Enhancement and validation of the antibiotic resistance prediction performance of a cloud-based genetics processing platform for Mycobacteria.

Jeremy Westhead[1], Catriona S Baker[1], Marc Brouard[1], Matthew Colpus[1], Bede Constantinides[1], Alexandra Hall[1], Jeff Knaggs[1], Marcela Lopes Alves[1], Ruan Spies[1], Hieu Thai[1], Sarah Surrell[5], Kumeren Govender[5], Timothy EA Peto[1,2,3], Derrick W Crook[1,2,3], Shaheed V Omar[4], Robert Turner[1], and Philip W Fowler*[1,2,3]

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

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

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

[4]Centre for Tuberculosis, National & WHO Supranational TB Reference Laboratory, National Institute for Communicable Diseases, a division of the National Health Laboratory Service, Johannesburg, South Africa

[5]Ellison Institute of Technology, Oxford, U.K.

Abstract

Tuberculosis remains a global health problem. Making it easier and quicker to identify which antibiotics an infection is likely to be susceptible to will be a key part of the solution. Whilst whole-genome sequencing offers many advantages, the processing of the genetic reads to produce the relevant public health and clinical information is, surprisingly, often the responsibility of the end user which inhibits uptake. Here we describe our Mycobacterial genetics processing pipeline and its deployment in a cloud-based platform. For antibiotic resistance prediction we have implemented the second edition of the WHO catalogue of resistance-associated variants. We validate the resistance prediction performance by constructing and processing a diverse dataset of 2,663 publicly-available *M. tuberculosis* samples with published drug susceptibility testing (DST) data and find that identifying a sample as resistant if it contains a minor allele known to be associated with resistance increases sensitivity. By only considering high confidence DST results we are able to show that both isoniazid and rifampicin achieve sensitivities and specificities in excess of 95%.

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

Introduction

In 2023 just under 11 million people became ill with tuberculosis (TB) worldwide and 1.25 million died[1]. The aetiological agent, *M. tuberculosis* , is difficult to cure with a single drug and therefore requires treating with multiple antibiotics. Susceptible *M. tuberculosis* is routinely treated with four antibiotics: rifampicin, isoniazid, pyrazinamide and ethambutol. Like other bacterial pathogens, resistance to all antibiotics has now been observed and infections resistant to both rifampicin and isoniazid are called multi-drug resistant (MDR) TB and require alternative treatment, such as the BPaLM (bedaquline, protonamid, linezolid and moxifloxacin) regimen that was recommended by the WHO in 2022[2]. MDR TB is increasingly recognised as a global health concern and as a result rifampicin-resistant *M. tuberculosis* was added to the WHO Bacterial Priority Pathogens List in 2024[3]. Positive identification of *M. tuberculosis complex* in a clinical sample and subsequent drug susceptibility testing (DST) is an important component in tackling this disease.

Due to the slow growth rate of *M. tuberculosis* culture-based DST takes weeks; whole genome sequencing (WGS) is an attractive alternative since it is faster, potentially more accurate and also yields epidemiological information. Over the last seven years[4], many high income countries have adopted WGS for *M. tuberculosis* (and in some cases the genus Mycobacteria). Predicting the antibiogram of a sample from its genetics has been boosted in recent years by the systematic sequencing and drug susceptibility testing of large numbers of clinical samples by projects such as the Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) which collected over 15,000 clinical samples through 14 laboratories based in 11 countries[5]. These and other publicly available datasets enabled the World Health Organization to release in 2021 the first catalogue of mutations in *M. tuberculosis complex* associated with drug resistance[6]. A second edition was released in late 2023[7] which we will call WHOv2. Both editions take the form of a traditional text-based report (which contains some expert rules) and an accompanying detailed Excel worksheet (with more rules), with the second edition also including a variant call file. Each edition used a single dataset to both infer the association of genetic variants with antibiotic resistance and estimate the performance of the resulting catalogue.

A key obstacle hindering the uptake of WGS is the requirement to write, host and maintain a pipeline that processes the raw genetic reads from a clinical sample, yielding the relevant public health and clinical information; this may include removing any human reads for data privacy, identifying contaminants and what species and/or mixtures are present before producing a consensus genome from which mutations in resistance genes can be derived and then looked up in a resistance catalogue. Cloud computing is an especially appealing solution since the entire pipeline and computing infrastructure required to run it can be hosted 'elsewhere' – all the user has to do is upload the samples, which admittedly, if the network is neither resilient or fast enough, can pose additional challenges. Other solutions include 'on device' computing or some variation on edge computing, but these all suffer from their own challenges and will not be discussed further.

In this paper we shall validate the antibiotic resistance prediction capabilities of a genetics pipeline we have developed that has been implemented in a cloud-based platform that is free for academics and LMICs. The

ability of the pipeline to detect different Mycobacterial subspecies and identify members of the *M. tuberculosis complex* that are putatively related and therefore could be part of the same outbreak lie outside the scope of this work. At present, the pipeline consumes short-reads and work is underway to validate long-reads. We shall specifically test our implementation of WHOv2, investigate the effect of some enhancements we have made and evaluate the performance of our Diverse Testset of 2,663 *M. tuberculosis* samples. Our hope is that cloud-deployed pipelines, such as show-cased here, will encourage the uptake of WGS by public health bodies for *M. tuberculosis* , even in LMICs which have the opportunity to 'leap-frog' several technologies in one go.

