A generalisable approach to drug susceptibility prediction for M. tuberculosis using machine learning and whole-genome sequencing

The CRyPTIC consortium*

* Members of the CRyPTIC consortium are listed at the end.

Corresponding author: Dr. Alexander S. Lachapelle, M.D.; Alexander.Lachapelle@eng.ox.ac.uk

Abstract

Rapid and up-to-date drug susceptibility testing is urgently needed to address the threat of multidrug resistant tuberculosis. We developed a composite machine learning system to predict susceptibility from whole-genome sequences for 13 anti-tuberculosis drugs. We trained, validated and externally tested the system, and assessed its performance against a previously validated mutation catalogue, existing molecular assays, and World Health Organization Target Product Profiles. 174,492 phenotypes and 26,328 isolates from 34 countries were studied. The sensitivity of the model was greater than 90% for all drugs except ethionamide, clofazimine and linezolid. Specificity was greater than 95% for all drugs except ethambutol, ethionamide, bedaquiline, delamanid and clofazimine. The machine learning system was more sensitive than the mutation catalogue and molecular assays. For rifampicin-resistant samples, it correctly predicted a pan-susceptible second-line regimen with 98% accuracy. The proposed system can help guide therapy and be updated automatically as new resistance determinants emerge.

Background

In 2019, 10 million individuals were diagnosed with a *Mycobacterium tuberculosis* infection and 1.4 million died[1]. The problem of multidrug resistant tuberculosis (MDR-TB) - defined as resistance to isoniazid and rifampicin – has been described by the World Health Organization (WHO) as a global health crisis[1]. Despite advances in diagnostics and treatment, MDR-TB remains under-detected and treatment success remains stubbornly below 60% globally[1,2]. The SARS-CoV-2 pandemic is expected to set back the progress that has been made by years[3].

The WHO has called for universal drug susceptibility testing (DST)[4]. Culture-based DST is too slow, expensive and technically challenging to offer a realistic solution. Molecular assays can rapidly detect resistance to rifampicin, isoniazid and a subset of second-line drugs, but are limited in the number of resistance-conferring mutations they can detect[5], constraining their sensitivity, although more for some drugs than for others. Some countries already rely on whole-genome sequencing (WGS) to identify susceptibility to first-line drugs[6], but nowhere are routine diagnostic algorithms advanced enough to dispense with culture-based DST where it is available for second-line, new and repurposed drugs[7]. Indeed, no algorithm has yet been demonstrated to meet the WHO Target Product Profile (TPP) thresholds for clinical application for drugs now recommended to treat MDR-TB[8–11].

Artificial intelligence and machine learning algorithms have been suggested as potential solutions where molecular determinants of resistance are either unknown or complex, such as gene-gene interactions, while allowing for real-time updating as new resistant samples are collected[12–19]. Here, we compare the performance for priority anti-tuberculosis agents of a previously validated mutation catalogue with a composite machine learning DST system using WGS data. We assess the extent to which machine learning can bridge the gap between the already good performance of catalogue-based predictions for some drugs and what is needed to dispense with routine phenotypic DST for anti-tuberculosis agents in general.

Results

Dataset characteristics

A total of 174,492 phenotypes from 26,328 isolates were studied across two large datasets. The CRyPTIC dataset derived phenotypes from 96-well broth microdilution plates for 10,859 isolates from 22 countries. Lineages 1 to 4, 6 and *Mycobacterium bovis* were represented, with lineage 4 (50%, 5,436/10,859) and lineage 2 (35%, 3,745/10,859) the most common. 28% of samples (3,033/10,859) were MDR. Phenotypes were available for three first-line antibiotics (isoniazid, rifampicin, ethambutol), plus rifabutin, and nine second-line antibiotics used against MDR-TB: two fluoroquinolones (moxifloxacin, levofloxacin), two injectable agents (amikacin, kanamycin), ethionamide, and four new or repurposed drugs (bedaquiline, linezolid, clofazimine, delamanid). Prevalence of resistance ranged from 1% for bedaquiline, to 47% for isoniazid (Table 1). For each new and repurposed drug, a minimum of 69 resistant samples were available. Pyrazinamide was not present on the microdilution plate for technical reasons. Where two drugs of the same class were studied, we report results for the one present in WHO guidelines or the most commonly prescribed in primary results (amikacin, rifampicin, moxifloxaxin), and for the other in the supplementary appendix (kanamycin, rifabutin and levofloxacin). A second, independent dataset used Mycobacteria Growth Indicator Tube (MGIT)-derived phenotypes and included 15,469 isolates from 22 countries, 21% of which were MDR (3,189/15,469). The independent set included phenotypes for all antibiotics except new/repurposed drugs and rifabutin (Table 1).

Machine learning in the CRyPTIC dataset

We developed a machine learning system comprising two complementary predictors (Figure 1), a kmer-based, hypothesis-free, genome-wide supervised machine learning algorithm and an algorithm associating mutations with phenotypic resistance (Methods). To assess the performance of the machine learning system on the widest possible set of antibiotics, it was initially trained on 75% of the CRyPTIC dataset (8,146 randomly selected isolates). Predictions were made for the remaining 25% (2,713 isolates) (Table 2). For first-line drugs, the sensitivity

of the machine learning system was 95% for isoniazid (1,119/1,173), 97% for rifampicin (906/931) and 95% for ethambutol (387/406), with specificity 99% (1,195/1,207), 98% (1,287/1,315) and 88% (1,326/1,501) respectively (Figure 2). For second-line drugs against MDR-TB, sensitivity was 96% for moxifloxacin (251/261), 92% for amikacin (146/158) and 88% for ethionamide (271/309), with specificity 96% (1,377/1,438), 99% (2,068/2,090) and 89% (1,694/1,901) respectively. Although there were comparatively few phenotypically resistant isolates, sensitivities for bedaquiline, delamanid, clofazimine and linezolid were 94% (15/16), 90% (18/20), 87% (20/23) and 57% (16/28), at the cost of low specificity (71%, 55%, 72% and 96% respectively). Importantly for clinical decisions on whether a drug should be given, the negative predictive value for resistance prediction was above 98% for all drugs except isoniazid, where it was 96%, noting that the low prevalence of resistance to new and repurposed drugs (0.7-1%) was a major contributor to high negative predictive value. We simulated the negative predictive value of the system for each drug for difference resistance prevalences (Figure 3). We assessed whether results were affected by the split of training and test data, batch effects, or training and testing on genetically-related samples from the same site, by repeating the experiment using a "leave-one-site-out" cross-validation approach, sequentially using each site as the test set, and training the model on the remaining 10 sites (Table S1). Performance was similar across all first- and second-line drugs with the exception of new and repurposed drugs, where sensitivity decreased (67-73%) and specificity increased (73-78%) using the leave-one-site-out approach. A high proportion of samples resistant to these agents were from the same South African site (31/69 resistant to bedaquiline, 55/105 to clofazimine) causing variability when this specific site was used to train or test.

Machine learning in the independent dataset

We further assessed the system's performance in a large independent dataset where phenotypic DST was based entirely on MGIT, thus testing the generalizability of the machine learning system trained on CRyPTIC broth microdilution plates. We re-trained the machine learning system using the entire CRyPTIC dataset; predictions on the independent dataset were similarly accurate to those above (Table 3). For first-line drugs, sensitivity was 95% for isoniazid

(3,216/3,397), 98% for rifampicin (2,957/3,021) and 94% for ethambutol (1,765/1,877), with specificity 99% (9,493/9,602), 98% (10,298/10,502) and 92% (10,758/10,502) respectively. For second-line drugs, sensitivity was 93% for moxifloxacin (288/311) and 88% for amikacin (266/302), with specificity of 96% (2,072/2,168) and 95% (2,403/2,535) respectively. Negative predictive value was greater than 98% for all drugs. There were no phenotypes to new or repurposed drugs in the independent dataset to make predictions for.

Comparison to the mutation catalogue, molecular assays and target product profiles

We next compared predictions from the machine learning system with those from a validated mutation catalogue[20] in the independent set (Table 3). For first-line drugs, sensitivity for the catalogue was 94% for isoniazid (3,177/3,397), 97% for rifampicin (2,936/3,021) and 89% for ethambutol (1,676/1,877), with specificity of 99% (9,525/9,602), 99% (10,394/10,502) and 96% (11,185/10,502) respectively. These results were consistent with the previously described performance of this catalogue that led to its clinical implementation for DST to first-line drugs in several countries[7]. Nevertheless, these sensitivities were lower than those from the machine learning system, which was superior by 1% for isoniazid and rifampin, and by 5% for ethambutol ($p<0.001$). The improved sensitivity of the machine learning system came at a small cost in specificity which was 1% lower for rifampin and 4% lower for ethambutol ($p<0.001$). The machine learning also system proved more sensitive than the catalogue for moxifloxacin (93% vs 86%, $p<0.001$), amikacin (88% vs 85%, $p<0.001$) and ethionamide (84% vs 50%, $p<0.01$), with 1% lower specificity for moxifloxacin, 3% for amikacin and 21% for ethionamide ($p=0.003$).

