random forest, WDNN = wide and deep neural network

GenTB-RF and TB-Profiler had the best and comparable performance with sensitivity 86% (95% CI = 85–87) and specificity 92% (95% CI 92–93).

GenTB-RF predictions for pyrazinamide using the original model (v1.0) had low sensitivity at 56% (95% CI 54–58) with adequate specificity (98% [95% CI = 98–99]) compared to the other tools when evaluated on the 19,880 isolates (2336 phenotypically resistant and 11,932 susceptible) [16]. Pyrazinamide resistance is known to be caused by a large number of individually rare variants in the gene *pncA* [44]. Given the large interval increase in available WGS data and recent implication of novel resistance loci (*panD*, *clpC1*, *clpP*) [36] since GenTB-RF was last trained, we assessed the number of rare variants in the four aforementioned genes linked to pyrazinamide resistance. In a random 75% subset of the 20,379 isolates, we detected a total of 393 different variants in *pncA*, *panD*, *clpC1*, and *clpP* with 40% (158/393) occurring only once. The majority of these variants, i.e., 73% (285/393) were not previously seen by the original model. As a result of these observations, we retrained a GenTB-RFv2.0, on 75% of the data using all 393 non-synonymous variants including singletons and insertion/deletion variants from *pncA*, *panD*, *clpC1*, and *clpP*. The retrained model, when benchmarked on an independent validation dataset of 5,098 isolates, offered a sensitivity (79% [95% CI 76–83]) and specificity (94% [95% CI 93–95]) similar to the other tools (Table 1).

#### Second-line drugs

For second-line drugs, larger discrepancies between genotype and resistance phenotype have been previously described compared with first-line drugs [14, 15]. Diagnostic sensitivity to the second-line injectable drugs amikacin and kanamycin ranged between 63 and 68% across the four tools, with the exception of a sensitivity of 55% by TB-Profiler for amikacin (Table 1). Specificities to



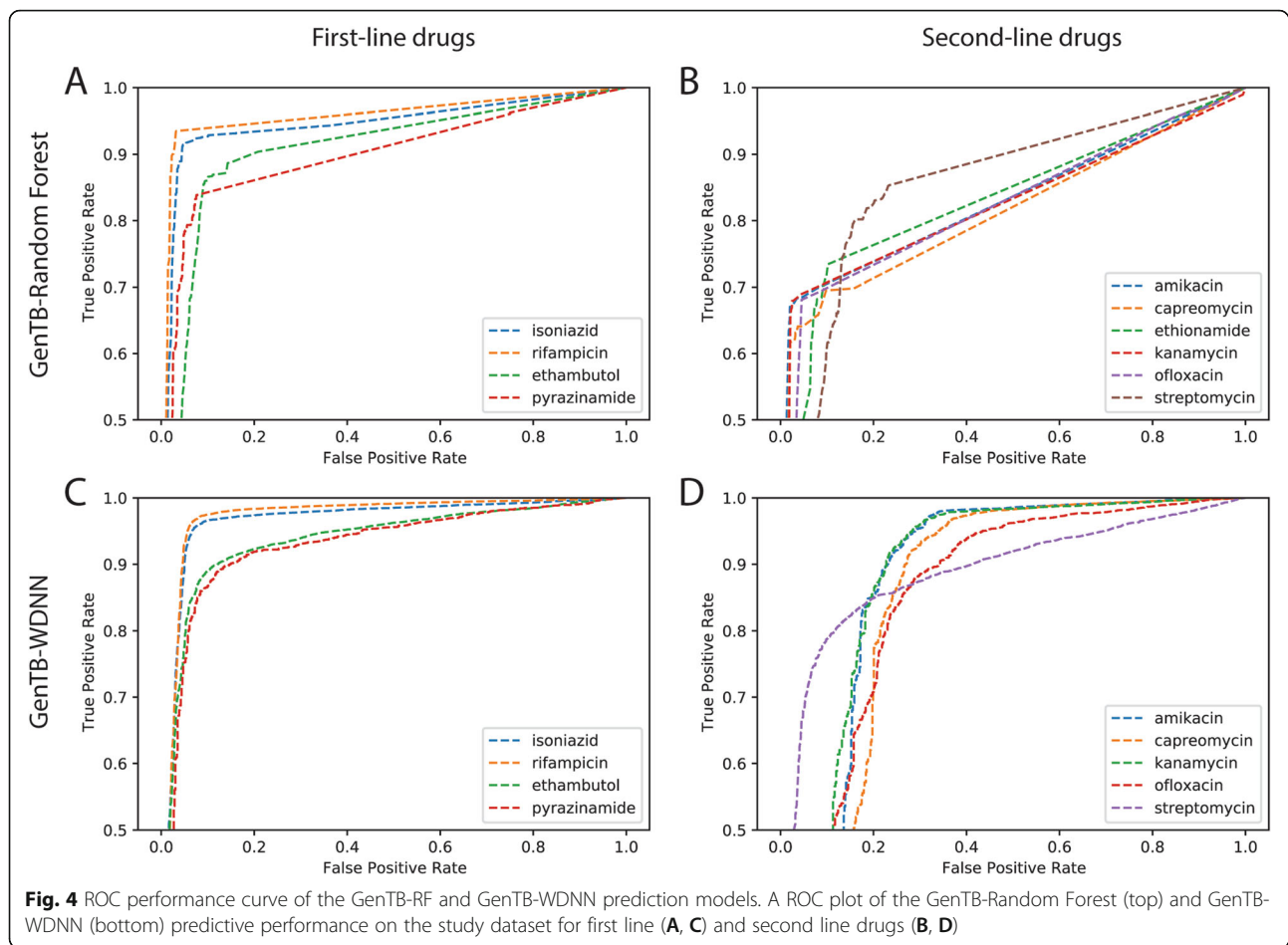
these drugs were  $\geq 98\%$  for all four tools. For the fluoroquinolone ofloxacin, sensitivity ranged from 62–68% and specificity from 96–99% across the four tools. Three drugs had too few isolates with known phenotypic resistance (ciprofloxacin [ $n = 63$ ], levofloxacin [ $n = 111$ ], and para-aminosalicylic acid [ $n = 46$ ]), and hence the tool's predictions had wide confidence intervals for these drugs (Additional file 2: Table S4 and Additional file 2: Table S5). For levofloxacin, GenTB's diagnostic sensitivity was 81% (95% CI 73–88) with a specificity of 77% (95% CI 66–87) (Additional file 2: Table S4).

