Drug resistance prediction for *Mycobacterium tuberculosis* with reference graphs

1.1 Author names

Michael B. Hall[1,2,*] (ORCiD: 0000-0003-3683-6208), Leandro Lima[1] (ORCiD: 0000-0001-8976-2762), Lachlan J. M. Coin[2] (ORCiD: 0000-0002-4300-455X), Zamin Iqbal[1,*] (ORCiD: 0000-0001-8466-7547)

1.2 Affiliation(s)

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, United Kingdom
[2]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Australia
[*]Corresponding authors

1.3 Corresponding author and email address

Michael B. Hall: michael.hall2 [at] unimelb.edu.au
Zamin Iqbal: zi [at] ebi.ac.uk

1.4 Keywords

genome graphs, reference graphs, drug resistance prediction, *Mycobacterium tuberculosis*, benchmark, software

2. Abstract

The dominant paradigm for analysing genetic variation relies on a central idea: all genomes in a species can be described as minor differences from a single reference genome. However, this approach can be problematic or inadequate for bacteria, where there can be significant sequence divergence within a species.

Reference graphs are an emerging solution to the reference bias issues implicit in the "single-reference" model. Such a graph represents variation at multiple scales within a population – e.g., nucleotide- and locus-level.

The genetic causes of drug resistance in bacteria have proven comparatively easy to decode compared with studies of human diseases. For example, it is possible to predict resistance to numerous anti-tuberculosis drugs by simply testing for the presence of a list of single nucleotide polymorphisms and insertion/deletions, commonly referred to as a catalogue.

We developed DrPRG (Drug resistance Prediction with Reference Graphs) using the bacterial reference graph method Pandora. First, we outline the construction of a *Mycobacterium tuberculosis* drug resistance reference graph, a process that can be replicated for other species. The graph is built from a global dataset of isolates with varying drug susceptibility profiles, thus capturing common and rare resistance- and susceptible-associated haplotypes. We benchmark DrPRG against the existing graph-based tool Mykrobe and the pileup-based approach of TBProfiler using 44,709 and 138 publicly available Illumina and Nanopore datasets with associated phenotypes. We find DrPRG has significantly improved sensitivity and specificity for some drugs compared to these tools, with no significant decreases. It uses significantly less computational memory than both tools, and provides significantly faster runtimes, except when runtime is compared to Mykrobe on Illumina data.

42   We discover and discuss novel insights into resistance-conferring variation for *M.*
43   *tuberculosis* - including deletion of genes *katG* and *pncA* – and suggest mutations that may
44   warrant reclassification as associated with resistance.
45

3. Impact statement

47   *Mycobacterium tuberculosis* is the bacterium responsible for tuberculosis (TB). TB is one of
48   the leading causes of death worldwide; before the coronavirus pandemic it was the leading
49   cause of death from a single pathogen. Drug-resistant TB incidence has recently increased,
50   making the detection of resistance even more vital. In this study, we develop a new software
51   tool to predict drug resistance from whole-genome sequence data of the pathogen using new
52   reference graph models to represent a reference genome. We evaluate it on *M. tuberculosis*
53   against existing tools for resistance prediction and show improved performance. Using our
54   method, we discover new resistance-associated variations and discuss reclassification of a
55   selection of existing mutations. As such, this work contributes to TB drug resistance
56   diagnostic efforts. In addition, the method could be applied to any bacterial species, so is of
57   interest to anyone working on antimicrobial resistance.

4. Data summary

59   **The authors confirm all supporting data, code and protocols have been provided within**
60   **the article or through supplementary data files.**
61   The software method presented in this work, DrPRG, is freely available from GitHub under
62   an MIT license at https://github.com/mbhall88/drprg. We used commit 9492f25 for all results
63   via a Singularity[1] container from the URI
64   `docker://quay.io/mbhall88/drprg:9492f25`.
65   All code used to generate results for this study are available on GitHub at
66   https://github.com/mbhall88/drprg-paper. All data used in this work are freely available from
67   the SRA/ENA/DRA and a copy of the datasheet with all associated phenotype information
68   can be downloaded from the archived repository at https://doi.org/10.5281/zenodo.7819984
69   or found in the previously mentioned GitHub repository.
70   The *Mycobacterium tuberculosis* index used in this work is available to download through
71   DrPRG via the command `drprg index --download mtb@20230308` or from
72   GitHub at https://github.com/mbhall88/drprg-index.

5. Introduction

74   Human industrialisation of antibiotic production and use over the last 100 years has led to a
75   global rise in prevalence of antibiotic resistant bacterial strains. The phenomenon
76   was even observed within patients in the first clinical trial of streptomycin as a drug for
77   tuberculosis (TB) in 1948[2], and indeed as every new drug class has been introduced, so has
78   resistance followed. Resistance mechanisms are varied, and can be caused by point mutations
79   at key loci (e.g., binding sites of drugs[3,4]), frame-shifts rendering a gene non-functional[5],
80   horizontal acquisition of new functionality via a new gene[6], or by up-regulation of efflux
81   pumps to reduce the drug concentration within the cell[7].
82
83   Phenotypic and genotypic methods for detecting reduced susceptibility to drugs play
84   complementary roles in clinical microbiology. Carefully defined phenotypic assays are used
85   to give (semi)quantitative or binary measures of drug susceptibility; these have the benefit of

86   being experimental, quantitative measurements, and are able to detect resistance caused by
87   hitherto unknown mechanisms. Prediction of drug resistance from genomic data has different
88   advantages. Detection of a single nucleotide polymorphism (SNP) is arguably more
89   consistent than a phenotypic assay, as it is not affected by whether the resistance it causes is
90   near some threshold defining a resistant/susceptible boundary. Additionally, combining
91   sequence datasets from different labs is more reliable than combining different phenotypic
92   datasets, and using sequence data allows one to detect informative genetic changes (e.g., a
93   tandem expansion of a single gene to form an array, thus increasing dosage). More subtly,
94   defining the cut-off to separate resistant from susceptible is only simple when the minimum
95   inhibitory concentration distribution is a simple bimodal distribution; in reality it is
96   sometimes a convolution of multiple distributions caused by different mutations, and genetic
97   data is sometimes needed to deconvolve the data and choose a threshold[8,9].
98
99   The key requirement for a genomic predictor is to have an encodable understanding of the
100   genotype-to-phenotype map. Research has focussed on clinically important pathogens,
101   primarily *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Pseudomonas*
102   *aeruginosa* and *Mycobacterium tuberculosis* (MTB). The challenges differ across species;
103   almost all bacterial species are extremely diverse, with non-trivial pan-genomes and
104   considerable horizontal gene transfer causing transmission of resistance genes[10]. In these
105   cases, species are so diverse that detection of chromosomal SNPs is affected heavily by
106   reference bias[11]. Furthermore, there is an appreciable proportion of resistance which is not
107   currently explainable through known SNPs or genes [12–14]. At the other extreme, MTB has
108   almost no accessory genome, and no recombination or plasmids[15]. Resistance appears to be
109   caused entirely by mutations, indels, and rare structural variants, and simple sets of rules ("if
110   any of these mutations are present, or any of these genes inactivated, the sample is resistant")
111   work well for most drugs[16]. MTB has an exceptionally slow growth rate, meaning culture-
112   based drug susceptibility testing (DST) is slow (2-4 weeks depending on media), and
113   therefore sequencing is faster[17]. As part of the end TB strategy, the WHO strives towards
114   universal access to DST[18], defining Target Product Profiles for molecular
115   diagnostics[19,20] and publishing a catalogue of high-confidence resistance mutations
116   intended to provide a basis for commercial diagnostics and future research[16]. There was a
117   strong community-wide desire to integrate this catalogue into software for genotypic
118   resistance prediction, although independent benchmarking confirmed there was still need for
119   improvement[12]. Hence, there is a continuing need to improve the understanding of the
120   genetic basis of resistance and integrate it into software for genotypic DST.
121
122   In this paper we develop and evaluate a new software tool for genotypic DST for MTB, built
123   on a generic framework that can be used for any bacteria. Several tools have been developed
124   previously[21–25]. Of these, only Mykrobe and TBProfiler work on Illumina and Nanopore
125   data, and both have been heavily evaluated previously[22,23,26,27] - so we benchmark
126   against these. Mykrobe uses de Bruijn graphs to encode known resistance alleles and thereby
127   achieves high accuracy even on indel calls with Nanopore data[27]. However it is unable to
128   detect novel alleles in known resistance genes, nor to detect gene truncation or deletion,
129   which would be desirable. TBProfiler is based on mapping and variant calling (by default
130   using Freebayes[28]), and detects gene deletions using Delly[29].
131
132   Our new software, called DrPRG (Drug resistance Prediction with Reference Graphs), builds
133   on newer pan-genome technology than Mykrobe[11] using an independent graph for each
134   gene in the catalogue, which makes it easier to go back-and-forth between VCF and the

135 graph. To build an index, it takes as input a catalogue of resistant variants (a simple 4-column
136 TSV file), a file specifying expert rules (e.g. any missense variant between codons X and Y in
137 gene Z causes resistance to drug W), and a VCF of population variation in the genes of
138 interest. This allows it to easily incorporate the current WHO-endorsed catalogue[16], which
139 is conservative, and for the user to update the catalogue or rules with minimal effort. Finally,
140 to provide resistance predictions, it takes a prebuilt index (an MTB one is currently provided)
141 and sequencing reads (FASTQ).
142
143 We describe the DrPRG method, and to evaluate it, gather the largest MTB dataset of
144 sequencing data with associated phenotype information and reveal novel insights into
145 resistance-determining mutations for this species.