Higher sensitivities can be obtained using the catalogue if predictions are only made on isolates containing genomic variation that is known to the catalogue[20] instead of making predictions for all isolates (Methods). However, returning “indeterminate” predictions where novel variation is seen in candidate genes in an isolate does not align with recent WHO TTP that require a minimum indeterminate rate of less than 10% for DST implementation[10]. This rate would have been 6% for isoniazid, 2% for rifampin, 10% for ethambutol, 9% for moxifloxacin, 14% for amikacin and 36% for ethionamide in the independent test set – all of which were counted as susceptible in the analysis above. The machine learning system has the advantage of providing

predictions for all isolates (Table S2). We examined phenotypically resistant samples where novel variation is seen in candidate genes. In these, the machine learning system correctly predicted 41/106 isolates phenotypically resistant to isoniazid that were missed by the catalogue, 81/94 for ethambutol, 12/15 for moxifloxacin, 2/11 for amikacin and 9/12 for ethambutol. The specificity of the machine learning system for isolates that would have been called "indeterminate" by the catalogue ranged from 88% for ethambutol to 99.7% for rifampin (Table S3).

As most patients in the world have little, or no access to phenotypic DST, we compared the performance of the machine learning system against the expected combined performance of Xpert MTB/RIF and Xpert XDR for the independent set in anticipation of its wider uptake to address the WHO's call for universal DST. The sensitivity of the machine learning system was 4% higher than Xpert for isoniazid (95% vs 91%, $p<0.001$) and rifampicin (98% vs 94%, $p<0.001$), 7% higher for moxifloxacin (93% vs 86%, $p<0.001$) and 3% higher for amikacin (88% vs 85%, $p=0.030$). Specificity was no more than 1% lower for each drug, with the exception of amikacin (95% vs 98%). Therefore, if 1,000 isolates were resistant to a second-line quinolone or an injectable drug, the machine learning system would accurately find between 30 and 113 phenotypically resistant isolates predicted as 'not resistant' by Xpert, at the cost of calling between 0 and 34 phenotypically susceptible isolates 'resistant' (Table 2).

The WHO TPP for rapid molecular DST assays require a minimum sensitivity of 95% for rifampicin, 90% for isoniazid and fluoroquinolones and 80% for other second-line agents; a specificity of 98% for all drugs; and a minimum indeterminate rate of less than 10%. In the CRyPTIC dataset, the machine learning system met the minimum TPP sensitivity threshold for all drugs, with the exception of linezolid (sensitivity 57%). Specificity thresholds were met for isoniazid, rifampicin, levofloxacin and amikacin, but were not met for ethambutol (88%), moxifloxacin (96%), ethionamide (89%) and new and repurposed drugs (55-72%) - although they still outperformed the specificity of the catalogue and existing molecular assays for each (Table S4). The machine learning system met the requirement for indeterminate results for all drugs as it provides predictions for all samples.

Full drug regimen prediction for rifampicin-resistant isolates

While most DST focuses on predicting susceptibility to individual drugs, clinicians are left with the task of assembling a full regimen themselves. This is especially challenging for rifampicin-resistant and MDR-TB, where new WHO guidelines recommend the inclusion of new and repurposed drugs like bedaquiline and delamanid for which there is no widely-used DST.

We therefore trained the machine learning system to predict an entire treatment regimen, designed according to the latest WHO guidance[20]. A total of 50 possible regimens were considered, including all combinations of group A, group B and group C drugs meeting WHO standards (Figure 1, Table S5)[21]. As only the CRyPTIC dataset included phenotypic DST data for new and repurposed drugs, we used the machine learning system trained on the original 75% of CRyPTIC to predict regimens for the rifampicin-resistant isolates in the original 25% test set.

Sufficient phenotypic data were available for 768/931 rifampicin-resistant isolates to assess at least one potential regimen for treatment of MDR-TB. The machine learning system predicted a fully susceptible regimen composed of 4 to 5 drugs of groups A to C[21] for 482 of these 768 isolates, and was correct in doing so for 472 (98%). In 8 of the 10 remaining regimens, only one drug in each regimen was phenotypically resistant (Table S6). The system predicted some phenotypic resistance in every potential regimen for the 296 other isolates, of which 139 (47%) isolates had a phenotypically susceptible regimen. Prevalence of bedaquiline, linezolid and clofazimine resistance was 1-2% (9, 7 and 14 samples respectively). Considering each drug individually in phenotypically rifampicin-resistant isolates, the sensitivity for moxifloxacin, levofloxacin and amikacin was respectively 98% (229/233), 96% (256/267) and 96% (133/139), and specificity 90% (357/398), 96% (393/408) and 99% (642/652). Sensitivity for bedaquiline and linezolid was 100% (9/9 and 7/7 respectively), and specificity was 78% and 51% (Table S7).

Discrepancy analysis

We reviewed individual cases where the machine learning system made an incorrect prediction in the CRyPTIC dataset. Where a phenotypically resistant isolate was predicted to be susceptible (false negative), we interrogated the predictions from the two subcomponents of

the machine learning system for evidence of a predicted increase in median inhibitory concentration (MIC), albeit still below the cutoff. For isoniazid, 54/1173 phenotypically resistant samples were predicted to be susceptible in the test set. One or other of the subcomponents of the machine learning system (ML or the algorithm) predicted an MIC above the baseline or near the epidemiological cut-off for 16 of these 54 false-negatives (30%), while it only predicted an elevated MIC for 46 of the 1,195 true negatives (4%). For ethambutol, 13/19 (68%) false negatives were predicted to have a higher MIC from at least one of the two subcomponents, compared to 437/1,287 (34%) true negatives. Higher predicted MIC in false-negative samples were similarly found in rifampicin (19/25), ethionamide (24/38), levofloxacin (1/22), moxifloxacin (1/10) and amikacin (3/12). For each drug, MIC increase occurred more in false negatives than true negatives, with the exception of levofloxacin (Table S8).

Discussion

We assessed the extent to which machine learning can bridge the gap between the good performance of catalogue-based predictions and what is needed to dispense with routine phenotypic DST not only for first-line drugs but for almost all other anti-tuberculosis drugs too. We trained a machine learning system to predict susceptibility to 13 antituberculosis agents using whole genome sequencing data, and tested its performance on a large independent test set. We followed best practice guidance for studies evaluating the accuracy of rapid tuberculosis drug-susceptibility testing (DST)[8]. The machine learning system fully met WHO target product profiles (TPP) for three priority drugs in the CRyPTIC dataset - rifampicin, isoniazid, and amikacin - and met sensitivity but not specificity targets for ethambutol, moxifloxacin, ethionamide and new and repurposed drugs. For linezolid, no targets were met.

For drugs where the WHO-endorsed molecular GeneXpert assay is available (rifampicin for Xpert MTB/RIF, and isoniazid, fluoroquinolones, aminoglycosides and ethionamide for Xpert MTB/XDR), our system significantly increased the sensitivity and negative predictive values on the independent test set compared to the expected performance of these assays, at a small cost to specificity. There are several explanations, including that the assays only look at eight

genes and promoter regions and exclude rare variants therein[5], while the machine learning system is able to explore genome-wide features, leverage interactions between features and assess lineage and genetic background through genome-wide features.

The WHO guidelines for MDR-TB management recommend giving all patients on long MDR-TB regimens bedaquiline, linezolid and clofazimine[21]. Our sensitivity and specificity for these three drugs fall below WHO TPP requirements. Sensitivity of 93%, 90% and 87% in the CRyPTIC set for bedaquiline, delamanid and clofazimine likely reflect the very low prevalence of resistance (15,18 and 20 resistant samples respectively), while also explaining the high negative predictive values of >99% for all three drugs. As more resistant isolates are collected, the sensitivity and specificity of the machine learning system will almost certainly increase, and negative predictive value decrease, as seen for other drugs. Nevertheless, a test with negative predictive value >99% and sensitivity 70% would still provide value to clinicians who currently have no other test for these new and repurposed drugs and hence treat their patients empirically in the absence of reliable, rapid and robust molecular or genotypic DST[22]. Even imperfect test performance could still play a key role in preventing the amplification and dissemination of resistance.