#### Predictive performance of GenTB-WDNN

We sought to determine the performance of GenTB-WDNN that was previously shown to outperform other statistical resistance prediction approaches [17]. To

determine the probability of phenotypic resistance, GenTB-WDNN combines multitask logistic regression to learn the effect of individual mutations with a three-layer perceptron to account for more complex epistatic effects on antibiotic resistance [45, 46]. Similar to GenTB-RF, the overall GenTB-WDNN performance was marked by high prediction accuracy of first-line drug resistance and lower accuracy of second-line resistance (Table 1). AUC 95% CI overlapped for all drugs between the two models except for ofloxacin and rifampicin for which the GenTB-RF AUC was higher (Table 2). For streptomycin, the GenTB-WDNN offered the best sensitivity and specificity of all four models (sensitivity 87%, 95% CI 85–88%, specificity 87% (95%CI 86–88%). Specificities were  $> 95\%$  for all drugs except for streptomycin (87%, 95% CI 85 to 88) and ethambutol (93%, 95% CI 93 to 94).





### Predictive performance depends on sequencing depth

We evaluated the need for quality control on sequencing depth as several tools do not currently implement this prior to resistance prediction [9, 14, 15]. We observed predictive performance to be highly dependent on sequencing depth as indicated by lower sensitivity to predict rifampicin or isoniazid resistance by all four tools for the 499 isolates that did not meet the threshold of  $\geq 10\times$  depth across  $> 95\%$  of the genome (median depth of  $21\times$ , IQR 17 to 26, Fig. 3E, F). Using GenTB-RF, the mean sensitivity of isoniazid and rifampicin prediction was 84.6% (SD 3.6) and 87.3% (SD 3.6) respectively among low-depth isolates, compared with 91% and 93%, respectively, on high-depth isolates (Additional file 2: Table S6, Fig. 3E, F). Loss of sensitivity due to low sequencing depth was comparable across the four tools.

### Discordant resistance predictions

To gain insight into model performance, we probed discrepancies between GenTB-RF's genotype-based prediction and the resistance phenotype. We focused on this model as it had the highest overall sensitivity. We examined specifically rifampicin and isoniazid as resistance to

these two drugs defines MDR-TB, and their genetic resistance mechanisms are well understood. We investigated isolates for which GenTB-RF predicted resistance while the phenotype was reported as susceptible (false positives) and isolates for which GenTB-RF predicted susceptibility with a resistant phenotype (false negatives). We confirmed that false negative predictions were not due to low sequencing depth in relevant drug resistance loci (i.e., that depth was  $\geq 10\times$  across all bases, Additional file 2: Fig. S6 and Additional file 2: Fig. S7).

