IDENTIFICATION OF GENOMIC MUTATIONS ASSOCIATED WITH DRUG-RESISTANT TUBERCULOSIS

Yuxiang Chen

Belarusian State University, Minsk;
c894424323@outlook.com ;
supervisor – A.V. Tuzikov, Doctor of physic-mathematical sciences, professor

Drug resistance in tuberculosis (TB) is a global public health issue, and resistance testing early in therapy can help prevent antibiotic abuse. Data used is from the NIAID TB Portals program (https://tbportals.niaid.nih.gov). Mtb whole genome sequences from 645 patients were utilized after quality inspection. To identify mutation sites associated with drug resistance, single-marker and multi-marker tests were used. Important mutation sites associated with TB drug resistance were discovered. On the one hand, these mutation sites can give important information for understanding TB resistance, and they can be utilized for a quick screening approach for various forms of TB drug resistance.

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

INTRODUCTION

Worldwide, it is estimated that between 1.7 billion and 1.8 billion individuals are infected with the causative agent, Mycobacterium tuberculosis (Mtb) [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 in Belarus[5].

MATERIALS AND METHODS

Data used is from the NIAID TB Portals program (https://tbportals.niaid.nih.gov). A total of 8 drugs were selected, four first-line drugs, Isoniazid, Rifampicin, Ethambutol Streptomycin (INH", "RIF", "EMB", "S") and four second-line drugs Ofloxacin, Amikacin, Kanamycin, Capreomycin ("OFX", "AM", "KM", "CM"). There are five drug resistance test systems: bactec, le, ipaother, hain, and geneexpert. At least one of the five tests for a drug for a sample is present, otherwise the test results for this sam-

ple are missing. TB whole genome sequences from 645 patients were utilized after quality inspection. Subsets of drug resistant and sensitive samples to the specified drugs are presented in Table 1.

Table 1

Subsets of drug resistant (R) and sensitive (S) case

Drug names	INH	RIF	EMB	S	OFX	AM	KM	CM
R	338	414	224	261	138	74	99	86
S	194	204	268	198	296	313	260	363

There are a total of 4,418,596 nucleotide sites in the whole genome of Mycobacterium tuberculosis. Due to the large amount of data, some unmutated sites need to be removed. Remove the nucleotide sites that have not been mutated in the sample subset. At this time, the sample's total number of sites is 253,195. Set the MAF (Minor Allele Frequency) to 0.01, and remove the sites whose mutation ratio is less than MAF. The number of mutations (SNPs) in the sample left after filtering out is 11,846.

Single-marker tests are used to test associations between observed drug resistance and individual mutations [3]. 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.

Table 2

Contingency tables considered in single-marker tests for finding mutations associated with resistance

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

Linear regression model

$$Y = X\beta + \varepsilon$$

Y - Phenotype vector, β- Estimate, 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. 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 is used to select SNP combinations with forward selection method. 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 useful 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.

RESULTS

We used the R software functions from the stats package: fisher test for Fisher's exact test and linear regression model. We calculated p-values for all corresponding mutations. The most significant mutation sites for resistance to each drug were identified.

Table 3

The most significant mutation sites for resistance for each drug

Drug names	Most significant Mutation sites
INH	C2155175G, C761158T
RIF	C2155175G, C761158T
EMB	C2155175G, A1473252G
S	C2155175G, C761158T, A781690G
OFX	C2155175G, C761158T
AM	G1473252A, T764844C, C2155175G
KM	G1473252A, C2155175G
CM	G1473252A, C2155175G

Result of Multi-marker test

We use the SVM function in the ‘e1071’ package to complete the calculation, and the parameters kernel, c, and γdefault values are respectively. Due to the huge amount of computation, the maximum number of combinations of mutation sites is set to 6. For each drug, we get some combination of mutation sites that can help improve classification accuracy.

Table 4

The combination of mutation sites with the highest classification accuracy for each drug

a.First-lines drugs

Combination	INH		RIF		EMB		S	
	Positions	ACC	Positions	ACC	Positions	ACC	Positions	ACC
1st	2155175	93.7%	2155175	90.8%	2155175	85.7%	2155175	87.6%
+2nd	2715379	94.3%	761158	93.5%	2945211	87.0%	3594340	89.0%
+3rd	3382091	95.0%	1673431	94.0%	2635600	87.8%	4247607	89.8%

Continuation Table 4

Combination	INH		RIF		EMB		S	
	Positions	ACC	Positions	ACC	Positions	ACC	Positions	ACC
+4th	3736070	95.6%	1473252	94.6%	3594340	88.4%	761143	9.05%
+5th	2187314	96.2%	1593235	95.1%	2059530	89.1%	2747161	9.12%
+6th			580451	95.7%			761158	91.2%

b.Second-lines drugs

Combination	OFX		AM		KM		CM	
	Positions	ACC	Positions	ACC	Positions	ACC	Positions	ACC
1st	2155175	78.3%	1473252	92.2%	1473252	86.9%	1473252	88.9%
+2nd	2626523	82.2%	104912	93.9%	2715379	90.7%	2155175	89.6%
+3rd	4353537	84.5%	2155175	94.8%	2715356	93.5%	837917	90.4%
+4th	7582	86.0%	3946824	95.7%	1789677	94.4%	761112	91.1%
+5th	2796141	88.4%			2340112	95.3%	1091972	91.9%
+6th	761098	89.9%					4053284	92.6%

DISCUSSION

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. At the same time, we found that among the same drugs, significant mutation sites are highly similar, which also reflects the existence of cross-resistance between drugs [6]. However, for some second-line drugs, the accuracy improvement of the classifier is larger with combination of mutations.

References

1. World Health Organization et al. World Health Organization Global Tuberculosis Report 2020 // World Health Organization. – 2020. – Vol. 232.
2. for the Meta T. C. G. et al. Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis // The Lancet. – 2018. – Vol. 392. – №. 10150. – P. 821-834.
3. Farhat M. R. et al. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value //American journal of respiratory and critical care medicine. – 2016. – Vol. 194. – №. 5. – P. 621-630.
4. Kouchaki S. et al. Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking //Frontiers in microbiology. – 2020. – T. 11. – P. 667.
5. Sergeev R. S. et al. Genome-wide analysis of MDR and XDR Tuberculosis from Belarus: Machine-learning approach //IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2017. – Vol. 16. – №. 4. – P. 1398-1408.
6. Rumiantsava K. et al. Search for Genomic Mutations Associated with Drug-resistant Tuberculosis. – 2021.