A key benefit of our genome-wide approach over molecular DST is the ability to update and train automatically as new resistant samples are added. This is critical as resistance to existing and new agents like bedaquiline emerge, avoiding the expensive multi-phase multi-year development times of molecular assays[10,23], or the need to update catalogues through expert review[20]. The U.S. Food and Drug Administration (FDA) recently released a regulatory framework for 'live' modifications to artificial intelligence and machine learning-based software as a medical device[24] and has recently provided clearance or approval for several such diagnostic devices[25], paving the way for clinical implementation and dissemination.

We note several further novelties and benefits of our systematic approach. First, by combining machine learning with algorithmic catalogue generation we leverage existing knowledge, including known genes associated with resistance, avoiding a common complaint against pure machine learning systems. Second, a prediction can be made for all isolates, while previous published catalogue-based methods that met clinical thresholds required the exclusion of 4-

10% of samples with unknown mutations in candidate genes[7]. Third, using kmers from sequencing reads allows for genome-wide analysis while being robust to potential errors or variability in genotype mapping or variant calling, known to affect prediction of transmission inferences and resistance prediction[26]. A *vcf* file filters out sites called with low confidence, while the machine learning approach uses all kmers from reads and therefore can use them as training features. Fourth, the model predicts MIC as an intermediate step. Although we have focussed on predicting binary DST results so that we can perform external validation on MGIT data, the predicted MICs would allow treatment to be individualized both in terms of drug selection and dosing, potentially improving outcomes given associations between sub-threshold MICs and outcome[27]. MIC predictions could also be used to assess confidence in a susceptibility prediction and mitigate future errors, with isolates without any predicted elevation less likely to be resistant than isolates with a sub-resistant increase in MIC. Fifth, by using an interpretable supervised machine learning algorithm, we provide a list of features used for prediction, which in turn can be used as hypotheses for potential causal mutations, when combined with protein analysis.

A study limitation is the use of a previously published literature-derived catalogue, rather than the more cutting-edge, recently published WHO-endorsed catalogue[20]. This was impossible as the WHO catalogue was developed using samples from both the CRyPTIC and independent datasets. Second, we were unable to access an external dataset with sufficient resistance to perform independent validation of predictions to bedaquiline, linezolid, delamanid and clofazimine. Consequently, we report performance only in the CRyPTIC dataset (which does contain DST for these compounds in large numbers) using MIC data and both a train-test approach and cross-validation of models tested on each site and trained on all other sites. Third, we report the performance of GeneXpert *in silico*, but clinical performance of the actual method might differ. Fourth, the use of kmers from raw sequencing reads, while extremely effective on existing datasets, might not translate directly to new sequencing methods in the future, such as third-generation long-read sequencing with Oxford Nanopore, given the different format of reads. Fifth, regimen prediction for RR-TB was calculated using a simple union of individual drug predictions; training a novel system explicitly for regimen prediction,

rather than individual drug prediction, would provide additional benefits, including accounting for potential gene-gene and drug-drug interactions influencing the efficacy of entire regimens.

In summary, this study demonstrates that WGS can now be used to provide clinically actionable susceptibility prediction for many drugs recommended for the treatment of susceptible and of MDR-TB, using an composite machine-learning system. This study shows how a machine learning system can be used to help guide therapy, and can be straightforwardly updated as increasing numbers of resistant samples to new and repurposed drugs are collected.

Methods

Study design

We performed a training, validation and external testing study of a mutation catalogue and a machine learning system to predict susceptibility to 13 anti-tuberculosis antibiotics using whole-genome sequencing (WGS). We trained and tested the system on 10,859 isolates from 11 laboratories in 22 countries collected by the CRyPTIC consortium. Phenotypes were determined using the UKMYC broth microdilution system[28]. We then assessed how this system, trained on UKMYC-derived phenotypes, would perform against a commonly used DST method in independent samples. For this we made predictions for an external set of isolates used to derive the WHO catalogue of drug resistant mutations[28]. We selected only those samples that had been phenotypically characterized by Mycobacteria Growth Indicator Tube (MGIT), namely 15,239 *M. tuberculosis complex* isolates from 22 countries (Table 1 for an overview and Table S9 for a detailed description of each dataset).

Whole-genome sequencing

All isolates were whole-genome sequenced using Illumina next-generation sequencing, with sequencing protocols varying between sites as previously described[28]. Sequencing reads were trimmed and mapped to the reference genome H37Rv, and variants called using Clockwork

(v0.8.3) a bespoke processing pipeline built for CRyPTIC and optimized to detect both single nucleotide polymorphisms (SNPs) and insertions and deletions (indels). Prior to mapping and calling, raw nucleotide kmers from sequencing reads were set aside for training the machine learning predictor.

Phenotypic drug-susceptibility testing

Phenotypic drug susceptibility testing (DST) for the CRyPTIC training and test set was performed across all sites using a standard protocol described elsewhere[28]. Briefly, samples were subcultured and inoculated into 96-well broth microdilution plates containing 13 drugs and designed by the CRyPTIC consortium and manufactured by Thermo Fisher Inc., U.K.. Between 5-10 doubling dilutions were used for each drug, and minimum inhibitory concentrations (MIC) for each were read after 14 days using three methods for quality assurance. MICs were converted to predictions of resistance or susceptibility using epidemiological cutoffs (ECOFFs)[28]. As the plate design was modified during the study, the intersect of both plates was used as the MIC phenotype, and concentrations outside both were right-censored or left-censored as appropriate (Table S10). Phenotypic DST for the external test set used the BACTEC MGIT 960 system.

Susceptibility prediction

DST for each sample was predicted using two methods: a mutation catalogue previously tested and validated in CRyPTIC[7], and a machine learning system. Although the catalogue had previously been tested on first-line drugs, here we used targets assayed by commercial molecular assays to expand the catalogue to cover some second-line drugs (Table S13). The machine learning system was itself a composite of two complementary predictors (Figure 1). The first predictor was a kmer-based, hypothesis-free, genome-wide supervised machine learning algorithm. Raw nucleotide kmers (k=31) from sequencing reads (i.e. prior to mapping or assembly) were used as features. A total of 1.9×10^9 individual kmers were considered. Where <5 kmers were identified for an isolate these were considered sequencing errors (Figure S1). We merged features across patterns[29], applied feature selection using the F-test applied to MICs, and trained an optimized tree-based extreme gradient boosting method to allow for

rapid training, testing and feature interpretation. After training, the top features relevant to each prediction were mapped to H37Rv using bowtie2 for detailed feature analysis (Figure S2, Table S11). The second predictor was an algorithm associating mutations with phenotypic resistance based on previously described approaches[30]. It focussed on the same pre-determined list of candidate genes and promotor sequences as used to generate the WHO *M. tuberculosis* drug-resistance mutations catalogue[28] (Table S12). After the masking of neutral mutations using the same process as described[28], the remaining genetic variation across candidate genes relevant to a drug was taken as a unique genetic signature. This included the absence of any remaining variation, and where there was just a single remaining mutation. The mode MIC from all isolates sharing that unique genetic signature was then taken to predict MICs in isolates that shared the same unique signature. If no exact match was made to a genetic signature (combination of variants), the highest mode MIC associated with any individual mutation in the genetic signature was used to predict the MIC. Where no match could be made to any genetic signature described in the training set, the test set phenotype prediction was left as 'U' (unknown). Both methods' outcomes were combined into a final joint prediction system using an "or" logic gate, in order to optimize sensitivity and negative predictive value. Youden's J statistic was applied to derive the operating threshold of the system. Performance on the 25% CRyPTIC test set was estimated by training the system on the 75% CRyPTIC samples not included in it. Performance on the independent test set was generated by training the system on the entire CRyPTIC dataset. P-values were calculated using McNemar chi-square test. To better assess the generalizability of the approach within the CRyPTIC dataset and minimize the risk of training and testing on genomically-related isolates, we compared main test set results to those from a leave-one-site-out approach, where each of the 11 sites was left out in turn for training, but correspondingly used for testing, with performance taken as the mean weighted by resistance prevalence. We benchmarked the performance of the mutation catalogue and machine learning system against the expected performance of Xpert® MDR/RIF and Xpert® XDR (Cepheid, Sunnyvale, U.S.), based on the targets they probe (Table S13). "Indeterminate" predictions by the catalogue where a novel variation is seen in a candidate gene were counted as susceptible for the purpose of the

analysis. Finally, we simulated negative predictive values for each drug for different prevalences of resistance. For each drug, we selected 138 samples at random to generate data sets with a percentage prevalence of resistance for every 1% between 1-49%, and repeated this 100 times. 138 corresponds to twice the number of isolates resistant to the drug with the smallest resistance prevalence, bedaquiline (69 resistant isolates).

