Identification of feature combinations in genome-wide association studies

Yuxiang Chen
Faculty of Applied Mathematics and Computer Science
Belarusian State University
Minsk, Belarus
c894424323@outlook.com

Alexander M. Andrianov
Institute of Bioorganic Chemistry
National Academy of Sciences of Belarus
Minsk, Belarus,
alexande.andriano@yandex.ru

Alexander V. Tuzikov
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Belarus,
tuzikov@newman.bas-net.by

***Abstract*—Association of single nucleotide polymorphisms (SNPs) with traits is the most popular method used in genome-wide association studies. SNPs with high association are often considered as a feasible locus for searching SNP combinations. However, this approach has a potential pitfall: correlated SNPs are usually not good partners to improve associations because their combinations do not enhance the quality of trait prediction. Therefore, a computational approach that could reduce the redundancy of SNPs is required. To solve this issue, an approach to reducing the SNP redundancy is proposed in this study. The feature relevance approach was used to select an optimized feature set which could generate the enhanced prediction performance. This approach was applied for the identification of mutations in *Mycobacterium tuberculosis* strains resistant to drugs. It was found that the combination of 2-4 SNPs could achieve an accuracy range from 65% to 90% to predict resistance for some drugs applied for the tuberculosis treatment.**

***Index Terms*—GWAS, SNPs, Feature Relevance, Feature Combinations, M.tuberculosis, Drug Resistance**

I. Introduction

Genome-wide association studies (GWAS) are conducted to identify single nucleotide polymorphisms (SNPs) associated with a phenotype [1]. In the general context of the SNP analysis, a prevalent method involves the individual testing of each SNP. This entails assessing the p-value associated with each SNP through statistical associations and subsequently comparing these p-values to a predefined threshold. SNPs with p-values falling below this threshold are deemed to be associated with the trait under investigation. However, it is imperative to acknowledge that the majority of traits are influenced by a complex interplay of multiple SNPs. Consequently, it becomes important to investigate the intricate relationships between combinations of SNPs and the phenotypic traits of interest.

Commonly employed methods encompass both traditional statistical methods and machine learning approaches. For instance, An et al. [2] employed a regression algorithm, utilizing the LASSO regression method, to discern correlations between SNPs and phenotypes. Mieth et al. [3] introduced an innovative two-step algorithm, COMBI which initially trains a support vector machine to identify a subset of candidate SNPs and subsequently conducts hypothesis testing on these SNPs, incorporating appropriate threshold corrections. Importantly, a significant computational burden is unavoidably incurred during the exploration of SNPs combinations in this search process. In our study, we present a less computationally demanding approach for identifying significant SNPs combinations associated with phenotype. The central concept of this approach focuses on employing the feature relevance to filter a subset of SNPs. Within this subset, SNPs exhibit the enhanced classification accuracy and reduced inter-SNP correlations. The method enhances the computational efficiency of the SNPs combination search by mitigating redundancy within the subset of candidate SNPs.

II. Materials and methods

A. Data

The original data sets used in the study included the drug susceptibility test data (DST) and genome-wide data corresponding to these cases. These data were taken from the tuberculosis (TB) portal [4] presenting an excellent platform for drug-resistant TB research. The DST data provide the verified information on the resistance or sensitivity of *Mycobacterium tuberculosis* (Mtb) samples to considered drugs. The data set C contained 3,178 samples and their resistance test results to 20 drugs, including 5 first-line drugs, namely Ethambutol (EMB), Isoniazid (INH), Pyrazinamide (PZA), Rifampin (RIF), Streptomycin (SM), and 15 second-line drugs, such as Amikacin (AMK), Amoxicillin-Clavulanate (AMX-CL), Capreomycin (CM), Cycloserine (CS), Ethionamide (ETO), Imipenem-Cilastatin (IMI), Kanamycin (KM), Linezolid (LZD), Levofloxacin (LFX), Mycobutin/Rifabutin (RFB), Moxifloxacin (MFX), Moxifloxacin 0.25 (MFX 0.25), Ofloxacin (OFX), Para-aminosalicylic acid (PAS), and Prothionamide (PTH).