6. Methods

147 DrPRG is a command-line software tool implemented in the Rust programming language.
148 There are two main subcommands: `build` for building a reference graph and associated
149 index files, and `predict` for producing genotypic resistance predictions from sequencing
150 reads and an index (from `build`).

6.1 Constructing a resistance-specific reference graph and index

152 The `build` subcommand of DrPRG requires a Variant Call Format (VCF) file of variants
153 from which to build a reference graph, a catalogue of mutations that confer resistance or
154 susceptibility for one or more drugs, and an annotation (GFF) and FASTA file of the
155 reference genome.
156 For this work, we used the reference and annotation for the MTB strain H37Rv (accession
157 NC_000962.3) and the default mutation catalogue from Mykrobe (v0.12.1)[12,26].
158 To ensure the reference graph is not biased towards a particular lineage or susceptibility
159 profile, we selected samples from a VCF of 15,211 global MTB samples[30]. We randomly
160 chose 20 samples from each lineage 1 through 4, as well as 20 samples from all other
161 lineages combined. In addition, we included 17 clinical samples representing MTB global
162 diversity (lineages 1-6)[31,32] to give a total of 117 samples. In the development phase of
163 DrPRG we also found it necessary to add some common mutations not present in these 177
164 samples; as such, we added 48 mutations to the global VCF (these mutations are listed in
165 archived repository – see Data summary). We did not add all catalogue mutations as there is a
166 saturation point for mutation addition to a reference graph, and beyond this point,
167 performance begins to decay[33].
168 The `build` subcommand turns this VCF into a reference graph by extracting a consensus
169 sequence for each gene and sample. We use just those genes that occur in the mutation
170 catalogue and include 100 bases flanking the gene. A multiple sequence alignment is
171 constructed for each gene from these consensus sequences with MAFFT (v7.505)[34,35] and
172 then a reference graph is constructed from these alignments with `make_prg` (v0.4.0)[11].
173 The final reference graph is then indexed with `pandora`[11].

6.2 Genotypic resistance prediction

175 Genotypic resistance prediction of a sample is performed by the `predict` subcommand of
176 DrPRG. It takes an index produced by the `build` command (see Constructing a resistance-
177 specific reference graph and index) and sequencing reads – Illumina or Nanopore are
178 accepted. To generate predictions, DrPRG discovers novel variants (`pandora`), adds these to
179 the reference graph (`make_prg` and MAFFT), and then genotypes the sample with respect

180 to this updated graph (pandora). The genotyped VCF is filtered such that we ignore any
181 variant with less than 3 reads supporting it and require a minimum of 1% read depth on each
182 strand. Next, each variant is compared to the catalogue. If an alternate allele has been called
183 that corresponds with a catalogue variant, resistance ('R') is noted for the drug(s) associated
184 with that mutation. If a variant in the VCF matches a catalogue mutation, but the genotype is
185 null ('.'), we mark that mutation, and its associated drug(s), as failed ('F'). Where an alternate
186 allele call does not match a mutation in the catalogue, we produce an unknown ('U')
187 prediction for the drug(s) that have a known resistance-conferring mutation in the relevant
188 gene.
189 DrPRG also has the capacity to detect minor alleles and call minor resistance ('r') or minor
190 unknown ('u') in such cases. Minor alleles are called when a variant (that has passed the
191 above filtering) is genotyped as being the susceptible (reference) allele, but there is also read
192 depth on the resistant (alternate) allele above a given minor allele frequency parameter (--
193 maf; default is 0.1 for Illumina data). Minor allele calling is turned off by default for
194 Nanopore data as we found it led to a drastic increase in the number of false positive calls
195 (this is also the case for Mykrobe and TBProfiler).
196 When building the index for DrPRG and making predictions, we also accept a file of "expert
197 rules" for calling variants of a certain class. A rule is associated with a gene, an optional
198 position range, a variant type, and the drug(s) that rule confers resistance to. Currently
199 supported variant types are missense, nonsense, frameshift, and gene absence.
200 The output of running predict is a VCF file of all variants in the graph and a JSON file of
201 resistance predictions for each drug in the index, along with the mutation(s) supporting that
202 prediction and a unique identifier to find that variant in the VCF file (see Supplementary
203 Section S1 for an example). The reference graph gene presence/absence (as determined by
204 pandora) is also listed in the JSON file.

6.3   Benchmark

206 We compare the performance of DrPRG against Mykrobe (v0.12.1)[26] and TBProfiler
207 (v4.3.0)[22] for MTB drug resistance prediction. Mykrobe is effectively a predecessor of
208 DrPRG; it uses genome graphs, in the form of de Bruijn graphs, to construct a graph of all
209 mutations in a catalogue and then genotypes the reads against this graph. TBProfiler is a more
210 traditional approach which aligns reads to a single reference genome and calls variants from
211 that alignment via a pileup.
212 A key part of such a benchmark is the catalogue of mutations, as this generally accounts for
213 the majority of differences between tools[26]. As such, we use the same catalogue for all
214 tools to ensure any differences are method-related - not catalogue disparities. The catalogue
215 we chose is the default one provided in Mykrobe[12]. It is a combination of the catalogue
216 described in Hunt *et al.* [26] and the category 1 and 2 mutation and expert rules from the
217 2021 WHO catalogue[16]. This catalogue contains mutations for 14 drugs: isoniazid,
218 rifampicin, ethambutol, pyrazinamide, levofloxacin, moxifloxacin, ofloxacin, amikacin,
219 capreomycin, kanamycin, streptomycin, ethionamide, linezolid, and delamanid.
220 We used Mykrobe and TBProfiler with default parameters, except for a parameter in each
221 indicating the sequencing technology of the data as Illumina or Nanopore and the TBProfiler
222 option to not trim data (as we do this in Quality control).
223 We compare the prediction performance of each program using sensitivity and specificity. To
224 calculate these metrics, we consider a true positive (TP) and true negative (TN) as a case
225 where a program calls resistance and susceptible, respectively, and the phenotype agrees; a
226 false positive (FP) as a resistant call by a program but a susceptible phenotype, with false

227 negatives (FN) being the inverse of FP. We only present results for drugs in the catalogue and
228 where at least 10 samples had phenotypic data available.
229 To benchmark the runtime and memory usage of each tool, we used the Snakemake
230 benchmark feature within our analysis pipeline[36].

6.4 Datasets

232 We gathered various MTB datasets where whole-genome sequencing data (Nanopore or
233 Illumina) were available from public repositories (ENA/SRA/DRA) and associated
234 phenotypes were accessible for at least one drug present in our catalogue[16,27,37–49].
235 All data was downloaded with `fastq-dl` (v1.1.1; https://github.com/rpetit3/fastq-dl).

6.5 Quality control

237 All downloaded Nanopore fastq files had adapters trimmed with `porechop` (v0.2.4;
238 https://github.com/rrwick/Porechop), with the option to discard any reads with an adapter in
239 the middle, and any reads with an average quality score below 7 were removed with `nanoq`
240 (v0.9.0)[50]. Illumina reads were pre-processed with `fastp` (v0.23.2)[51] to remove adapter
241 sequences, trim low quality bases from the ends of the reads, and remove duplicate reads and
242 reads shorter than 30bp.
243 Sequencing reads were decontaminated as described in Hall *et al.*[27] and Walker *et al.*[16].
244 Briefly, sequenced reads were mapped to a database of common sputum contaminants and the
245 MTB reference genome (H37Rv; accession NC_000962.3)[52] keeping only those reads
246 where the best mapping was to H37Rv.
247 After quality control, we removed any sample with average read depth less than 15, or where
248 more than 5% of the reads mapped to contaminants.
249 Lineage information was extracted from the TBProfiler results (see Benchmark).

6.6 Statistical Analysis

251 We used a Wilcoxon rank-sum paired data test from the Python library SciPy[53] to test for
252 significant differences in runtime and memory usage between the three prediction tools.
253 The sensitivity and specificity confidence intervals were calculated with a Wilson's score
254 interval with a coverage probability of 95%.

