Rapidly and reproducibly building a comprehensive catalogue of resistance-associated variants for *M. tuberculosis*

Dylan Adlard[1], Kerri M Malone[2], Jeremy Westhead[1], Martin Hunt[1,2], Hieu Thai[1], Matthew Colpus[1], Robert D Turner[1], Shaheed V Omar[3], David W Eyre[1,4,5], Nazir Ismail[6], Timothy M Walker[1], Timothy EA Peto[1,4,5], Derrick W Crook[1,4,5], Zamin Iqbal[2,7], and Philip W Fowler[*1,4,5]

[1]Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Headley Way, Oxford, U.K.

[2]European Bioinformatics Institute, Cambridge, UK

[3]Centre for Tuberculosis, National Institute for Communicable Diseases a Division of the National Health Laboratory Service, Johannesburg, South Africa

[4]Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, U.K.

[5]National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, UK

[6]Department of Clinical Microbiology and Infectious Diseases, University of the Witwatersrand, Johannesburg, South Africa

[7]Milner Centre for Evolution, University of Bath, UK

[*]To whom correspondence should be addressed: philip.fowler@ndm.ox.ac.uk, @philipwfowler.bsky.social

Abstract

Background

Catalogues of genetic variants associated with resistance underpin whole-genome sequencing (WGS)-based predictions of drug susceptibility in *Mycobacterium tuberculosis*, and are essential for molecular diagnostics and surveillance. The current gold standard catalogues are those released by the WHO but the underlying data are not fully released and they are difficult to interpret. Open and reproducible methods would help address these problems, extending the important work already done.

Methods

We have developed an automated method, `catomatic`, that uses a binomial test to associate informative isolates with resistance or susceptibility, and built a catalogue (`catomatic-1`) from the same 39,358 samples used to construct the first edition of the WHO catalogue (`WHOv1`). We performed a sensitivity analysis to optimise statistical and bioinformatic parameters for each drug, and benchmarked `catomatic-1` against `WHOv1` using an independent Validation Dataset of 14,380 isolates.

Findings

By using simpler statistics, `catomatic-1` algorithmically classified 1,329 genetic variants, ranging from five for linezolid to 440 for pyrazinamide. `WHOv1` included generalisable rules added by a panel of experts, increasing its predictive coverage, but at the cost of reproducibility. Despite not including such expert rules, `catomatic-1` achieves comparable performance for all drugs, with sensitivities for first-line agents above 88% on the independent Validation Dataset. The automated process allowed us to efficiently explore parameter space; for instance, detecting resistant variants with low read support improved the sensitivity for all drugs.

Interpretation

Performant resistance catalogues for *M. tuberculosis* can be built automatically using transparent and reproducible statistical methods. As more data are collected, catalogue content and performance will evolve, highlighting the need for proper versioning, machine/human readability, and open access. This approach demonstrates resistance catalogues used in surveillance and diagnostics can be rapidly and reproducibily updated.

Funding

The National Institute for Health and Care Research (NIHR), Engineering and Physics Sciences Research Council (EPSRC) and ORACLE Corporation.

Research in context

Evidence before this study

We searched PubMed and preprint servers (bioRxiv, medRxiv), and publicly available mutation catalogues for studies linking *Mycobacterium tuberculosis* genomic variants with drug resistance using whole-genome or targeted sequencing and phenotypic drug-susceptibility testing (pDST). Search terms combined "Mycobacterium tuberculosis", "genome sequencing", "mutation catalogue", "mutation effects", "drug resistance", and individual drug names, with no language or date restriction. We included studies providing paired, clinical genomic and pDST or MIC data, excluding purely in-silico or case-only reports. This work directly builds on methodologies and data published by five prior studies, and makes primary comparisons with the First (`WHOv1`) and Second (`WHOv2`) Editions of the WHO Catalogue of mutations in *Mycobacterium tuberculosis*.

Added value of this study

We developed `catomatic`, a transparent, reproducible tool for building catalogues of resistance- and susceptibility-associated genetic variants. Trained on the same samples used to build `WHOv1` and benchmarked on an independent Validation Dataset, `catomatic` achieves comparable sensitivity, specificity, and definitive prediction rates to `WHOv1` without expert-rule augmentation and despite using simpler statistics. It optimises parameters per drug, produces machine-readable outputs (CSV/JSON), and demonstrates that adjusting read-support thresholds can improve detection of minor resistance subpopulations.

Implications of all the available evidence

Catalogues of resistance-associated variants for *M. tuberculosis* can be rapidly and transparently constructed. Making catalogues available in human/machine-readable formats with uncertainty estimates will improve uptake of WGS for *M. tuberculosis* surveillance and diagnostics; using a reproducible process permits diagnostic test manufacturers, researchers, clinical and public health laboratories to select the level of statistical support necessitated by their specific use-case, Policymakers should balance the benefits of expert rules against loss of reproducibility. Future work will expand the size of the datasets used, integrate minimum inhibitory concentration data, and establish consensus workflows for routine, transparent catalogue updates.

Introduction

Tuberculosis is the poster child for the translation of whole genome sequencing (WGS) into clinical microbiology; genetics is faster, potentially more accurate than phenotypic drug susceptibility testing (pDST), more comprehensive than molecular tests and also yields epidemiological insights. The key to predicting from a *Mycobacterium tuberculosis (Mtb)* genome whether a tuberculosis infection is susceptible or resistant to a range of antibiotics is the *catalogue* – a list of known genetic variants and their effects on individual antibiotics[1]. Catalogues currently have two primary uses: to guide the development of molecular tests that detect specific genetic variants, and application in WGS-based surveillance and diagnostics.

