Feature selection and aggregation for antibiotic resistance GWAS in *Mycobacterium tuberculosis*: a comparative study.

Reshetnikov K.O.[1,2,3], Bykova D.I.[1,4], Kuleshov K.V.[1,5], Chukreev K.[1], Guguchkin E.P.[2,4], Akimkin V.G.[1], Neverov A.D.[1] and Fedonin G.G.[1,3,6 *]

1. Central Research Institute of Epidemiology, Moscow, Russian Federation
2. Ivannikov Institute for System Programming of the Russian Academy of Sciences (ISP RAS), Moscow, Russian Federation
3. Moscow Institute of Physics and Technology (National Research University), Moscow, Russian Federation
4. Moscow State University, Moscow, Russian Federation
5. Federal State Budget Scientific Institution “Federal Scientific Center VIEV”, Moscow, Russian Federation
6. Institute for Information Transmission Problems of the Russian Academy of Sciences (IITP RAS), Moscow, Russian Federation

* Corresponding author: gennady.fedonin@gmail.com

Abstract

Drug resistance (DR) remains a global healthcare concern. In contrast to other human bacterial pathogens, acquiring mutations in the genome is the main mechanism of drug resistance for *Mycobacterium tuberculosis* (MTB). For some antibiotics resistance of a particular isolate can be predicted with high confidence knowing whether specific mutations occurred, but for some antibiotics our knowledge of resistance mechanism is moderate. Statistical machine learning (ML) methods are used in attempts to infer new genes implicated in drug resistance. These methods use large collections of isolates with known whole-genome sequences and resistance status for different drugs. However, high correlations between the presence or absence of resistance to drugs that are used together in one treatment regimen complicate inference of causal mutations by traditional ML. Recently, several new methods were suggested to deal with the problem of correlations of response variables in training data. In this study, we applied the following methods to tackle the confounding effect of resistance co-occurrence in a dataset of approximately 13 000 complete genomes of MTB with characterized resistance status for 13 drugs: logistic regression with different regularization penalty functions, a polynomial-time algorithm for best-subset selection problem (ABESS), and “Hungry, Hungry SNPos” (HHS) method. We compared these methods by the ability to select known causal mutations for the resistance to each particular drug and not to select mutations in genes that are known to be associated with resistance to other drugs. ABESS significantly outperformed the others selecting more relevant sets of mutations. We also showed that aggregation of rare mutations into features

indicating changes of PFAM domains increased the quality of prediction and these features were majorly selected by ABESS.

Keywords

M. tuberculosis, drug resistance, feature selection, machine learning, PFAM domains, drug resistance.

Data Summary

The dataset unifies characterized whole-genome sequences of *M. tuberculosis* from multiple studies [1–10]. Short Illumina reads are available in public repositories (SRA or ENA). Sample ids, phenotypes and links to the source papers are summarized and listed in Table S1.

Abbreviations

MTB, *Mycobacterium tuberculosis*; DR, drug resistance; MDR, multidrug resistant; XDR, extensively drug resistant; WHO, World Health Organization; NGS, Next Generation Sequencing; PCR, polymerase chain reaction; GWAS, genome-wide association study; DA, direct association; ML, machine learning; PCA, principal component analysis; HMM, hidden Markov model; MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation; HHS, ‘hungry-hungry SNPs‘; ABESS, polynomial-time algorithm for best-subset selection problem

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (MTB), remains the leading cause of death from single infectious agents in spite of decreasing annual numbers of TB deaths in recent years [11]. TB is a curable disease, which, in most cases, could be effectively treated by a complex of drugs, there is a great concern about increasing fraction of isolates resistant to the most potent anti-TB agents rifampicin (RIF) and isoniazid (INH) called multidrug resistant (MDR) as well as extensively resistant (XDR) isolates which are additionally resistant to any aminoglycosides and fluoroquinolones [11]. According to the WHO estimates, the numbers of multidrug-resistance tuberculosis (MDR-TB) and rifampicin resistance tuberculosis (RR-TB) cases detected had increased from 2018 to 2019 by 10% [11]. The average efficiency of treatment of MDR-TB over the world is about 60% and for XDR-TB is about 20%, and the prognosis of the life span for treatment failures is negative [11]. Today, the two basic schemas of TB therapy are applied [12], the naive patients are effectively treated by a combination of first-line drugs rifampicin, isoniazid, pyrazinamide (PZA), ethambutol (EMB), and streptomycin (STR), and relapse cases are treated by the second-line drugs fluoroquinolones: ofloxacin (OFX), levofloxacin (LEV), moxifloxacin (MOX) and ciprofloxacin (CIP) and injectable drugs kanamycin (KAN), amikacin (AMK) and capreomycin (CAP). There are other less effective second-line anti-TB drugs: ethionamide (ETH)/prothionamide (PTH), cycloserine (CS), P-aminosalicylic acid (PAS)

Early detection of resistant cases is necessary to correct therapy and prevent transmission of resistant strains because patients infected with drug resistant strains are harder to treat and easier to transmit disease [13]. Acquisition of mutations in genes employed in processes which are disturbed by drugs is a main mechanism of drug resistance in MTB [14]. Moreover, if MTB is resistant to one drug, it acquires resistance to other drugs easier [15].

Measuring the minimal concentrations of a substance inhibiting bacterial growth is the 'gold standard' for assessing resistance/susceptibility levels for drugs [16]. Because MTB is a slow-growing bacteria, phenotypic tests are time-consuming, expensive, and could be performed in laboratories with high biosafety levels only. Thus, the genotypic tests have been gradually replacing it. PCR-based molecular tests could be used for MTB DNA extracted directly from clinical samples. MTB GeneXpert MTB/RIF, line probe assays MTBDRplus and MTBDRsl, detect the presence of frequent mutations in a few drug associated genes within two hours and have high DR prediction accuracy for RIF and INH, but lower for other drugs [17–19]. Meanwhile, molecular methods based on sequencing allow for detection of rare mutations and have higher accuracy [20]. Despite the fact that sequencing tests usually require the MTB culturing stage, they are less time and cost consuming than standard laboratory diagnostics [21]. In Great Britain [22] and in some other high-income countries [23] WGS-based tests are adopted for use for drug susceptibility testing before prescription of therapy. WHO recently released the technical guide for use of Next Generation Sequencing (NGS) for the detection of mutations associated with MTB drug resistance [24].

Most methods of MTB phenotype prediction by genotype utilize the direct association (DA) approach based on the interpretation of mutations as benign or causal [20]. Prior knowledge of genes, which products are likely implicated in drug action, allows the identification of drug-associated causal mutations. A pioneering DA approach was developed by Walker et. al. [9]: for each drug the method iteratively identifies mutations in a relevant subset of preselected genes which explain resistant phenotype of at least one previously unexplained isolate in the training set. Some other catalogs of resistance causal mutations were developed based on thorough literature reviews [23, 25, 26], machine learning methods [27] and comparative genomics [28].

Direct association models perform well for some drugs, e.g. rifampicin and isoniazid, which action mechanisms are well known, however, for some drugs, e.g. pyrazinamide, ethambutol, fluoroquinolones, mutations in known genes do not always explain phenotypes well [20, 29].

Many approaches were used to identify new loci in the MTB genome associated with drug resistance by genome-wide association studies (GWAS) that process datasets of almost complete-genome genotypes with characterized drug phenotypes. Some methods are based on the detection of sites in the genome with recurrent mutations in phylogenetically unrelated strains [1–3, 30, 31], others use machine learning (ML) approaches: k nearest neighbors [32], regularized logistic regression [32, 33], support vector machine [4, 33, 34], gradient-boosted trees [32, 35], random forest [4, 32], mixture models [4] and linear mixed models that account for the clonal structure of MTB population [3, 5, 36]. New methods are constantly being invented, for example, the one based on boolean compressed sensing [37]

Although a few loci identified in GWAS studies were experimentally confirmed as causal [2, 30, 31] others require verification and may be spuriously associated. Likely, the number of potential DR causal loci is large [5], e. g. in experiments with *E. coli*, high levels of resistance could be achieved by modulation of expression of a large number of functionally diverse genes [38], therefore, discovering numerous causal mutations is a hard objective.

Compared to the DA approach, the accuracy of drug phenotype prediction could be significantly improved by the application of ML for mutations occurring in known drug-associated genes [6, 20, 27, 34, 39, 40]. The improvement is partly achieved due to high correlation of DR phenotypes for drugs usually prescribed in common [6, 39], so identification in an isolate a specific mutation that causes resistance to one drug informs model about high probability of co-resistance to other drugs, e.g. the mutation katG S315T conferring resistance to INH also usually arises before acquisition of RIF resistance [41, 42] and is a good predictor for MDR [43, 44].