7. Results

256 To benchmark DrPRG, Mykrobe, and TBProfiler, we gathered an Illumina dataset of 45,702
257 MTB samples with a phenotype for at least one drug. After quality control (see Quality
258 control), this number reduced to 44,709. In addition, we gathered 142 Nanopore samples, of
259 which 138 passed quality control. In Figure 1 we show all available drug phenotypes for
260 those interested in the dataset, yet our catalogue does not offer predictions for all drugs listed
261 (see Benchmark). Lineage counts for all samples that passed quality control and have a
262 single, major lineage call can be found in Table 1.
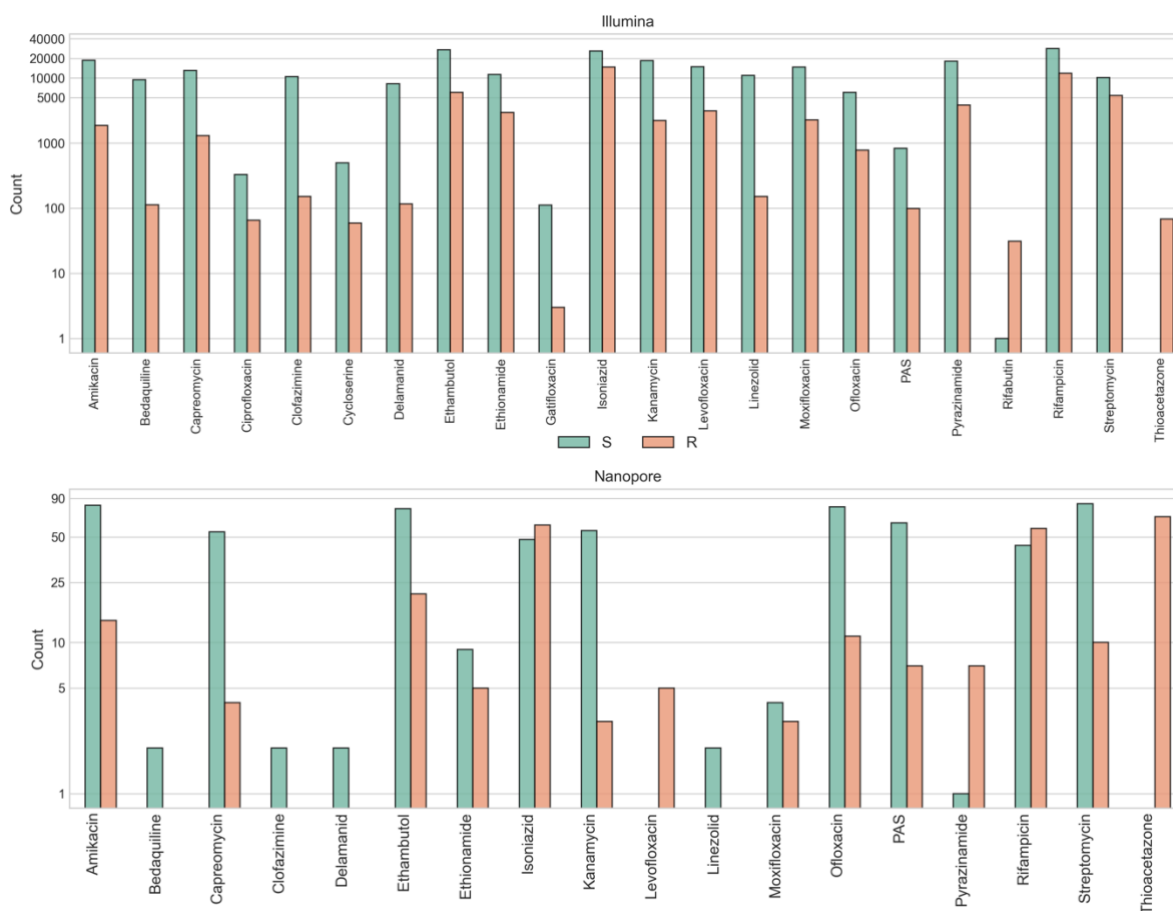
Figure 1: Drug phenotype counts for Illumina (upper) and Nanopore (lower) datasets. Bars are stratified and coloured by whether the phenotype is resistant (R; orange) or susceptible (S; green). Note, the y-axis is log-scaled. PAS=para-aminosalicylic acid

Table 1: Lineage counts from the Illumina and Nanopore datasets, covering main lineages 1-9 (L1-L9) and the three livestock-associated lineages (La1-La3) as defined in [54]

| Lineage	Illumina	Nanopore
La1	239	0
La2	7	0
La3	71	0
L1	3907	32
L2	12870	38
L3	5803	9
L4	20731	59
L5	63	0
L6	78	0
L7	3	0
L9	1	0

7.1 Sensitivity and specificity performance

We present the sensitivity and specificity results for Illumina data in Figure 2 and Suppl. Table S1 and the Nanopore data in Figure 3 and Suppl. Table S2.

273 When comparing DrPRG's performance to that of Mykrobe and TBProfiler, we look for
274 instances where the confidence intervals do not overlap; indicating a significant difference.
275 With Illumina data (Figure 2 and Suppl. Table S1), DrPRG achieves significantly greater
276 sensitivity than Mykrobe for rifampicin (96.4% [96.0-96.7] vs. 95.6% [95.2-95.9]),
277 streptomycin (85.3% [84.4-86.3] vs. 83.1% [82.1-84.1]), amikacin (85.6% [83.9-87.1] vs.
278 80.8% [78.9-82.5]), capreomycin (77.5% [75.2-79.7] vs. 71.8% [69.3-74.1]), kanamycin
279 (83.7% [82.1-85.2] vs. 79.9% [78.2-81.5]), and ethionamide (75.2% [73.7-76.8] vs. 71.4%
280 [69.7-73.0]), with no significant difference for all other drugs. In terms of sensitivity, there
281 was no significant difference between DrPRG and TBProfiler except for ethionamide, where
282 DrPRG was significantly more sensitive (75.2% [73.7-76.8] vs. 71.5% [69.8-73.1]). For
283 specificity, there was no significant difference between the tools except that DrPRG and
284 Mykrobe were significantly better than TBProfiler for rifampicin (97.8% [97.6-98.0] vs.
285 97.2% [97.0-97.4]). There was no significant difference in sensitivity or specificity for any
286 drug with Nanopore data.
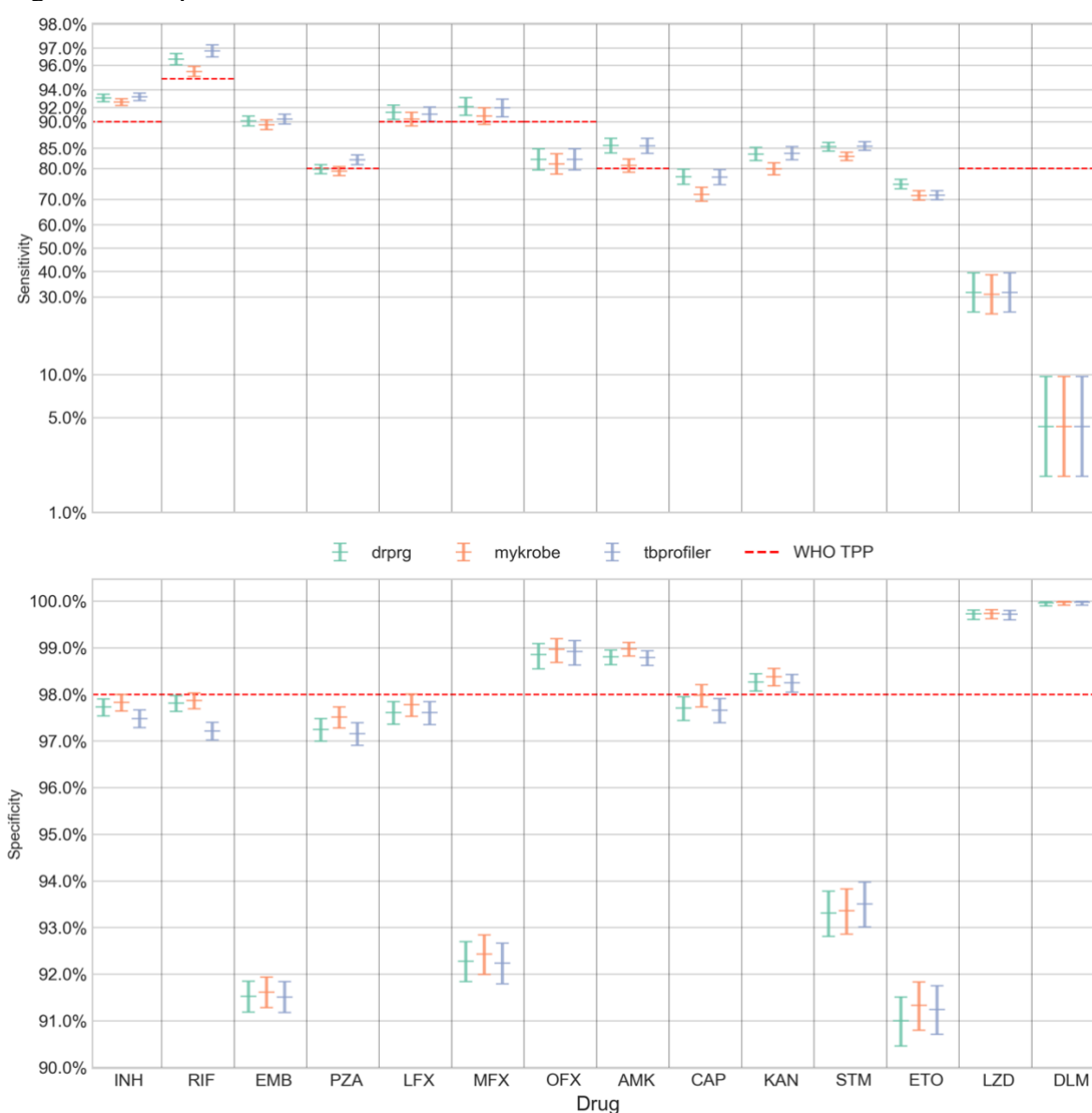
287
288 Figure 2: Sensitivity (upper panel; y-axis) and specificity (lower panel; y-axis) of resistance predictions for different
289 drugs (x-axis) from Illumina data. Error bars are coloured by prediction tool. The central horizontal line in each
290 error bar is the sensitivity/specificity and the error bars represent the 95% confidence interval. Note, the sensitivity
291 panel's y-axis is logit-scaled. This scale is similar to a log scale close to zero and to one (100%), and almost linear

