УДК 004.048+57.087.1

IDENTIFICATION OF GENOMIC MUTATIONS ASSOCIATED WITH DRUG-RESISTANCE

Y. Chen[a], A. Tuzikov[b]

[a]*Belarusian State University, 4 Niezalieznasci Avenue, Minsk 220030, Belarus,* c894424323@outlook.com

[b]*United Institute of Informatics Problems of the National Academy of Sciences of Belarus Surganova 6, 220012 Minsk, Belarus*

Corresponding author: c894424323@outlook.com

Drug resistance in tuberculosis (TB) is a global public health problem, and resistance testing early in treatment can help prevent antibiotic misuse. The data used were from the NIAID TB Portals project (https://tbportals.niaid.nih.gov). Mtb whole genome sequences of 3178 patients and resistance testing to 27 drugs were utilized after quality checks. To identify mutant loci associated with drug resistance, single and multiple marker tests were used. Important mutant loci associated with drug resistance in TB were identified. On the one hand, these mutant loci can provide important information for understanding drug resistance in TB, and on the other hand, they can be used as a rapid screening method for various forms of Mtb resistance.

Keywords: drug-resistance; tuberculosis; single-marker tests; multi-marker tests.

Introduction

Worldwide Globally, it is estimated that about 10 million (range 8.9-11 million) people have the diseases with TB in 2020, and this number has been declining very slowly in recent years [1]. More cases of drug resistance have emerged, the appearance including Monoresistance (MonoDR), resistance to one first-line anti-TB drug only; multi-drug resistance (MDR-TB), resistance to isoniazid and rifampicin; and extensive drug resistance (XDR-TB), one fluoroquinolone, and one second- line injectable drug [2]. For the most effective first-line drug, rifampicin, the proportion of new cases of resistance is higher [1]. Therefore, TB drug resistance is a global public health issue. Various machine learning models have been applied to determine drug resistance, e.g., logistic regression (LR) [3], and random forest (RF) [4]. The Genome-wide association analysis (GWAS) method was applied for TB drug resistance analysis [5–7].

1. Methodology

Dataset

The data set contained 3178 samples and their resistance test results for 27 drugs. There are 4418596 nucleotide loci in the whole genome of Mycobacterium tuberculosis. Due to the large amount of data, some unmutated loci need to be removed. The unmutated nucleotide loci in the sample subset are deleted. At this point, the total number of loci in this sample is 294153. The MAF (minor allele frequency) was set to 0.01 and remove loci with mutation rates smaller than the MAF were removed. After filtering out, the number of mutations (SNPs) left in the samples was 20,976.

Single-marker test

Single-marker tests are used to test associations between observed drug resistance and individual mutations. Fisher's exact test and the linear regression model were used as single-marker tests. Fisher's exact test needs constructing the drug sensitivity test and mutation 2D contingency table of cases. Contingency tables considered in single-marker tests for finding mutations associated with resistance

Table 1 – Contingency table considered in single-marker tests for finding mutations associated with resistance

Drug susceptibility	Presence of mutation		
Present	Absent	Present	Total
Sensitive	n_{00}	n_{01}	n_{0*}
Resistant	n_{10}	n_{11}	n_{1*}
Total	n_{*0}	n_{*1}	n_{**}

The contingency table used for testing correlation of mutation in position 2155175 of Mtb genome and resistance to the isoniazid drug for our data is the following:

Table 2 – Isoniazid susceptibility and SNP(2155175) contingency table

Present	Absent	Present
Sensitive	435	15
Resistant	151	844

Application of the Fisher's exact test to this table results in the probability p=1.31e-212 which characterize a statistical significance of the mutation for the resistance to isoniazid. In this case null hypothesis assumes that there is no correlation between the mutation in the considered position and resistance to isoniazid. The probability achieved strongly rejects the null hypothesis that this mutation and isoniazid drug resistance are independent.

Linear regression model

$$Y = \beta X + \varepsilon \quad (1)$$

Here Y - phenotype vector, β - parameters to be estimated, X – genotype vector, ε – residual vector.