The release of the first edition of the catalogue of mutations in *Mtb* and their association with resistance by the World Health Organization (WHO) in 2021[2,3] marked a milestone in the adoption of WGS in tuberculosis surveillance and diagnostics; we call this catalogue `WHOv1`. A dataset of 41,137 clinical isolates with WGS and pDST data donated by the CRyPTIC and Seq&Treat consortia was analysed[3], resulting in 1,486 genetic variants associated with either resistance or susceptibility to one of 13 antituberculars. Following the endorsement of the oral six-month BPaLM regimen for multidrug-resistant TB[4], a second edition (`WHOv2`) was published in late 2023 based on 61,986 clinical isolates[5], and notably added resistance-associated variants (RAVs) for bedaquiline (BDQ) and clofazimine (CFZ). Both catalogues were constructed by associating binary phenotypes (resistant or susceptible) to solitary mutations in resistance genes, followed by confidence grading and the addition of rules derived from the scientific literature as agreed by an expert panel[2,5]. Together these catalogues mark a significant step towards standardising how the genetics of tuberculosis infections are interpreted clinically.

They suffer, however, from several shortcomings, especially a lack of transparency and reproducibility: although the donated dataset for `WHOv1` is available via the CRyPTIC project[6], the `WHOv2` dataset is not public, and, while some code is available[7], it is not sufficient to repeat the entire process. Secondly, despite strong interest in integrating the WHO catalogues into pipelines and software[1,8], their complexity makes them difficult to read and parse[8]. For example, the catalogue is not a single artefact – most of the rules are contained in an Excel spreadsheet but some must be parsed from the report, which has led to disagreement between validation studies[9]. Thirdly, since the release of `WHOv2` in November 2023, minor errors in the Excel spreadsheet have been corrected, but these changes are only flagged via the GitHub page[7], leading to the risk that people continue to use deprecated versions. Finally, rebuilding the WHO catalogues is resource-intensive and time-consuming and hence updating the catalogue as new data becomes available is a major undertaking.

In this paper, we describe how we have reproducibly built a catalogue of *Mtb* genetic variants associated with resistance or susceptibility to 15 antibiotics using the same dataset that was used to build `WHOv1`. We then

demonstrate comparable performance of our catalogue to `WHOv1` using an independent Validation Dataset of 14,380 samples. This work is enabled by (i) the continued release of versioned data by the CRyPTIC consortium[10], (ii) the use of an online cloud platform for rapid processing of the raw genetics files[8], (iii) a publicly-available software tool, `catomatic`, that automatically classifies the effect of individual solitary mutations in resistance genes[11] resulting in (iv) a catalogue described by a single file that is both human- and computer-parsable that (v) can be ingested by an existing freely-available prediction tool, `gnomonicus`[8,12].

Methods

Construction of independent Training and Validation Datasets.

We downloaded v3.4.0 of the CRyPTIC dataset[10], containing 53,897 samples that have whole genome sequencing data with translated variant calls and at least one pDST result. This is referred to as the Entire Dataset. Genetic subpopulations supported by three or more short reads were called, here termed minor alleles. The pDST methods are varied; the two most common are 96-well broth microdilution plates (288,302 R/S measurements) and the BD Mycobacterial Growth Indicator Tube (190,909 R/S measurements). A set of published epidemiological cut-off (ECOFF) values[13] had been used to convert the MICs into R/S labels. Plate readings are subjective and we only took forward those that were concordant with a machine learning model, `TMAS`[14]; this minimised exclusions to 159.

The Training Dataset (Table 1) was defined as all 41,130 samples found in v1.1.1 of the CRyPTIC dataset[6]. This is seven samples fewer than used to build `WHOv1`[3]; the reason for this is unknown. The remaining samples from v3.4.0 of the CRyPTIC dataset therefore formed The Validation Dataset (Table 1). It is likely that some of these samples were also used to build `WHOv2`.

Identification of genetic variants statistically associated with resistance or susceptibility

We applied an algorithm originally termed 'definitive defectives'[15], later repurposed for generating *Mtb* catalogues[16], together with a two-tailed binomial test (Fig. S1), to identify and classify benign variants, as implemented in `catomatic` (v0.1.9)[11,17] and described in the Supplement. This naturally leads to a ternary classification system: variants in resistance genes are classified as conferring either resistance ('R') or susceptibility ('S') or having an unknown effect ('U'). We focussed on genes implicated in conferring resistance because the univariate method can only use samples with a single mutation; including non-relevant genes reduces statistical power by limiting informative isolates. Accordingly, we only considered the highly-confident ('Tier 1') candidate genes listed in `WHOv2` that contain at least one RAV[5]. Synonymous and phylogenetic mutations (defined in *Merker et al.*[18]) also reduce our power: all the former were excluded, and only a few of the latter (S95T, E21Q, G668D in *gyrA*, M291I in *gyrB*) impaired performance and were removed via a customisable filter.

The background rate (H_0), confidence level (p) and the minimum fraction of read support (FRS_{min}) must also be defined - the latter is the minimum proportion of reads at a genetic loci required to support a variant call and is specified separately for building and applying catalogues. The process generated a catalogue (a CSV file) for each drug in a format that is compatible with `gnomonicus`[8] to maximise readability while enabling rapid application to new datasets.

Evaluation and benchmarking of catalogue performance

When assessing catalogue performance, we predicted any sample containing a RAV for a particular drug to be Resistant (R) to that drug. If the sample contained only susceptible variants, we predicted it to be Susceptible (S), whilst samples containing one or more variants with unknown effects (but no RAVs) were classified as Unknown (U). Since it is not reasonable to assume that all samples classified as having an Unknown effect are Susceptible for all drugs, we defined the Definite Prediction Rate (DPR) as the proportion of samples for which a definite prediction, $(n_R + n_S)/(n_R + n_S + n_U)$, can be returned[11], where n_R is the number of samples predicted to be resistant, etc. The sensitivity and specificity of the samples with a definite prediction (R or S) is then calculated. Usually one aims to minimise the very major error (VME) rate, which is the proportion of resistant samples incorrectly predicted to be susceptible (i.e. $1 -$ sensitivity), whilst also balancing the proportion of susceptible samples incorrectly predicted to be resistant, the major error (ME = 1 – specificity) rate. We captured this in an arbitrary cost function: the weighted sum of sensitivity (weight = 0·5), specificity (0·3), and DPR (0·2). To benchmark our catalogue against `WHOv1`, and to compare the performance of `WHOv1` on our data vs its reported results, we used a two-proportions Z-test on the Validation Dataset.