The model trained on data with certain proportions of MDR/XDR could be less accurate in predicting resistance for the data with different proportions of MDR/XDR [6] which may lead to the poor prediction quality of the models trained on clinical isolates for in-vitro selection experiments or for regions with different proportions of resistant TB. Therefore, such models will have high prediction accuracies only if regimens of usage of drugs will be the same in the future. But the WHO recommendations for treatment are constantly changing and so does the level of drug resistance in the population. Also, models based on resistance co-occurrence are unable to capture rare causal mutations because covariates would mask weak signals [39]. Finally, such models hardly produce new knowledge about the mechanisms of drug resistance.

Along with the resistance co-occurrence, the curse of dimensionality is one of the major problems in GWAS. Yang et al. among the others [6, 32] applied principal component analysis (PCA) to MTB DR drug prediction reducing dimensions, but separability between INH-resistant and susceptible classes was poor [39]. Artificial neural networks, including deep learning and autoencoders, are widely treated as automatic feature generation and were successfully applied to MTB DR prediction [20, 40]. These models, potentially, account for complex interactions between mutations. However, the improvement of prediction accuracy of such complex models in comparison to linear regression models is modest or insignificant [20] while interpretation of such models is nontrivial. Matrix factorization methods are another example of feature generation for dimensionality reduction. This approach could be unified with feature selection, as it was done recently in [45] for MTB DR prediction.

Aggregation of mutations with the subsequent exclusion of rare mutations is another way of dimensionality reduction by feature generation. It is typically done by constructing features indicating the presence of mutations in a particular gene or in some regulatory region [20]. Such indicators reduce dimensions without loss of interpretability and allow accounting for mutations that were not present in the training set. In this study, we tested HMM models from the PFAM database as a mechanism of aggregation.

Feature selection techniques try to select a subset of initial features, optimal in some sense. For large dimensions, however, enumeration of subsets is computationally intractable. Different methods vary in heuristics they use to evade the full enumeration. In this paper, we concentrate on feature selection with linear models. We compare multiple selection methods both in terms of prediction accuracy and biological relevance. We have chosen logistic regression with L1, minimax concave penalty (MCP) and smoothly clipped absolute deviation (SCAD) regularization, 'Hungry, Hungry SNPos' (HHS), and a polynomial-time algorithm for best-subset selection problem (ABESS). Below is a brief description of these methods.

The logistic regression model searches for additive coefficients $\overline{\beta}$ of genomic variations to maximize the likelihood of the dataset - the probability to observe the dataset with given $\overline{\beta}$. Wald and likelihood statistics allow estimating statistical significance of coefficients $\overline{\beta}$ [46].

Likelihood maximization tends to assign nonzero weights to all features, underestimating the impact of causative features and overfitting the noise. Also, due to correlations with benign features, the causative ones may be lost in the process of the selection. Lasso regularization [47] adds a linear regularization term penalizing large coefficients which are also known as the L1 norm regularization. As a result, only one feature is chosen from a subset of highly correlated ones. Ridge regression [48, 49] penalizes the squares of $\overline{\beta}$ coefficients (also known as the L2 norm), while elastic net combines both penalties. Lasso was shown to outperform unregularized stepwise feature selection both in the quality of prediction and in the percentage of causative features in the selected set on simulated data [47, 50]. The lasso achieves better or even performance in simulated bacterial GWAS compared to LMM models even in the presence of high LD (strong population structure) [51].

Penalizing large coefficients comes at a price: large true coefficients will be punished. It's obvious that any regularization will have its own drawbacks and limitations, but new regularization terms are constantly being invented. Fan and Li developed SCAD regularization which has a nonlinear decreasing penalty for large coefficients [52]. A similar effect is achieved by MCP regularization [53]. MCP and SCAD feature selection outperformed lasso on all types of simulated data except for non-sparse data with a low signal-to-noise ratio, in which case the elastic net was the leader [54].

Some methods penalize the number of nonzero coefficients (L0 norm regularization). The functional being optimized is no longer differentiable: enumeration of feature subsets is required. ABESS goes further and can optimize various information criteria (AIC, BIC, cross-validation, and an original asymptotic criterion - SIC) searching for the best subset [55]. For each feature set size from 1 to the user-defined maximum, the best subset is constructed and the final subset is selected based on the given information criterion. The best subset of a given size is constructed by starting from the features which are the most correlated with the target variable and iteratively improving the solution by scoring all the features with the score dependent on the currently selected set and exchanging the worst features to the best ones. This procedure is shown to converge in a finite number of steps [55]. The authors showed that ABESS outperformed SCAD, MCP, and Lasso regularization on simulated data.

Authors also provide a deep theoretical analysis of the ABESS algorithm studying the sufficient conditions of correctness and proving that if such conditions hold the algorithm's running time is polynomial. Like LASSO, SCAD, and MCP the ABESS algorithm finds the correct model if the correlations between features are not too big as well as the true number of causative features, the true coefficients are not too small while the data is not too sparse. All the methods have difficulties when there are a lot of small effects [55].

Along with well-known and theoretically grounded approaches described above, new heuristic methods are being invented for GWAS analysis which try to cope with correlations between features by taking into account the population structure [56]. HHS algorithm starts with assigning a score reflecting the correlation with the phenotype corrected to the population structure to each SNP. The score is adjustable to the class imbalance in the phenotype and allele frequencies. Similar to ABESS the currently selected set of features (all the features with positive scores) is optimized iteratively by decreasing the normalized scores of mutually correlated features in the selected set until the convergence [56].

Methods

Dataset preparation

The dataset unifies characterized whole-genome sequences of *M. tuberculosis* from multiple studies [1–10]. Illumina reads were trimmed and mapped onto the H37Rv reference genome. Variants were called with GATK HaplotypeCaller and multiple filtering steps were applied.

Nucleotide sequences of each protein-coding gene were translated and aligned to the reference protein sequences with extreme care to take into account possible frameshifts, start or stop codon loss, etc. The mutated gene was considered nonfunctional if it was 50% shorter or 30% longer than the reference protein: no mutations were considered in such genes. Finally, we obtained a list of mutations relative to the reference genome: mutations in noncoding regions, missense mutations, loss of function mutations, and aggregated features (see below). Synonymous mutations were excluded from the analysis.

The final dataset was split into smaller sets containing only samples with known phenotypes for each antibiotic. Both aggregation and machine learning were performed separately for each dataset.

See "Dataset and raw data processing" in Supplementary Methods for details.

Aggregation of mutations

We examined three types of aggregated features: an indicator of the presence of any mutation in a gene, an indicator that the gene is broken, i.e. it has a loss of function mutation, and features indicating alternation of PFAM domains.

If there is any mutation in a gene then the indicator of the presence of mutations in this gene takes the value of “1” and “0” otherwise. Further, this approach is called “gene aggregation”. “Broken gene” is another binary feature. Gene is considered to be ‘broken’ if the length of its open reading frame (ORF) is decreased due to mutations by half or increased by 30% relative to the annotated ORF length.

To generate PFAM domain features we used pre-trained models available for 3369 *M. tuberculosis* genes in the PFAM database. Using these models for each genome sequence we computed domain scores which were probabilities that domain sequences had been generated by corresponding domain HMM models. We then converted scores into binary variables selecting for each domain and for each drug separately a threshold optimizing separation of resistant and susceptible sequences in the training subset of data. If a domain score is over the threshold then the domain is considered to be changed and the corresponding binary variable takes the value of “1” and it takes the value “0” otherwise (see “PFAM domains based feature generation” in Supplementary methods).

Estimating the impact of aggregation

To measure the effect of aggregation we evaluated results of classification of genotypes into resistant and susceptible phenotypic groups by logistic regression with L1 regularization using different feature set combinations. The basic feature set comprised SNPs and Indels without aggregation. Three other sets contain aggregating features and mutations that happened more often than 2 times: 1) SNPs and Indels with PFAM features and excluding rare mutations; 2) SNPs and Indels with “broken gene” features and excluding rare mutations; 3) SNPs and Indels with “gene aggregation” features and excluding rare mutations.

ROC AUC and F1 scores were computed by 5-fold cross-validation: the dataset was split randomly into 5 folds, each fold was used as a testing set one at a time while the rest 4 were used for training of logistic regression, the scores computed on test folds were averaged. Statistical significance of the differences of scores between the models trained on different feature sets was estimated by repeating this procedure 100 times and calculating Wilcoxon paired rank test between two vectors of length 100 of corresponding scores. The null model assumes that there is no difference between the performance of logistic regression on different feature sets and the difference of scores obtained on 5-fold cross-validation is explained only by the lucky dataset split. We considered the differences to be significant if the p-value with Bonferroni correction to the number of drugs was lower than 5%.