B. GWAS problem

For each drug d, the samples can be divided into drug-resistant and sensitive ones. Therefore, a benchmark data set C can be presented as

$$C = C_d^+ \bigcup C_d^-,$$

where C_d^+ denotes a subset of resistant samples to drug d and C_d^- represents a subset of sensitive samples.

Each sample $S_i,\ i = 1, 2, ..., m$, in the GWAS is a genome consisting of four nucleotides $s_{ij} \in \{A, T, C, G\}$:

$$S_i = (s_{i,1}, s_{i,2}, ..., s_{i,n}).$$

The size n varies for different organisms. In particular, for *Mycobacterium tuberculosis*, this value is equal to 4,418,596 nucleotides. Genome sequence can contain various SNPs, which are present in a sufficiently large part of the population and mean a substitution of a single nucleotide at a specific position with another nucleotide. Some SNPs are important for the organism life and also relate to its phenotype (or trait), in our case, a microorganism resistance to some drug. In this case, it is assumed that a phenotype vector $\mathbf{y}_d = (y_1, y_2, ..., y_m)^t$ is given and $y_i = 1$ if sample S_i is resistant to the considered drug d. Otherwise, $y_i = 0$, and sample S_i is sensitive to this drug. Here, m is the size of the sample set C.

The GWAS problem consists in finding genome SNPs associated with phenotypes, if there are such SNPs. In this case, the problem size can be reduced to analysis of the SNPs sequence only instead of the whole genome. Suppose that genome data set C contains a sequence $\mathbf{x} = (x_1, x_2, ..., x_p)$ of p SNPs, which are obtained comparing genome samples with a reference genome, each $x_i,\ i = 1, 2, ..., p$, corresponds to some SNP in genome samples. Then vector $\mathbf{x}_{i,0} = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ describes the SNPs values for the genome sample $S_i,\quad x_{i,k} = 1$ if SNP number $k \le p$ exists in the sample and $x_{i,k} = 0$, otherwise. All SNPs in the sample set C for a drug d are defined by the SNPs matrix:

$$X_d = (x_{i,j})_{m \times p}. \qquad (1)$$

The following problem is considered in the paper.

Problem 1: Given a SNPs matrix X_d (1) and a phenotype vector $\mathbf{y}_d$ find SNPs associated with the phenotype.

First, we investigate the association between separate SNPs, presented in the sample set and given by vectors $\mathbf{x}_{0,j} = (x_{1,j}, x_{2,j}, ..., x_{m,j})^t$ for $j = 1, 2, ..., p$, i.e. by the columns of the SNPs matrix, and the phenotype vector $\mathbf{y}_d$. This approach is called a single-marker test. Secondly, the association between combinations of several SNPs and the phenotype vector called a multi-marker test will be investigated as well. For the second case, the problem gets a combinatorial nature and some heuristics should be used to verify various combinations of SNPs. One can unity column vectors corresponding to a combination of several SNPs using various logical operations and introduce a combined SNP into the SNPs matrix (1). In this study, we used the logical ”or” for every coordinate of the column vectors when uniting the corresponding column vectors of the SNPs matrix.

C. Prediction measures

The quality of prediction of phenotype based on a SNP in a considered genome position can be evaluated using several measures:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \qquad (2)$$

Prediction of drug resistance is based on the presence of a SNP under consideration. If a sample is resistant to drug d and is predicted correctly, then the prediction is considered as true positive (TP). Similarly, if a sample is sensitive to drug d and is predicted to be susceptible, the prediction is considered as true negative (TN). However, there is no perfect prediction, and, if a sample is resistant but is predicted to be sensitive, then the prediction is considered as false negative (FN). In analogous, if a sample is sensitive to drug but is predicted to be resistant, the prediction result is false positive (FP). The precision, recall and accuracy values for single SNPs can be directly computed from the SNPs matrix and phenotype vector.