### Rifampicin false positives

Most of the variants linked to rifampicin resistance are concentrated in a 81-bp window in the *rpoB* gene *a.k.a* the rifampicin resistance determining region (RRDR, H37Rv coordinates 761081 to 761162, accession AL123456) [47]. For rifampicin, we observed 254 false positive predictions (phenotypically susceptible isolates predicted resistant). GenTB-RF detected one or more non-silent RRDR variants in 198 of these 254 isolates (78%). The most common RRDR variants were S450L (occurred in 49/254 isolates), L430P (in 33/254), and H445N (in 31/254) (Additional file 2: Table S7). The

remaining 56 of 254 isolates, harbored non-RRDR variants, the two most common were *rpoB* I491F (occurred in 29/56) and *rpoB* V695L (occurred in 24/56). Twenty-eight of the 56 isolates (50%) were phenotypically resistant to isoniazid and a further 16 (29%) were resistant to ethambutol.

#### **Rifampicin false negatives**

Among the 333 false negative rifampicin predictions (phenotypically resistant isolates predicted susceptible), 96 (29%) isolates harbored a variant in *rpoB* and of these 75 (23% of the 333) were in the RRDR (Additional file 2: Table S7). These included most commonly three base pair insertion in *rpoB* codon 433 (occurred in 14/333 isolates) and *rpoB* codon 443 (occurred in 9/333 isolates) and *rpoB* substitution Q432L (in 9/333) [48]. These *rpoB* variants were not previously seen by the GenTB-RF model when initially trained. For the remaining 237 of 333 isolates (71%), phenotypic resistance remained unexplained.

#### **Isoniazid false positives**

For isoniazid, we observed 315 false positive predictions (phenotypically susceptible isolates predicted resistant by GenTB-RF). Among these isolates, 119/315 (38%) had a total of 40 unique non-silent non-lineage variants in genes linked to isoniazid resistance (*inhA*, *katG*, *ahpC*, *fabG1*) (Additional file 2: Table S8). Most variants, 36/40, were rare, occurring in only 2 or fewer isolates. Five out of the 40 unique mutations detected in 75/315 (24%) isolates are considered important for isoniazid resistance prediction by GenTB-RF [16]. The most frequent INH resistance variants were the canonical isoniazid resistance mutation *katG* S315T [49] (occurred in 56/315 isolates) and non-silent variants at *inhA* codon 94 (occurred in 14/315 isolates). Seventy-six of the 315 (24%) apparent false positive isolates were phenotypically resistant to rifampicin and 189 (60%) isolates had a phenotypic resistance to at least one other drug.

#### **Isoniazid false negatives**

Among the 518 false negative isoniazid predictions (phenotypically resistant isolates predicted susceptible by GenTB-RF), 194/518 (37%) harbored non-silent variants in isoniazid resistance-associated genes (Additional file 2: Table S8). Only 13 of the 139 unique variants observed in the 518 isolates were seen before by GenTB-RF and none of these were considered important isoniazid resistance mutations. *KatG* W328L was the variant detected most frequently (occurred in 10/518 isolates predicted false negative) and although not previously seen by GenTB-RF was described to occur in 0.2% of isoniazid resistance in one study [50]. Most variants linked to isoniazid resistance

observed in these isolates were rare, i.e., 134/139 (96%) occurred in  $\leq 3$  isolates.

#### **Output comparison across the three tools**

All four tools are accessible to the non-experienced user via either an online interface (GenTB, TB-Profiler) or via a Desktop application. We compared each tool's output using the criteria specified in the "Implementation" section (Table 3). GenTB-RF provides a heatmap indicating the probability of resistance including the models' error rate with all prediction and intermediary files available for download. TB-Profiler and Mykrobe present binary (resistant or susceptible) predictions in overview tables with download options in CSV or JSON formats, respectively. TB-Profiler and GenTB present resistance causing variants and variants not associated with resistance. All tools provide the main- and sub-lineage call made but GenTB also specifies the lineage typing schemes used.