If resistance to the corresponding drug or drug combination is observed, Y_i=1; otherwise, it is equal to 0. If the genotype of this site is '0/0', means no mutation, then X_i = 0, otherwise if its genotype is '1/1', then X_i = 2. For example, consider testing mutations at position 2155175 of the Mtb genome and isoniazid resistance. The regression function is $\hat{Y} = 0.25641 + 0.363074X$. The chi-square test probability for the parameter β is p = 1.456125e-283. In this case null hypothesis assumes that there is no correlation between the mutation in the considered position and resistance to isoniazid. The probability achieved strongly rejects the null hypothesis that this mutation and isoniazid drug resistance are independent. By calculating the estimate and its negative logarithm of p-values of all SNPs, and sorting them, we can finally get the relevant mutation sites for drugs.

Multi-marker test

Multi-marker test is used to select SNP combinations with forward selection method (greedy algorithm). The ratio of training set to test set is 7 versus 3. The classification model is SVM. The evaluation indicator is accuracy.

First, the p-value of a single SNP can be obtained according to the linear regression model. In order to reduce the amount of calculation, SNPs with p-values smaller than 0.05 are used for classification. The number of useful SNPs depends on the type of drug. Then, in the second step, based on the selection of the first SNP, each SNP is re-evaluated to participate in the classification together with the first SNP, and the combination with the greatest improvement in accuracy is selected. Finally, keep iterating to add new SNPs until the accuracy no longer improves.

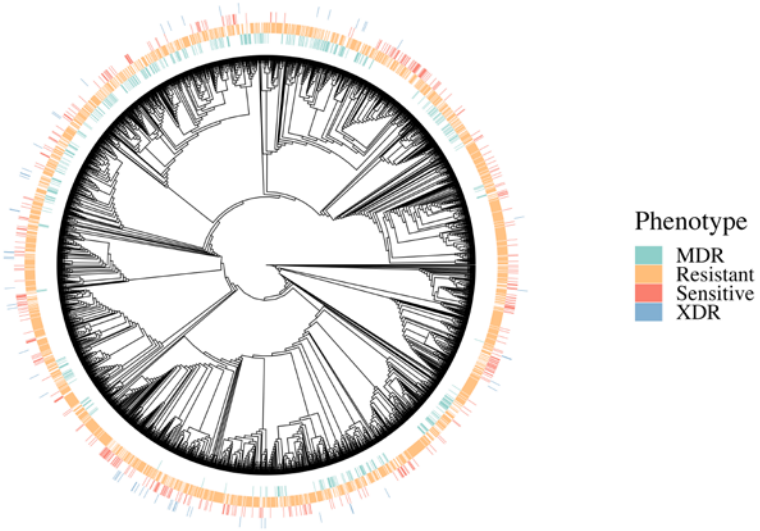
2. Results and discussion

Characterization of the dataset

Each isolate was tested for resistance to at least one of 27 anti-TB drugs: four first-line drugs, isoniazid, rifampicin, ethambutol, and pyrazinamide, some second-line drugs, and other drugs.

A phylogenetic tree of the samples was constructed using all genome-wide SNPs (Picture1). Phenotypic analysis of anti-TB drug susceptibility revealed that 71.1% of the isolates were resistant to at least one drug, of which 13.8% were classified as MDR-TB and 1.9% as XDR-TB

A phylogenetic tree was constructed using all genome-wide SNPs (Picture 1). Phenotypic analysis of anti-TB drug susceptibility revealed that 71.1% of the isolates were resistant to at least one drug, of which 13.8% were classified as MDR-TB and 1.9% as XDR-TB Because of the small sample size or severe data imbalance for bedaquiline, clarithromycin, aminoglycosides injectible agents, and fluoroquinolone, these four drugs are not involved in training the model