292   around 0.5 (50%). The red dashed line in each panel represents the minimal standard WHO target product profile
293   (TPP; where available) for next-generation drug susceptibility testing for sensitivity and specificity. INH=isoniazid,
294   RIF=rifampicin, EMB=ethambutol, PZA=pyrazinamide, LFX=levofloxacin, MFX=moxifloxacin, OFX=ofloxacin,
295   AMK=amikacin, CAP=capreomycin, KAN=kanamycin, STM=streptomycin, ETO=ethionamide, LZD=linezolid,
296   DLM=delamanid.
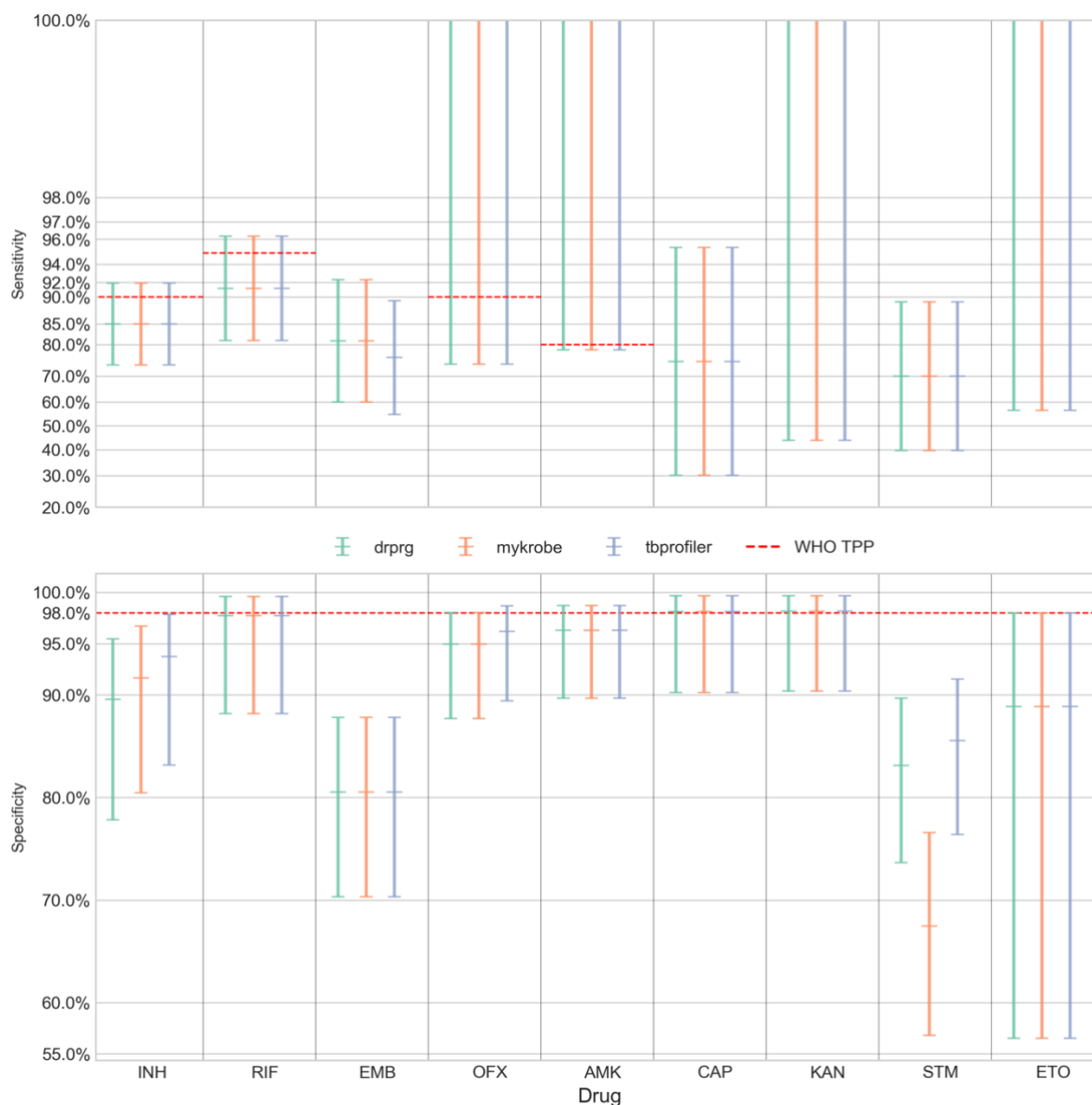
297
298   Figure 3: Sensitivity (upper panel; y-axis) and specificity (lower panel; y-axis) of resistance predictions for different
299   drugs (x-axis) from Nanopore data. Error bars are coloured by prediction tool. The central horizontal line in each
300   error bar is the sensitivity/specificity and the error bars represent the 95% confidence interval. Note, the sensitivity
301   panel's y-axis is logit-scaled. This scale is similar to a log scale close to zero and to one (100%), and almost linear
302   around 0.5 (50%). The red dashed line in each panel represents the minimal standard WHO target product profile
303   (TPP; where available) for next-generation drug susceptibility testing for sensitivity and specificity. INH=isoniazid,
304   RIF=rifampicin, EMB=ethambutol, OFX=ofloxacin, AMK=amikacin, CAP=capreomycin, KAN=kanamycin,
305   STM=streptomycin, ETO=ethionamide.

306   In both figures, we show the minimal requirements from the WHO target product profiles for
307   sensitivity and specificity of genotypic drug susceptibility testing[19] as red dashed lines.
308   Note, a sensitivity target is not specified by the WHO for ethambutol (EMB), capreomycin
309   (CAP), kanamycin (KAN), streptomycin (STM), or ethionamide (ETO). For Illumina data, all
310   tools' predictions for rifampicin, isoniazid, levofloxacin, moxifloxacin and amikacin are

311 above the sensitivity minimal requirement target. TBProfiler also exceeds the target for
312 pyrazinamide, which DrPRG misses by 0.2%. No drug's sensitivity target was achieved with
313 Nanopore data. For specificity, the tools are all very similar and either exceed or fall below
314 the threshold together (see Figure 2). The target of >98% is met by all tools on Illumina data
315 only for ofloxacin, amikacin, linezolid, and delamanid. Mykrobe also exceeds the target for
316 capreomycin. As such, amikacin is the only drug where both sensitivity and specificity
317 performance exceed the minimal requirement of the WHO target product profiles. Only
318 capreomycin and kanamycin specificity targets are exceeded (by all tools) with Nanopore
319 data.
320 However, for Illumina data, we did find that likely sample-swaps or phenotype instability[55]
321 could lead to some drugs being on the threshold of the WHO target product profiles. If we
322 excluded samples where all three tools make a FP call for the strong isoniazid and rifampicin
323 resistance-conferring mutations *katG* S315T (*n*=152) and *rpoB* S450L (*n*=119) [16]
324 respectively, all three tools would exceed the isoniazid specificity target of 98% - thus
325 meeting both sensitivity and specificity targets for isoniazid. In addition, DrPRG and
326 Mykrobe would meet the rifampicin specificity target of 98% – leading to both targets being
327 met for rifampicin for these two tools. As previously reported [55,56], we also found a lot of
328 instability in the ethambutol result caused by *embB* mutations M306I (*n*=827) and M306V
329 (*n*=519) being called for phenotypically susceptible samples (FP) by all three tools. Other
330 frequent consensus FP calls included: *fabG1* c-15t, which is associated with ethionamide
331 (*n*=441) and isoniazid (*n*=241) resistance; *rrs* a1401g, which is associated with resistance to
332 capreomycin (*n*=241), amikacin (*n*=70), and kanamycin (*n*=48). In addition there were
333 common false positives from *gyrA* mutations A90V and D94G, which are associated with
334 resistance to the fluoroquinolones levofloxacin (*n*=108 and *n*=70, respectively), moxifloxacin
335 (*n*=419 and *n*=349) and ofloxacin (*n*=19 and *n*=17), and are known to cause heteroresistance
336 and minimum inhibitory concentrations (MIC) close to the critical concentration
337 threshold[57–59].