#### **Discussion**

The increasing affordability of WGS and our improving comprehension of mycobacterial drug resistance mechanisms has placed sequencing at the forefront of *M. tuberculosis* resistance diagnosis in clinical and public health laboratories (e.g., Public Health England in the UK and the Centers for Disease Control and Prevention in the USA) [7, 51]. Yet, the complexity of resistance biology is such that large and diverse bacterial isolate datasets are needed to confirm the accuracy of genotype-based resistance prediction and its generalizability. Further, the required computational resources and knowledge to conduct sequencing analysis prohibit both the access to and confidence in WGS based resistance prediction in clinics in both low- and high-incidence settings. High confidence automated tools that are systematically benchmarked on diverse datasets are needed to facilitate adoption, and to act as the standard for future tool development and regulation by oversight agencies such as the World Health Organization (WHO).

GenTB is an automated open tool for resistance prediction from WGS. Here we benchmarked its two prediction models against two other leading TB prediction tools. Both GenTB models predicted resistance and susceptibility against first-line drugs with high accuracy. Predictive performance for second line drugs showed lower sensitivity, and this may be helped by studying a larger number of isolates with ethionamide, amikacin, capreomycin, kanamycin, and fluoroquinolone resistance in the future. Specificity was high for several second line drugs, i.e., capreomycin, kanamycin, and ofloxacin. This high specificity may be used to rule out resistance when no resistance-conferring variant for these drugs was

**Table 3** Output comparison across tools

Criteria	GenTB	TB-Profiler	Mykrobe
1) Output			
Type	Heatmap and barplot	Overview tables	Overview table
Download	All intermediate and output files (JSON)	Yes (CSV)	Yes (JSON)
2) Genotypic predictions			
	Probability	Binary	Binary
3) Error rate			
	Yes	N.A.	N.A.
4) Resistance variants			
	Variant by drug	Variant by drug incl. fraction of mutant/wild-type allele	Variant by drug incl. depth of mutant and wild-type alleles
5) Unknown variants			
	Yes, in all genes	Yes, in candidate resistance genes	No
6) <i>M. tuberculosis</i> Lineage			
Main lineage	Yes	Yes	Yes
Sublineage	Yes	Yes	Yes
Typing scheme	Yes	No	No
7) Quality metrics			
	Trimming and contamination report	No. of reads, Percentage of reads mapped	No

found. A detailed analysis of discrepant predictions made by GenTB-RF illustrated that a number of false positive predictions were supported by canonical resistance variants, e.g., non-silent mutation in the *rpoB* RRDR in case of rifampicin, suggesting that their phenotypes were erroneously labeled as susceptible. Similarly, nearly half (48%) of the variants found in isoniazid false positive predictions are canonical resistance variants. These isoniazid resistance variants, the large proportion (60%) of phenotypic resistance to another drug among these isolates, and the knowledge that isoniazid is usually a gateway drug resistance, suggest that some phenotypes were erroneously characterized as susceptible [52]. Accordingly, specificity of genotype-based prediction in practice maybe even higher than reported here (Table 1). Given the estimated 2% prevalence of rifampicin resistance among new TB cases in the USA in 2019 [1], GenTB-RF's diagnostic accuracy translates to a positive predictive value of 49.5% and a negative predictive value of 99.9%.

For isolates with a resistant rifampicin phenotype that were predicted susceptible by GenTB-RF, we found a mutation in the *rpoB* RRDR in a nearly a quarter (23%) of isolates that reasonably accounts for the resistance phenotype, but had not been seen by the model previously. For the remaining majority of false negatives (71% for rifampicin) no relevant resistance variant was found. In these cases, phenotypic resistance remained unexplained and could be due to erroneous phenotypes or yet unknown resistance mechanisms. For isolates with a resistant isoniazid phenotype predicted susceptible, no important resistance-conferring mutations were found.

In these cases, phenotypic resistance could be due to rare and yet undescribed resistance variants. A substantial proportion of false negative predictions to isoniazid or rifampicin had genotypic resistance to at least another drug (48% of rifampicin false negatives and 40% of isoniazid false negatives). These observations overall suggest that a viable option to reduce false negative predictions by current models would be to leverage genotypic predictions to other drugs and flag such isolates for complementary phenotypic DST. In the future as new larger datasets of paired genotype and resistance phenotype are curated, e.g. by efforts sponsored by the WHO [29], retraining existing resistance prediction models will improve diagnostic sensitivity.