Role of the funding source

The funders played no role in the design of this study.

Drug	Samples in Training Dataset			Samples in Validation Dataset		
	R	S	Total	R	S	Total
Rifampicin, RIF	10,928	26,053	36,981	4,283	7,199	11,482
Isoniazid, INH	13,662	23,498	37,160	5,225	6,155	11,380
Ethambutol , EMB	5,525	28,980	34,505	2,033	5,003	7,036
Pyrazinamide, PZA	3,190	18,521	21,711	1,797	5,266	7,063
Levofloxacin, LEV	2,192	9,583	11,775	466	2,979	3,445
Moxifloxacin, MXF	2,983	14,207	17,190	829	6,576	7,405
Bedaquiline, BDQ	144	11,082	11,226	838	3,476	4,314
Clofazimine, CFZ	451	12,293	12,744	106	3,476	3,582
Linezolid, LZD	205	14,369	14,574	138	4,331	4,469
Delamanid, DLM	147	10,311	10,458	44	2,789	2,833
Amikacin, AMI	1,497	17,715	19,212	742	7,062	7,804
Streptomycin, STM	4,543	9,837	14,380	1,630	1,954	3,584
Ethionamide, ETH	2,887	11,596	14,483	1,053	3,806	4,859
Kanamycin, KAN	1,839	16,360	18,199	991	7,029	8,020
Capreomycin, CAP	956	10,582	11,538	595	5,461	6,056

Table 1: The number of samples, stratified by drug, in The Training Dataset (CRyPTICv1.1.1) and the independent Validation Dataset. All samples have both drug susceptibility data and whole-genome sequences. Drugs are ordered as listed in `WHOv2`[2]. The first four compounds (RIF, INH, EMB and PZA) are the first-line treatment for drug-susceptible tuberculosis.

	RIF	INH	EMB	PZA	LEV	MXF	BDQ	CFZ	LZD	DLM	AMI	STM	ETH	KAN	CAP
R	42	62	20	440	13	11	34	46	1	18	11	115	298	11	43
S	16	23	58	0	19	22	0	0	4	0	6	8	0	6	2
Total	58	85	78	440	32	33	34	46	5	18	17	123	298	17	45
U	352	626	1056	143	330	489	172	172	1	29	383	542	325	366	234

Table 2: The number of unique genetic variants by drug classified as resistant (R) or susceptible (S) in the `catomatic-1` catalogue. Those unable to be classified i.e. have an unknown (U) phenotype are also listed. These have been stratified by gene in Table S2. The full drug names are given in Table 1

Results

Setting parameters by drug improves performance

Anti-tuberculosis drugs act on a range of protein targets in *Mtb* and a range of resistance mechanisms have subsequently evolved. Accordingly, each drug's catalogue should be built with tailored statistical parameters to maximise predictive performance. We therefore performed parameter grid searches on the Training Dataset using `catomatic` allowing the background threshold ($0{\cdot}05 \leq H_0 \leq 0{\cdot}25$) and confidence level ($p \in 0{\cdot}9, 0{\cdot}95$) to vary independently for each drug (Fig. S2). Drugs with many resistant samples in the Training Dataset, such as the first-line agents, consistently perform well on the Training Dataset, with the DPR, sensitivity, and specificity always exceeding 80%. In contrast, drugs with fewer resistant samples, including bedaquline, clofazimine, linezolid, and delamanid, exhibit both lower sensitivity (typically $< 50\%$) and a stronger negative relationship with increasing background H_0 value. We programmatically selected optimal parameters for each drug using our arbitrary cost function (Table S1), thereby creating our catalogue. Rifampicin performs best, with a sensitivity of 96.1%, a specificity 98.5%, and a DPR 97.1%..

The number of variants associated with resistant or susceptibility varies by drug (Table 2) and gene (Table S2) and ranges from five variants classified for linezolid to 440 for pyrazinamide. Most genetic variants occur too infrequently to be classified, leaving many unclassified; however, their combined impact is usually minimal as most resistance is typically (but not always) conferred by a few, highly 'penetrant' mutations. In the accompanying repository, this catalogue is versioned as `CRyPTICv1.1.1-2025.8`; throughout this manuscript, we refer to it as `catomatic-1`.

Genetic subpopulations contribute to resistance and must be considered

Genetic sub-populations were detected in all key resistance genes. Investigating the effect of these minor alleles is complicated by (assumed) neutral lineage-associated variants, such as R463L in *katG*, as these can be highly prevalent in heterogenous samples. The gene *embB* harbours a disproportionately high number of benign minor alleles — 28.8% of all identified variants — in particular, R24P. Excluding these reveals that, as one might expect, non-essential genes have a higher proportion of minor alleles compared with essential genes (8.91% vs 6.44%, $p < 0.001$, Fig. 1A).

Irrespective of essentiality, minor alleles are more likely to be found at a lower FRS of 10-40% (Fig. S3A), in contrast with the more even distributions observed for *resistant* sub-populations (Fig. S3B). Notably, the proportion of resistant samples that are heterogeneous exceeds 10% for several genes, including *gyrA* (12.2%), *gyrB* (27.7%), and *rplC* (17.7%), as well as the non-essential genes *Rv0678* (41.0%) and *ahpC* (16.4%).