Feature selection and model quality evaluation

The performances of methods: of regularized logistic regression (L1, MCP and SCAD), HHS and ABESS, were estimated by 5-fold cross-validation. Again, the dataset was divided into five subsets referred further as folds. Union of four folds was used as a training set, while the fifth fold was used as a testing set. Each fold was used once as a testing set, as a result,

different lists of mutations were selected for five splits of the dataset into training and testing sets.

After the training, a list of features that have non-zero coefficients as a result of the training of each method on each training subset was obtained. For each feature, the number of times that this feature has been selected by each particular method was computed (Table S9-S13).

We defined a feature to be majorly selected if it was selected in three or more dataset splits. For each such feature Fisher exact test p-value was computed on the testing part of the first dataset split (Table S11-S15). Ordinary logistic regression was trained using these features and ROC AUC and F1 scores were computed (Table S16-S20).

To search for additional features associated with the drug phenotypes that were missed by the described procedure of application of ABESS for the five-fold dataset splits we performed the second iteration of ABESS to the subset of data unexplained at the first iteration. For the second iteration of ABESS, we removed from the dataset resistant isolates correctly classified on the previous iteration and some randomly selected susceptible isolates to maintain the class balance. All features which were majorly selected for at least one drug with positive coefficients at the first iteration were completely removed, enforcing the method to search for the new associations. After the training of ABESS, the feature selection process was repeated (Table S20).

See Supplementary Methods for details.

Results

The most of mutations are rare

Short reads and phenotypes for 13 drugs were obtained from publicly available sources (TableS1) [1–10]. After subsequent data preprocessing our dataset comprised 12815 whole-genome sequences for which binary phenotype (resistance or susceptibility) for at least one of 13 antibiotics was known.

For all antibiotics, the number of unique mutations significantly exceeds the number of samples. Moreover, most of the mutations are rare: they appear in the dataset less than for three isolates. Such mutations comprise more than 67% of all unique mutations for all drugs (Table 1).

Table 1. *The descriptive statistics of the dataset. Columns from left to right represent the drug name, classification by line, pharmacological group of the drug, the number of samples in the dataset for which it is known from the literature that they are resistant or susceptible to the given drug, the fraction of samples which are known to be resistant to the given drug, the number of unique mutations, the fraction of indels among mutations, the fraction of mutations which occur less than 3 times in the dataset.*

Drug name	Classification by line	Pharmacological group	Number of samples with known phenotype	Resistance fraction	Number of mutations	Fraction of indels in mutations	Fraction of the rare mutations
Rifampicin	First line		11,913	24.2%	301,227	3.6%	77.1%
Isoniazid		Nicotinamide derivative	11,828	29.5%	300,368	3.6%	77.1%
Pyrazinamide			9,329	12.0%	261,081	3.6%	77.6%
Ethambutol			10,346	14.9%	287,143	3.6%	77.9%
Streptomycin		Aminoglycoside	4,805	35.6%	149,347	3.8%	77.6%
Kanamycin	Second line		1,209	28.7%	45,512	4.6%	74.8%
Amikacin			1,878	21.7%	47,794	4.5%	72.8%
Capreomycin			1,959	21.0%	54,308	4.6%	75.2%
Ofloxacin		Fluoroquinolones	1,974	24.1%	55,591	4.5%	76.4%
Moxifloxacin			1,223	21.4%	36,115	4.8%	74.1%
Ciprofloxacin			471	22.3%	30,313	4.2%	74.8%
Ethionamide		Nicotinamide derivative	700	40.3%	27,284	4.5%	71.0%
Prothionamide			442	40.7%	8,231	5.1%	67.9%

Most of the drugs demonstrate coresistance patterns

To illustrate and quantitatively characterize the coresistance we have computed Pearson's correlations for all pairs of drugs. The correlation was computed on the samples phenotypically characterized for both drugs in a pair. The resulting heatmap is presented in Fig. 1. Many pairs of phenotypes turn out to be strongly correlated (Fig. 1). High correlations are expected for drugs from one pharmacological group such as aminoglycosides (Amikacin, Kanamycin, Streptomycin, Capreomycin) and fluoroquinolones (Moxifloxacin, Ofloxacin, and

Ciprofloxacin). Aminoglycosides share three known genes associated with resistance (rrs, tlyA, gidB) and fluoroquinolones share two (gyrA, gyrB) (Table 2). Another cause of high correlation is joint antibiotic treatment, which is, probably, the case for Rifampicin and Isoniazid, as well as for all pairs of the first-line drugs.

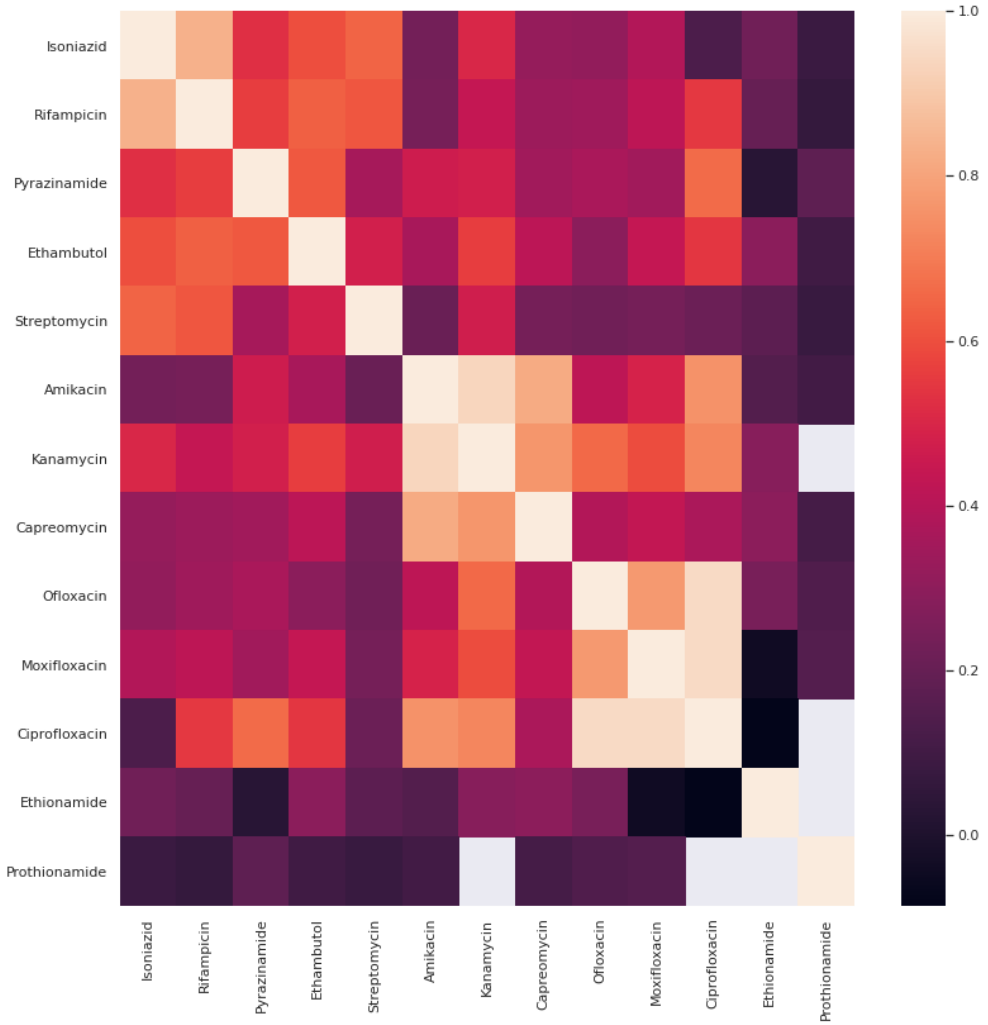

Fig 1. *The heatmap of Pearson correlation between each drug phenotype pair. For each pair only isolates with both phenotypes being known were used.*

Table 2. *Genes associated with antibiotic resistance to different drugs according to previous studies.*