However, selection of the best combination of SNPs according to the above criterion is a complicated combinatorial problem. It is therefore suggested to impose the constrains on a subset of SNPs combinations to find the appropriate solutions The maximum number of SNPs to be tested for the combination of SNPs will be limited by parameter $q \le p$ to reduce the computational complexity of the algorithm. Let parameter $l \le p$ limits the maximum number of SNPs to form a combined SNP.

D. Feature set reduction

Max-Relevance-Max-Distance method [5] was applied to introduce significance of pairs of SNPs allowing one to reduce the redundancy between features. Given two SNPs x_i and x_j define a relevance value $f_r(x_i, x_j)$ between them as follows:

$$f_r(x_i, x_j) = (1 - k_p(i, j))(a(x_i) + a(x_j)), \qquad (3)$$

where $k_p(x_i, x_j)$ is the Pearson correlation coefficient between vectors $\mathbf{x}_{0,i}$ and $\mathbf{x}_{0,j}$, corresponding to SNPs x_i and x_j, and $a(x_i), a(x_j)$ are the accuracy values of these SNPs calculated by formula (2).

The relevance values of SNP pairs are used here to select a subsequence $\mathbf{x}_q$ of $q \le p$ relevant SNPs from some initial sequence $\mathbf{x} = (x_1, x_2, ..., x_p)$. This subsequence will be used further to find combinations of SNPs associated with a phenotype. A selection procedure is organized as follows. Let us choose a pair of SNPs (x_i, x_j) from the sequence $\mathbf{x}$ with the maximum relevance value $f_r(x_i, x_j)$. If several pairs with the maximum relevance value exist, a pair with the smallest value $\min\{i, j\}$ of their indexes should be taken. SNPs x_i and x_j are then removed from the sequence $\mathbf{x}$ and added to the sequence $\mathbf{x}_q$. A new pair of SNPs (x_{i_1}, x_{j_1}) in the updated sequence $\mathbf{x}$ with the highest relevance value is selected and the sequence $\mathbf{x}_q$ is updated. After $q/2$ similar steps, one gets the sequence $\mathbf{x}_q$ containing q SNPs.

The following Algorithm 1 is proposed for identification of combinations of SNPs associated with the resistance to a considered drug. Note that parameter $l \leq p$ limits the maximum number of initial SNPs to form a combined SNP.

Algorithm 1 Combination of SNPs

```
Given a sequence of SNPs $\mathbf{x} = (x_1, x_2, ..., x_p)$ sorted
in non-ascending order of their accuracy values $a(x_1) \geq$
$a(x_2) \geq ... \geq a(x_p)$, compute the relevance values for each
SNP pair.

Select a subsequence $\mathbf{x}_q$ of $q \leq p$ relevant SNPs from the
sequence $\mathbf{x}$ .

Calculate accuracy values for all combined SNPs from the
sequence $\mathbf{x}_q$ consisting from up to $l$ initial SNPs.
```

III. Results

A. Characterization of the dataset

Table I presents the distribution of sensitive and resistant samples for each drug in the dataset.

TABLE I
Characterization of the dataset

Drug	Sensitive	Resistant	Drug	Sensitive	Resistant
EMB	839	617	INH	438	977
PZA	447	498	RIF	627	933
SM	539	875	AMK	942	1302
AMX-CL	626	297	MFX 0.25	254	137
CS	810	576	ETO	155	158
IMI	384	180	KM	861	767
LFX	777	1188	LZD	719	766
MFX	1159	413	CM	943	1185
OFX	589	385	PAS	705	1264
PTH	605	1127	RFB	197	169

B. Comparison of SNP combinations

The accuracy values were computed for all drugs and SNPs associated with drug resistance (p-values computed by the Fisher exact test were $\leq 10^{-5}$). Combinations of SNPs for $q = 50$ and $l = 5$ were found using Algorithm 1. Fig. 1 shows how the accuracy values depend on the number of SNPs in the combined mutations.