Ethics

Approval for CRyPTIC study was obtained by Taiwan Centers for Disease Control IRB No. 106209, University of KwaZulu Natal Biomedical Research Ethics Committee (UKZN BREC) (reference BE022/13) and University of Liverpool Central University Research Ethics Committees (reference 2286), Institutional Research Ethics Committee (IREC) of The Foundation for Medical Research, Mumbai (Ref nos. FMR/IEC/TB/01a/2015 and FMR/IEC/TB/01b/2015), Institutional Review Board of P.D. Hinduja Hospital and Medical Research Centre, Mumbai (Ref no. 915-15-CR [MRC]), scientific committee of the Adolfo Lutz Institute (CTC-IAL 47-J / 2017) and in the Ethics Committee (CAAE: 81452517.1.0000.0059) and Ethics Committee review by Universidad Peruana Cayetano Heredia (Lima, Peru) and LSHTM (London, UK).

Members of the CRyPTIC consortium (in alphabetical order)

Correspondence to: Alexander S Lachapelle (alexander.lachapelle@eng.ox.ac.uk)

Ivan Barilar[29], Simone Battaglia[1], Emanuele Borroni[1], Angela P Brandao[2,3], Alice Brankin[4], Andrea Maurizio Cabibbe[1], Joshua Carter[5], Daniela Maria Cirillo[1], Pauline Claxton[6], David A Clifton[4], Ted Cohen[7], Jorge Coronel[8], Derrick W Crook[4], Viola Dreyer[29], Sarah G Earle[4], Vincent Escuyer[9], Lucilaine Ferrazoli[3], Philip W Fowler[4], George Fu Gao[10], Jennifer Gardy[11], Saheer Gharbia[12], Kelen T Ghisi[3], Arash Ghodousi[1,13], Ana Luíza Gibertoni Cruz[4], Louis Grandjean[33], Clara Grazian[14], Ramona Groenheit[44], Jennifer L Guthrie[15,16], Wencong He[10], Harald Hoffmann[17,18], Sarah J Hoosdally[4], Martin Hunt[19,4], Zamin Iqbal[19], Nazir Ahmed Ismail[20], Lisa

Jarrett[21], Lavania Joseph[20], Ruwen Jou[22], Priti Kambli[23], Rukhsar Khot[23], Jeff Knaggs[19,4], Anastasia Koch[24], Donna Kohlerschmidt[9], Samaneh Kouchaki[4,25], Alexander S Lachapelle[4], Ajit Lalvani[26], Simon Grandjean Lapierre[27], Ian F Laurenson[6], Brice Letcher[19], Wan-Hsuan Lin[22], Chunfa Liu[10], Dongxin Liu[10], Kerri M Malone[19], Ayan Mandal[28], Mikael Mansjö[44], Daniela Matias[21], Graeme Meintjes[24], Flávia F Mendes[3], Matthias Merker[29], Marina Mihalic[18], James Millard[30], Paolo Miotto[1], Nerges Mistry[28], David AJ Moore[31,8], Kimberlee A Musser[9], Dumisani Ngcamu[20], Nhung N Hoang[32], Stefan Niemann[29, 48], Kayzad Soli Nilgiriwala[28], Camus Nimmo[33], Nana Okozi[20], Rosangela S Oliveira[3], Shaheed Vally Omar[20], Nicholas I Paton[34], Timothy EA Peto[4], Juliana MW Pinhata[3], Sara Plesnik[18], Zully M Puyen[35], Marie Sylvianne Rabodoarivelo[36], Niaina Rakotosamimanana[36], Paola MV Rancoita[13], Priti Rathod[21], Esther Robinson[21], Gillian Rodger[4], Camilla Rodrigues[23], Timothy C Rodwell[37,38], Aysha Roohi[4], David Santos-Lazaro[35], Sanchi Shah[28], Thomas Andreas Kohl[29], E Grace Smith[21,12], Walter Solano[8], Andrea Spitaleri[1,13], Philip Supply[39], Utkarsha Surve[23], Sabira Tahseen[40], Nguyen Thuy Thuong Thuong[32], Guy Thwaites[32,4], Katharina Todt[18], Alberto Trovato[1], Christian Utpatel[29], Annelies Van Rie[41], Srinivasan Vijay[42], Timothy M Walker[4,32], A Sarah Walker[4], Robin M Warren[43], Jim Werngren[44], Maria Wijkander[44], Robert J Wilkinson[45,46,26], Daniel J Wilson[4], Penelope Wintringer[19], Yu-Xin Xiao[22], Yang Yang[4], Zhao Yanlin[10], Shen-Yuan Yao[20], Baoli Zhu[47]

Institutions

1 IRCCS San Raffaele Scientific Institute, Milan, Italy
2 Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
3 Institute Adolfo Lutz, São Paulo, Brazil
4 University of Oxford, Oxford, UK
5 Stanford University School of Medicine, Stanford, USA
6 Scottish Mycobacteria Reference Laboratory, Edinburgh, UK
7 Yale School of Public Health, Yale, USA
8 Universidad Peruana Cayetano Heredia, Lima, Perú
9 Wadsworth Center, New York State Department of Health, Albany, USA
10 Chinese Center for Disease Control and Prevention, Beijing, China
11 Bill & Melinda Gates Foundation, Seattle, USA
12 UK Health Security Agency, London, UK
13 Vita-Salute San Raffaele University, Milan, Italy
14 University of Sydney, Australia
15 The University of British Columbia, Vancouver, Canada
16 Public Health Ontario, Toronto, Canada
17 SYNLAB Gauting, Munich, Germany
18 Institute of Microbiology and Laboratory Medicine, IMLred, WHO-SRL Gauting, Germany
19 EMBL-EBI, Hinxton, UK
20 National Institute for Communicable Diseases, Johannesburg, South Africa
21 Public Health England, Birmingham, UK

22 Taiwan Centers for Disease Control, Taipei, Taiwan
23 Hinduja Hospital, Mumbai, India
24 University of Cape Town, Cape Town, South Africa
25 University of Surrey, Guildford, UK
26 Imperial College, London, UK
27 Université de Montréal, Canada
28 The Foundation for Medical Research, Mumbai, India
29 Research Center Borstel, Borstel, Germany
30 Africa Health Research Institute, Durban, South Africa
31 London School of Hygiene and Tropical Medicine, London, UK
32 Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam
33 University College London, London, UK
34 National University of Singapore, Singapore
35 Instituto Nacional de Salud, Lima, Perú
36 Institut Pasteur de Madagascar, Antananarivo, Madagascar
37 FIND, Geneva, Switzerland
38 University of California, San Diego, USA
39 Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France
40 National TB Reference Laboratory, National TB Control Program, Islamabad, Pakistan
41 University of Antwerp, Antwerp, Belgium
42 University of Edinburgh, Edinburgh, UK
43 SAMRC Centre for Tuberculosis Research, Stellenbosch University, Cape Town, South Africa
44 Public Health Agency of Sweden, Solna, Sweden
45 Wellcome Centre for Infectious Diseases Research in Africa, Cape Town, South Africa
46 Francis Crick Institute, London, UK
47 Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
48 German Center for Infection Research (DZIF), Hamburg-Lübeck-Borstel-Riems, Germany

Additional authors contributing to the CRyPTIC consortium (in alphabetical order)

Irena Arandjelovic[1], Anna Barbova[2], Maxine Caws[3], Iñaki Comas[4], Roland Diel[5], Carla Duncan[6], Sarah Dunstan[7], Maha Farhat[8], Margaret M Fitzgibbon[9], Victoria Furió[10], Jennifer Gardy[11], Jennifer Guthrie[6], Dang Thi Minh Ha[12], Kathryn Holt[13], Michael Inouye[14], Frances B Jamieson[6], SM Mostofa Kamal[15], Julianne V Kus[6], Vanessa Mathys[16], Rick Twee-Hee Ong[17], Youwen Qin[7,14], Thomas R Rogers[9,19], Gian Maria Rossolini[20], Emma Roycroft[9], Vitali Sintchenko[21], Alena Skrahina[22], Yik Ying Teo[17], Phan Vuong Khac Thai[12], Dick van Soolingen[23], Mark Wilcox[24], Matteo Zignol[25]