Despite this, catalogues have historically been built assuming homogenous genetics, typically by requiring a high FRS_{min} (0.75 or 0.9) to support variant calls, which also has the effect of mitigating short-read sequencing errors. To evaluate the impact of including these minor alleles, we built catalogues in the range $0.1 \leq FRS_{min} \leq 0.9$ and assessed their performance on the Training Dataset (Fig. S4). Lowering the FRS_{min} when *building* the catalogues had a negligible impact on sensitivity or specificity ($< 1\%$, $p > 0.05$). This suggests that resistance is usually conferred by mutations that can be identified and classified using a high FRS_{min}, provided sufficient resistant samples are available.

A second FRS_{min} threshold is set when *applying* the catalogue. Aside from linezolid and delaminid, lowering this FRS_{min} increases the sensitivity for all drugs ($p < 0.001$), enabling the detection of clinically relevant RAVs in subpopulations that would otherwise be missed (Fig. S4). For example, the number of classified RAVs for amikacin remains largely constant at 11 regardless of the value of FRS_{min}, yet sensitivity improves by up to six percentage points as the threshold used for detection is reduced (Fig. 1B). Even for *rpoB*, which is highly essential and has a low incidence of minor RAVs (3.7%), excluding these variants from predictions decreases sensitivity by 2.1% (Fig. S4)[19]. The FRS_{min} should therefore be lowered when applying catalogue as this will reduce the VME rate.

Our automatically generated `catomatic-1` catalogue performs similarly to `WHOv1`

Neither `WHOv1` nor `catomatic-1` were trained on the 14,380 samples in the Validation Dataset (Table 1) and therefore these can be used to assess their performance. The performance of `WHOv1`, when evaluated using a value for FRS_{min} of 0.75, was largely consistent with that reported[2], indicating that `WHOv1` generalises well (Fig. S5, Table S3). We then evaluated the performance of `catomatic-1` on the Validation Dataset and compared its performance to `WHOv1` (Fig. 2, Table S3)

Both catalogues achieved comparable performance across most drugs. Ethambutol showed the largest improvement over `WHOv1`, with a 10% increase in DPR (88.6% v 78.6%,$p < 0.001$) and a +3.6% boost to specificity (82.3% v 78.8%, $p < 0.001$), balanced by a 2% reduction in sensitivity (91.5% v 93.5%, $p = 0.021$). This reflects the classification of 47 susceptible variants and eight RAVs in *embB* absent from `WHOv1`, partly offset by eight susceptible variants and three RAVs unique to `WHOv1` (Fig. 3). Isoniazid also performs better ($p < 0.001$), albeit the increases in DPR (95.7% v 94.1%) and sensitivity (94.6% v 93.7%) are more modest. Notably, `catomatic-1` makes fewer definite predictions for kanamycin and capreomycin (91.6% v 94.2% and 93.9 v 95.6%, respectively) than `WHOv1` and this is not offset by corresponding gains in sensitivity or specificity (Fig. 2, Table S3).

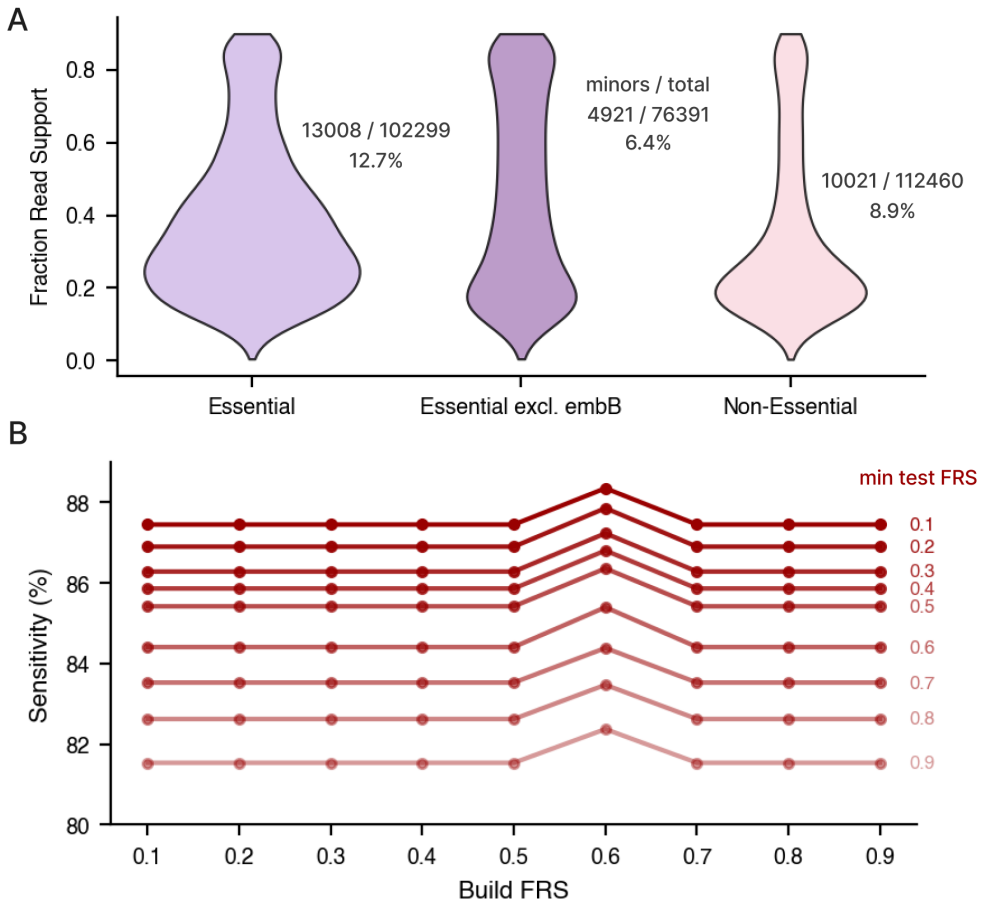

Figure 1: A. Distributions of the number of pooled minor alleles ($FRS < 0.9$) in The Training Dataset minus phylogenetic mutations, scaled against the total number of variants, from $0.1 \leq \text{FRS} < 0.9$, for essential genes (lilac) and non-essential genes (pink). The dark purple violin excludes *embB*. The absolute fraction of minor alleles over the total number of alleles in The Training Dataset are shown - the area of each violin is proportional to this fraction. Distributions stratified by gene are shown in Figure S3. B. Sensitivity for amikacin at varying FRS_{min} thresholds used to train the catalogue (Build FRS) and when making predictions for samples (min test FRS). The 'spike' is due the loss of a false positive mutation at FRS_{min} 0.6. Complete sensitivity, specificity, and DPR results when varying the FRS_{min} threshold at the build and test steps are shown for all drugs in Fig. S4. The full drug names are given in Table 1