7.2 Evaluation of potential additions to the WHO catalogue

339 False negatives are much harder to investigate as it is not known which mutation(s) were
340 missed as they are presumably not in the catalogue if all tools failed to make a call. However,
341 looking through those FNs where DrPRG makes an "unknown" resistance call, we note some
342 potential mutations that may need reclassification or inclusion in the WHO catalogue. For
343 delamanid FNs, we found five different nonsense mutations in the *ddn* gene in seven samples
344 – W20* (*n*=2), W27* (*n*=1), Q58* (*n*=1), W88* (*n*=2), and W139* (*n*=1) – none of which
345 occurred in susceptible samples. We also found 13 pyrazinamide FN cases with a nonstop
346 (stop-loss) mutation in *pncA* – this mutation type was also seen in two susceptible samples.
347 Another *pncA* mutation, T100P, was also observed in 10 pyrazinamide FN samples and no
348 susceptible samples. T100P only appears once in the WHO catalogue data ("solo" in a
349 resistant sample). As such, it was given a grading of uncertain significance. As our dataset
350 includes those samples in the WHO catalogue dataset, this means an additional nine isolates
351 have been found with this mutation - indicating this may warrant an upgrade to 'associated
352 with resistance'. We found an interesting case of allele combinations, where nine ethambutol
353 FN samples have the same two *embA* mutation c-12a and c-11a and *embB* mutation P397T -
354 this combination is only seen in two susceptible samples. Interestingly, *embB* P397T and
355 *embA* c-12a don't appear in the WHO catalogue, but have been described as causing
356 resistance previously[60]. Three *katG* mutations were also detected in isoniazid FN cases.
357 First, G279D occurs in eight missed resistance samples and no susceptible cases. This
358 mutation is graded as 'uncertain significance' in the WHO catalogue and was seen solo in

359 four resistant samples in that data. Singh *et al.* performed a protein structural analysis caused
360 by this mutation and found it produced "an undesirable effect on the functionality of the
361 protein"[61]. Second, G699E occurs in eight FN samples and no susceptible cases, but has a
362 WHO grading of 'uncertain significance' based on six resistant isolates; thus, we add two
363 extra samples to that count. And third, N138H occurs in 14 FN samples and one susceptible.
364 In seven of these cases, it co-occurs with *ahpC* mutations t-75g ($n=2$) and t-76a ($n=5$). This
365 mutation occurs in only three resistant isolates in the WHO catalogue dataset, giving it an
366 uncertain significance, but we add a further 11 cases. This mutation has been found to cause a
367 high isoniazid MIC and be associated with resistance[62,63].

7.3 Detection of large deletions

369 There are expert rules in the WHO catalogue which treat gene loss-of-function (any
370 frameshift or nonsense mutation) in *katG*, *ethA*, *gid*, and *pncA* as causing resistance for
371 isoniazid, ethionamide, streptomycin, and pyrazinamide, respectively[16]. Although
372 examples of resistance caused by gene deletion are rare[64–68], with a dataset of this size
373 ($n=44,709$), we can both evaluate these rules, and compare the detection power of DrPRG
374 and TBProfiler for identifying gene deletions (Mykrobe does not, although in principle it
375 could). In total we found 206 samples where DrPRG and/or TBProfiler identified deletions of
376 *ethA*, *katG*, or *pncA*. Although many of these isolates did not have phenotype information for
377 the associated drug ($n=100$), the results are nevertheless striking (Figure 4). Given the low
378 false-positive rate of `pandora` for gene absence detection[11], these no-phenotype samples
379 provide insight into how often gene deletions are occurring in clinical samples.
380 Of the 34 isolates where *katG* was identified as being absent, and an isoniazid phenotype was
381 available, all 34 were phenotypically resistant. DrPRG detected all 34 (100% sensitivity) and
382 TBProfiler identified 26 (76.5% sensitivity). Deletions of *pncA* were detected in 56 isolates,
383 of which 49 were phenotypically resistant. DrPRG detected 47 (95.9% sensitivity) and
384 TBProfiler detected 46 (93.9% sensitivity). Lastly, *ethA* was found to be missing in 16
385 samples with an ethionamide phenotype, of which 10 were phenotypically resistant. Both
386 DrPRG and TBProfiler correctly predicted all 10 (100% sensitivity). No *gid* deletions were
387 discovered. We note that the TP calls made by Mykrobe were due to it detecting large
388 deletions that are present in the catalogue, which is understandable given the whole gene is
389 deleted.
390 We conclude that DrPRG is slightly more sensitive at detecting large deletions than
391 TBProfiler (and Mykrobe) for *katG*, and equivalent for *pncA* and *ethA*. However we note that
392 the WHO expert rule which predicts resistance to isolates missing specific genes appears
393 more accurate for *katG* (100% of isolates missing the gene are resistant) than for *pncA* (87%
394 resistant) and *ethA* (62.5% resistant).
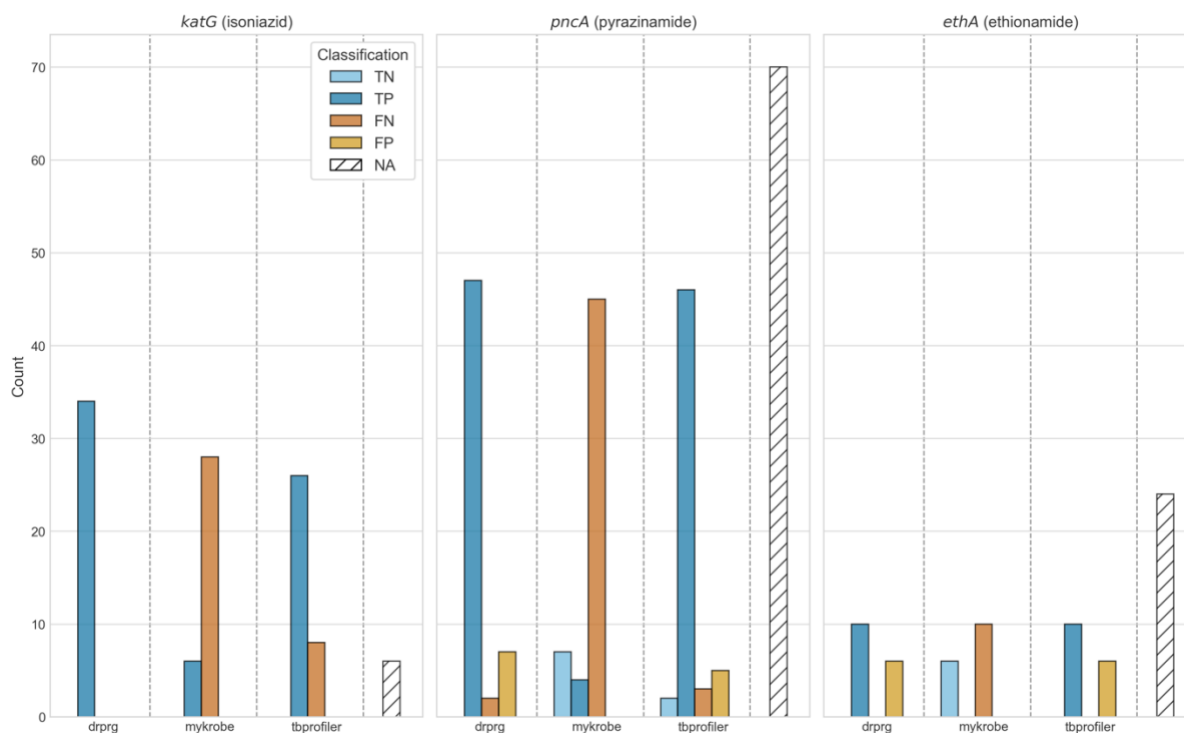
Figure 4: Impact of gene deletion on resistance classification. The title of each subplot indicates the gene and drug it effects. Bars are coloured by their classification and stratified by tool. Count (y-axis) indicates the number of gene deletions for that category. The NA bar (white with diagonal lines) indicates the number of samples with that gene deleted but no phenotype information for the respective drug. TP=true positive; FN=false negative; TN=true negative; FP=false positive; NA=no phenotype available.

7.4   Runtime and memory usage benchmark

The runtime and peak memory usage of each program was recorded for each sample and is presented in Figure 5. DrPRG (median 161 seconds) was significantly faster than both TBProfiler (307 seconds; $p \leq 0.0001$) and Mykrobe (230 seconds; $p \leq 0.0001$) on Illumina data. For Nanopore data, DrPRG (250 seconds) was significantly faster than TBProfiler (290 seconds; $p \leq 0.0001$), but significantly slower than Mykrobe (213 seconds; $p=0.0347$). In terms of peak memory usage, DrPRG (Illumina median peak memory 58MB; Nanopore 277MB) used significantly less memory than Mykrobe (1538MB; 1538MB) and TBProfiler (1463MB; 1990MB) on both Illumina and Nanopore data ($p \leq 0.0001$ for all comparisons).
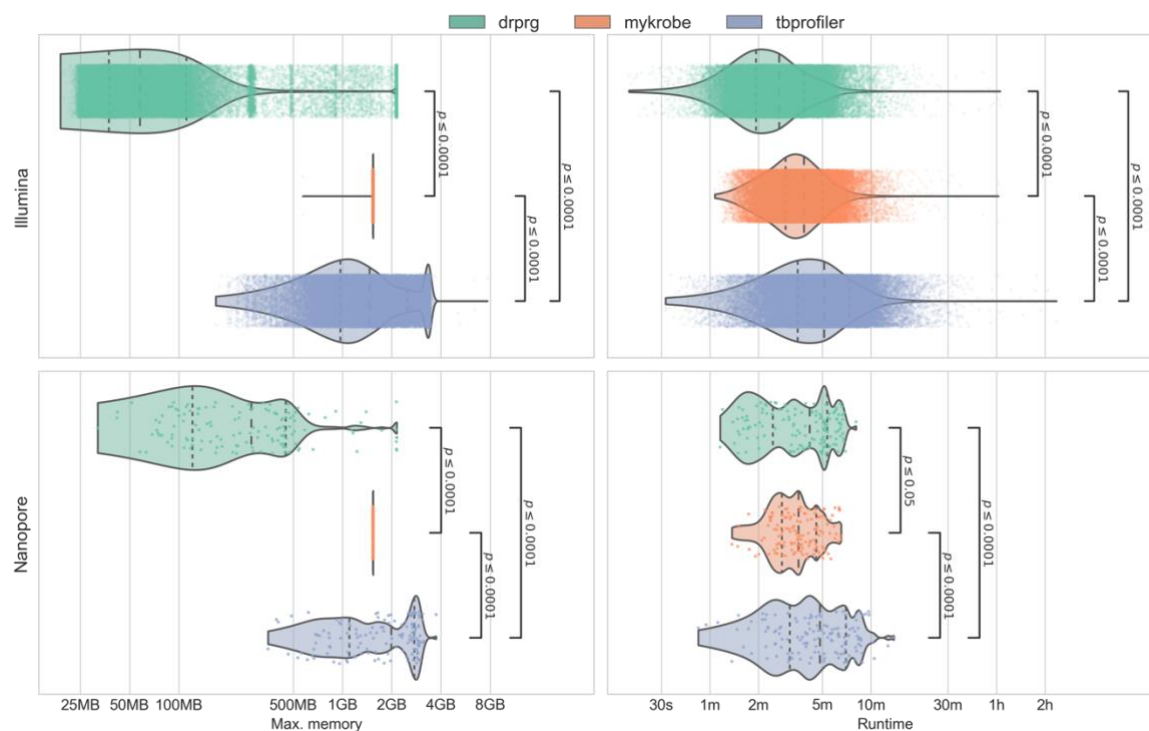
Figure 5: Benchmark of the maximum memory usage (left panels) and runtime (right panels) from Illumina (upper row) and Nanopore (lower row) data. Each point and violin is coloured by the tool, with each point representing a single sample. Statistical annotations are the result of a Wilcoxon rank-sum paired data test on each pair of tools. Dashed lines inside the violins represent the quartiles of the distribution. Note, the x-axis is log-scaled.