The most significant combinations with up to $l = 5$ SNPs were further compared using prediction measures (2). The results of this evaluation for the accuracy measure for the first-line and second-line drugs are shown in Tables II and III, respectively. In these tables, the combined SNPs with the maximum prediction accuracy for drug resistance are presented.

TABLE II
Results for the first-line drugs

Drug	SNP Combination	Accuracy
EMB	rs4248003 & rs4247429 & rs4247431 & rs1473246 & rs764817	0.80
INH	rs2155168 & rs761155	0.896
PZA	rs2155168	0.722
RIF	rs2155168 & rs761155	0.885
SM	rs2155168 & rs761155	0.805

TABLE III
Results for second-line drugs

Drug	Combination of SNPs	Accuracy
AMK	rs2155168	0.782
AMX-CL	rs1473246	0.838
CM	rs2155168	0.686
CS	rs7582 & rs7570	0.792
ETO	rs761155 & rs1673425	0.645
IMI	rs3380439 & rs1637255 & rs2196858	0.700
KM	rs7582 & rs1473246 & rs7570	0.749
LFX	rs2155168	0.745
LZD	rs2155168 & res761155	0.864
MFX	rs1473246	0.805
MFX 0.25	rs3738503 & rs7582 & rs7570 & rs1473246	0.747
OFX	rs2030634 &rs7582 & rs7570	0.754
PAS	rs2155168 & rs761155	0.851
PTH	rs2155168	0.775
RFB	rs2715344 & rs4247429 & rs1161026	0.708

IV. Conclusion

Genome-wide association studies confront formidable challenges, primarily stemming from the high dimensionality of data and the substantial computational burden, notably in the exploration of SNPs combinations. To address these challenges, it becomes imperative to reduce the number of SNPs under consideration for testing their combinations. The Max-Relevance-Max-Distance method offers a valuable approach for streamlining the set of SNPs, focusing solely on the relevant features. While this approach may not yield optimal SNP combinations, it does enhance the associations between SNP combinations and phenotypes when compared to individual SNPs in certain scenarios.

We introduced an effective approach for constructing feature combinations adapted for GWAS, which was tested in the context of drug resistance of *Mycobacterium tuberculosis*. For each of the 20 examined drugs, we obtained combinations of no more than 4 SNPs associated with drug resistance. Our results indicate accuracy levels ranging from 65% to 90%, testifying to the efficacy of the proposed approach to identifying SNPs combinations for GWAS.