The final output produced by the four tools varies in terms of detail and type of variants reported with GenTB providing the most detail. In addition to resistance-associated variants, GenTB outputs a description of novel variants in resistance genes that have not been previously seen by GenTB's models. The phylogenetic lineage calling procedure implemented in GenTB [25] uses several validated typing schemes to facilitate comparisons across lineage schemes.

Unlike other published resistance prediction tools that rely on a curated list of resistance-conferring mutations that call resistance when a specific variant is present, GenTB-RF and GenTB-WDNN use multi-variable statistical models to predict resistance phenotype. These models are better suited to account for the complex relationships between resistance genotype and phenotype. Among the advantages of multivariate prediction models is that relationships between

variables are taken into account as both individual variants and gene-gene interactions cause phenotypic drug resistance [45, 46]. As such, the two models provide a probability value that a given isolate is resistant or susceptible rather than a binary classification. This is relevant in case of variants that, if present alone, confer only weak to no resistance, but may confer complete resistance if present in combination. Also, each variable in a multivariable model has different weights depending on the strength of association with resistance in the training data, reflecting the biological reality where variants cause differing levels of resistance. The benchmarking data presented here confirm that these multivariate models offer gains in sensitivity over the other two tools that use curated mutation lists; however, this comes at a small decrease in specificity overall. Given its higher overall performance GenTB-RF is currently implemented as the default prediction model. As larger and more diverse data will become available for model training, especially for prediction of resistance more quantitatively, i.e., to predict minimum inhibitory concentrations or MICs, we anticipate multivariate models including the more complex GenTB-WDNN architecture to have an even bigger advantage over direct association of mutation lists.

This study was not without limitations. An important prerequisite for reliable genotypic resistance prediction is the quality of the raw sequencing data. Variants and small indels in resistance-conferring genes can be accurately and confidently called from Illumina raw sequence data if the genes are adequately covered at an acceptable sequencing depth [53]. However, short-read sequencing data is recognized to have lower sensitivity for detecting more complex genomic variants including long indels or structural variation and these may have been missed in this study. But these latter types of variants are expected to be rare. Our finding of “apparent” false positive predictions (i.e., resistance call by GenTB-RF while susceptible phenotype) in isolates harboring canonical resistance variants portends some erroneous phenotypes in our ground truth dataset. None of the three tools studied predict resistance to recently introduced or repurposed drugs, such as bedaquiline or linezolid, due to limited phenotypic resistance data which is partially driven by limited clinical experience with them thus far [54]. Due to the scale and public nature of the dataset used for benchmarking in this study, we were unable to retest the laboratory-based drug susceptibility profiles of isolates with discordant predictions, but hope that it provides a test closer to a “real-world” scenario for these tool’s application. We also note that the models’ performance as benchmarked here is an average across the currently publicly available datasets that represent

isolates from different countries. In the future, large country-specific datasets will further validate their diagnostic accuracy based on the local epidemiology.

## Conclusions

The rapid emergence and affordability of sequencing of *M. tuberculosis* along with the herein confirmed high accuracy of several genotypic resistance prediction tools supports the use of informatically assisted treatment design in the clinical setting. Independent benchmarking efforts will facilitate regulatory reviews and assessments and build confidence in the tools’ performances. As genotypic resistance predictions will accompany and increasingly replace laboratory-based resistance phenotyping performance criteria will need to be defined to guide clinical and public health laboratories in their use. Lastly, it will be important to communicate the confidence and uncertainty that is inherent to all genotypic predictions to clinicians and provide clear diagnostic algorithms in case of genotype-phenotype discordances.

## Availability and requirements

Project name: GenTB - Translational Genomics of Tuberculosis

Project home page: <https://gentb.hms.harvard.edu>

Operating system(s): Platform independent

Programming language: Python 2 and 3, R, Perl

License: AGPLv3

Any restrictions to use by non-academics: None.

## Abbreviations

DST: Drug susceptibility testing; Indels: Insertions and deletions; MDR-TB: Multi-drug-resistant TB; SNP: Single nucleotide polymorphism; TB: Tuberculosis; WGS: Whole genome sequencing; XDR-TB: Extensively drug-resistant TB

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-021-00953-4>.

**Additional file 1.** Accessions and phenotypic drug susceptibility profiles of isolates included in the benchmarking dataset.

**Additional file 2: Fig S1.** Characteristics of variables used for resistance prediction by Gentb-RF for isoniazid and rifampicin. **Fig S2.** Probability distribution and selected threshold for Gentb-RF resistance predictions.