Materials and Methods

Sample selection

A total of 11,887 samples were identified from the publicly available CRyPTIC dataset[5] that (i) had been whole genome sequenced using short-read (Illumina) approaches and (ii) had minimum inhibitory concentrations (MICs) to 13 different antibiotics measured using a bespoke 96-well broth microdilution (BMD) plate. Each sample was sequenced and incubated on a BMD plate as described previously[5,8]. In addition to the MICs measured visually by the laboratory scientist, photographs of each plate were image processed[9] and classified by at least 11 volunteers as part of a citizen science project[10]. All MICs where two or three of these independent measurements agreed were annotated as high confidence MICs with the intention that this subset would have reduced measurement error. Two plate designs were used (UKMYC5 & UKMYC6), each of which included 13 antibiotics: amikacin, bedaquline, clofazimine delamanid, ethambutol, ethionamide, isoniazid, kanamycin, levofloxacin, linezolid, moxifloxacin, rifabutin and rifampicin. All MICs were converted to a binary `Resistant`/`Susceptible` classification using a set of research ECOFFs[8]. All 13 antibiotics bar rifabutin are included in the WHOv2 catalogue. The WHOv2 catalogue also includes three drugs not present on the plates: pyrazinamide, capromycin and streptomycin. We therefore also identified a further 10,606 publicly available samples[5] which had at least one binary phenotype measured via the MGIT960 system.

Construction of the Diverse Testset

The Diverse Testset has two competing aims: (i) have as close to 50% resistance / 50% susceptibility for all drugs to maximise resolution whilst also (ii) being as small as possible to enable rapid, repeated testing. The difficulty being, of course, that each sample chosen brings phenotypic information for more than one drug which makes it difficult to achieve the first criterion. Finally there is a risk of introducing bias if all samples have some degree of resistance. We therefore arbitrarily decided to create an initial dataset of 1,000 samples with phenotypes for the 13 drugs on the UKMYC plate designs ensuring that 200 of these were pan-susceptible.

Since the CRyPTIC project collected very few samples resistant to the new and repurposed drugs, we therefore first selected all 284 samples which were assessed as resistant to one or more of bedaquline (65), linezolid (117) and delamanid (140). Additional samples were then randomly selected if they (i) were resistant to the next drug with fewest resistant samples, (ii) were not resistant to any drugs that had already reached 50% penetration in the dataset and (iii) were drawn from the remaining 100 samples with the greatest number of high confidence MICs. This process was repeated until 800 samples had been chosen whereupon a further 200 pan-susceptible samples were chosen from the 1,000 pan-susceptible samples with the greatest number of high confidence MICs.

Despite our efforts, there are fewer than 250 resistant samples for bedaquline, linezolid, and delamanid in the UKMYC dataset (Table 1). We therefore repeated the process of iteratively selecting individual samples on the dataset of 10,606 samples with MGIT DST data, the main differences being (i) we only considered pyrazinamide, capreomycin, streptomycin, linezolid and bedaquline and (ii) each sample data for a variable

	DST method	%R	R	S	TOTAL
Bedaquiline	UKMYC	6.5	65	935	1000
Linezolid	UKMYC	11.7	117	883	1000
Delamanid	UKMYC	14.0	140	860	1000
Moxifloxacin	UKMYC	28.8	288	712	1000
Amikacin	UKMYC	28.8	288	712	1000
Ethambutol	UKMYC	28.8	288	712	1000
Ethionamide	UKMYC	28.8	288	712	1000
Clofazimine	UKMYC	28.8	288	712	1000
Kanamycin	UKMYC	28.9	289	711	1000
Levofloxacin	UKMYC	32.3	323	677	1000
Rifampicin	UKMYC	47.6	476	524	1000
Isoniazid	UKMYC	50.0	500	500	1000
Linezolid	MGIT	13.5	105	675	780
Streptomycin	MGIT	41.0	252	362	614
Capreomycin	MGIT	30.8	256	575	831
Pyrazinamide	MGIT	46.4	291	336	627
Bedaquiline	MGIT	50.9	394	380	774

Table 1: The proportion of resistance by drug in the 1,000 UKMYC and 1,663 MGIT samples. Due to having the highest prevalence of resistance in the CRyPTIC dataset, only isoniazid reached 50% resistance. A key difference between the two datasets is that each UKMYC samples has an MIC and thence a binary phenotype for 13 drugs whereas the MGIT samples have between one and four binary phenotypes, with two being the most common (702 samples).

number of these drugs and therefore we tried to maximise the number of antibiotics per sample. After de-duplication with the UKMYC dataset and verification that all samples had genetics available in the European Nucleotide Archive (ENA), this led to a dataset of 1,663 samples. The aggregated dataset of 2,663 samples therefore contains at least 250 resistant samples for all drugs (Table 1), except linezolid (117) and delamanid (140).

Processing of the Diverse Testset by EIT Pathogena

Raw genetics files (FASTQ files) were downloaded from the ENA and then uploaded to a test environment of EIT Pathogena using v1.0.3 of the command-line upload tool. This decontaminates the samples prior to upload by mapping reads to a custom human genome using the `hostile` algorithm[11]. Upon arrival in the cloud all samples were processed by our Mycobacterial Pipeline (v1.22.29). A complete description of the pipeline can be found in the Supplementary Information; we are focussing on the performance of resistance prediction here and therefore will not discuss in detail here the methods underlying speciation or detection of related samples.

In brief, all samples undergo a second round of decontamination with `hostile`[11] then poor quality/short reads are discarded before the number of reads belonging to the Mycobacterium genus is assessed using `kraken2` in conjunction with the `Standard` index from June 2023[12]. If a sample has over 10,000 Mycobacterial (and unclassified) reads it is progressed. These reads are then competitively mapped to a curated list of 186 Mycobacterial genomes using `minimap2`[13] to provide fine-grained speciation. Lastly, in the case of complexes, such as *M. tuberculosis complex* , the reads are then examined by `mykrobe`[14] which classifies reads down to the level of lineage (in the case of *M. tuberculosis*) or subspecies (for *Mycobacterium avium* complex and *Mycobacterium abscessus* complex). Reads are then mapped to version 3 of the *M. tuberculosis* H37Rv reference genome[15–17] using `clockwork`[18] which in turn uses `minimap2`[13] to build a pile-up and then both `samtools`[19] and `cortex`[20] to call variants, the former being better at identifying SNPs and the latter insertions and deletions; `minos`[21] adjudicates when there is overlap. `Clockwork` was used by the CRyPTIC project[5] and also for all variant calling in the first edition of the WHO catalogue of resistance-associated variants[6]. The variant call file is used by the resistance prediction process, which is described below, whilst the genome is passed to a novel algorithm, `FindNeighbour5`, which rapidly returns a list of samples within 20 SNPs that therefore could be epidemiologically related. The entire pipeline is deployed within the EIT Pathogena cloud platform that, through the use of industry standard technologies like `kubernetes` and object storage, is able to scale with demand and all data is resiliently stored and encrypted at rest.