Institutions

1 University of Belgrade, Belgrade, Serbia
2 National Institute of phthisiology and pulmonology NAMS Ukraine, Kyiv
3 Liverpool School of Tropical Medicine,United Kingdom
4 Biomedicine Institute of Valencia IBV-CSIC, Spain
5 University Medical Hospital Schleswig-Holstein, Germany
6 Public Health Ontario, Toronto, Canada
7 University of Melbourne, Australia
8 Harvard Medical School, Boston, USA
9 Irish Mycobacteria Reference Laboratory, Dublin, Ireland

10 Universitat de València, Spain
11 Bill & Melinda Gates Foundation, Seattle, USA
12 Pham Ngoc Thach Hospital, Ho Chi Minh City, Vietnam
13 Monash University, Melbourne, Australia
14 Baker Institute, Melbourne, Australia
15 National Institute of Diseases of the Chest and Hospital, Dhaka, Bangladesh
16 Sciensano, Belgian reference laboratory for M. tuberculosis
17 National University of Singapore, Singapore
19 Trinity College Dublin, Ireland
20 Careggi University Hospital, Florence, Italy
21 University of Sydney, Australia
22 Republican Scientific and Practical Centre for Pulmonology and TB, Minsk, Belarus
23 National Institute for Public Health and the Environment, Bilthoven, The Netherlands
24 Leeds Teaching Hospital NHS Trust, Leeds, United Kingdom
25 World Health Organization, Geneva

Author contributions

DAC, DMC, DWC, HH, SJH, NAI, NM, DM, SN, TEAP, CR, GS, PS, GT, ASW, TMW, DJW, ZY contributed to high-level CRyPTIC study design. ASL, TMW, DWC, TEAP, ASW, PWF, DAC designed the specifics of this study. MH, JK, ZI and PWF retrieved and processed genotypic data including kmers. PWF retrieved and processed phenotypic data. ASL, DAC, TMW, SK, YY, PWF, developed the machine learning system. All other authors contributed to the generation of data. ASL and TMW performed all the analysis. ASL and TMW wrote the manuscript with all authors offering feedback.

Acknowledgements & funding

This work was supported by Wellcome Trust/Newton Fund-MRC Collaborative Award (200205/Z/15/Z); and Bill & Melinda Gates Foundation Trust (OPP1133541). Oxford CRyPTIC consortium members are funded/supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), and the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, a partnership between Public Health England and the University of Oxford. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, Public Health England or the Department of Health and Social Care. J.M. is supported by the Wellcome Trust (203919/Z/16/Z). Z.Y. is supported by the National Science and Technology Major Project, China Grant No. 2018ZX10103001. K.M.M. is supported by EMBL's EIPOD3 programme funded

by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska Curie Actions. T.C.R. is funded in part by funding from Unitaid Grant No. 2019-32-FIND MDR. R.S.O. is supported by FAPESP Grant No. 17/16082-7. L.F. received financial support from FAPESP Grant No. 2012/51756-5. B.Z. is supported by the National Natural Science Foundation of China (81991534) and the Beijing Municipal Science & Technology Commission (Z201100005520041). N.T.T.T. is supported by the Wellcome Trust International Intermediate Fellowship (206724/Z/17/Z). G.T. is funded by the Wellcome Trust. R.W. is supported by the South African Medical Research Council. J.C. is supported by the Rhodes Trust and Stanford Medical Scientist Training Program (T32 GM007365). T.C. has received grant funding and salary support from US NIH, CDC, USAID and Bill and Melinda Gates Foundation. L.G. was supported by the Wellcome Trust (201470/Z/16/Z), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number 1R01AI146338, the GOSH Charity (VC0921) and the GOSH/ICH Biomedical Research Centre (www.nihr.ac.uk). A.L. is supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Respiratory Infections at Imperial College London. S.G.L. is supported by the Fonds de Recherche en Santé du Québec. C.N. is funded by Wellcome Trust Grant No. 203583/Z/16/Z. A.V.R. is supported by Research Foundation Flanders (FWO) under Grant No. G0F8316N (FWO Odysseus). G.M. was supported by the Wellcome Trust (098316, 214321/Z/18/Z, and 203135/Z/16/Z), and the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa (Grant No. 64787). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of this report. The opinions, findings and conclusions expressed in this manuscript reflect those of the authors alone. A.B. is funded by the NDM Prize Studentship from the Oxford Medical Research Council Doctoral Training Partnership and the Nuffield Department of Clinical Medicine. D.J.W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant No. 101237/Z/13/B) and by the Robertson Foundation. A.S.W. is an NIHR Senior Investigator. T.M.W. is a Wellcome Trust Clinical Career Development Fellow (214560/Z/18/Z). A.S.L. is supported by the Rhodes Trust. R.J.W. receives funding from the Francis Crick Institute which is supported by Wellcome Trust, (FC0010218), UKRI

(FC0010218), and CRUK (FC0010218). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. Parts of the work were funded by the German Center of Infection Research (DZIF). The Scottish Mycobacteria Reference Laboratory is funded through National Services Scotland. The Wadsworth Center contributions were supported in part by Cooperative Agreement No. U60OE000103 funded by the Centers for Disease Control and Prevention through the Association of Public Health Laboratories and NIH/NIAID grant AI-117312. Additional support for sequencing and analysis was contributed by the Wadsworth Center Applied Genomic Technologies Core Facility and the Wadsworth Center Bioinformatics Core. SYNLAB Holding Germany GmbH for its direct and indirect support of research activities in the Institute of Microbiology and Laboratory Medicine Gauting. N.R. thanks the Programme National de Lutte contre la Tuberculose de Madagascar.

Competing Interest

E.R. is employed by Public Health England and holds an honorary contract with Imperial College London. I.F.L. is Director of the Scottish Mycobacteria Reference Laboratory. S.N. receives funding from German Center for Infection Research, Excellenz Cluster Precision Medicine in Chronic Inflammation, Leibniz Science Campus Evolutionary Medicine of the LUNG (EvoLUNG)tion EXC 2167. P.S. is a consultant at Genoscreen. T.R. is funded by NIH and DoD and receives salary support from the non-profit organization FIND. T.R. is a co-founder, board member and shareholder of Verus Diagnostics Inc, a company that was founded with the intent of developing diagnostic assays. Verus Diagnostics was not involved in any way with data collection, analysis or publication of the results. T.R. has not received any financial support from Verus Diagnostics. UCSD Conflict of Interest office has reviewed and approved T.R.'s role in Verus Diagnostics Inc. T.R. is a co-inventor of a provisional patent for a TB diagnostic assay (provisional patent #: 63/048.989). T.R. is a co-inventor on a patent associated with the processing of TB sequencing data (European Patent Application No. 14840432.0 & USSN 14/912,918). T.R. has agreed to "donate all present and future interest in and rights to royalties from this patent" to UCSD to ensure that he does not receive any financial benefits from this

patent. S.S. is working and holding ESOPs at HaystackAnalytics Pvt. Ltd. (Product: Using whole genome sequencing for drug susceptibility testing for Mycobacterium tuberculosis). G.F.G. is listed as an inventor on patent applications for RBD-dimer-based CoV vaccines. The patents for RBD-dimers as protein subunit vaccines for SARS-CoV-2 have been licensed to Anhui Zhifei Longcom Biopharmaceutical Co. Ltd, China. No other authors declare a conflict of interest.

Acknowledgements – people

We thank Faisal Masood Khanzada and Alamdar Hussain Rizvi (NTRL, Islamabad, Pakistan), Angela Starks and James Posey (Centers for Disease Control and Prevention, Atlanta, USA), and Juan Carlos Toro and Solomon Ghebremichael (Public Health Agency of Sweden, Solna, Sweden).

Data and code availability

All data used in this manuscript are publicly available on the European Bioinformatics Institute (http://ftp.ebi.ac.uk/pub/databases/cryptic/). No custom code submitted for peer review.

Wellcome Trust Open Access

This research was funded in part, by the Wellcome Trust/Newton Fund-MRC Collaborative Award [200205/Z/15/Z]. 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.

This research was funded, in part, by the Wellcome Trust [214321/Z/18/Z, and 203135/Z/16/Z]. 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.We thank Faisal Masood Khanzada and Alamdar Hussain Rizvi (NTRL, NTP Islamabad, Pakistan), and Angela Starks and James Posey (Centers for Disease Control and Prevention, Atlanta, USA).