**Fig S3.** Probability distribution and thresholds according to [17] of Gentb-WDNN resistance predictions. **Fig S4.** User-friendliness evaluation of the GenTB tool. **Fig S5.** Diagnostic performance of the four prediction tools across antituberculosis drugs. **Fig S6.** Sequencing depth of resistance-conferring genes in isolates falsely predicted susceptible to first line agents. **Fig S7.** Sequencing depth of resistance-conferring genes in isolates falsely predicted susceptible to second line agents. **Table S1.** Genetic loci used for random forest model training. **Table S2.** Phenotypic drug susceptibility testing methods used by studies included in this benchmarking dataset. **Table S3.** Frequencies and percentages of available drug susceptibility data per drug. **Table S4.** Diagnostic accuracy comparison of tools for drugs with insufficient phenotype data and pyrazinamide performance on all isolates. **Table S5.** Area under the Receiver Operating Characteristic curve for GenTB-RF and GenTB-WDNN. **Table S6.** Diagnostic accuracy to rifampicin and isoniazid across low-depth and

passed-depth isolates. **Table S7.** Non-silent variants in the gene *ropB* among isolates with discordant phenotype and genotype predictions for the drug rifampicin. **Table S8.** Non-silent variants in the genes *inhA*, *katG*, *ahpC*, or *fabG1* among isolates with discordant phenotype and genotype predictions for the drug isoniazid.

**Additional file 3.** Questionnaire to evaluate the user-friendliness of the GenTB tool.

### Acknowledgements

We would like to thank Megan Murray and the Institute for Quantitative Social Sciences at Harvard, in particular Christine Choirat, Raman Prasad, James Honaker, Merce Crosas, and Gary King for their invaluable support when GenTB was launching. We thank members of the Research Computing group at Harvard Medical School for continued support of the GenTB webservice. We thank the anonymous reviewers for helpful feedback that allowed us to improve this manuscript.

### Authors' contributions

MIG and MRF conceived and designed the study. MIG, MO, RV, LF, AD, and MRF wrote scripts for the website and sequence data analysis. JP and ZI provided help and advice throughout the project. MIG and MRF wrote the first version of the manuscript and the final manuscript contained contributions from all authors. All authors read and approved the final manuscript.

### Funding

This work was funded by National Institutes of Health (K01 ES026835 and R01 AI55765 to MRF). MIG is supported by the German Research Foundation (GR5643/1-1). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

### Availability of data and materials

All raw *M. tuberculosis* sequence data used in this study to benchmark the performance of the genotypic resistance prediction tools is available online from public repositories such as the National Center for Biotechnology Information or the European Nucleotide Archive. Accessions are given in Additional file 1. GenTB can be accessed online at <https://gentb.hms.harvard.edu>. The source code for the GenTB website is available in the repository <https://github.com/farhat-lab/gentb-site> [19] and the source code for the *snakemake* implementation of GenTB is available in the repository <https://github.com/farhat-lab/gentb-snakemake> [27].

### Declarations

#### Competing interest

The authors declare that they have no competing interests.

#### Ethics approval and consent to participate

Not applicable

#### Consent for publication

Not applicable

#### Author details

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>2</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, USA. <sup>3</sup>Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. <sup>4</sup>European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK. <sup>5</sup>Division of Infectious Diseases, Boston Children's Hospital, Boston, MA, USA. <sup>6</sup>Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA.

Received: 29 March 2021 Accepted: 12 August 2021

Published online: 30 August 2021

### References

- World Health Organization. Global Tuberculosis Report 2020. World Health Organization; 2020. Available from: <https://www.who.int/publications/i/item/9789240013131>