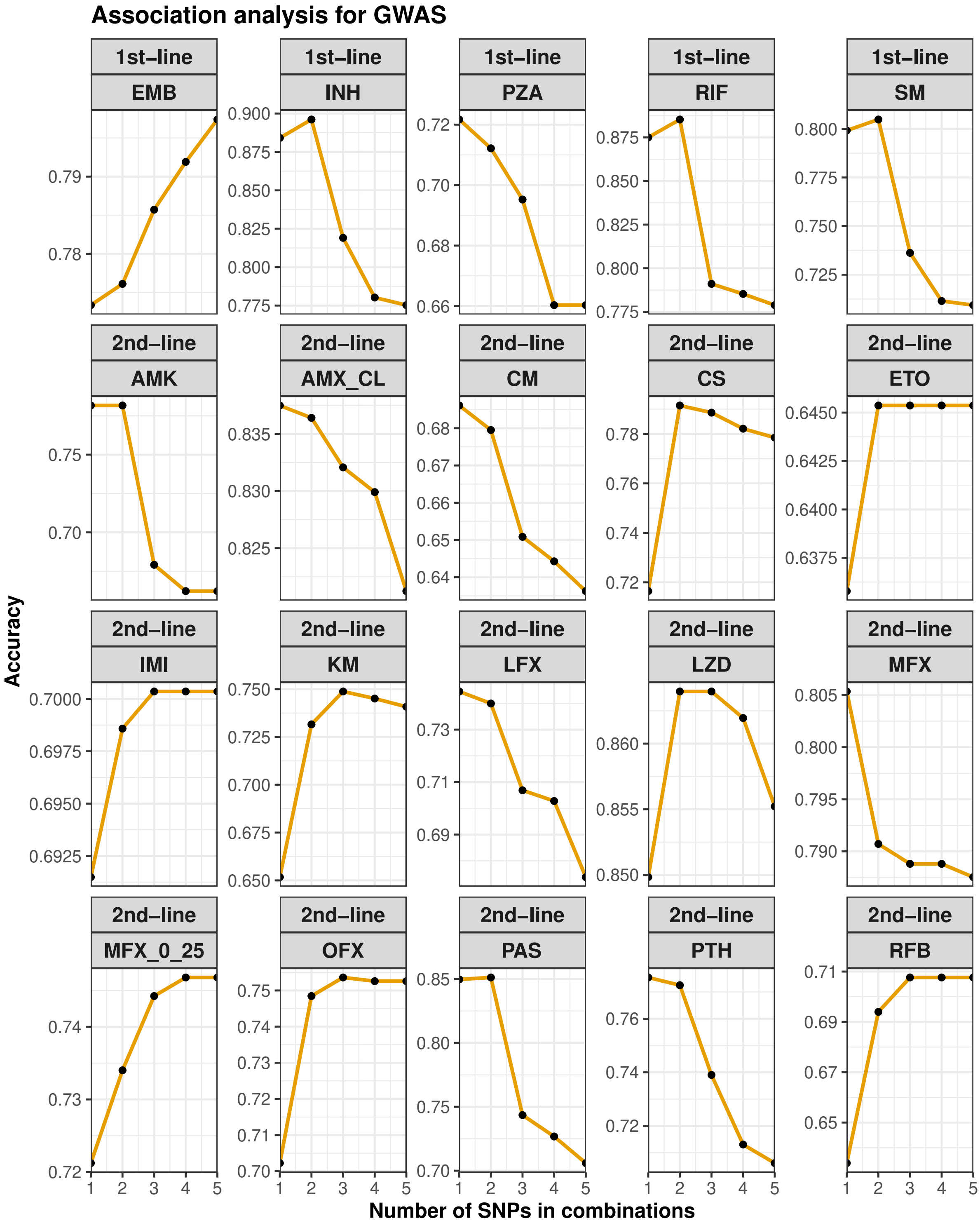

Fig. 1. For each of the 20 different drugs, combinations of SNPs most significantly associated with drug resistance were identified. Each subfigure represents a drug and the prediction accuracy of combinations consisting from $l \in \{1, 2, 3, 4, 5\}$ SNPs.

V. Acknowledgment

The authors are very grateful to the Consortium and the Drug Resistant Tuberculosis Portal Program (https://tbportals.niaid.nih.gov) for support and providing M.tuberculosis data.

References

[1] D. O. Enoma, J. Bishung, T. Abiodun, O. Ogunlana, and V. C. Osamor, "Machine learning approaches to genome-wide association studies," *Journal of King Saud University-Science*, vol. 34, no. 4, p. 101847, 2022.

[2] B. An, X. Gao, T. Chang, J. Xia, X. Wang, J. Miao, L. Xu, L. Zhang, Y. Chen, J. Li *et al.*, "Genome-wide association studies using binned genotypes," *Heredity*, vol. 124, no. 2, pp. 288–298, 2020.

[3] B. Mieth, M. Kloft, J. A. Rodríguez, S. Sonnenburg, R. Vobruba, C. Morcillo-Suárez, X. Farré, U. M. Marigorta, E. Fehr, T. Dickhaus *et al.*, "Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies," *Scientific reports*, vol. 6, no. 1, p. 36671, 2016.

[4] A. Rosenthal, A. Gabrielian, E. Engle, D. E. Hurt, S. Alexandru, V. Crudu, E. Sergueev, V. Kirichenko, V. Lapitskii, E. Snezhko *et al.*, "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *Journal of clinical microbiology*, vol. 55, no. 11, pp. 3267–3282, 2017.

[5] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.