Figures and Tables

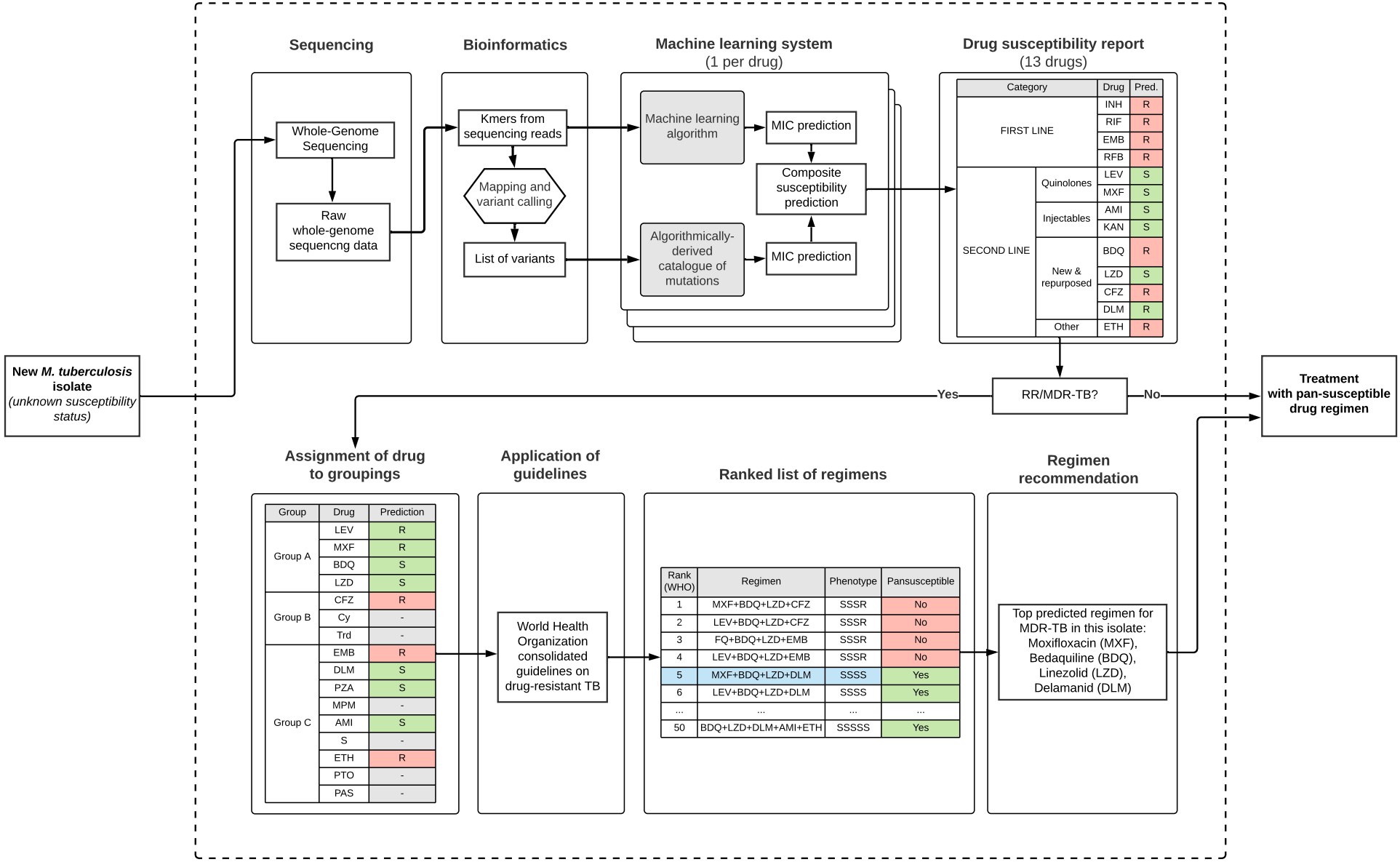
All primary and supplementary figures and tables are presented in separate files (*Figures.pdf* and *Tables.xlsx*).

References

1. World Health Organization. *Global Tuberculosis Report*. https://www.who.int/publications/i/item/9789240013131 (2020).
2. Dheda, K. *et al.* The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir. Med.* **5**, 291–360 (2017).
3. World Health Organization. *Impact of the COVID-19 pandemic on TB detection and mortality in 2020*. https://www.who.int/publications/m/item/impact-of-the-covid-19-pandemic-on-tb-detection-and-mortality-in-2020 (2021).
4. World Health Organization. *The End TB Strategy*. https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19 (2015).
5. Cao, Y. *et al. Xpert MTB/XDR: A ten-color reflex assay suitable for point of care settings to detect isoniazid-, fluoroquinolone-, and second line injectable drug-resistance directly from* Mycobacterium tuberculosis *positive sputum*. http://biorxiv.org/lookup/doi/10.1101/2020.09.08.288787 (2020).
6. Quan, T. P. *et al.* Evaluation of Whole-Genome Sequencing for Mycobacterial Species Identification and Drug Susceptibility Testing in a Clinical Setting: a Large-Scale Prospective Assessment of Performance against Line Probe Assays and Phenotyping. *J. Clin. Microbiol.* **56**, e01480-17 (2018).
7. The CRyPTIC Consortium and the 100, 000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).
8. Georghiou, S. B. *et al.* Guidance for Studies Evaluating the Accuracy of Rapid Tuberculosis Drug-Susceptibility Tests. *J. Infect. Dis.* **220**, S126–S135 (2019).
9. Kontsevaya, I. *et al.* Perspectives for systems biology in the management of tuberculosis. *Eur. Respir. Rev.* **30**, 200377–200377 (2021).
10. World Health Organization. Target product profile for next-generation tuberculosis drug-susceptibility testing at peripheral centres. https://www.who.int/publications-detail-redirect/9789240032361 (2021).
11. Denkinger, C., Schito, M. & Pai, M. *The Journal of Infectious Diseases Tuberculosis Diagnostics in 2015: Landscape, Priorities, Needs, and Prospects*. http://jid.oxfordjournals.org/.
12. Kouchaki, S. *et al.* Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* **35**, 2276–2282 (2019).
13. Chen, M. L. *et al.* Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine* **43**, 356–369 (2019).
14. Yang, Y. *et al.* Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34**, 1666–1671 (2018).
15. Deelder, W. *et al.* Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. *Front. Genet.* **10**, (2019).
16. Davis, J. J. *et al.* Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.* **6**, 27930–27930 (2016).
17. Santerre, J. W., Davis, J. J., Xia, F. & Stevens, R. Machine Learning for Antimicrobial

Resistance. *arXiv* **48**, 18972A-18972C (2016).
18. Drouin, A. *et al.* Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* **17**, 1–15 (2016).
19. Nguyen, M. *et al.* Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal Salmonella. *J. Clin. Microbiol.* **57**, 1–15 (2018).
20. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. https://www.who.int/publications-detail-redirect/9789240028173.
21. World Health Organization. *WHO consolidated guidelines on drug-resistant tuberculosis treatment*. (2019).
22. Kaniga, K. *et al.* Validation of bedaquiline phenotypic drug susceptibility testing methods and breakpoints: A multilaboratory, multicountry study. *J. Clin. Microbiol.* **58**, 1677–1696 (2020).
23. van Belkum, A. *et al.* Developmental roadmap for antimicrobial susceptibility testing systems. *Nat. Rev. Microbiol.* **17**, 51–62 (2019).
24. United States Food and Drug Administration. *Artificial Intelligence and Machine Learning in Software as a Medical Device*. (2021).
25. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 118 (2020).
26. Walter, K. S. *et al.* Genomic variant-identification methods may alter mycobacterium tuberculosis transmission inferences. *Microb. Genomics* **6**, 1–16 (2020).
27. Colangeli, R. *et al.* Bacterial Factors That Predict Relapse after Tuberculosis Therapy. *N. Engl. J. Med.* **379**, 823–833 (2018).
28. The CRyPTIC Consortium. Epidemiological cutoffs for a 96-well broth microtitre plate for high-throughput research antibiotic susceptibility testing of M . tuberculosis. *medRxiv* (2021) doi:10.1101/2021.02.24.21252386.
29. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 1–8 (2016).
30. Walker, T. M. *et al.* Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

Figure 1: Illustration of the machine learning system and regimen prediction workflow