Drug name	Classification by line	Pharmacological group	Gene name	References
Rifampicin	First line		rpoB, ponA1	[57]
Isoniazid		Nicotinamide derivative	ahpC, fabG1, inhA, katG, ndh	[9, 57]
Pyrazinamide			pncA, rpsA, panD	[57]
Ethambutol			embA, embB, embC, embR, iniA, iniC, manB, rmlD, ubiA	[9, 57]
Streptomycin		Aminoglycosides	gidB, rrs, tlyA, rpsL	[9, 57]
Kanamycin	Second line		gidB, rrs, tlyA, eis	[9, 57]
Amikacin			gidB, rrs, tlyA	[9, 57]
Capreomycin			gidB, rrs, tlyA	[9, 57]
Ofloxacin		Fluoroquinolones	gyrA, gyrB	[9]
Moxifloxacin			gyrA, gyrB	[57]
Ciprofloxacin			gyrA, gyrB	[9]
Ethionamide		Nicotinamide derivative	ethA, ethR, inhA	[57]
Prothionamide			ethA, ethR, inhA	[57]

Aggregation of mutations improves the prediction quality

To measure the effect of aggregation we evaluated results of classification of genotypes into resistant and susceptible phenotypic groups by logistic regression with L1 regularization using different feature set combinations (see “Aggregation of mutations” and “Estimating the impact of aggregation” in “Methods”). Four feature sets were involved: SNPs and Indels; SNPs and Indels with PFAM features and excluding rare mutations; SNPs and Indels with “broken gene” features and excluding rare mutations; SNPs and Indels with “gene aggregation” features and excluding rare mutations. Table S2-S7 presents the results of pairwise ROC AUC and F1 scores comparison of sets with aggregation to the basic set comprising SNPs and Indels. We performed comparisons for each drug separately and for drug groups (first-line, second-line, aminoglycosides, fluoroquinolones and all drugs together). For groups, the values of scores are arithmetic means of the corresponding scores obtained for drugs within groups. Sets with aggregated features demonstrated higher

prediction quality for some drugs and the difference was usually statistically significant (Fig. 2, Table S2-S7). The differences of values of ROC AUC are moderate, but, frequently, have large statistical significance due to small variances for all feature sets (Table S8). The feature sets including the PFAM domain features have higher AUC scores for all drug groups except fluoroquinolones and for six drugs individually (Table S2). “Broken genes” feature wins in pairwise comparisons for 9 from 13 drugs (Table S3) and “gene aggregation” - only for 5 from 13, while for 7 drugs the result is the opposite (Table S4). The results for F1 score are consistent with ROC AUC in general (Table S5-S7). We compared PFAM domain features with gene aggregation: PFAM domains-based aggregation outperforms gene aggregation for most drugs both in terms of ROC AUC and F1 score (Table S9-S10). We kept PFAM domain features and “broken genes” features in all further experiments.

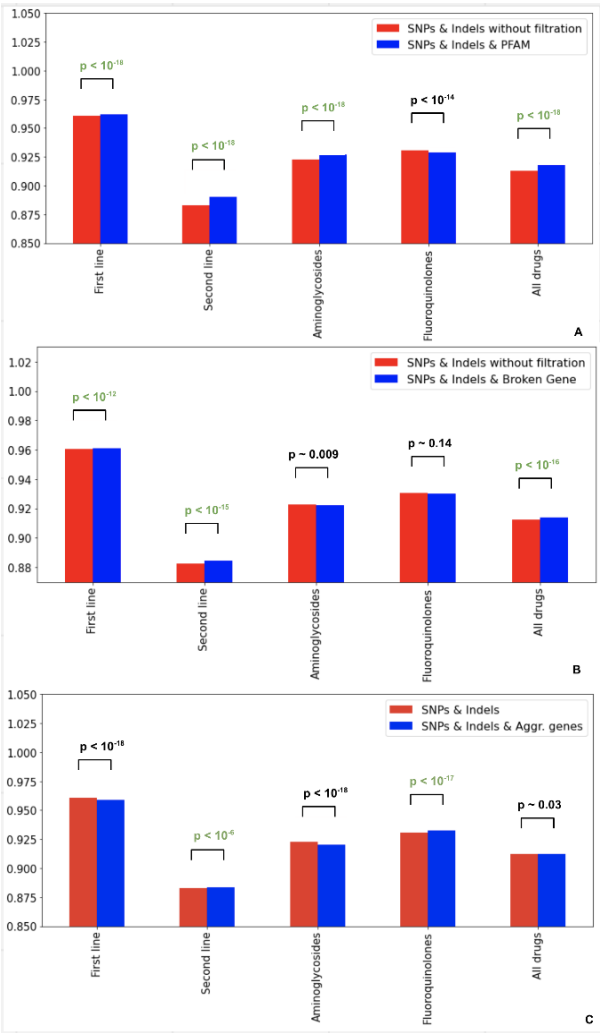

Fig 2. *The comparison of mean ROC AUC scores of logistic regression evaluated on different combinations of features sets for drug groups between: A. “SNPs & Indels” and “SNPs & Indels & PFAM domains excluding rare mutations”; B. “SNPs & Indels” and “SNPs & Indels & “broken genes” excluding rare mutations”; C. “SNPs & Indels” and “SNPs & Indels & “gene aggregation” excluding rare mutations”. Wilcoxon signed-rank test was used for evaluating statistical significance. P-values corresponding to drug groups that demonstrated statistically significant increase in AUC due to aggregation are colored in green.*

Comparison of the feature selection algorithms

In this study, we explored five methods for the selection of causal DR mutations: logistic regression with L1, MCP, and SCAD regularization, HHS, and ABESS. To test the performances of methods the datasets were split into 5-fold for cross-validation (see Methods). The list of mutations selected in at least one split by each method is listed in Table S11-S15 along with the numbers of splits in which it was chosen.

ROC AUC was computed to compare prediction quality of feature sets selected by different methods. Selected features obtained from logistic regression with different regularization techniques get predictions of similar quality, whereas HHS shows worse ones except for Capreomycin and Prothionamide (Fig. 3, Table S16-S20).

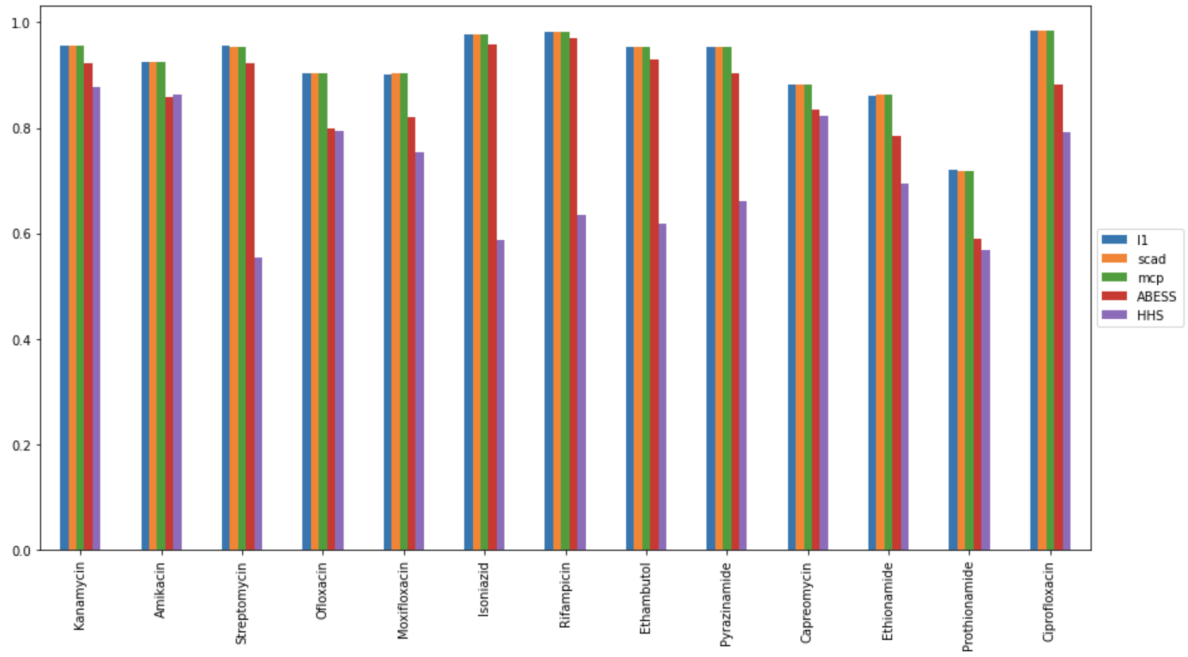

Fig 3. *Average ROC-AUC score of logistic regression models trained on features majorly selected by regression with L1 (blue), MCP (green) and SCAD (orange) regularization, HHS (purple) and ABESS (red). The feature selection models were trained on training parts of 5 dataset splits and features selected on at least 3 parts were chosen. The logistic regression was trained using these features on the training part and AUC was estimated on the testing part of the same splits. The AUC scores were averaged.*

We also compared the number of selected features and the stability of selection by different methods (Table S21). The numbers of selected features vary between the algorithms: logistic regression with different regularization techniques typically pick more features than ABESS and HHS. We compared the number of features selected in at least one dataset split with the number of majorly selected features (selected in at least three splits): surprisingly L1 regularization demonstrated the largest fraction of features selected in 3 and more splits, thus demonstrating relatively high stability, but these sets are still too large to interpret, while the second leader - ABESS - majorly selects the small number of features. (Table S21).