Translation of the second edition of the WHO catalogue of resistance-associated variants in *M. tuberculosis* and resistance prediction

The second edition of the WHO catalogue of resistance-associated variants comprises three artefacts: a report, an Excel file and a variant call file (VCF). No single artefact contains all the rules in the catalogue and only the last two are parsable by computer code. The Excel file adopts the HGVS nomenclature for describing genetic variants but this is unable to encode some of the broader rules found in the catalogue such as *"any frameshift in gene X"* which is a key component of the new Loss of Function rules. The CRyPTIC project developed a grammar, GARC, specifically for this purpose[8] and we therefore translated the Excel file into the GARC format[22] that can be read and understood by `piezo`[23]. Code changes were required to incorporate the new epistatic rules introduced in WHOv2. Our catalogue maps Groups 1 & 2 onto `Resistant` (`R`) and Groups 4 & 5 onto `Susceptible` (`S`) as suggested by the WHO, with Group 3 being labelled `Unknown` (`U`).

There are several important differences in our implementation of WHOv2; our catalogue will report any mutation in a gene known to be associated with resistance but not in the catalogue as `Unknown`. If there are two or fewer reads at a genetic locus known to be associated with resistance our catalogue will call this a `Fail` since there is not enough information to know if the sample is resistant or not and it probably needs re-sequencing. Neither scenario is covered by WHOv2 and therefore both would be predicted susceptible without these rules. Finally, we call any resistance-associated variant listed in the catalogue if it is supported by three or more reads, regardless of how many other (usually wild-type) reads are also found at that locus. Minor alleles with resistance-associated variants are therefore detected; by contrast 75% of reads were required to support a ge-

netic variant for it to be considered for the WHOv2 catalogue but there is no guidance on what threshold should be applied when detecting variants listed in WHOv2.

Our translation of the WHOv2 catalogue[24], along with the variant call file produced by `clockwork` and version 3 of the reference H37Rv genome (as a GenBank file) are then ingested by `gnomonicus`[25] which produces lists of which genetic variants are detected in the sample (translated into amino acid mutations where appropriate), their predicted effects on antibiotics as specified by the catalogue and finally the resulting predicted antibiogram.

Data availability and reproducibility

Once processing was complete, two output files for each sample were downloaded using the same command line tool that handled uploading; one of these contains general information about the sample whilst the second contains more detailed information about resistance prediction. Both are structured using the javascript object notation (JSON) standard and all 5,326 files are stored in an attendant GitHub repository [*]. This repository contains a series of Juypter notebooks containing Python3 code that allows the user to discover, parse and save as data tables the relevant information in the JSON files. Other notebooks allow the user to reproduce all the analysis underlying this work, including reproducing the figures and many of the tables. The repository also lists the 2,663 samples that make up our Diverse Testset and includes bash scripts allowing all the FASTQ files to be downloaded from the ENA with the intention that people can either reproduce our results, or use the same dataset for other analyses.

[*] https://github.com/fowler-lab/validate-myco-amr

Results

Some antibiotics perform slightly better, some slightly worse if we apply a Simple implementation of the WHOv2 catalogue.

Since the WHOv2 catalogue does not provide any guidance on whether to detect minor alleles or what to do if there are insufficient reads at a loci associated with resistance, we shall start by ignoring both of these effects and simply calculate the performance on our Diverse Testset – we shall call this the Simple implementation (Fig. 1 and Table 2). The only basis for comparison we have available is the published performance of WHOv2 on its own Training Set[7] which is an imperfect comparison since the datasets are different. Comparing to the Simple implementation applied to on our Diverse Testset (Fig. 1) we find six antibiotics have sensitivities within ±1.5% (rifampicin, isoniazid, streptomycin, amikacin, kanamycin, capreomycin) that reported by the WHO, two antibiotics (pyrazinamide, ethambutol) have higher sensitivities (>1.5%) and the remaining seven drugs (bedaquline, linezolid, moxifloxacin, levofloxacin, clofazimine, delamanid, ethionamide) all record lower sensitivities (< 1.5%) than seen on the WHOv2 Training Dataset. The variation in specificity is lower: two antibiotics (ethionamide, kanamycin) record higher specificities (> 1.5%) than reported for WHOv2 with one drug (ethambutol) lower. We cannot conclude what is responsible for the differences and hence all options – from implementation errors to differences between the datasets – remain valid.

Failing samples with insufficient reads at loci associated with resistance improves performance

If there are insufficient reads at a genetic locus one does not know if the nucleotide is the same as the reference, a SNP or even a deletion, perhaps leading to a frame-shift. If the genetic locus is also one associated with resistance then the likelihood is such that even in the absence of genetic information the sample has a reasonable probability of being resistant to the relevant antibiotic. This is a qualitatively different outcome to `R`, `S` or `U` since re-sequencing is likely to result in more reads at that position, allowing a definite result to be returned. In our implementation of the WHOv2 catalogue, any genetic locus in a sample which has two or fewer (including zero) reads but could be a resistance-associated variant (RAV) is described as a `Fail` (`F`).