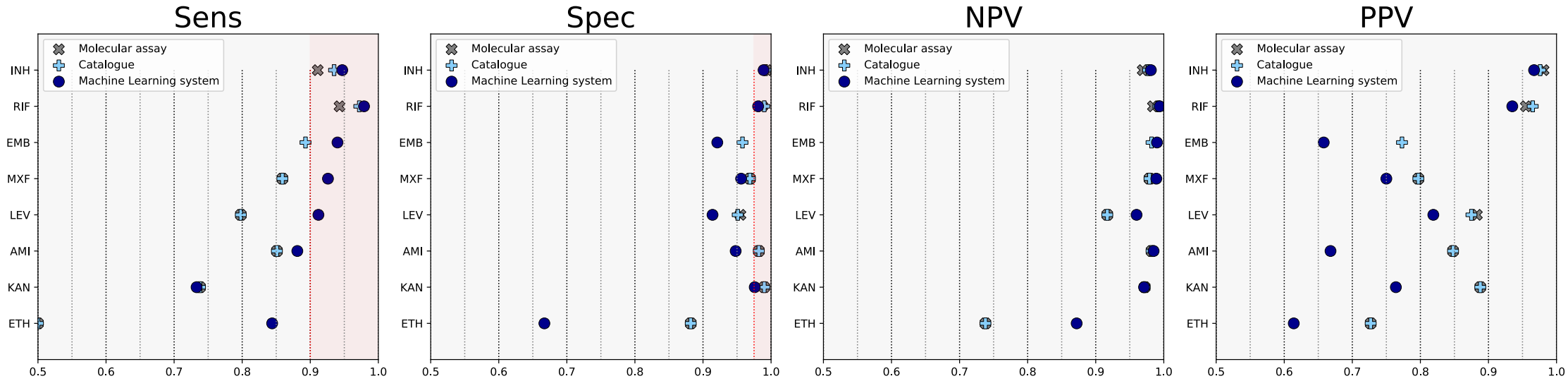


*Drug names: INH: Isoniazid, RIF: Rifampin, EMB: Ethambutol, MXF: Moxifloxacin, LEV: Levofloxacin, AMI: Amikacin, KAN: Kanamycin, ETH: Ethionamid, RFB: Rifabutin, BDQ: Bedaquilin, CFZ: Clofazimine, DLM: Delamanid, LZD: Linezolid

†Other acronyms: RR-TB: Rifampin-resistant tuberculosis; MDR-TB: Multidug-resistant tuberculosis; MIC: Minimum Inhibitory Concentration

†† Guidelines and drug groupings refer to World Health Organization Consolidated Guidelines on Tuberculosis (2020); "In multidrug- or rifampicin-resistant tuberculosis (MDR/RR-TB) patients on longer regimens, all three Group A agents and at least one Group B agent should be included […] If only one or two Group A agents are used, both Group B agents are to be included. If the regimen cannot be composed with agents from Groups A and B alone, Group C agents are added to complete it."

Figure 2: Performance of the machine learning prediction on the independent test set, and comparison to the catalogue and molecular assay



*Drug names: INH: Isoniazid, RIF: Rifampin, EMB: Ethambutol, MXF: Moxifloxacin, LEV: Levofloxacin, AMI: Amikacin, KAN: Kanamycin, ETH: Ethionamid, RFB: Rifabutin; new and repurposed drugs are not included as they are not present in the independent set

†Other acronyms: Sens: Sensitivity; Spec: Specificity; NPV: Negative Predictive Value; PPV: Positive Predictive Value

†† Areas shaded in red correspond to a sensitivity of 90% and a specificity of 98%; this correspond to the target specificity for all drugs and sensitivity for isoniazid and quinolones per the “Target product profile for next-generation tuberculosis drug-susceptibility testing at peripheral centres” of the World Health Organization (2021); sensitivity targets are 95% for rifampicinand 80% for other second-line agents

Figure 3: Simulated negative predictive values of the machine learning system for different resistance prevalences