Next, we examined the correctness of selected features. Accounting that most of the phenotype could be explained by mutations in a relatively small number of genes that are involved in actions of drugs we compared the methods on their ability to correctly identify known associations of drugs to genes. Below, we refer to a gene as a “gene associated with a given drug” if this gene was associated with resistance to given drug in previous studies (Table 2). We refer to a gene as a “gene selected by the method for a given drug” if at least one mutation in this gene was majorly selected by this method for this drug. We compared methods by their ability to select for each drug features corresponding to genes whose associations with resistance with this particular drug are known and avoid selecting features associated with other drugs. The first comparison criterion is the Jaccard index (Table 3). In our case, it is the number of genes associated with a given drug selected by the method divided by the number of all genes associated with resistance to any drug selected for the particular drug. The ABESS method had the highest Jaccard index for 10 antibiotics. HHS was the best for Ethambutol, Ciprofloxacin, Kanamycin and Capreomycin while for Prothionamide logistic regression with SCAD and MCP regularization were the best choice. Features from logistic regression with L1 regularization showed the worse result.

Table 3. *The comparison of genes whose mutations were selected by different methods for all drugs to the set of genes associated with drug resistance in previous studies. Below, we refer to a gene as a “gene associated with a given drug” if this gene was associated with resistance to given drug in previous studies (Table 2). We refer to a gene as a “gene selected by the method for given drug” if at least one mutation in this gene was majorly selected by this method for this drug. In the first two rows the method name and the evaluation metric name are listed. Jaccard index is the number of genes, which are associated with a given drug selected by the method divided by a total number of genes selected by the method for this drug. True gene fraction is the number of genes associated with a given drug selected by the method divided by the number of genes associated with this drug. False gene fraction is a number of genes associated with the other drugs selected by the method divided by the number of genes associated with the other drugs. The rest of the rows present these metrics' values for all drugs. Maximal values for all drugs and all metrics are marked in bold.*

Metric	**Jaccard index**					**True gene percentage**					**False gene percentage**				
Method	**L1**	**SCAD**	**MCP**	**ABESS**	**HHS**	**L1**	**SCAD**	**MCP**	**ABESS**	**HHS**	**L1**	**SCAD**	**MCP**	**ABESS**	**HHS**
Rifampicin	0.01	0.01	0.01	**0.13**	0.02	0.50	0.50	0.50	0.50	0.50	**0.29**	**0.29**	**0.29**	0.03	0.06
Isoniazid	0.02	0.03	0.03	**0.38**	0.17	**1.00**	0.80	0.80	0.60	0.20	**0.25**	0.18	0.18	0.04	0.00
Pyrazinamide	0.02	0.02	0.02	**0.09**	0.01	0.33	0.33	0.33	0.33	0.33	0.23	0.23	0.23	0.03	**0.40**
Ethambutol	0.01	0.02	0.02	0.07	**0.08**	**0.33**	**0.33**	**0.33**	0.11	0.11	**0.38**	0.33	0.33	0.04	0.08
Streptomycin	0.04	0.04	0.04	**0.29**	0.06	**0.50**	**0.50**	**0.50**	**0.50**	0.25	**0.21**	**0.21**	**0.21**	0.07	0.14
Kanamycin	0.01	0.02	0.02	0.29	**0.33**	0.50	**0.75**	**0.75**	0.50	0.50	0.34	**0.38**	**0.38**	0.03	0.00
Amikacin	0.00	0.00	0.00	**0.33**	0.11	0.33	0.33	0.33	0.33	0.33	**0.37**	**0.37**	**0.37**	0.00	0.03

Capreomycin	0.01	0.01	0.01	**0.25**	**0.25**	**0.67**	**0.67**	**0.67**	0.33	0.67	**0.40**	**0.40**	**0.40**	0.03	0.00
Ofloxacin	0.00	0.00	0.00	**0.25**	0.08	**1.00**	**1.00**	**1.00**	0.50	0.50	**0.48**	**0.48**	**0.48**	0.00	0.06
Moxifloxacin	0.01	0.00	0.00	**0.25**	0.17	**1.00**	**1.00**	**1.00**	0.50	0.50	0.42	**0.45**	**0.45**	0.03	0.00
Ciprofloxacin	0.02	0.02	0.02	**0.33**	**0.33**	0.50	0.50	0.50	0.50	0.50	**0.23**	**0.23**	**0.23**	0.00	0.00
Ethionamide	0.02	0.02	0.08	**0.40**	0.00	**0.67**	**0.67**	**0.67**	**0.67**	0.00	**0.30**	**0.30**	0.20	0.03	0.03
Prothionamide	**0.01**	**0.01**	**0.01**	0.00	0.00	**0.67**	**0.67**	**0.67**	0.00	0.00	**0.60**	0.57	0.57	0.03	0.03

We also examine two additional metrics to emphasize the strength and weakness of each method: true gene fraction and false gene fraction. We define true gene fraction as the number of genes associated with a given drug selected by the algorithm for a given drug divided by the total number of genes associated with resistance to a given drug (Table 2). We define false gene fraction as the number of genes associated with the other drugs selected by the algorithm for the given drug divided by the total number of genes associated with other drugs (Table 2). The values of true gene fractions were similar for all methods, but they were slightly better for logistic regression methods. Whereas the false gene fraction values were also consistently higher for logistic regression methods (Table 3).

Genes that are selected by ABESS generally are in agreement with the literature (Table 2 and 4). Also, it's worth noting that PFAM domain features are majorly selected by ABESS: about 53% of all majorly selected features are PFAM-domain features (Table 4).

Table 4. *The mutations which are majorly selected by the ABESS algorithm. Columns from left to right represent drug name, gene name, position of event, event type (SNV, domain mutation or indel), reference subsequence in the position, mutated subsequence in the position and number of dataset splits in which the feature was selected.*

Drug name	Gene	Position	Type of event	From	To	Number of splits in which the feature was selected
Rifampicin	eccC2	258	SNV	R	P	5
	rpoB	435	SNV	D	V	5
	rpoB	450	SNV	S	L	5
	rpoB	-	PF04565 is changed	-	-	5
	fadA	-	PF00108 is changed	-	-	4
	katG	-	PF00141 is changed	-	-	4
	rpoB	435	SNV	D	Y	3

Isoniazid	katG	-	PF00141 is changed	-	-	5
	katG	315	SNV	S	T	5
	fabG1	-15	SNV	G	A	5
	rpoB	-	PF04565 is changed	-	-	4
	inhA	-	PF13561 is changed	-	-	4
	pepC	-	PF02127 is changed	-	-	3
Pyrazinamide	pncA	-	PF00857 is changed	-	-	5
	fadA	-	PF00108 is changed	-	-	5
	Rv2017	262	SNV	A	E	4
	Rv2017	-	PF06114 is changed	-	-	4
	ctpB	-	PF00403 is changed	-	-	4
	Rv0108c	-73	Insertion	A	AC	4
	nadD	-44	SNV	A	C	4
	rpoB	-	PF04565 is changed	-	-	4
	Rv2585c	462	SNV	C	S	3
Ethambutol	embB	406	SNV	G	A	5
	embB	-	PF04602 is changed	-	-	5
	rpoB	-	PF04565 is changed	-	-	5
	embB	406	SNV	G	S	4
	ctpB	-	PF00403 is changed	-	-	3
	eccC2	258	SNV	R	P	3
	fadA	-	PF00108 is changed	-	-	3
Streptomycin	inhA	-	PF13561 is changed	-	-	5
	katG	315	SNV	S	T	5
	katG	-	PF00141 is changed	-	-	5
	rpsL	-	PF00164 is changed	-	-	5
	rrs	514	SNV	T	G	5
	gyrA	-	PF00521 is changed	-	-	5