Adding these rules to our Simple implementation yields the Nulls implementation. In 2,663 samples 142 `Fails` are generated from 125 genetic loci which affect the antibiogram of 33 (1.2%) samples. The majority of `Fail` calls occur in the *rrs* gene (82) followed by *rpoB* (19), *pncA* (8) and *embB* (4). The resulting changes to the performance of streptomycin, rifampicin, pyrazinamide and ethambutol are small and lie within error and are therefore not shown. Note that the majority of loci in *rrs* with `Fails` do not contribute to amikacin and kanamycin resistance according to WHOv2 and therefore these antibiotics are, perhaps surprisingly, not affected.

Identifying minor alleles containing resistance-associated variants improves performance

Thus far all genetic variants have been identified using the default filter in the `clockwork` pipeline which requires 90% of reads to support a variant for it to be called. Note that this threshold is the same as used

	Simple		Nulls+Minors		Nulls+Minor (High confidence)	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Isoniazid	91.2 ±0.5	95.4 ±0.4	92.5 ±0.5	94.6 ±0.4	95.2 ±0.4	95.6 ±0.4
Rifampicin	93.7 ±0.4	96.0 ±0.3	95.3 ±0.4	94.9 ±0.3	96.7 ±0.3	95.5 ±0.3
Pyrazinamide*	81.7 ±0.7	97.4 ±0.3	85.1 ±0.7	96.7 ±0.4		
Ethambutol	85.5 ±0.9	84.9 ±0.5	86.7 ±0.8	84.1 ±0.5	88.7 ±0.8	81.4 ±0.6
Bedaquiline*	41.2 ±1.0	98.6 ±0.2	46.1 ±1.0	98.6 ±0.2		
Linezolid	22.5 ±1.8	99.9 ±0.0	28.1 ±1.9	99.8 ±0.1	42.5 ±1.9	99.8 ±0.0
Moxifloxacin	81.2 ±1.0	94.6 ±0.3	85.8 ±0.8	93.8 ±0.4	90.0 ±0.5	93.3 ±0.4
Levofloxacin	77.6 ±0.9	96.9 ±0.3	82.3 ±0.7	96.4 ±0.3	87.3 ±0.7	96.6 ±0.3
Clofazimine	7.3 ±0.6	98.2 ±0.2	8.1 ±0.6	98.0 ±0.2	4.9 ±0.7	97.6 ±0.2
Delamanid	12.3 ±1.1	99.9 ±0.0	12.3 ±1.1	99.9 ±0.0	21.7 ±2.0	100.0 ±0.0
Amikacin	71.7 ±1.0	99.6 ±0.1	74.3 ±1.0	99.4 ±0.1	83.3 ±0.8	99.3 ±0.1
Streptomycin*	79.9 ±0.8	95.6 ±0.3	80.7 ±0.8	95.3 ±0.4		
Ethionamide	71.3 ±1.0	87.5 ±0.5	72.2 ±1.0	87.0 ±0.5	75.4 ±1.0	86.0 ±0.5
Kanamycin	74.2 ±1.1	99.1 ±0.2	76.7 ±1.0	98.7 ±0.2	84.3 ±0.7	98.3 ±0.2
Capreomycin*	65.1 ±1.1	98.3 ±0.2	75.0 ±0.9	98.3 ±0.2		

Table 2: The performance of the WHOv2 catalogue on the Diverse Testset of 2,663 *M. tuberculosis* samples. Nulls+Minors includes rules that will Fail a drug if there are two or fewer reads at a locus associated with resistance but will call Resistant if there are three or more reads at the same locus. Simple excludes these rules. Nulls+Minors (High confidence) applies the same catalogue, but to a subset of the Diverse Testset with only MICs supported by two or more independent reading methods concur (high confidence). This is only possible for antibiotics with UKMYC DST data – antibiotics marked with an asterisk used simple binary MGIT DST and therefore have no value for this last category. Confidence limits were estimated by bootstrapping.