- World Health Organization. Guidelines for surveillance of drug resistance in tuberculosis 5th Edition. WHO; 2015. Available from: <https://apps.who.int/iris/bitstream/handle/10665/174897?jsessionid=52537DA4A0B0E19A10382076AC23874?sequence=1>
- Cabibbe AM, Trovato A, De Filippo MR, Ghodousi A, Rindi L, Garzelli C, et al. Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. *Eur Respir J*. 2018;51. Available from: <https://doi.org/10.1183/13993003.00387-2018>
- Pankhurst LJ, del Ojo EC, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med*. Elsevier BV. 2016;4(1):49–58. [https://doi.org/10.1016/S2213-2600\(15\)00466-X](https://doi.org/10.1016/S2213-2600(15)00466-X).
- Cirillo DM, Miotto P, Tortoli E. Evolution of Phenotypic and Molecular Drug Susceptibility Testing. *Adv Exp Med Biol*. 2017;1019:221–46. [https://doi.org/10.1007/978-3-319-64371-7\\_12](https://doi.org/10.1007/978-3-319-64371-7_12).
- Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nat Rev Microbiol*. 2019; Available from: <https://doi.org/10.1038/s41579-019-0214-5>
- CRyPTIC Consortium and the 100,000 Genomes Project, Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med*. 2018;379:1403–15.
- McNerney R, Zignol M, Clark TG. Use of whole genome sequencing in surveillance of drug resistant tuberculosis. *Expert Rev Anti Infect Ther*. 2018; 16(5):433–42. <https://doi.org/10.1080/14787210.2018.1472577>.
- Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. *PeerJ*. 2018;6:e5895. <https://doi.org/10.7717/peerj.5895>.
- Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*. 2014;15(1):881. <https://doi.org/10.1186/1471-2164-15-881>.
- 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. <https://doi.org/10.1128/JCM.00025-15>.
- Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T. 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*. 2015;95:843–844.
- Sezikuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total Genotyping Solution for Mycobacterium tuberculosis Using Short-Read Whole-Genome Sequencing. *PLoS One*. 2015;10(11):e0142951. <https://doi.org/10.1371/journal.pone.0142951>.
- 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. <https://doi.org/10.1186/s13073-019-0650-x>.
- Hunt M, Bradley P, Lapiere SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res*. 2019;4:191. <https://doi.org/10.12688/wellcomeopenres.15603.1>.
- Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *Am J Respir Crit Care Med*. 2016;194(5):621–30. <https://doi.org/10.1164/rccm.201510-2091OC>.
- 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. <https://doi.org/10.1016/j.ebiom.2019.04.016>.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2. <https://doi.org/10.1093/bioinformatics/bts480>.
- Martin Owens, Raman Prasad, Maha R Farhat, Corinne Bintz, Davey Hughes, Jimmy Royer, Patrick Hanaj, Christine Choirat, Vladislav Doster, Mohib Javri. gentb-site. 2021. Available from: <https://github.com/farhat-lab/gentb-site>
- Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90. <https://doi.org/10.1093/bioinformatics/bty560>.

21. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46. <https://doi.org/10.1186/gb-2014-15-3-r46>.
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100. <https://doi.org/10.1093/bioinformatics/bty191>.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
24. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963. <https://doi.org/10.1371/journal.pone.0112963>.
25. Freschi L, Vargas R Jr, Hussain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. *bioRxiv.* bioRxiv; 2020. Available from: <https://doi.org/10.1101/2020.09.29.293274>
26. Grünig B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods.* 2018;15:475–6.
27. Matthias Gröschel and Martin Owens. *gentb-snake*. 2021. Available from: <https://github.com/farhat-lab/gentb-snake>
28. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* 2017;45:D535–42.
29. Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, et al. Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci Rep.* 2018;8:15382.
30. Zignol M, Cabibbe AM, Dean AS, Glaziou P, Alikhanova N, Ama C, et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect Dis.* 2018;18(6):675–83. [https://doi.org/10.1016/S1473-3099\(18\)30073-2](https://doi.org/10.1016/S1473-3099(18)30073-2).
31. Wollenberg KR, Desjardins CA, Zalutskaya A, Slodovnikova V, Oler AJ, Quiñones M, et al. Whole-genome sequencing of *Mycobacterium tuberculosis* provides insight into the evolution and genetic composition of drug-resistant tuberculosis in Belarus. *J Clin Microbiol.* 2017;55(2):457–69. <https://doi.org/10.1128/JCM.02116-16>.
32. Phelan JE, Lim DR, Mitarai S, de Sessions PF, Tujan MAA, Reyes LT, et al. *Mycobacterium tuberculosis* whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci Rep.* 2019;9:9305.
33. Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, et al. Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol.* 2018;3(9):1032–42. <https://doi.org/10.1038/s41564-018-0218-3>.
34. 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. <https://doi.org/10.1038/s41588-017-0029-0>.
35. Dheda K, Limberis JD, Pietersen E, Phelan J, Esmail A, Lesosky M, et al. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet Respir Med.* 2017;5(4):269–81. [https://doi.org/10.1016/S2213-2600\(16\)30433-7](https://doi.org/10.1016/S2213-2600(16)30433-7).
36. Gopal P, Sarathy JP, Yee M, Ragunathan P, Shin J, Bhushan S, et al. Pyrazinamide triggers degradation of its target aspartate decarboxylase. *Nat Commun.* 2020;11:1661.
37. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. *SciPy*; 2010. Available from: <https://doi.org/10.25080/majora-92bf1922-00a>
38. Waskom M, Botvinnik O, O’Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. *Mwaskom/Seaborn: V0.8.1* (September 2017). Zenodo; 2017. Available from: <https://doi.org/10.5281/zenodo.883859>
39. Adler D, Kelly ST. *vioplot: violin plot.* 2020. Available from: <https://github.com/TomKellyGenetics/vioplot>
40. Team RC, Others. R: A language and environment for statistical computing. Vienna, Austria; 2013. Available from: <http://cran.univ-paris1.fr/web/packages/dplR/vignettes/intro-dplR.pdf>
41. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. *JOSS.* 2019;4(43):1686. <https://doi.org/10.21105/joss.01686>.
42. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics.* 2018;34:867–8.
43. King G. An introduction to the dataverse network as an infrastructure for data sharing. *Social Methods Res.* 2007;36(2):173–99. <https://doi.org/10.1177/0049124107306660>.
44. Yadon AN, Maharaj K, Adamson JH, Lai Y-P, Sacchetti JC, Iøerger TR, et al. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat Commun.* 2017;8:588.
45. Farhat MR, Freschi L, Calderon R, Iøerger T, Snyder M, Meehan CJ, et al. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat Commun.* 2019;10(1):2128. <https://doi.org/10.1038/s41467-019-10110-6>.
46. Vargas R Jr, Freschi L, Spitaleri A, Tahseen S, Barilar I, Niemann S, et al. The role of epistasis in amikacin, kanamycin, bedaquiline, and clofazimine resistance in *Mycobacterium tuberculosis* complex. *bioRxiv.* bioRxiv; 2021 [cited 2021 May 14]. p. 2021.05.07.443178. Available from: <https://www.biorxiv.org/content/10.1101/2021.05.07.443178v1>
47. Donnabella V, Martiniuk F, Kinney D, Bacerdo M, Bonk S, Hanna B, et al. Isolation of the gene for the beta subunit of RNA polymerase from rifampicin-resistant *Mycobacterium tuberculosis* and identification of new mutations. *Am J Respir Cell Mol Biol.* 1994;11(6):639–43. <https://doi.org/10.1165/ajrcmb.1116.7946393>.
48. Miotto P, Cabibbe AM, Borroni E, Degano M, Cirillo DM. Role of disputed mutations in the *rpoB* gene in interpretation of automated liquid MGIT culture results for rifampin susceptibility testing of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 2018;56. Available from: <http://jcm.asm.org/cgi/pmidlookup?view=long&pmid=29540456>
49. Heym B, Alzari PM, Honoré N, Cole ST. Missense mutations in the catalase-peroxidase gene, *katG*, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Mol Microbiol.* Wiley. 1995;15(2):235–45. <https://doi.org/10.1111/j.1365-2958.1995.tb02238.x>.
50. Seifert M, Catanzaro D, Catanzaro A, Rodwell TC. Genetic mutations associated with isoniazid resistance in *Mycobacterium tuberculosis*: a systematic review. *PLoS One.* 2015;10:e0119628.
51. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J.* 2017;50. Available from: <https://doi.org/10.1183/13993003.01354-2017>
52. Ektefaie Y, Dixit A, Freschi L, Farhat MR. Globally diverse *Mycobacterium tuberculosis* resistance acquisition: a retrospective geographical and temporal analysis of whole genome sequences. *Lancet Microbe.* Elsevier BV. 2021;2(3):e96–104. [https://doi.org/10.1016/S2666-5247\(20\)30195-6](https://doi.org/10.1016/S2666-5247(20)30195-6).
53. Marin M, Vargas R Jr, Harris M, Jeffrey B, Epperson LE, Durbin D, et al. Genomic sequence characteristics and the empiric accuracy of short-read sequencing. *bioRxiv.* bioRxiv; 2021. Available from: <https://doi.org/10.1101/2021.04.08.438862>
54. Kadura S, King N, Nakhoul M, Zhu H, Theron G, Köser CU, et al. Systematic review of mutations associated with resistance to the new and repurposed *Mycobacterium tuberculosis* drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. *J Antimicrob Chemother.* 2020;75:2031–43.

## Publisher’s Note

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

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