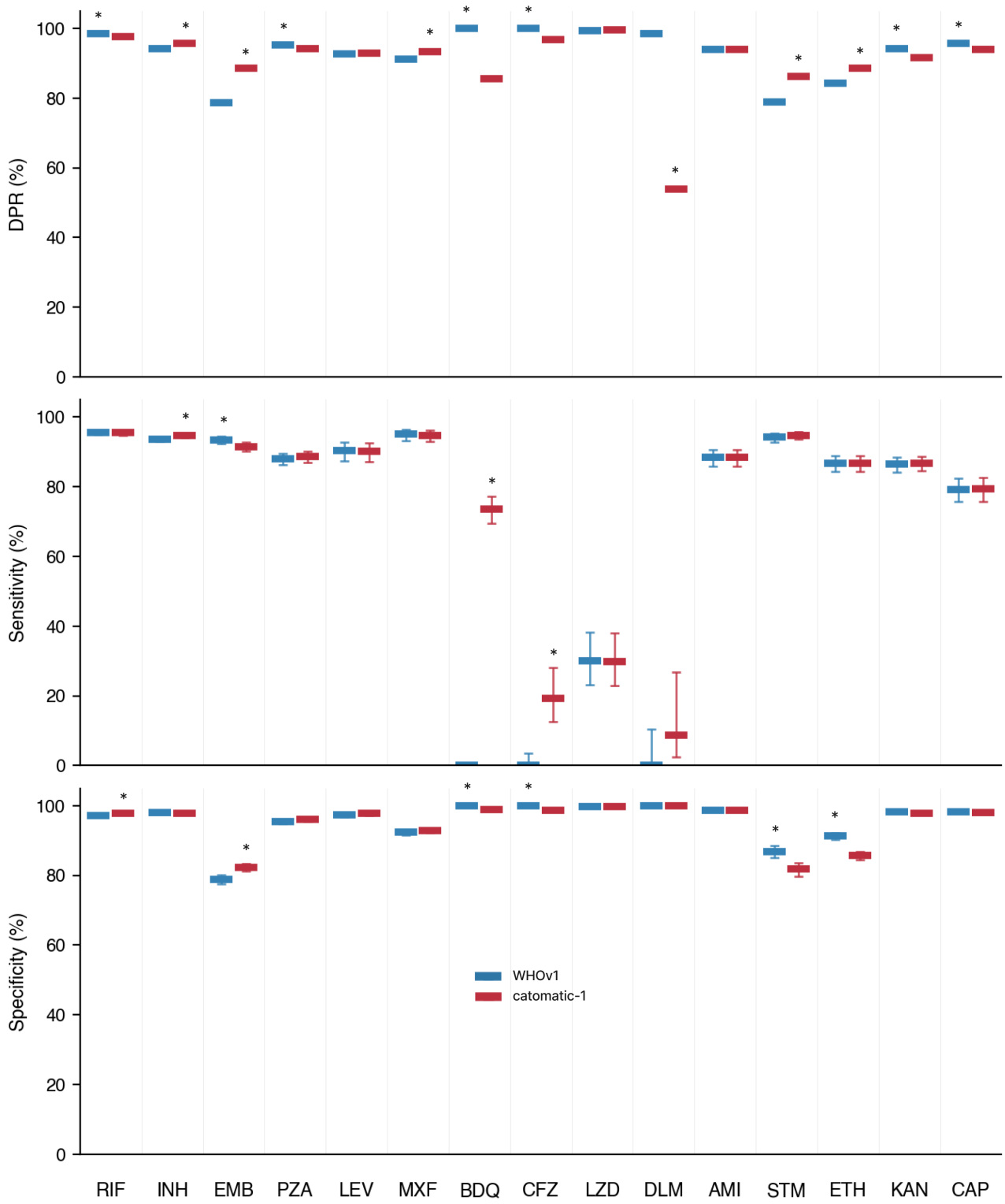

Figure 2: The definitive prediction rate (DPR), sensitivity and specificity, and for `WHOv1` (blue) and `catomatic-1` (red) catalogues when evaluated on The Validation Dataset of 14,380 samples using a minimum FRS of 0.1 to call genetic variants. Both `WHOv1` and `catomatic-1` were built on The Training Dataset and therefore these results are an independent assessment of performance. Asterisks show the significantly greater value at the 95% confidence level.

Expert rules can improve performance, but make reproducibility harder

Although the performance of `WHOv1` and `catomatic` is similar, the nummber of genetic variants for which `WHOv1` returns a definite classification is greater due to generic rules which were added post-hoc to `WHOv1` by an Expert Panel for rifampicin, isoniazid, pyrazinamide, ethionamide, the fluoroquinolones, and the aminoglycosides[2]. For example, an expert rule in `WHOv1` states that any genetic variant within the *rpoB* "Rifampicin Resistance Determining Region" are resistant and this allows `WHOv1` to classify 87 RAVs in the Training Dataset that `catomatic` does not, and 105 RAVs compared to its base algorithm; representing 71% of all mutations classified across both catalogues (Fig. 3A, S6A). The net effect these rules have on performance is small, particularly for rifampicin and isoniazid (Fig. S6B), although they do increase the DPR values achieved by the `WHOv1` catalogue on the Validation Dataset for rifampicin, pyrazinamide, streptomycin, ethionamide (all $p < 0.001$), and amikacin ($p = 0.030$), albeit at costs to specificity ($p < 0.01$) for the first four drugs (Fig. S7, Table S3).