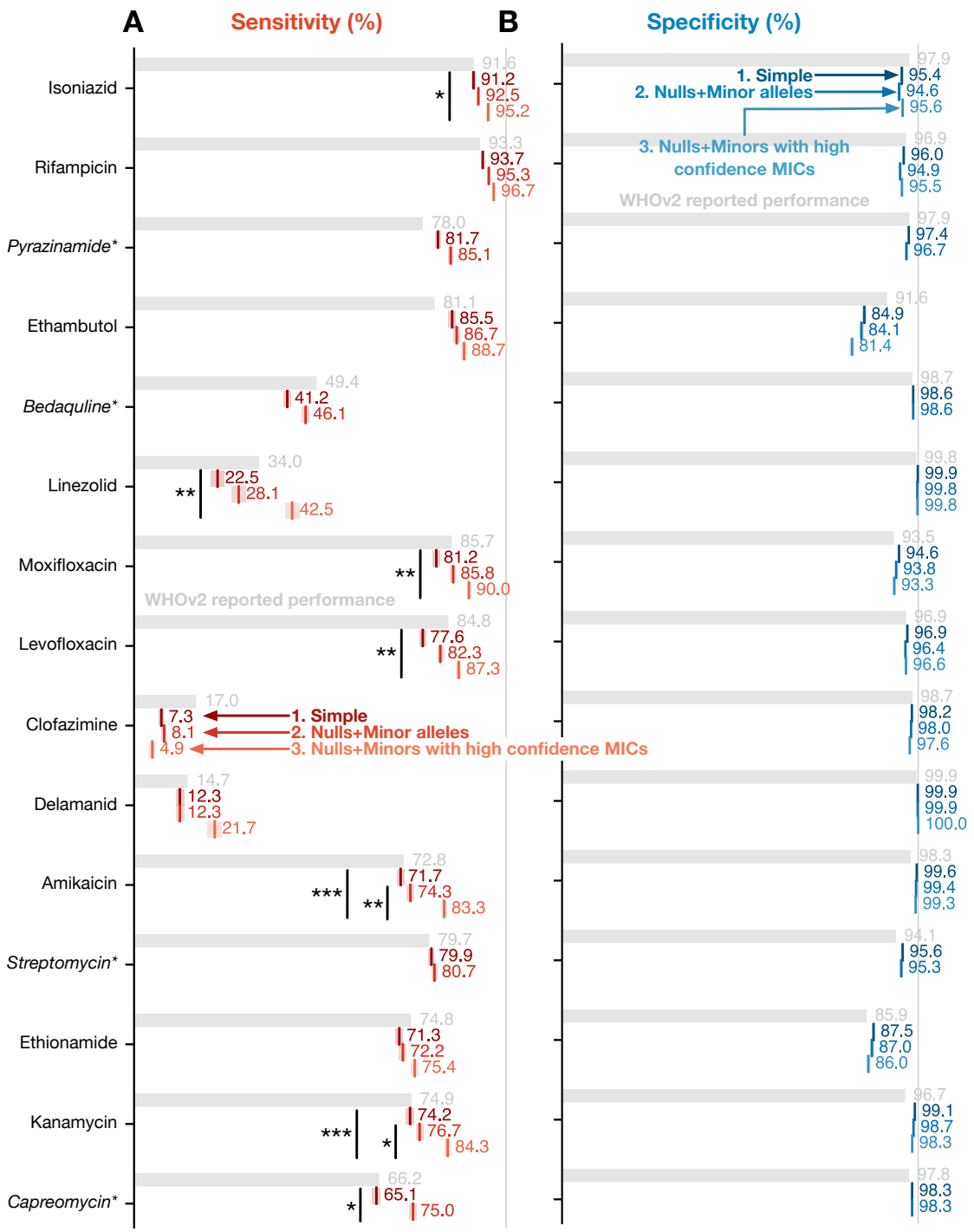

Figure 1: When tested on the 2,663 validation samples on the Mycobacterial Genetics Pipeline it produces comparable AMR prediction performance as reported for the WHOv2 catalogue. The (A) sensitivity and (B) susceptibility values for the 15 antibiotics included in the WHOv2 catalogue. The grey bars are the performance reported on the WHO Training Set and since are calculated on a different, larger dataset are not directly comparable. Four drugs (marked with an asterisk) have phenotypes from MGIT testing. Three values are reported for the Mycobacterial Genetics Pipeline: (1) a simple translational of the WHOv2 catalogue, (2) then allowing resistant minor alleles with three or more reads to be called and, lastly, (3) only comparing to high confidence MICs in addition to calling minor alleles. Since the latter can only be calculated for the antibiotics with UKMYC MICs, no values can be reported for the four drugs measured by MGIT. Mantel-haentzsel statistics were calculated for all comparisons.

in WHOv1[6] but higher than the 75% used in WHOv2[7]. Let us now add rules to our implementation of the WHOv2 catalogue such that any RAV supported by three or more short reads, regardless of how many reads are in support of different (usually wild-type) alleles are called `Resistant`: we shall call this the Nulls+Minors implementation. The sensitivity of every antibiotic increases (Fig. 1 and Table 2) with values ranging from +0.8% (streptomycin) to +9.9% (capreomycin). Eight antibiotics experienced an increase of sensitivity of 2% or more (pyrazinamide +3.4%, bedaquline +4.9%, linezolid +5.1%, moxifloxacin +4.6%, levofloxacin +4.7%, amikacin +2.6%, kanamycin +2.5% and capreomycin +9.9%). Only isoniazid experienced a reduction in specificity of more than 1% (-1.1%). Including minor alleles therefore brings the performance of our implementation of the WHOv2 catalogue on our 2,663 sample Diverse Testset closer to that reported by the WHO[7], but that is a spurious comparison because neither the dataset nor the implementation are the same.

Discrepancy analysis on UKMYC dataset

There are many reasons why we cannot perfectly predict the antibiogram for each sample; these include sample mislabelling, not all genes or genetic variants having been classified, DST measurement errors and genetics not being a perfect predictor of phenotype. We are fortunate in that the UKMYC dataset of 1,000 samples had images of all the 96-well plates from when they were read by the laboratory scientist after two weeks incubation. As mentioned in the Methods, this allowed the CRyPTIC project to produce high-confidence MICs where two or more independent methods agreed on the value[5,8], thereby reducing measurement error. We will therefore first examine the effect of only using these high-confidence MICs.

Only using high confidence UKMYC phenotypes improves performance

The sensitivities of ten of the 11 antibiotics for which we have UKMYC data increases (clofazimine being the exception) compared to the Nulls+Minors dataset, with increases ranging from +1.4% for rifampicin to +14.4% for linezolid (Fig. 1 and Table 2). Specificity is largely unchanged with only ethambutol (-2.7%) changing by more than an absolute percentage point. This indicates that measurement error is partly constraining the measured performance of the WHOv2 catalogue. Notably both rifampicin and isoniazid now achieve sensitivities and specificities above 95% which is a requirement to pass ISO 20776-2:2021[26]. The number of MICs discarded varies by antibiotic with isoniazid having 906 high-confidence measurements and clofazimine only 676 which tells us something about the relative difficulties of reading MICs for the different drugs, which could, in part be due to the differing levels of growth observed on the UKMYC 96-well plates.

Discrepants tend to have lower growth and/or MICs near the ECOFF/ECV.

Even after subsetting down to only consider high-confidence MICs, our hypothesis is that discrepant samples are more likely to have poor bacterial growth on the UKMYC plates after two weeks incubation as this would affect all measurement methods. Examining the distributions of bacterial growth (averaged from the two positive control wells) shows that this is not true (Fig. 2) as in all cases, whilst there is large variation between samples, there is no significant difference in the distributions for e.g. the True Positives (RR) and False Positives (RS).