Kanamycin	rrs	1401	SNV	T	C	5
	eis	-10	SNV	G	A	3
	eis	-14	SNV	C	T	3
	htrA	-33	Insertion	A	AC	3
Amikacin	rrs	1401	SNV	T	C	5
Capreomycin	rrs	1401	SNV	T	C	5
	fabG1	-15	SNV	G	A	4
Ofloxacin	gyrA	-	PF00521 is changed	-	-	5
	Rv1897c	-	PF02580 is changed	-	-	3
	gyrA	91	SNV	S	P	3
Moxifloxacin	gyrA	-	PF00521 is changed	-	-	5
	katG	-	PF00141 is changed	-	-	4
	gyrA	91	SNV	S	P	3
Ciprofloxacin	pks8	1357	SNV	A	T	5
	gyrA	-	PF00521 is changed	-	-	3
Ethionamide	fabG1	-15	SNV	G	A	4
	inhA	-	PF13561 is changed	-	-	4
	ethA	-	PF00743 is changed	-	-	3
	fbpC	-	PF00756 is changed	-	-	3
Prothionamide	pncA	-	PF00857 is changed	-	-	3

New associations can be found by repeating the search on the unexplained resistant isolates

Features majorly selected by ABESS do not fully explain resistance for several drugs: the significant percentage of resistant samples is not explained by the mutations selected in three or more folds, especially for the second-line drugs (Table 5). We tried to improve these results and performed the second iteration of ABESS on a reduced dataset to find additional mutations which can explain the resistance (Table 6) (see Methods). We removed from the dataset sequences with resistant phenotypes which were classified as resistant on the previous iteration. We also removed features which were majorly selected for at least one drug with positive coefficients at the first iteration enforcing ABESS to search for new

associations. For six drugs the second iteration allowed to slightly increase the number of explained samples (from 0.3% to 5.8%) (Table 5). Among 12 features selected at the second iteration 10 features were known from the literature as DR-associated mutations and domains in the DR-associated genes, 8 of them are correctly predicted for corresponding drugs. Two drug-associated genes which were missed at the first iteration gyrB for Ofloxacin and gid for Streptomycin were found at the second iteration.

Table 5. *The statistics of explained resistant isolates by the first and the second iteration of the ABESS algorithm. We defined a resistant isolate to be explained by ABESS on a given dataset split if logistic regression trained on the training part of this split using the features majorly selected by ABESS classifies this isolate as resistant. Columns from left to right represents the number of isolates known to be resistant to the given drug, the fraction of resistant samples explained by the first ABESS iteration from the total number of resistant isolates averaged by all dataset splits, the number of resistant isolates unexplained after the first ABESS iteration averaged by all dataset splits, the fraction of resistant isolates explained by the first and the second ABESS iteration from the total number of resistant isolates averaged by all dataset splits, the number of resistant isolates unexplained after the second ABESS iteration averaged by all dataset splits. In case of increase of the fraction of explained isolates for the given drug the values after the first and the second iteration are marked in bold.*

Drug name	**Number of resistant isolates**	**Number of samples participated in the first ABESS iteration**	**Fraction of resistant samples explained by the first ABESS iteration**	**Number of resistant samples unexplained after the first ABESS iteration**	**Number of samples participated in the second ABESS iteration**	**Fraction of resistant samples explained by the first and the second ABESS iteration**	**Number of resistant samples unexplained after the second ABESS iteration**
Rifampicin	2276	11565	**89.9%**	229	1163	**91.0%**	204
Isoniazid	2754	11649	**90.4%**	264	1116	**90.7%**	255
Pyrazinamide	884	9167	**55.5%**	393	4075	**60.9%**	346
Ethambutol	1223	10179	74.7%	309	2571	74.7%	309
Streptomycin	1344	4711	86.2%	186	651	86.2%	186
Kanamycin	275	1201	**79.6%**	56	244	**82.2%**	49
Amikacin	322	1852	72.0%	90	517	72.0%	90
Capreomycin	323	1934	**63.5%**	118	706	**69.3%**	99
Ofloxacin	376	1949	**69.4%**	115	596	**71.3%**	108
Moxifloxacin	205	1205	62.0%	78	458	62.0%	78
Ciprofloxacin	80	457	55.0%	36	205	55.0%	36

Ethionamide	221	691	67.9%	71	221	67.9%	71
Prothionamide	144	439	29.2%	102	310	29.2%	102

Table 6. *The additional mutations majorly selected by ABESS on the second iteration. Columns from left to right represent drug name, gene name, position of event, event type (SNV, domain mutation or indel), reference subsequence in the position, mutated subsequence in the position and number of dataset splits in which the feature was selected. The pairs of drug and genes which were associated with each other in previous studies (Table 2) are marked with bold, while the pairs in which the gene was associated with another drug are marked with italic.*

Drug name	Gene	Position	Type of event	From	To	Number of splits in which the feature was selected
Rifampicin	**rpoB**	170	SNV	V	F	5
Isoniazid	**katG**	-	Broken	-	-	4
Pyrazinamide	**pncA**	-11	SNV	A	G	3
Pyrazinamide	*rpsL*	43	SNV	K	R	3
Ethambutol	**embB**	306	SNV	M	V	5
Kanamycin	**rrs**	1484	SNV	C	A	3
Streptomycin	**gid**	-	PF02527 is changed	-	-	5
Streptomycin	Rv1321	144	SNV	S	R	3
Streptomycin	atsF (Rv3077)	310	SNV	G	R	3
Capreomycin	*embA*	-11	SNV	G	T	3
Capreomycin	**rrs**	1484	SNV	C	A	3
Ofloxacin	**gyrB**	-	PF01751 is changed	-	-	3

Discussion

We analyzed 12815 full genome MTB samples for which drug resistance status was known for at least one of 13 antibiotics being studied. For most of the samples, drug resistance status was known only for the first-line drugs. Typically, if a sample is resistant to one of the

first line drugs it is resistant to the others due to stepwise acquisition of resistance-conferring mutations during long ineffective treatment by multiple drugs [58, 59]. This co-resistance decreases the power of statistical methods to find new genes and pathways weakly associated with resistance. Even the best of the algorithms being studied according to our metrics ABESS was unable to completely untangle cross-associations of features to different drugs (Table 4): rpoB is selected for Isoniazid, Ethambutol and Pyrazinamide while katG is selected for Rifampicin. Also if a sample is resistant to some second-line drug it is typically resistant to one or more first-line drugs. Again, such co-resistance leads to cross associations of genes of first and second line drugs: katG was selected for Streptomycin and Moxifloxacin, pncA was selected for Prothionamide, fabG1 was selected for Capreomycin and Ethionamide. Cross-associations between second-line drug features are also observed: the change of gyrA domain is selected for Kanamycin.

ABESS is the most appropriate method among the tested to solve the problem of correlation, yet at the same time, it has provided the least amount of novel genes involved in drug resistance. Totally, 14 novel genes are selected as markers of any drug resistance: 12 of them are identified after the first iteration and 2 more genes are found after the second iteration of ABESS. Some of the findings are likely to be feasible.

EccC2 is predicted to contribute to Ethambutol and Rifampicin resistance (single-residue replacement in site 258 is selected by ABESS), fadA is associated with Ethambutol, Rifampicin and Pyrazinamide resistance (alteration of N-terminal domain is selected as important resistance predictor). The metal-binding domain of ctpB is selected as probable Ethambutol and Pyrazinamide resistance determinant. EccC2 belongs to the ESX-2 type VII secretion system and may be involved in virulence of the pathogen and its survival under hypoxic conditions [60].

FadA gene product is a possible acyl-CoA thiolase involved in lipid degradation and its expression is shown to be altered upon Isoniazid exposure as a compensatory response [61]. FadA was also associated with Ofloxacin and Kanamycin resistance in one of the early GWAS studies in MTB [1].

CtpB is a plasma membrane copper (I) transporting P-type ATPase. The ctpB gene transcription was activated during hypoxia, nitrosative and oxidative stress which are prevalent in the host phagosomes environment [62].

Pks8 is associated with Ciprofloxacin resistance. Pks8 gene product is a probable polyketide synthase and involved in fatty acids biosynthesis process [63]. Other polyketide synthases, pks2 and pks12, are shown to be associated with MDR in other studies [64].

FbpC is found to be associated with Ethionamide resistance and its gene product is secreted antigen 85-C. Proteins of the antigen 85 complex are responsible for the high affinity of mycobacteria to fibronectin. FbpC possesses a mycolyltransferase activity required for the biogenesis of trehalose dimycolate (cord factor), a dominant structure necessary for maintaining cell wall integrity [65, 66].

Mutations in **pepC** and **htrA** are selected by ABESS as markers of resistance to Isoniazid and Kanamycin, respectively. Both of these gene products are proteases, pepC is shown to be expressed during long pellicle growth [67].

SNP in **nadD** upstream region is associated with Pyrazinamide resistance. NadD is involved in NAD biosynthesis, which is targeted by Isoniazid, Pyrazinamide and some other promising antimycobacterial drugs [68].