Excluding Expert Rules from `WHOv1` enables a like-for-like comparisons of the two algorithms and demonstrates that the simpler statistical framework used by `catomatic` is more liberal, leading to many more *algorithmically* classified variants across all drugs compared to `WHOv1` (Fig. 3B, S6A). The Expert Rules therefore improve the ability of `WHOv1` to make a prediction at the price of preventing the final catalogue from being independently reproduced, updated, or audited unless they are first removed. We argue this constrains usability and reliability for a modest performance gain that could alternatively be achieved with adequate but less stringent statistical testing.

Updating an algorithmically generated, version controlled catalogue is straightforward

We used `catomatic` to construct a final, larger catalogue on the Entire Dataset, versioned `MBTC-CRyPTICv3.4.0-2025.8` and referred to here as `catomatic-2`. This catalogue cannot be independently validated so is primarily for interest. Incorporating the 14,380 additional samples (Table 1) enabled 415 additional variants to be classified, boosting the DPR of 11 drugs and increasing bedaquiline sensitivity by 10.3% (69.0% v 79.2%), bringing its performance close to a previously published catalogue[11] (Fig. S8, Table S5). We also calculated the performance of `WHOv1` and `WHOv2` on this dataset for completeness (Table S5).

One must expect that when the evidence base is updated, classifications may change. For example, the net gain in classified variants by `catomatic-2` is 387 (not 415), as some mutations change classification. For example S441V in *rpoB* is no longer designated as an RAV since it now occurs in a susceptible isolate, while A686T is no longer considered susceptible following its appearance in a resistant isolate (Fig. S9). Such changes are to be expected when applying a statistical test and can be readily identified when using version-controlled catalogues downloadable in a standardized, machine-readable format. However having variants change clas-

sification, as above, is undesirable when designing a molecular test, one of the use cases for such catalogues. The obvious solution is to provide the statistical certainty and evidence alongside each RAV in the catalogue so classifications based on fewer samples (with therefore a greater probability of changing classification in future) can be screened out when using the catalogue in this way.