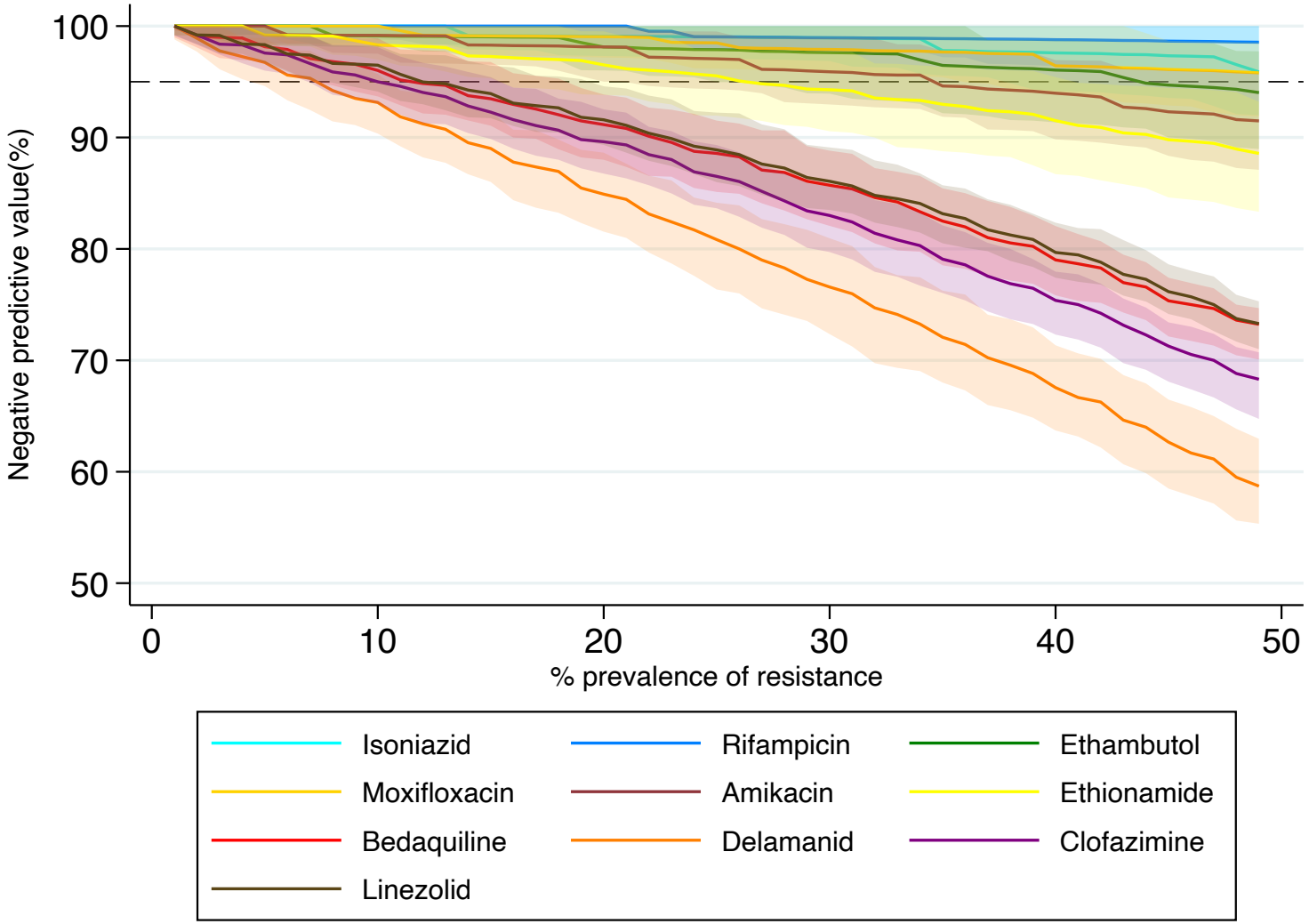
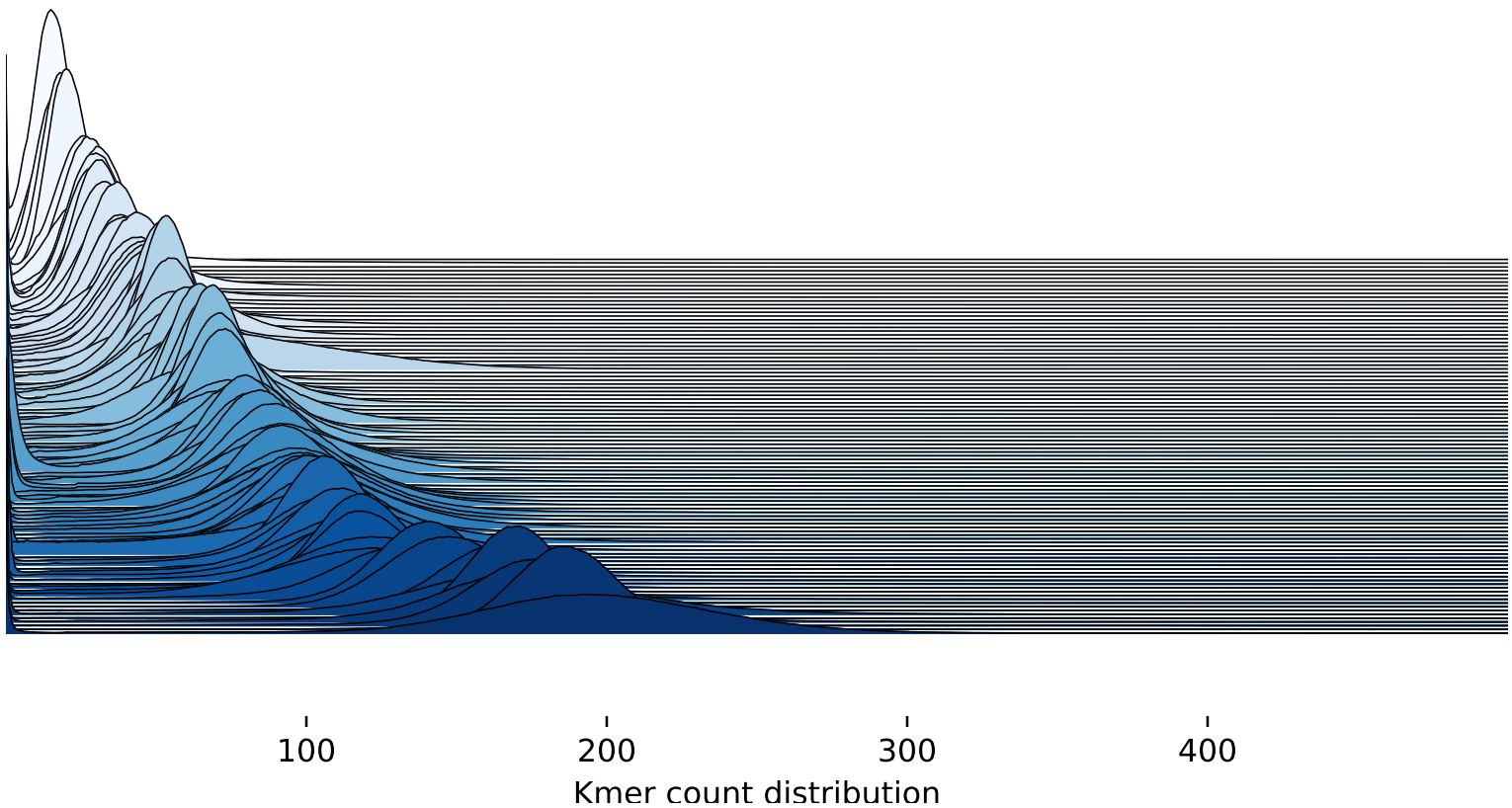
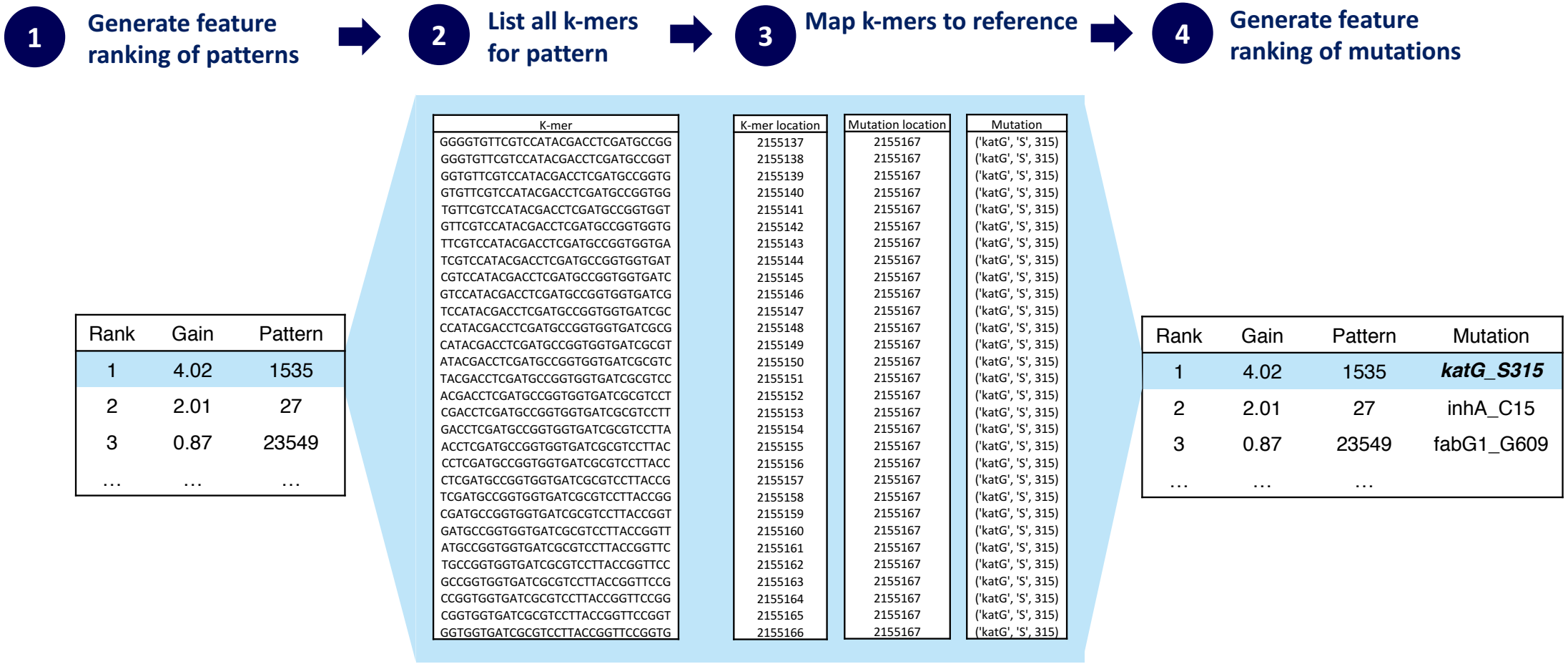


Figure S1: Illustration of kmer distributions across isolates used as features for the machine learning model



This figure illustrates kmer frequency distribution for a subset of 100 kmers for a single isolate. After the generation of whole genome sequencing data, instead of the alignment and variant calling process, each read was analyzed, and decomposed into a series of 31-mers. In total, each isolate could be described by 3 to 5 million unique 31-mers, present an average of 50 to 200 times each. Some kmers are present <5 times - as seen on this figure. This is likely the result of a sequencing mistakes. One of the disadvantages of using kmers from reads, as opposed to assembled genomes, is the lack of any error processing. To reduce the influence of sequencing errors on our analysis and on the machine learning system, all kmers present five times or fewer were removed from the dataset using methodology presented in Earle et al. (2016).

Figure S2: Illustration of the pipeline for kmer feature analysis after machine learning system training



After the machine learning system was trained, a feature analysis was performed to identify potential mutations of interest. Since our features are patterns of kmers (see Figure S3), we follow four steps to achieve this: (1) rank pattern features by importance, (2) list all kmers corresponding to each pattern, (3) map each kmer to the reference MTB H37Rv reference genome to find its location, and (4) identify mutation for each pattern, and generate final feature list. The steps are illutrated in this figure. Step 1 is performed using feature ranking methods based on weight. In step 2, we use a custom-built pattern-k-mer mapper to find all kmers corresponding to each pattern that was ranked. Step 3 - the mapping of all kmers to areference genome, requires several tools. We start by converting text kmers to a.fastafile format, using the open-source BioPython package version 1.73. We proceed to mapping the fastafile to the H37Rv reference genome (GenBank NC000962.2) using BowTie2 version 2.3.5. From the newly generated samfile containing the alignment results, we rank each kmer coordinates in location order, identify the mutation location (if present), and finally look-up the gene and protein corresponding to that mutation. We finally compile all the computed genes into the original feature table, which constitute our final feature table. We note that in several cases, kmers without a mutation -i.e. mapping perfectly to the reference genome - will be used by the models. These are features used by the model to predict susceptibility, rather than resistance, but remain important features to consider for feature analysis. Results of this analysis are presented in table S11.