Lastly, whilst the original measurement was an MIC, this was binarised using an ECOFF/ECV[8], and discrepants could arise if their MIC was close to the ECOFF/ECV such that it was within measurement error i.e. on another day the same sample could have grown such that the MIC was now on the other side of the threshold leading to a different binary classification. Results vary by drug (Fig. 2), but this appears at least partly true for ethambutol, moxifloxacin and delamanid. It is clear that the less bimodal the MIC distribution, the more likely this effect will lead to misclassification and thereby discrepants.

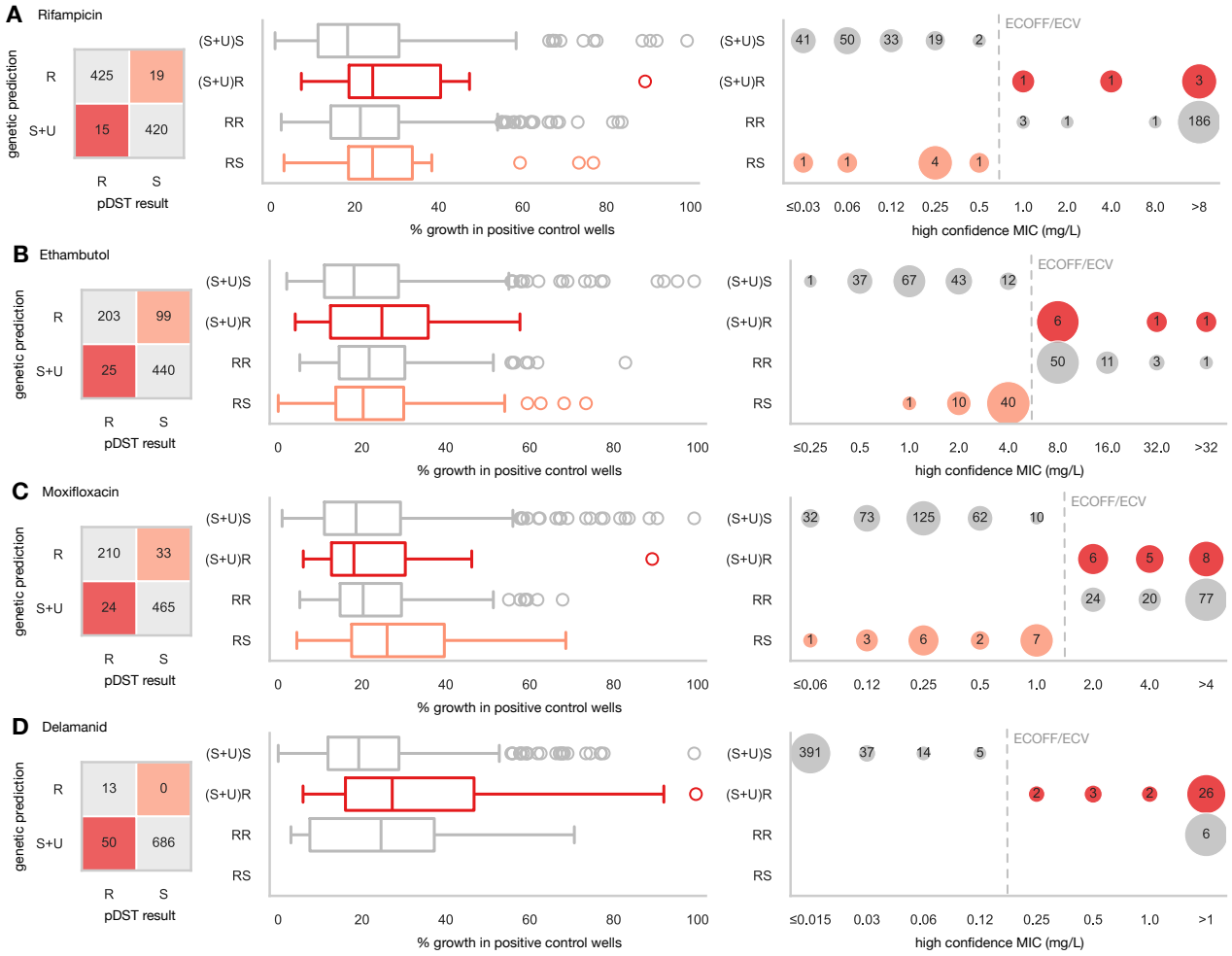

Figure 2: Examining the discrepancies for (**A**) rifampicin, (**B**) ethambutol, (**C**) moxifloxacin and (**D**) delamanid shows that it is not because these samples were growing less well on the UKMYC plates and were therefore more difficult to measure but for some drugs (e.g. ethambutol) a majority of the discrepant samples have MICs close to the ECOFF/ECV whilst for other drugs (e.g. delamanid) it appears not all the genetic basis for resistance is yet understood. All genetic predictions used the Nulls+Minor Alleles implementation of the WHOv2 catalogue and only high confidence MICs were used. MIC distributions are shown for just one of the two 96-well plate designs used.

Discussion

We have validated the antimicrobial resistance prediction functionality of our Mycobacterial genetics processing pipeline that has been implemented in a cloud-based platform. Crucially, the two main antitubercular drugs, rifampicin and isoniazid, achieve sensitivities and specificities above 95% which is required by ISO 20776-2:2021[26]. Whilst the resistance prediction applies an accepted and well-known catalogue – the second edition of the WHO catalogue of *M. tuberculosis* RAVs[7] – we have had to, in effect, translate the catalogue since the published catalogue is not a single artefact that is computer-parsable and therefore we are testing two distinct statements: (1) that our translation is correct and (2) that the reported performance for WHOv2[7] is representative of what would be expected in the clinic.