8. Discussion

In this work, we have presented a novel method for making drug resistance predictions with reference graphs. The method, DrPRG, requires only a reference genome and annotation, a catalogue of resistance-conferring mutations, a VCF of population variation from which to build a reference graph, and (optionally) a set of rules for types of variants in specific genes which cause resistance. We apply DrPRG to the pathogen *M. tuberculosis*, for which there is a great deal of information on the genotype/phenotype relationship, and a great need to provide good tools which implement and augment current and forthcoming versions of the WHO catalogue. We illustrate the performance of DrPRG against two existing methods for drug resistance prediction – Mykrobe and TBProfiler.

We benchmarked the methods on a high-quality Illumina sequencing dataset with associated phenotype profiles for 44,709 MTB genomes; the largest known dataset to-date[16]. All tools used the same catalogue and rules, and for most drugs, there was no significant difference between the tools. However, DrPRG did have a significantly higher specificity than TBProfiler for rifampicin predictions, and sensitivity for ethionamide predictions. DrPRG's sensitivity was also significantly greater than Mykrobe's for rifampicin, streptomycin, amikacin, capreomycin, kanamycin, and ethionamide. Evaluating detection of gene loss, we found DrPRG was more sensitive to *katG* deletions than TBProfiler.
We also benchmarked using 138 Nanopore-sequenced MTB samples with phenotype information, but found no significant difference between the tools. This Nanopore dataset was quite small and therefore the confidence intervals were large for all drugs. Increased Nanopore sequencing over time will provide better resolution of the overall sensitivity and specificity values and improve the methodological nuances of calling variants from this emerging, and continually changing, sequencing technology.

441 DrPRG also used significantly less memory than Mykrobe and TBProfiler on both Nanopore
442 and Illumina data. In addition, the runtime of DrPRG was significant faster than both tools on
443 Illumina data and faster than TBProfiler on Nanopore data. While the absolute values for
444 memory and runtime for all tools mean they could all easily run on common computers found
445 in the types of institutions likely to run them, the differences for the Nanopore data warrant
446 noting. As Nanopore data can be generated "in the field", computational resource usage is
447 critical. For example, in a recent collaboration of ours with the National Tuberculosis
448 program in Madagascar[27], Nanopore sequencing and analysis are regularly performed on a
449 laptop, meaning memory usage is sometimes a limiting factor. DrPRG's median peak
450 memory was 277MB, meaning it can comfortably be run on any laptop and other mobile
451 computing devices[69].
452 It is clear from the Illumina results that more work is needed to understand resistance-
453 conferring mutations for delamanid and linezolid. However, we did find that nonsense
454 mutations in the *ddn* gene appear likely to be resistance-conferring for delamanid – as has
455 been noted previously[39,70–72]. We also found a novel (likely) mechanism of resistance to
456 pyrazinamide - a nonstop mutation in *pncA*. Phenotype instability in *embB* at codon 306 was
457 also found to be the main driver in poor ethambutol specificity, as has been noted
458 elsewhere[55,56], indicating the need to further investigate cofactors that may influence the
459 phenotype when mutations at this codon are present.
460 Gene absence/deletion detection allowed us to confirm that the absence of *katG* – a
461 mechanism which is rare in clinical samples[64–67,73] - is highly likely to confer resistance
462 to isoniazid. Additionally, we found that the absence of *pncA* is likely to cause resistance to
463 pyrazinamide, as has been noted previously[68]. One finding that requires further
464 investigation is the variability in ethionamide phenotype when *ethA* is absent. We found that
465 only 63% of the samples with *ethA* missing, and an ethionamide phenotype, were resistant.
466 An *et al.* have suggested that *ethA* deletion alone does not always cause resistance and there
467 might be an alternate pathway via *mshA*[74].
468 Given the size of the Illumina dataset used in this work, the results provide a good marker of
469 Illumina whole-genome sequencing's ability to replace traditional phenotyping methods.
470 With the catalogue used in this study, DrPRG meets the WHO's target product profile for
471 next-generation drug-susceptibility testing for both sensitivity and specificity for amikacin,
472 and sensitivity only for rifampicin, isoniazid, levofloxacin, and moxifloxacin. However, if we
473 exclude cases where all tools call *rpoB* S450L or *katG* S315T for phenotypically susceptible
474 samples (these are strong markers of resistance[16] and therefore we suspect sample-swaps or
475 phenotype error[75]), DrPRG also meets the specificity target product profile for rifampicin
476 and isoniazid. For the other first-line drugs ethambutol and pyrazinamide, ethambutol does
477 not have a WHO target and DrPRG's sensitivity is 0.2% below the WHO target (although the
478 confidence interval spans the target), while the specificity target is missed by 0.8%.
479 The primary limitation of the DrPRG method relates to minor allele calls. DrPRG uses
480 `pandora` for novel variant discovery, which combines a graph of known population variants
481 (which can be detected at low frequency) with *de novo* detection of other variants if present at
482 above ~50% frequency. Thus, it can miss minor allele calls if the allele is absent from its
483 reference graph. While this issue did not impact most drugs, it did account for the majority of
484 cases where DrPRG missed pyrazinamide-resistant calls (in *pncA*), but the other tools
485 correctly called resistance. Unlike most other genes, where there are a relatively small
486 number of resistance-conferring mutations, or they're localised to a specific region (e.g. the
487 rifampicin-resistance determining region in *rpoB*), resistance-conferring mutations are
488 numerous - with most being rare - and distributed throughout *pncA*[16,76,77]. Adding all of
489 these mutations will, and does, lead to decreased performance of the reference graph[33], and

490 so improving minor allele calling for pyrazinamide remains a challenge we need to revisit in
491 the future.
492 One final limitation is the small number of Nanopore-sequenced MTB isolates with
493 phenotypic information. In order to get a clearer picture of the sensitivities and specificities
494 this sequencing technology can provide, we need much larger and more diverse data.
495
496 In conclusion, DrPRG is a fast, memory frugal software program that can be applied to any
497 bacterial species. We showed that on MTB, it performs as well as, or better than two other
498 commonly used tools for resistance prediction. We also collected and curated the largest
499 dataset of MTB Illumina-sequenced genomes with phenotype information and hope this will
500 benefit future work to improved genotypic drug susceptibility testing for this species. While
501 we applied DrPRG to MTB in this study, it is a framework that is agnostic to the species.
502 MTB is likely one of the bacterial species with the least to gain from reference graphs given
503 its relatively conserved (closed) pan-genome compared to other common species[78]. As
504 such, we expect the benefits and performance of DrPRG to improve as the openness of the
505 species' pan-genome increases[11]; especially given its good performance on a reasonably
506 closed pan-genome.

9. Author statements

9.1 Author contributions

509 M.B.H: conceptualisation, data curation, formal analysis, investigation, methodology,
510 resources, software, visualisation, writing – original draft, writing – review & editing. L.L:
511 resources, software, writing – review & editing. L.J.M.C: funding acquisition, methodology,
512 supervision, writing – review & editing. Z.I: conceptualisation, funding acquisition,
513 methodology, supervision, writing – original draft, writing – review & editing.