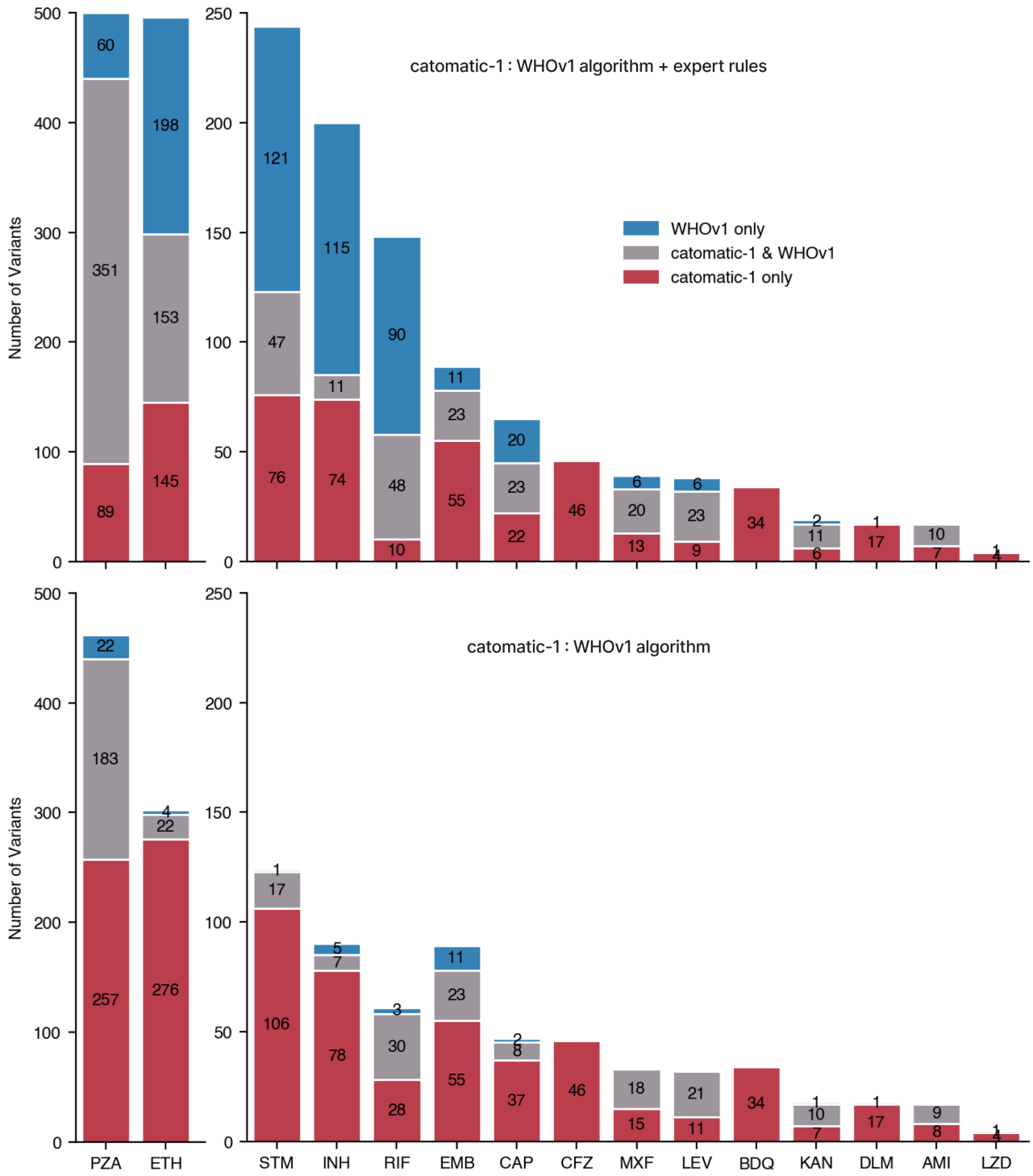

Figure 3: The number of unique genetic variants in The Training Dataset classified as either resistant or susceptible differs between `catomatic-1` and `WHOv1`. Variants only classified by `catomatic-1` or `WHOv1` are shown in red and blue, respectively, whilst variants classified by both catalogues are shown in grey. (A) Naively comparing the two catalogues shows that, for many drugs, the number of variants that both catalogues classify is in the minority. (B) Removing the Expert Rules from `WHOv1` to enable a like-for-like comparison clearly shows how the more permissive statistics used by `catomatic` has allowed more definite classifications to be made. Bars for pyrazinamde (PZA) and ethionamide (ETH) are plotted at half scale for clarity. Unclassified mutations are not shown. Fig. S6 illustrates the effect each group has on predictive coverage. The full drug names are given in Table 1

Discussion

We have reproducibly built a comprehensive catalogue of genetic variants in *M. tuberculosis* associated with resistance or susceptibility to 15 anti-tuberculosis drugs (`catomatic-1`) using the same dataset that was used to build `WHOv1`.[2]. Evaluating the performance of both catalogues on an independent Validation Dataset of 14,380 samples showed that, despite the process being automated and no Expert Rules being added, using a simpler statistical framework achieved similar performance to `WHOv1` for all drugs.

Regardless of method, building a catalogue is more straightforward if (i) the resistance gene(s) is (are) essential (as this tends to lead to the resistance mechanisms being dominated by just a few genetic variants), (ii) variant effects are large (i.e. many-fold MIC increases), (iii) there is good concordance between differrent pDST methods, and there being (iv) many clinical samples to train on with (v) a high prevalence of resistance. These conditions are met here for most drugs and hence performance is generally satisfactory, except for clofazimine, delamanid and linezolid which all have fewer than 1,000 resistant samples in the Entire Dataset.

It is unsurprising that a significant proportion of *Mtb* complex samples show heterogeneous genetic variation, given that individuals may be infected for years before disease develops, allowing ample time for secondary infections and/or *in vivo* evolution[19]. Heterogeneity is likely to be detected by WGS since it is usual to culture samples in liquid broth (MGIT tubes) and then extract DNA from multiple harvested 'crumbs'. Rapidly and automatically building catalogues with different values of FRS_{min} enabled us to examine the effect these subpopulations had on performance. Reducing FRS_{min} had little effect when building catalogues, therefore we conclude most RAVs can be discovered and classified in homogonous samples (in this dataset at least). When we allowed minor alleles to be identified in samples, by reducing FRS_{min}, the VME rate was reduced for all drugs (Fig. S4), consistent with other work[1,11,20,21]. We observed increases in performance down to an FRS_{min} threshold of 10%, however, we note that properly one should set this threshold statistically but there are insufficient hetereogenous samples in this dataset to do this yet.

Like other approaches, our automated method based on `catomatic` requires users to specify the resistance-associated genes for each drug, meaning the resulting catalogue is only as reliable as this prior knowledge. A second assumption is that silent mutations and frequent phylogenetic variants (of which we filtered out four in the DNA gyrase) in the candidate genes have no effect. Using `catomatic`, one could perform a sensitivity analysis on the candidate genes and phylogenetic variants to justify their inclusion/exclusion. Furthermore, dichotomising MIC distributions presupposes a clear bimodal separation between resistant and susceptible phenotypes, which is not always true, instead supporting the use of mixed-effects models trained directly on MICs[22]. Moreover, a univariate approach cannot capture interactions or additive effects, whereas multivariate regression can[20,22]. Whilst the components of the pipeline we used are freely-available[8] and it has been de-

ployed in a cloud-based platform it is not yet straightforward for other researchers to reprocess all samples we have used here due to the computational costs involved.

Performance is key but should not come at the expense of usability. The recent proposal to use large language models to simplify interpretating the WHO catalogues[23] is testament to their unwieldiness; to build trust, catalogues *must* be easily readable by both human and computer. The catalogues output by `catomatic` are described using the GARC grammar[8] and can be stored in either JSON or CSV formats, both of which are readable using standard programming libraries, while CSVs can also be opened in spreadsheet software. The CSV files follow a simple tabular structure and are compatible with `gnomonicus`[8,12], a freely-available tool that returns a list of antibiotiocs and whether a sample is susceptible or resistant to each, and the effect of individual genetic variants, given a specified catalogue, a GenBank file and a variant call file.

The fast and reproducible method for generating catalogues demonstrated here, when coupled to a reliable genetic processing pipeline, could enable an international responsive surveillance system that in turn regularly and automatically updates a tuberculosis knowledge base. This would be particularly valuable for newly approved drugs (assuming they are accompanied by the rapid roll-out of pDST methods) by minimising the delay in detecting the first resistance-associated variants. As more samples are collected, catalogues could even be built using samples originating from a specific country or continent or by lineage, offering greater resolution in specific settings.

Enormous strides have brought us to this point, and it is vital we recognise that automation and reproducibility are both required to get us further and are also inevitable. By adopting best practises now we lay the groundwork for broader uptake and interoperability. This is feasible right now, and the ability to generate catalogues of tuberculosis resistance-associated genetic mutations at the press of a button is now a matter of will and implementation.