Unfortunately the samples that make up the WHOv2 Training Set are not (yet) publicly available. There are likely very good reasons for this, for example data owners may be willing to share with the WHO but not willing to make their data public, but it does prevent researchers, such as ourselves, from reproducing the analysis that led to WHOv2 which is important to find errors and gain trust. We cannot therefore formally disentangle whether there are differences due to discordant results also being found in the WHO Training datatset as well as our Diverse Testset, problems with parsing the catalogue, or assembly and variant calling differences but using our diverse dataset of 2,663 *M. tuberculosis* samples we can achieve similar performance for many of the 15 antibiotics covered by WHOv2. The specificities of bedaquline, linezolid, clofazimine and delamanid, however, remain low and variable and clearly more resistant samples are needed for the next iteration of the WHO catalogue.

Since there is no guidance on how to apply the WHOv2 to a WGS sample we have chosen to make some enhancements; the first is to explicitly flag genetic loci which are associated with resistance when there are insufficient reads to know what the allele is. These we call `Fails` and, whilst small in number (1.2% of samples had at least one `F`), they correlate with poor sequencing quality and may indicate a sample needs to be re-sequenced. The second improvement is to allow any resistance-associated variant to be identified if it is supported by at least three short reads, thereby allowing minor alleles to contribute to resistance prediction. This has been shown to boost the sensitivity for fluoroquinolones[27] and the WHOv2 report also showed how lowering the fraction of reads supporting a call often increased the sensitivity[7]. The increase in sensitivity in some cases is marked, notably for capreomycin, and usually at little or not loss in specificity. Detecting minor alleles is not yet commonplace in WGS, but happens as a matter of course in targeted NGS approaches and nucleic acid amplification tests. In a further enhancement to the WHOv2, we also classify mutations in genes known to be associated with resistance as `Unknown` even if they are not listed in the catalogue. This is possible because our grammar allows wildcards such that our catalogues can contain a single rule encoding logic like "any missense mutation in the coding region of gene X" which, in addition to a way of prioritising rules, is necessary to enable this functionality.

Reducing measurement error by only using MICs in which two or more methods concur further improves sensitivity, suggesting that our Diverse Testset (like all others) contains some measurement errors. Finally we

note that some of the remaining discrepancies are likely a natural consequence of thresholding MICs which have a level of natural variability and this is particularly pronounced for antibiotics where the MIC distribution is not bimodal, such as ethambutol and the fluoroquinolones.

To permit comparison with the stated performance of WHOv2, we have grouped samples predicted to be `Susceptible` to an antibiotic along with those which contain a genetic variant with no definite classification: we call these `U`, for `Unknown`. Aggregating `Susceptible` and `Unknown` results like this is reasonable for antibiotics where the majority of the genetic determinants have been discovered, but for drugs like bedaquiline or even pyrazinamide it breaks down because the probability of a novel mutation has a reasonable probability of being associated with resistance. For these drugs at least, it seems sensible to report an `Unknown` result for the drug, rather than assume it is `Susceptible`. Other studies have shown how, in some cases, one can take advantage of the correlations between different drugs to infer that they are likely `Susceptible`[28].

Thus far we have employed a four-valued logic: `Resistant`, `Susceptible`, `Unknown` and `Fail`. The `Unknown` value in this logic is, in fact, composed of two distinct cases. These are delineated by whether the genetic variant has been seen in sufficient clinical samples to have adequate statistical support or not. In the latter case the label is therefore transitory; collecting more samples will improve the statistics and its effect will become associated with a definite label – we suggest this retains the `Unknown` label. The former corresponds to genetic variants, such as M306V/I in *embB*, where the distribution of minimum inhibitory concentration straddles the ECOFF/ECV. Collecting more samples changes nothing and one can argue another definite value, distinct from `Resistant` or `Susceptible`, is therefore needed. Naming such a value is controversial (e.g. `Intermediate`) but its existence is unavoidable and undeniable. One can argue that it would be confusing or inappropriate to adopt a five-valued logic – we have kept here to a four-valued logic (`Resistant`, `Susceptible`, `Unknown` and `Fail`). Fortunately, the requirement to minimise the very major error rate naturally gives us the rules for these logical systems; `Resistant` $>$ `Fail` $>$ `Unknown` $>$ `Susceptible`.

Clearly it would have been preferable to be able to isolate the implementation of the WHOv2 catalogue in our Mycobacterial pipeline from the performance of the WHOv2 catalogue but this was not possible due to the lack of publicly available datasets. Also, whilst we have described our 2,663 samples as a Diverse Testset is likely not truly independent since at least the 1,000 UKMYC samples form part of the WHO Training Set. Also, despite our best efforts, there are insufficient samples resistant to bedaquline, linezolid and delamanid in the Diverse Testset; additional, targeted sample collection with better sharing is the only answer here.

Despite these shortcomings, we hope that this dataset of 2,663 *M. tuberculosis* samples will be of use to other researchers and could even form the kernel for a standard testset upon which new tools or implementations of catalogues could be tested and preliminary performance reported. All samples and their results can be downloaded using the attendant GitHub repository* which also contains code allowing all analysis to be repeated and all figures to be redrawn.

*https://github.com/fowler-lab/validate-myco-amr

Acknowledgements

We are grateful to EIT Pathogena for helpful discussions and deploying our pipeline in their cloud platform and to ORACLE Corporation for access to their cloud.

Funding

The authors would like to acknowledge funding from the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [HPRU-2012-10041], the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Ellison Institute of Technology, Oxford. 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. The findings and conclusions in this report are solely the responsibility of the authors and do not necessarily represent the official views of the NHS, the NIHR, the Department of Health or the Ellison Institute of Technology, Oxford.

Ethics

All samples (both genetics and drug susceptibility data) were downloaded from public repositories. Ethics approval was previously obtained by the CRyPTIC project[5].

Author contributions

TEAP, DWC and PWF conceived of the design. JW, CB, MB, MC, BC, AH, JK, MLA, RS, HT, TEAP, DWC, RT and PWF built and tested the Mycobacterial pipeline. SS and KG contributed to the co-ordination and development of EIT Pathogena software.