9.2 Conflicts of interest

515 The authors declare no conflicts of interest.

9.3 Funding information

9.4 Acknowledgements

10. References

525 1. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of
526 compute. PLOS ONE. 2017;12: e0177459. doi:10.1371/journal.pone.0177459

527 2. Medical Research Council. Streptomycin Treatment of Pulmonary Tuberculosis: A
528 Medical Research Council Investigation. Br Med J. 1948;2: 769–782.
529 doi:10.1136/bmj.2.4582.769

3. Wengenack NL, Todorovic S, Yu L, Rusnak F. Evidence for Differential Binding of Isoniazid by Mycobacterium tuberculosis KatG and the Isoniazid-Resistant Mutant KatG(S315T). Biochemistry. 1998;37: 15825–15834. doi:10.1021/bi982023k

4. Hackbarth CJ, Kocagoz T, Kocagoz S, Chambers HF. Point mutations in Staphylococcus aureus PBP 2 gene affect penicillin-binding kinetics and are associated with resistance. Antimicrob Agents Chemother. 1995;39: 103–106. doi:10.1128/AAC.39.1.103

5. Esposito EP, Cervoni M, Bernardo M, Crivaro V, Cuccurullo S, Imperi F, et al. Molecular Epidemiology and Virulence Profiles of Colistin-Resistant Klebsiella pneumoniae Blood Isolates From the Hospital Agency "Ospedale dei Colli," Naples, Italy. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.01463

6. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis. 2016;16: 161–168. doi:10.1016/S1473-3099(15)00424-7

7. Maira-Litrán T, Allison DG, Gilbert P. An evaluation of the potential of the multiple antibiotic resistance operon (mar) and the multidrug efflux pump acrAB to moderate resistance towards ciprofloxacin in Escherichia coli biofilms. J Antimicrob Chemother. 2000;45: 789–795. doi:10.1093/jac/45.6.789

8. Werngren J, Sturegård E, Juréen P, Ängeby K, Hoffner S, Schön T. Reevaluation of the Critical Concentration for Drug Susceptibility Testing of Mycobacterium tuberculosis against Pyrazinamide Using Wild-Type MIC Distributions and pncA Gene Sequencing. Antimicrob Agents Chemother. 2012;56: 1253–1257. doi:10.1128/AAC.05894-11

9. Schön T, Miotto P, Köser CU, Viveiros M, Böttger E, Cambau E. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. Clin Microbiol Infect. 2017;23: 154–160. doi:10.1016/j.cmi.2016.10.022

10. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2: 17040. doi:10.1038/nmicrobiol.2017.40

11. Colquhoun RM, Hall MB, Lima L, Roberts LW, Malone KM, Hunt M, et al. Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. Genome Biol. 2021;22: 267. doi:10.1186/s13059-021-02473-1

12. Hall MB, Coin LJM. Assessment of the 2021 WHO Mycobacterium tuberculosis drug resistance mutation catalogue on an independent dataset. Lancet Microbe. 2022. doi:10.1016/s2666-5247(22)00151-3

13. Mahfouz N, Ferreira I, Beisken S, von Haeseler A, Posch AE. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. J Antimicrob Chemother. 2020;75: 3099–3108. doi:10.1093/jac/dkaa257

14. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health. 2019;7. doi:10.3389/fpubh.2019.00242

569   15. Godfroid M, Dagan T, Kupczok A. Recombination Signal in Mycobacterium tuberculosis
570   Stems from Reference-guided Assemblies and Alignment Artefacts. Genome Biol Evol.
571   2018;10: 1920–1926. doi:10.1093/gbe/evy143

572   16. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO
573   catalogue of Mycobacterium tuberculosis complex mutations associated with drug
574   resistance: a genotypic analysis. Lancet Microbe. 2022. doi:10.1016/s2666-
575   5247(21)00301-3

576   17. Votintseva AA, Bradley P, Pankhurst L, Elias C del O, Loose M, Nilgiriwala K, et al.
577   Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome
578   Sequencing of Direct Respiratory Samples. J Clin Microbiol. 2017;55: 1285–1298.
579   doi:10.1128/jcm.02483-16

580   18. The end TB strategy. Geneva: World Health Organization; 2015. Report No.:
581   WHO/HTM/TB/2015.19. Available: https://www.who.int/publications/i/item/WHO-
582   HTM-TB-2015.19

583   19. Target product profile for next-generation drug-susceptibility testing at peripheral centres.
584   Geneva: World Health Organization; 2021. Available:
585   https://www.who.int/publications/i/item/9789240032361

586   20. MacLean EL-H, Miotto P, Angulo LG, Chiacchiaretta M, Walker TM, Casenghi M, et al.
587   Updating the WHO target product profile for next-generation Mycobacterium
588   tuberculosis drug susceptibility testing at peripheral centres. PLOS Glob Public Health.
589   2023;3: e0001754. doi:10.1371/journal.pgph.0001754

590   21. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct
591   variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15: 881.
592   doi:10.1186/1471-2164-15-881

593   22. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al.
594   Integrating informatics tools and portable sequencing technology for rapid detection of
595   resistance to anti-tuberculous drugs. Genome Med. 2019;11: 41. doi:10.1186/s13073-
596   019-0650-x

597   23. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-
598   resistance predictions from genome sequence data for Staphylococcus aureus and
599   Mycobacterium tuberculosis. Nat Commun. 2015;6: 10063. doi:10.1038/ncomms10063

600   24. Kohl TA, Utpatel C, Schleusener V, Filippo MRD, Beckert P, Cirillo DM, et al. MTBseq:
601   a comprehensive pipeline for whole genome sequence analysis of Mycobacterium
602   tuberculosis complex isolates. PeerJ. 2018;6: e5895. doi:10.7717/peerj.5895

603   25. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE:
604   a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage
605   from Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53: 1908–1914.
606   doi:10.1128/JCM.00025-15

26. Hunt M, Bradley P, Lapierre 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. doi:10.12688/wellcomeopenres.15603.1

27. Hall MB, Rabodoarivelo MS, Koch A, Dippenaar A, George S, Grobbelaar M, et al. Evaluation of Nanopore sequencing for Mycobacterium tuberculosis drug susceptibility testing and outbreak investigation: a genomic analysis. Lancet Microbe. 2022;0. doi:10.1016/S2666-5247(22)00301-9

28. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv; 2012. doi:10.48550/arXiv.1207.3907

29. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28: i333–i339. doi:10.1093/bioinformatics/bts378

30. The CRyPTIC Consortium and the 100,000 Genomes Project. A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. PLOS Biol. 2022;20: e3001721. doi:10.1371/journal.pbio.3001721

31. Chiner-Oms Á, Berney M, Boinett C, González-Candelas F, Young DB, Gagneux S, et al. Genome-wide mutational biases fuel transcriptional diversity in the Mycobacterium tuberculosis complex. Nat Commun. 2019;10: 3994. doi:10.1038/s41467-019-11948-6

32. Letcher B, Hunt M, Iqbal Z. Gramtools enables multiscale variation analysis with genome graphs. Genome Biol. 2021;22: 259. doi:10.1186/s13059-021-02474-0

33. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. Genome Biol. 2018;19: 220. doi:10.1186/s13059-018-1595-x

34. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics. 2012;28: 3144–3146. doi:10.1093/bioinformatics/bts578

35. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066. doi:10.1093/nar/gkf436

36. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10: 33. doi:10.12688/f1000research.29032.2

37. Gröschel MI, Owens M, Freschi L, Vargas R, Marin MG, Phelan J, et al. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. Genome Med. 2021;13: 138. doi:10.1186/s13073-021-00953-4

38. Trisakul K, Nonghanphithak D, Chaiyachat P, Kaewprasert O, Sakmongkoljit K, Reechaipichitkul W, et al. High clustering rate and genotypic drug-susceptibility screening for the newly recommended anti-tuberculosis drugs among global extensively

645   drug-resistant Mycobacterium tuberculosis isolates. Emerg Microbes Infect. 2022;11:
646   1857–1866. doi:10.1080/22221751.2022.2099304

647   39. Battaglia S, Spitaleri A, Cabibbe AM, Meehan CJ, Utpatel C, Ismail N, et al.
648   Characterization of Genomic Variants Associated with Resistance to Bedaquiline and
649   Delamanid in Naive Mycobacterium tuberculosis Clinical Strains. J Clin Microbiol.
650   2020;58. doi:10.1128/jcm.01304-20

651   40. Huang H, Ding N, Yang T, Li C, Jia X, Wang G, et al. Cross-sectional Whole-genome
652   Sequencing and Epidemiological Study of Multidrug-resistant Mycobacterium
653   tuberculosis in China. Clin Infect Dis Off Publ Infect Dis Soc Am. 2019;69: 405–413.
654   doi:10.1093/cid/ciy883

655   41. Bainomugisa A, Lavu E, Hiashiri S, Majumdar S, Honjepari A, Moke R, et al. Multi-
656   clonal evolution of multi-drug-resistant/extensively drug-resistant Mycobacterium
657   tuberculosis in a high-prevalence setting of Papua New Guinea for over three decades.
658   Microb Genomics. 2018;4: e000147. doi:10.1099/mgen.0.000147

659   42. Smith C, Halse TA, Shea J, Modestil H, Fowler RC, Musser KA, et al. Assessing
660   Nanopore sequencing for clinical diagnostics: A comparison of NGS methods for
661   Mycobacterium tuberculosis. J Clin Microbiol. 2020. doi:10.1128/jcm.00583-20

662   43. Peker N, Schuele L, Kok N, Terrazos M, Neuenschwander SM, Beer J de, et al.
663   Evaluation of whole-genome sequence data analysis approaches for short- and long-read
664   sequencing of Mycobacterium tuberculosis. Microb Genomics. 2021;7.
665   doi:10.1099/mgen.0.000695

666   44. Merker M, Rasigade J-P, Barbier M, Cox H, Feuerriegel S, Kohl TA, et al.
667   Transcontinental spread and evolution of Mycobacterium tuberculosis W148
668   European/Russian clade toward extensively drug resistant tuberculosis. Nat Commun.
669   2022;13: 5105. doi:10.1038/s41467-022-32455-1