Amino-acid replacement in **NucS** (Rv1321, R144S) was selected by ABESS as a marker of resistance to Streptomycin and Ethambutol during the second iteration. In *Mycobacterium tuberculosis,* NucS was previously shown to be expressed in stressed cells [69]. NucS is the DNA repair endonuclease involved in a non-canonical mismatch repair (MMR) pathway [70]. The absence of NucS as well as introducing some NucS polymorphisms including R144S leads to an increased mutation rate in *Mycobacterium smegmatis* [70]. The evolutionary rate acceleration may be advantageous under drug pressure and independent data suggests that nucS-associated hypermutability is, indeed, under positive selection in clinical *M. tuberculosis* strains [71]. Overall, novel genes selected by ABESS are likely to be compensatory and actually associated with survival in the host and drug tolerance.

HHS identified 112 novel genes. After additional filtration by raw p-value (p-value < 0.05) 35 genes remained, 20 of them are associated with resistance to Pyrazinamide. There is no overlap in the selected sets of novel genes between HHS and ABESS. However, HHS output also contains some interesting findings.

HHS method recovers **rpoC** association with Streptomycin and Rifampicin resistance. Interestingly, associated variants are not the same for these drugs: single-residue replacement in site 402 of rpoC is selected as Rifampicin resistance marker while Streptomycin resistance is associated with replacement in site 516 of the same gene. Mutations in several sites in rpoC were shown to have a compensatory role for Rifampicin resistance [72] and Q1126K replacement in RpoC was beneficial in Rifampicin and Streptomycin double-resistant but not in sensitive strains [15].

RpoA, another RNA polymerase subunit, variant is associated with Pyrazinamide resistance. However, this variant (V183G) is also known to be compensatory for Rifampicin resistance [72].

Novel genes set for Pyrazinamide also contains 4 membrane transport proteins: **cycA** (involved in D-alanine, D-serine and glycine transport), **dppC** (probable ABC transporter of dipeptide across the membrane), **sugI** (involved in sugar transport) and **cysT** (probable sulfate-transport ABC transporter) [66]. CycA and sugI variants were previously associated with D-cycloserine resistance, the drug which is commonly used to treat MDR and XDR-TB [73, 74].

NusG variant is associated with Rifampicin resistance. NusG gene product is a probable transcription antitermination protein which interacts with rho factor and RNA polymerase. Thus, the selected variant is likely to be compensatory for rpoB mutations and nusG regulatory mutation was previously shown to be compensatory in Rifampicin and Streptomycin double-resistant genotypes [15].

RplU, which encodes 50S ribosomal protein L21, is associated with Streptomycin resistance while another ribosomal protein, rpsL, and ribosomal RNA rrs are known to be the Streptomycin targets.

PntAa is also found to be associated with Streptomycin resistance and functions as a proton pump suggesting its possible role in MDR acquisition.

Logistic regression methods selected the largest set of novel genes (more than 500 novel genes remained after raw p-value filtration for each regularization technique). These large sets lack biological interpretation proving this group of methods is not suitable for the task discussed, despite the fact that these methods provide the highest prediction quality (Fig. 3).

The prediction quality of GWAS algorithms is limited by the quality of datasets. One of the reasons is samples infected by multiple strains which may lead to errors both in genotyping and phenotyping. Assigning discrete phenotypes is problematic in case of intermediate levels of drug resistance: resistant isolates with a low growth rate are interpreted as susceptible in some cases [75]. Differences in culturing and measurement protocols lead to inconsistency in phenotyping and lead to problems with the reproducibility of results. There is growing evidence that MICs are more informative than binary labels for GWAS analysis [76].

Overall, our study demonstrates the usefulness of advanced feature selection combined with feature generation by aggregation with subsequent filtering by mutation frequency. Aggregated features were majorly selected and constituted the majority of selected features. Another advantage of the aggregation is the possibility to take into account the effects of rare mutations without including them directly into the model increasing both the robustness and training speed.

Authors contributions

Reshetnikov K.O.: Conceptualisation, Methodology, Software, Investigation, Writing – Original Draft Preparation, Visualization
Bykova D.I.: Conceptualisation, Methodology, Software, Investigation, Writing – Original Draft Preparation
Kuleshov K.V.: Methodology, Data Curation, Writing – Review and Editing
Chukreev K.: Software, Investigation
Guguchkin E.P.: Investigation
Akimkin V.G.: Project Administration, Funding
Neverov A.D.: Conceptualisation, Methodology, Writing – Original Draft Preparation, Writing – Review and Editing, Supervision, Project Administration
Fedonin G.G.: Conceptualisation, Methodology, Software, Investigation, Data Curation, Writing – Original Draft Preparation, Supervision

Competing Interests

The authors declare that there are no conflicts of interest.

References

1. **Zhang H, Li D, Zhao L, Fleming J, Lin N, *et al.*** Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet* 2013;45:1255–1260.
2. **Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, *et al.*** Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* 2013;45:1183–1189.
3. **Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, *et al.*** Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat Genet* 2018;50:307–316.
4. **Yang C, Luo T, Shen X, Wu J, Gan M, *et al.*** Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis* 2017;17:275–284.
5. **Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, *et al.*** GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* 2019;10:2128.
6. **Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, *et al.*** Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* 2019;35:2276–2282.
7. **Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, *et al.*** Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 2014;46:279–286.
8. **Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, *et al.*** Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 2016;4:49–58.
9. **Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, *et al.*** Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis*. Epub ahead of print June 2015. DOI: 10.1016/S1473-3099(15)00062-6.
10. **Johnsen CH, Clausen PTLC, Aarestrup FM, Lund O**. Improved Resistance Prediction in Mycobacterium tuberculosis by Better Handling of Insertions and Deletions, Premature Stop Codons, and Filtering of Non-informative Sites. *Front Microbiol*;10. Epub ahead of print 2019. DOI: 10.3389/fmicb.2019.02464.
11. World Health Organisation. Global Tuberculosis Report; 2020.
12. **Jnawali HN, Ryoo S**. First– and Second–Line Drugs and Drug Resistance. In: *Tuberculosis - Current Issues in Diagnosis and Management*. 2013. Epub ahead of print 2013. DOI: 10.5772/54960.
13. **Hang NTL, Hijikata M, Maeda S, Thuong PH, Ohashi J, *et al.*** Whole genome sequencing, analyses of drug resistance-conferring mutations, and correlation with transmission of Mycobacterium tuberculosis carrying katG-S315T in Hanoi, Vietnam. *Sci Rep* 2019;9:15354.
14. **Trauner A, Liu Q, Via LE, Liu X, Ruan X, *et al.*** The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome Biol* 2017;18:71.
15. **Sousa JM de, Balbontín R, Durão P, Gordo I**. Multidrug-resistant bacteria compensate for the epistasis between resistances. *PLOS Biol* 2017;15:e2001741.
16. **Andrews JM**. Determination of minimum inhibitory concentrations. *J Antimicrob*

Chemother 2001;48:5–16.
17. **Tomasicchio M, Theron G, Pietersen E, Streicher E, Stanley-Josephs D, *et al.*** The diagnostic accuracy of the MTBDRplus and MTBDRsl assays for drug-resistant TB detection when performed on sputum and culture isolates. *Sci Rep* 2016;6:17850.
18. **Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, *et al.*** Rapid molecular detection of tuberculosis and rifampin resistance. *N Engl J Med* 2010;363:1005–1015.
19. **Meaza A, Kebede A, Yaregal Z, Dagne Z, Moga S, *et al.*** Evaluation of genotype MTBDRplus VER 2.0 line probe assay for the detection of MDR-TB in smear positive and negative sputum samples. *BMC Infect Dis* 2017;17:280.
20. **Chen ML, Doddi A, Royer J, Freschi L, Schito M, *et al.*** Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine* 2019;43:356–369.
21. **Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, *et al.*** Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 2016;4:49–58.
22. **Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, *et al.*** Tuberculosis is changing. *Lancet Infect Dis* 2017;17:359–361.
23. **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.
24. **World Health Organization**. *The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide*. Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/274443 (2018).
25. **Domínguez J, Boettger EC, Cirillo D, Cobelens F, Eisenach KD, *et al.*** Clinical implications of molecular drug resistance testing for Mycobacterium tuberculosis: a TBNET/RESIST-TB consensus statement. *Int J Tuberc Lung Dis* 2016;20:24–42.
26. **Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, *et al.*** A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *Eur Respir J* 2017;50:1701354.
27. **Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, *et al.*** Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *Am J Respir Crit Care Med* 2016;194:621–630.
28. **Ghosh A, N. S, Saha S**. Survey of drug resistance associated gene mutations in Mycobacterium tuberculosis, ESKAPE and other bacterial species. *Sci Rep* 2020;10:8957.
29. **Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S**. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci Rep* 2017;7:46327.
30. **Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, *et al.*** Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol* 2018;3:1032–1042.
31. **Hicks ND, Carey AF, Yang J, Zhao Y, Fortune SM**. Bacterial Genome-Wide Association Identifies Novel Factors That Contribute to Ethionamide and Prothionamide Susceptibility in Mycobacterium tuberculosis. *mBio*. Epub ahead of print 23 April 2019. DOI: 10.1128/mBio.00616-19.
32. **Sergeev RS, Kavaliou IS, Sataneuski UV, Gabrielian A, Rosenthal A, *et al.*** Genome-Wide Analysis of MDR and XDR Tuberculosis from Belarus: Machine-Learning Approach. *IEEE/ACM Trans Comput Biol Bioinform* 2019;16:1398–1408.
33. **Niehaus KE, Walker TM, Crook DW, Peto TEA, Clifton DA**. Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2014. pp. 618–621.
34. **Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, *et al.*** Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic

signatures of antibiotic resistance. *Nat Commun* 2018;9:1–9.
35. **Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, *et al.*** Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. *Front Genet*;10. Epub ahead of print 2019. DOI: 10.3389/fgene.2019.00922.
36. **Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, *et al.*** Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 2016;1:16041.
37. **Zabeti H, Dexter N, Safari AH, Sedaghat N, Libbrecht M, *et al.*** INGOT-DR: an interpretable classifier for predicting drug resistance in M. tuberculosis. *Algorithms Mol Biol* 2021;16:17.
38. **Palmer AC, Chait R, Kishony R**. Nonoptimal Gene Expression Creates Latent Potential for Antibiotic Resistance. *Mol Biol Evol* 2018;35:2669–2684.
39. **Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, *et al.*** Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* 2018;34:1666–1671.
40. **Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, *et al.*** DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. *Bioinformatics* 2019;35:3240–3249.
41. **Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, *et al.*** Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLoS Med* 2015;12:e1001880–e1001880.
42. **Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, *et al.*** Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat Genet* 2017;49:395–402.
43. **Van Rie Annelies, Warren Robin, Mshanga Idris, Jordaan Annemarie M, van der Spuy Gian D., *et al.*** Analysis for a Limited Number of Gene Codons Can Predict Drug Resistance of Mycobacterium tuberculosis in a High-Incidence Community. *J Clin Microbiol* 2001;39:636–641.
44. **Hazbón Manzour Hernando, Brimacombe Michael, Bobadilla del Valle Miriam, Cavatore Magali, Guerrero Marta Inírida, *et al.*** Population Genetics Study of Isoniazid Resistance Mutations and Evolution of Multidrug-Resistant Mycobacterium tuberculosis. *Antimicrob Agents Chemother* 2006;50:2640–2649.
45. **Burkina M, Nazarov I, Panov M, Fedonin G, Shirokikh B**. Inductive Matrix Completion with Feature Selection. *Comput Math Math Phys* 2021;61:719–732.
46. **Richardson A**. Logistic Regression: A Self-Learning Text, Third Edition by David G. Kleinbaum, Mitchel Klein. *Int Stat Rev* 2011;79:296–296.
47. **Tibshirani R**. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Ser B Methodol* 1996;58:267–288.
48. **Hoerl AE, Kennard RW**. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970;12:55–67.
49. **Hoerl AE, Kennard RW**. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* 1970;12:69–82.
50. **Grogan TR, Elashoff DA**. A simulation based method for assessing the statistical significance of logistic regression models after common variable selection procedures. *Commun Stat Simul Comput* 2017;46:7180–7193.
51. **Saber MM, Shapiro BJ**. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genomics* 2020;6:e000337.
52. **Fan J, Li R**. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J Am Stat Assoc* 2001;96:1348–1360.
53. **Cun-Hui Zhang**. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010;38:894–942.
54. **Kumar A, Bhattacharyya S, Bouchard K**. Numerical Characterization of Support Recovery in Sparse Regression with Correlated Design. http://arXiv.org/abs/ (2021).

55. **Zhu J, Wen C, Zhu J, Zhang H, Wang X**. A polynomial algorithm for best-subset selection problem. *Proc Natl Acad Sci* 2020;117:33117.
56. **Libiseller-Egger J, Phelan J, Campino S, Mohareb F, Clark TG**. Robust detection of point mutations involved in multidrug-resistant Mycobacterium tuberculosis in the presence of co-occurrent resistance markers. *PLOS Comput Biol* 2020;16:e1008518.
57. **Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM**. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. *Genome Med*;11. Epub ahead of print December 2019. DOI: 10.1186/s13073-019-0660-8.
58. **Mnyambwa NP, Kim D-J, Ngadaya ES, Kazwala R, Petrucka P, *et al.*** Clinical implication of novel drug resistance-conferring mutations in resistant tuberculosis. *Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol* 2017;36:2021–2028.
59. **Hjort K, Jurén P, Toro JC, Hoffner S, Andersson DI, *et al.*** Dynamics of Extensive Drug Resistance Evolution of Mycobacterium tuberculosis in a Single Patient During 9 Years of Disease and Treatment. *J Infect Dis* 2020;jiaa625.
60. **Bose T, Das C, Dutta A, Mahamkali V, Sadhu S, *et al.*** Understanding the role of interactions between host and Mycobacterium tuberculosis under hypoxic condition: an in silico approach. *BMC Genomics* 2018;19:555.
61. **Abo-Kadoum M a., Dai Y, Asaad M, Hamdi I, Xie J**. Differential Isoniazid Response Pattern Between Active and Dormant Mycobacterium tuberculosis. *Microb Drug Resist* 2021;27:768–775.
62. **León-Torres A, Arango E, Castillo E, Soto CY**. CtpB is a plasma membrane copper (I) transporting P-type ATPase of Mycobacterium tuberculosis. *Biol Res* 2020;53:6.
63. **Dubey VS, Sirakova TD, Cynamon MH, Kolattukudy PE**. Biochemical Function of msl5 (pks8 plus pks17) in Mycobacterium tuberculosis H37Rv: Biosynthesis of Monomethyl Branched Unsaturated Fatty Acids. *J Bacteriol* 2003;185:4620–4625.
64. **Philalay JS, Palermo CO, Hauge KA, Rustad TR, Cangelosi GA**. Genes Required for Intrinsic Multidrug Resistance in Mycobacterium avium. *Antimicrob Agents Chemother* 2004;48:3412–3418.
65. **Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ, *et al.*** Role of the major antigen of Mycobacterium tuberculosis in cell wall biogenesis. *Science* 1997;276:1420–1422.
66. **Kapopoulou A, Lew JM, Cole ST**. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberc Edinb Scotl* 2011;91:8–13.
67. **Kerns PW, Ackhart DF, Basaraba RJ, Leid JG, Shirtliff ME**. Mycobacterium tuberculosis pellicles express unique proteins recognized by the host humoral response. *Pathog Dis* 2014;70:347–358.
68. **Rohde KH, Sorci L**. The Prospective Synergy of Antitubercular Drugs With NAD Biosynthesis Inhibitors. *Front Microbiol* 2021;11:3589.
69. Structure and function of a novel endonuclease acting on branched DNA substrates. *EMBO J* 2009;28:2479–2489.
70. **Castañeda-García A, Prieto AI, Rodríguez-Beltrán J, Alonso N, Cantillon D, *et al.*** A non-canonical mismatch repair pathway in prokaryotes. *Nat Commun* 2017;8:14246.
71. **Cebrián-Sastre E, Martín-Blecua I, Gullón S, Blázquez J, Castañeda-García A**. Control of Genome Stability by EndoMS/NucS-Mediated Non-Canonical Mismatch Repair. *Cells* 2021;10:1314.
72. **Comas I, Borrell S, Roetzer A, Rose G, Malla B, *et al.*** Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* 2012;44:106–110.
73. **Chen JM, Uplekar S, Gordon SV, Cole ST**. A Point Mutation in cycA Partially Contributes to the D-cycloserine Resistance Trait of Mycobacterium bovis BCG Vaccine Strains. *PLOS ONE* 2012;7:e43467.
74. **Chen J, Zhang S, Cui P, Shi W, Zhang W, *et al.*** Identification of novel mutations associated with cycloserine resistance in Mycobacterium tuberculosis. *J Antimicrob*

Chemother 2017;72:3272–3276.
75. **Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, *et al.*** Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun* 2015;6:10063.
76. **Ahmad S, Mokaddas E, Al-Mutairi N, Eldeen HS, Mohammadi S**. Discordance across Phenotypic and Molecular Methods for Drug Susceptibility Testing of Drug-Resistant Mycobacterium tuberculosis Isolates in a Low TB Incidence Country. *PLOS ONE* 2016;11:e0153563.