Contributors

PWF, TMW, TEAP, DWC and ZI conceptualised the study with input from SO and NI. KMM, MH and DA identified additional samples to add to the Validation Dataset. JW, HT, MC, RDT and PWF processed all the samples creating the data tables. DA analysed these data, with supervision from PWF, DWE and TEAP. DA wrote the first draft of the manuscript which was reviewed by all authors prior to submission. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

ZI, DWC & PWF work as consultants for the Ellison Institute of Technology, Oxford Ltd.

Data sharing

The versions of the CRyPTIC datasets used in this study are publicly available[24] and contain all phenotypic drug susceptibility testing measurements and all genetic variants detected in each sample; the datasets also includes the run accession numbers so the raw FASTQ files can be downloaded from the European Nucleotide Archive if required. An attendant GitHub repository[1] contains data and Python code in the form of juypter notebooks that allows all results, including many of the figures, to be reproduced. The key dependency, `catomatic`, is installable via PyPI[2].

Acknowledgments

This research is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR207397), a partnership between the UK Health Security Agency (UKHSA) and the University of Oxford and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). DA is supported by the EPSRC Sustainable Approaches to Biomedical Science: Responsible & Reproducible Research CDT and by ORACLE Corporation. The views expressed are those of the authors and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

[1] https://github.com/fowler-lab/cryptic-catalogues-2025

[2] https://pypi.org/project/catomatic/

References

[1] Hall MB, Lima L, Coin LJM, Iqbal Z. Drug resistance prediction for Mycobacterium tuberculosis with reference graphs. Microbial Genomics. 2023 8;9.

[2] World Health Organisation. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance. 2021. Available from: `https://www.who.int/publications/i/item/9789240028173`.

[3] 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. The Lancet Microbe. 2022;3:e265-73.

[4] World Health Organisation. Rapid communication: Key changes to the treatment of drug-resistant tuberculosis; 2022. Available from: `https://www.who.int/publications/i/item/WHO-UCN-TB-2022-2`.

[5] World Health Organization. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance. Second edition; 2023. Available from: `https://www.who.int/publications/i/item/9789240082410`.

[6] The CRyPTIC Consortium Dataset. Version 1.1.1, doi:10.5281/zenodo.15679731;.

[7] Laurent S, Chindelevitch L;. `https://github.com/GTB-tbsequencing/mutation-catalogue-2023`.

[8] Westhead J, Baker CS, Brouard M, Colpus M, Constantinides B, Hall A, et al.. Characterising the performance of an antibiotic resistance prediction tool, gnomonicus, using a diverse testset of 2,663 *Mycobacterium tuberculosis* samples; 2024. BioRxiv preprint, doi:10.1101/2024.11.08.622466.

[9] Laurent S, Phelan JE, Chindelevitch L, Walker TM, Cirillo DM, Suresh A, et al. All parts of the WHO *Mycobacterium tuberculosis* mutation catalog need to be applied when evaluating its performance. Microbiology Spectrum. 2025 Apr:e02157-24.

[10] The CRyPTIC Consortium Dataset. Version 3.4.0, doi:10.5281/zenodo.15680920;.

[11] Adlard D, Joseph L, Webster H, O'Reilly A, Knaggs J, Peto TEA, et al. An improved catalogue for whole-genome sequencing prediction of bedaquiline resistance in Mycobacterium tuberculosis using a reproducible algorithmic approach. Microbial Genomics. 2025 6;11.

[12] Westhead J, Fowler PW. `https://github.com/oxfordmmm/gnomonicus`; 2025.

[13] Fowler PW, Barilar I, Battaglia S, Borroni E, Brandao AP, Brankin A, et al. Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of M. tuberculosis. European Respiratory Journal. 2022 10;60.

[14] Vo HAT, Nguyen S, Tran AQT, Nguyen H, Ho HB, Fowler PW, et al. Deep learning-based framework for Mycobacterium tuberculosis bacterial growth detection for antimicrobial susceptibility testing. Computational and Structural Biotechnology Journal. 2025 1;27:2208-18.

[15] Dorfman R. The Detection of Defective Members of Large Populations. The Annals of Mathematical Statistics. 1943 12;14:436-40.

[16] Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. The Lancet Infectious Diseases. 2015 10;15:1193-202.

[17] Adlard D, Fowler PW. catomatic; 2024. doi: 10.5281/zenodo.14917986. Available from: `https://github.com/fowler-lab/catomatic`.

[18] Merker M, Kohl TA, Barilar I, Andres S, Fowler PW, Chryssanthou E, et al. Phylogenetically informative mutations in genes implicated in antibiotic resistance in Mycobacterium tuberculosis complex. Genome Medicine. 2020 12;12:27.

[19] Brunner VM, Fowler PW. Subpopulations in clinical samples of M. tuberculosis can give rise to rifampicin resistance and shed light on how resistance is acquired. J Antimicrob Chemother AMR. 2025. Available from: `https://doi.org/10.1093/jacamr/dlaf175`.

[20] Kulkarni SG, Laurent S, Miotto P, Walker TM, Chindelevitch L, Nathanson CM, et al. Multivariable regression models improve accuracy and sensitive grading of antibiotic resistance mutations in Mycobacterium tuberculosis. Nature Communications. 2025 3;16:2149.

[21] Brankin AE, Fowler PW. Inclusion of minor alleles improves catalogue-based prediction of fluoroquinolone resistance in *Mycobacterium tuberculosis*. JAC-Antimicrobial Resistance. 2023 3;5.

[22] The CRyPTIC Consortium. Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach. Nature Communications. 2024 12;15.

[23] Moreno-Molina M, Suresh A, Colman RE, Rodwell TC. Facilitating User Interaction with the Tuberculosis Mutation Catalogue using AI Tools. bioRxiv; 2025. Available from: `https://www.biorxiv.org/content/10.1101/2025.04.25.650567v1`.

[24] The CRyPTIC Consortium Datasets. doi:10.5281/zenodo.15679730;.