670   45. Finci I, Albertini A, Merker M, Andres S, Bablishvili N, Barilar I, et al. Investigating
671   resistance in clinical Mycobacterium tuberculosis complex isolates with genomic and
672   phenotypic antimicrobial susceptibility testing: a multicentre observational study.
673   Lancet Microbe. 2022;3: e672–e682. doi:10.1016/s2666-5247(22)00116-1

674   46. Roberts LW, Malone KM, Hunt M, Joseph L, Wintringer P, Knaggs J, et al. Repeated
675   evolution of bedaquiline resistance in Mycobacterium tuberculosis is driven by
676   truncation of mmpR5. bioRxiv; 2022. p. 2022.12.08.519610.
677   doi:10.1101/2022.12.08.519610

678   47. Di Marco F, Spitaleri A, Battaglia S, Batignani V, Cabibbe AM, Cirillo DM. Advantages
679   of long- and short-reads sequencing for the hybrid investigation of the Mycobacterium
680   tuberculosis genome. Front Microbiol. 2023;14. doi:10.3389/fmicb.2023.1104456

681   48. Lempens P, Decroo T, Aung KJM, Hossain MA, Rigouts L, Meehan CJ, et al. Initial
682   resistance to companion drugs should not be considered an exclusion criterion for the
683   shorter multidrug-resistant tuberculosis treatment regimen. Int J Infect Dis. 2020;100:
684   357–365. doi:10.1016/j.ijid.2020.08.042

49. Lempens P, Meehan CJ, Vandelannoote K, Fissette K, de Rijk P, Van Deun A, et al. Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations. Sci Rep. 2018;8: 3246. doi:10.1038/s41598-018-21378-x

50. Steinig E, Coin L. Nanoq: ultra-fast quality control for nanopore reads. J Open Source Softw. 2022;7: 2991. doi:10.21105/joss.02991

51. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

52. Hunt M, Letcher B, Malone KM, Nguyen G, Hall MB, Colquhoun RM, et al. Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. Genome Biol. 2022;23: 147. doi:10.1186/s13059-022-02714-x

53. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

54. Zwyer M, avusoglu C, Ghielmetti G, Pacciarini M, Scaltriti E, Van Soolingen D, et al. A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics. Open Res Eur. 2021;1. doi:10.12688/openreseurope.14029.2

55. Chen Y, Takiff HE, Gao Q. Phenotypic instability of Mycobacterium tuberculosis strains harbouring clinically prevalent drug-resistant mutations. Lancet Microbe. 2023;0. doi:10.1016/S2666-5247(23)00007-1

56. Sirgel FA, Warren RM, Streicher EM, Victor TC, Helden PD van, Böttger EC. embB306 Mutations as Molecular Indicators to Predict Ethambutol Susceptibility in Mycobacterium tuberculosis. Chemotherapy. 2013;58: 358–363. doi:10.1159/000343474

57. Huo F, Ma Y, Li S, Xue Y, Shang Y, Dong L, et al. Specific gyrA Gene Mutations Correlate with High Prevalence of Discordant Levofloxacin Resistance in Mycobacterium tuberculosis Isolates from Beijing, China. J Mol Diagn. 2020;22: 1199–1204. doi:10.1016/j.jmoldx.2020.06.010

58. Brankin AE, Fowler PW. Inclusion of minor alleles improves catalogue-based prediction of fluoroquinolone resistance in Mycobacterium tuberculosis. JAC-Antimicrob Resist. 2023;5: dlad039. doi:10.1093/jacamr/dlad039

59. Nimmo C, Brien K, Millard J, Grant AD, Padayatchi N, Pym AS, et al. Dynamics of within-host Mycobacterium tuberculosis diversity and heteroresistance during treatment. EBioMedicine. 2020;55: 102747. doi:10.1016/j.ebiom.2020.102747

60. Perdigão J, Gomes P, Miranda A, Maltez F, Machado D, Silva C, et al. Using genomics to understand the origin and dispersion of multidrug and extensively drug resistant tuberculosis in Portugal. Sci Rep. 2020;10: 2600. doi:10.1038/s41598-020-59558-3

61. Singh A, Singh A, Grover S, Pandey B, Kumari A, Grover A. Wild-type catalase peroxidase vs G279D mutant type: Molecular basis of Isoniazid drug resistance in

724   Mycobacterium tuberculosis. Gene. 2018;641: 226–234.
725   doi:10.1016/j.gene.2017.10.047

726   62. Vaziri F, Kohl TA, Ghajavand H, Kargarpour Kamakoli M, Merker M, Hadifar S, et al.
727   Genetic Diversity of Multi- and Extensively Drug-Resistant Mycobacterium
728   tuberculosis Isolates in the Capital of Iran, Revealed by Whole-Genome Sequencing. J
729   Clin Microbiol. 2019;57: e01477-18. doi:10.1128/JCM.01477-18

730   63. de Lourdes do Carmo Guimarães Diniz J, von Groll A, Unis G, Dalla-Costa ER, Rosa
731   Rossetti ML, Vianna JS, et al. Whole-genome sequencing as a tool for studying the
732   microevolution of drug-resistant serial Mycobacterium tuberculosis isolates.
733   Tuberculosis. 2021;131: 102137. doi:10.1016/j.tube.2021.102137

734   64. Altamirano M, Marostenmaki J, Wong A, FitzGerald M, Black WA, Smith JA. Mutations
735   in the catalase-peroxidase gene from isoniazid-resistant Mycobacterium tuberculosis
736   isolates. J Infect Dis. 1994;169: 1162–1165. doi:10.1093/infdis/169.5.1162

737   65. Ferrazoli L, Palaci M, da Silva Telles MA, Ueki SY, Kritski A, Marques LRM, et al.
738   Catalase Expression, katG, And MIC Of Isoniazid For Mycobacterium tuberculosis
739   Isolates From SãO Paulo, Brazil. J Infect Dis. 1995;171: 237–240.
740   doi:10.1093/infdis/171.1.237

741   66. Ramaswamy SV, Reich R, Dou S-J, Jasperse L, Pan X, Wanger A, et al. Single
742   Nucleotide Polymorphisms in Genes Associated with Isoniazid Resistance in
743   Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2003;47: 1241–1250.
744   doi:10.1128/AAC.47.4.1241-1250.2003

745   67. Zhang Y, Heym B, Allen B, Young D, Cole S. The catalase—peroxidase gene and
746   isoniazid resistance of Mycobacterium tuberculosis. Nature. 1992;358: 591–593.
747   doi:10.1038/358591a0

748   68. Martinez E, Holmes N, Jelfs P, Sintchenko V. Genome sequencing reveals novel deletions
749   associated with secondary resistance to pyrazinamide in MDR Mycobacterium
750   tuberculosis. J Antimicrob Chemother. 2015;70: 2511–2514. doi:10.1093/jac/dkv128

751   69. Samarakoon H, Punchihewa S, Senanayake A, Hammond JM, Stevanovski I, Ferguson
752   JM, et al. Genopo: a nanopore sequencing analysis toolkit for portable Android devices.
753   Commun Biol. 2020;3: 1–5. doi:10.1038/s42003-020-01270-z

754   70. Gómez-González PJ, Perdigao J, Gomes P, Puyen ZM, Santos-Lazaro D, Napier G, et al.
755   Genetic diversity of candidate loci linked to Mycobacterium tuberculosis resistance to
756   bedaquiline, delamanid and pretomanid. Sci Rep. 2021;11: 19431. doi:10.1038/s41598-
757   021-98862-4

758   71. Schena E, Nedialkova L, Borroni E, Battaglia S, Cabibbe AM, Niemann S, et al.
759   Delamanid susceptibility testing of Mycobacterium tuberculosis using the resazurin
760   microtitre assay and the BACTEC^{TM} MGIT^{TM} 960 system. J Antimicrob Chemother.
761   2016;71: 1532–1539. doi:10.1093/jac/dkw044

762   72. Gomes LC, Campino S, Marinho CRF, Clark TG, Phelan JE. Whole genome sequencing
763   reveals large deletions and other loss of function mutations in Mycobacterium

tuberculosis drug resistance genes. Microb Genomics. 2021;7: 000724. doi:10.1099/mgen.0.000724

73. De Maio F, Cingolani A, Bianco DM, Salustri A, Palucci I, Sanguinetti M, et al. First description of the katG gene deletion in a Mycobacterium tuberculosis clinical isolate and its impact on the mycobacterial fitness. Int J Med Microbiol. 2021;311: 151506. doi:10.1016/j.ijmm.2021.151506

74. Ang MLT, Zainul Rahim SZ, de Sessions PF, Lin W, Koh V, Pethe K, et al. EthA/R-Independent Killing of Mycobacterium tuberculosis by Ethionamide. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.00710

75. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;379: 1403–1415. doi:10.1056/NEJMoa1800474

76. Köser CU, Cirillo DM, Miotto P. How To Optimally Combine Genotypic and Phenotypic Drug Susceptibility Testing Methods for Pyrazinamide. Antimicrob Agents Chemother. 2020;64: e01003-20. doi:10.1128/AAC.01003-20

77. Yadon AN, Maharaj K, Adamson JH, Lai Y-P, Sacchettini JC, Ioerger TR, et al. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. Nat Commun. 2017;8: 588. doi:10.1038/s41467-017-00721-2

78. Park S-C, Lee K, Kim YO, Won S, Chun J. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.00834