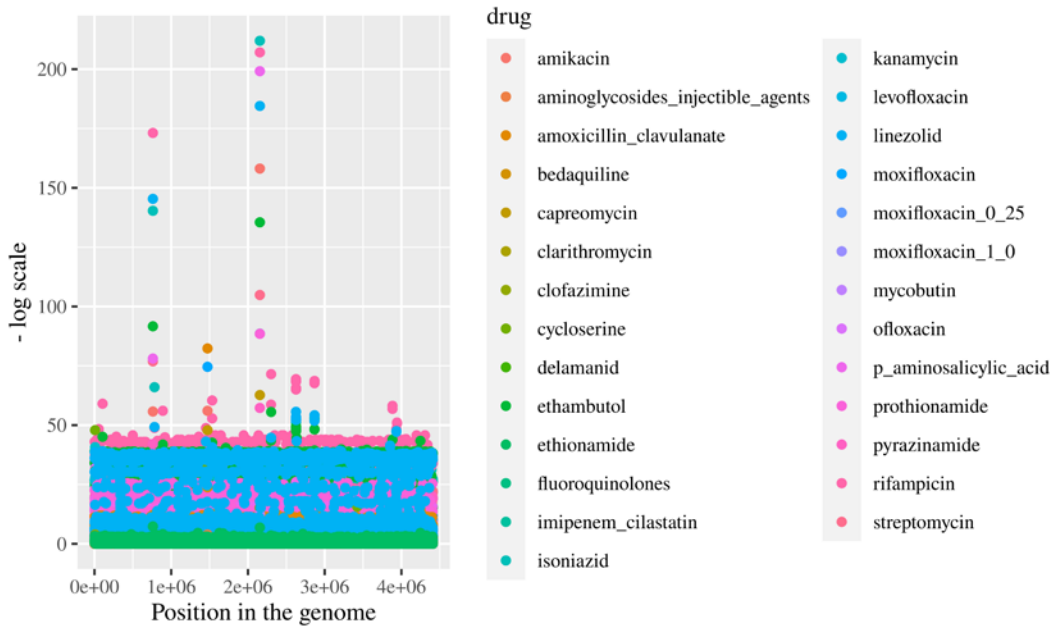


Picture 1 – Whole-genome phylogeny of the 3178 Mycobacterium tuberculosis isolates used for association study

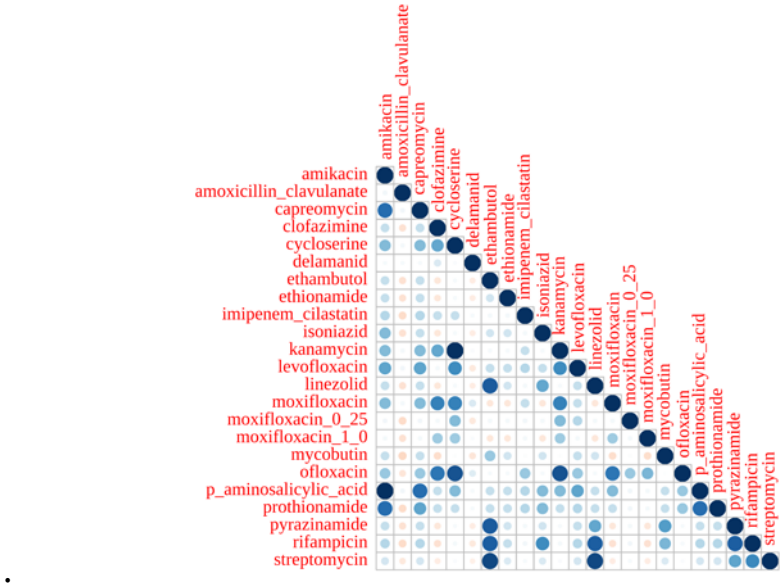
Single-marker test results.

We calculated p-values for the corresponding mutations for all SNPs. To visualize the test results, a Manhattan plot (Picture 2) was used. To allow a more visual representation of the results, the p-values were converted to -log10 (p-value). The height of the SNP locus on the Y-axis corresponds to the degree of association with a certain drug resistance, the stronger the association (i.e., the lower the p-value) the higher it is. These SNPs with strong associations with drug resistance were also of most interest throughout the research.

Then, the corresponding genes were found based on the five most significant mutant loci for each drug. A correlation analysis was then performed. From Picture 3, it was found that there was a correlation between drug resistance genes of different drugs