Conflict of Interest

SS and KG are employed by the Ellison Institute of Technology, Oxford. TEAP, DWC, and PWF receive consultancy fees from the Ellison Institute of Technology, Oxford.

References

[1] World Health Organization. *Global tuberculosis report*. 2024. ISBN: 978-92-4-010153-1.

[2] World Health Organization. Rapid communication: key changes to the treatment of drug-resistant tuberculosis. Technical report, World Health Organization, 2022.

[3] World Health Organization. *WHO Bacterial Priority Pathogens List: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance*. 2024. ISBN: 978-92-4-009346-1.

[4] Timothy M Walker, Ana Louisa Gibertoni Cruz, Tim E Peto, E Grace Smith, Hanif Esmail, and Derrick W Crook. Tuberculosis is changing. *Lancet Infec Disease*, 17(4):359–361, April 2017.

[5] The CRyPTIC Consortium. A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. *PLOS Biology*, 20(8):e3001721, 2022. Publisher: Public Library of Science.

[6] World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. Technical report, 2021. ISBN: 978-92-4-002817-3.

[7] World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd ed. Technical report, 2023. ISBN: 978-92-4-008241-0.

[8] The CRyPTIC Consortium. Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of M . tuberculosis. *Eur Respir J*, 60:2200239, 2022.

[9] Philip W Fowler, Ana Luíza Gibertoni Cruz, Sarah J Hoosdally, Lisa Jarrett, Emanuele Borroni, Matteo Chiacchiaretta, Priti Rathod, Sarah Lehmann, Nikolay Molodtsov, Clara Grazian, Timothy M Walker, Esther Robinson, Harald Hoffmann, Timothy E. A. Peto, Daniela Maria Cirillo, Grace E. Smith, and Derrick W Crook. Automated detection of bacterial growth on 96-well plates for high-throughput drug susceptibility testing of Mycobacterium tuberculosis. *Microbiology*, 164(12):1522–1530, 2018.

[10] Philip W Fowler, Carla Wright, Helen Spiers-bowers, Tingting Zhu, Elisabeth M L Baeten, W Hoosdally, Ana Lu, Gibertoni Cruz, Aysha Roohi, Samaneh Kouchaki, Timothy M Walker, Timothy E A Peto, Grant Miller, Chris Lintott, David Clifton, Derrick W Crook, and A Sarah Walker. A crowd of BashTheBug volunteers reproducibly and accurately measure the minimum inhibitory concentrations of 13 antitubercular drugs from photographs of 96-well broth microdilution plates. *eLife*, 11:e75046, 2022.

[11] Bede Constantinides, Martin Hunt, and Derrick W Crook. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics*, page btad728, 2023.

[12] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, 2019.

[13] Heng Li. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, 37(23):4572–4574, 2021.

[14] Martin Hunt, Phelim Bradley, Simon Grandjean Lapierre, Simon Heys, Mark Thomsit, Michael B. Hall, Kerri M. Malone, Penelope Wintringer, Timothy M. Walker, Daniela M. Cirillo, Iñaki Comas, Maha R. Farhat, Phillip Fowler, Jennifer Gardy, Nazir Ismail, Thomas A. Kohl, Vanessa Mathys, Matthias Merker, Stefan Niemann, Shaheed Vally Omar, Vitali Sintchenko, Grace Smith, Dick van Soolingen, Philip Supply, Sabira Tahseen, Mark Wilcox, Irena Arandjelovic, Tim E. A. Peto, Derrick W. Crook, and Zamin Iqbal. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. *Wellcome Open Research*, 4:191, December 2019.

[15] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, June 1998.

[16] Jean-Christophe Camus, Melinda J. Pryor, Claudine Médigue, and Stewart T. Cole. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*, 148(10):2967–2973, 2002. Publisher: Microbiology Society,.

[17] Jocelyne M. Lew, Adamandia Kapopoulou, Louis M. Jones, and Stewart T. Cole. TubercuList – 10 years after. *Tuberculosis*, 91(1):1–7, January 2011.

[18] Martin Hunt. Clockwork: Pipelines for processing bacterial sequence data (Illumina only) and variant calling, 2021.

[19] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, 2021.

[20] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232, 2012. Publisher: Nature Publishing Group.

[21] Martin Hunt, Brice Letcher, Kerri M. Malone, Giang Nguyen, Michael B. Hall, Rachel M. Colquhoun, Leandro Lima, Michael C. Schatz, Srividya Ramakrishnan, Zamin Iqbal, and CRyPTIC consortium. Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *Genome Biology*, 23(1):147, 2022.

[22] Jeremy Westhead and Philip W Fowler. Conversion of the WHO TB catalogue to GARC, August 2024. `https://github.com/fowler-lab/who_catalogue_conversion`.

[23] Jeremy Westhead and Philip William Fowler. piezo: predicting the effect of a genetic mutation on an antibiotic, 2021. https://github.com/oxfordmmm/piezo.

[24] Jeremy Westhead and Philip William Fowler. Tuberculosis AMR catalogues in a standard grammar, 2021. https://github.com/oxfordmmm/tuberculosis_amr_catalogues.

[25] Jeremy Westhead and Philip W Fowler. gnomonicus, 2023. https://github.com/oxfordmmm/gnomonicus.

[26] International Organization for Standardization. Clinical laboratory testing and in vitro diagnostic test systems - Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices. Technical report, 2021.

[27] Alice E Brankin and Philip W Fowler. Inclusion of minor alleles improves catalogue-based prediction of fluoroquinolone resistance in Mycobacterium tuberculosis. *JAC-Antimicrobial Resistance*, 5(2), March 2023. Publisher: Oxford Academic ISBN: 039/7103433.

[28] The CRyPTIC Consortium and 100000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *New Eng J Med*, 379(15):1403–1415, October 2018.