Picture 2 – Single-marker test result

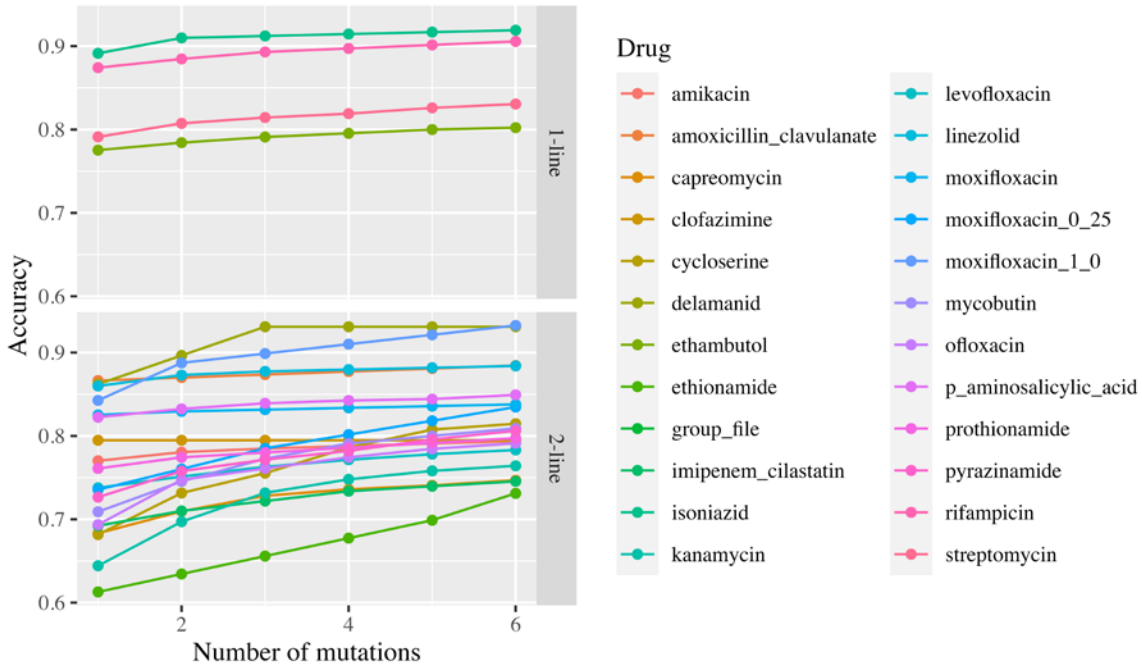


Picture 3 – Correlation between drugs with mutated loci

Multi-marker test results.

We used the SVM function in the ‘e1071’ package to complete the calculations, and the parameters kernel, c, with default values are respectively. Due to the huge amount of computation, the maximum number of combinations of mutation sites was set to 6. For each drug, some combination of mutation sites that can help improve classification accuracy were obtained.

It can be seen from Picture 4, that the classification accuracy of the model increases as the number of SNPs as classification features increases.



Picture 4 – Multi-marker test result

Conclusions

In this paper, we used single-marker and multi-marker tests to identify mutations associated with TB drug resistance. The results of the single marker test reflect the association of a single mutation site with resistance to each drug. We found that the mutation sites highly associated with first-line drug resistance were different from those of second-line drugs. We have found that mutations at some loci were highly associated with resistance to several drugs, reflecting the presence of cross-resistance between drugs. In addition, for some second-line drugs, the accuracy improvement of the classifier is larger with combination of mutations.

Acknowledgment

We acknowledge support from TB Portals Consortium and the TB Portals Program (https://tbportals.niaid.nih.gov) [8].

References

1. World Health Organization et al. World Health Organization Global Tuberculosis Report 2021. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2021
2. for the Meta T C G, Ahmad N, Ahuja S D, et al. Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis // J. The Lancet. 2018. Vol. 392, № 10150. P. 821–834.

3. Farhat M R, Sultana R, Iartchouk O, et al. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value // J. American journal of respiratory and critical care medicine, 2016. Vol. 194, № 5. P. 621–630.
4. Kouchaki S, Yang Y, Lachapelle A, et al. Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking // J. Frontiers in microbiology. 2020. Vol. 11. P. 667.
5. Sergeev R S, Kavaliou I S, Sataneuski U V, et al. Genome-wide analysis of MDR and XDR Tuberculosis from Belarus: Machine-learning approach // J. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2017. Vol. 16, № 4. P. 1398–1408.
6. Crook D W, Rodrigues C, Ismail N A, et al. Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms // J. PLoS biology. 2022. Vol. 20, № 8. P. e3001755.
7. Conkle-Gutierrez D, Kim C, Ramirez-Busby S M, et al. Distribution of Common and Rare Genetic Markers of Second-Line-Injectable-Drug Resistance in Mycobacterium tuberculosis Revealed by a Genome-Wide Association Study // J. Antimicrobial Agents and Chemotherapy, 2022. P. e02075-21.
8. Rosenthal A., Gabrielian A., Engle E, et. al. The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. // Journal of Clinical Microbiology, 2017. № 77(1). P. 3261–3282. doi.org/10.1128/JCM.01013-17.