DOI 10.15622/sp.58.1

L.V. UTKIN, I.L. UTKINA

A HIGH-PERFORMANCE GENOME-WIDE ASSOCIATION STUDY ALGORITHM BASED ON ANALYSIS OF PAIRS OF INDIVIDUALS

Utkin L.V., Utkina I.L. **A High-Performance Genome-Wide Association Study Algorithm based on Analysis of Pairs of Individuals.**

Abstract. An extremely simple and high-performance genome-wide association study (GWAS) algorithm for estimating the main and epistatic effects of markers or single nucleotide polymorphisms (SNPs) is proposed. The main idea underlying the algorithm is based on comparison of genotypes of pairs of individuals and comparison of the corresponding phenotype values. It is used the intuitive assumption that changes of alleles corresponding to important SNPs in a pair of individuals lead to a large difference of phenotype values of these individuals. In other words, the algorithm is based on considering pairs of individuals instead of SNPs or pairs of SNPs. The main advantage of the algorithm is that it weakly depends on the number of SNPs in a genotype matrix. It mainly depends on the number of individuals, which is typically very small in comparison with the number of SNPs. Another important advantage of the algorithm is that it can detect the epistatic effect viewed as gene-gene interaction without additional computations. The algorithm can also be used when the phenotype takes only two values (the case-control study). Moreover, it can be simply extended from the analysis of binary genotype matrices to the microarray gene expression data analysis. Numerical experiments with real data sets consisting of populations of double haploid lines of barley illustrate the outperformance of the proposed algorithm in comparison with standard GWAS algorithms from the computation point of view especially for detecting the gene-gene interactions. The ways for improving the proposed algorithm are discussed in the paper.

Keywords: GWAS, ANOVA, machine learning, epistasis, SNP, quantitative trait, distance metric.

1. Introduction. A genome-wide association study (GWAS) aims to discover genetic factors underlying phenotypic traits, i.e., GWAS examines the association between phenotypes and genetic variants or genotypes across the entire genome. It can be regarded as one of the methods for the well-known feature selection problem where features are the so-called single nucleotide polymorphisms (SNPs). SNPs are typically used as markers of a genomic region and can be defined as a DNA sequence variation where a single nucleotide (A, T, C, G) in the genomic sequence differs among the individuals of a biological species. It should be noted that most SNPs have no effect on the phenotype values or their effect is very insignificant. However, there are SNPs which might be very important in associations between SNPs and the phenotypes. Therefore, another formulation of the main aim of GWAS is to identify or select the most relevant SNPs which differentiate one group of individuals from another or which contribute to the phenotypic differences among the individuals.

From the machine learning point of view, a GWAS is one of the supervised classification or regression problems, where each individual can be

regarded as an example in terms of machine learning. It is defined by many SNPs which can be viewed as features in terms of machine learning. Therefore, many machine learning methods, including Lasso and ridge regressions, support vector machines, random forests, neural networks, have been used for GWAS. It should be noted that GWAS problem can be referred to as the well-known feature selection methods which are an important part of the machine learning approaches. In contrast to many standard statistical approaches underlying GWAS, machine learning models allows us to get a solution by taking into account the information of the whole genotype, and thus implicitly consider all possible correlations. Moreover, several variable importance measures can be derived from the machine learning models [1].

We point out some difficulties of solving the GWAS problem mentioned by many authors. First of all, the number of SNPs p is usually very large. It is typically 10-100 times the number of individuals n in the training sample. This is the so called $p > n$ (or large p small n) problem. Second, genetic mechanisms might involve complex interactions among genes and between genes and environmental conditions which are not fully captured by additive models [2, 3]. SNPs may interact in their effects on phenotype, i.e., there is the so-called epistatic effect. Third, many genetic variants are not genotyped, i.e., there are missing data in the genotype information. Fourth, GWAS is applied to find the association between SNPs and different kinds of the trait. It is mentioned by Korte and Farlow [4] in their interesting review of the GWAS methods that the successful GWAS methods applied to identifying SNPs contributing a disease (the two-valued or case-control phenotype) may have problems in finding SNPs associated with complex traits (quantitative or continuous phenotype).

A huge amount of the statistical procedures and methods solving the GWAS problem have been developed the last decades. A part of methods can be referred to as filter methods [5] which use statistical properties of SNPs to filter out poorly informative ones. The Fisher criterion, Pearson χ^2-test, Cochran-Armitage test are the well-known statistical methods for detecting differential SNPs between two samples. These methods can be joined as the so-called single-locus association tests because the tests are performed separately for each SNP when the case-control phenotypes are analyzed. For quantitative phenotypes, a standard tool is the one-way ANOVA [6]. Another part of methods uses various kinds of regression models which can be referred to as embedded methods [7-10]. One of the pioneering papers devoted to the use of regression models in SNP selection has been written by Lander and Botstein [11]. The regression models mainly include the Ridge regression and Lasso techniques, their combination called the elastic nets [12]. Comprehensive reviews of the methods and al-

gorithms using the regression models and their various modifications for solving the GWAS problems are provided by Wray et al. [13], Hayes [14], Visscher et al. [15], Bühlmann [16].

It has been mentioned that the standard GWAS analyzes each SNP separately in order to identify a set of significant SNPs showing genetic variations associated with the trait. However, an important challenge in the analysis of genome-wide data sets is taking into account the so-called epistatic effect when different epistatic loci interact in their association with phenotype. The epistatic effect can be viewed as gene-gene interaction when the action of one locus depends on the genotype of another locus. At the same time, there are different interpretations of the epistatic effect. A fundamental critical review of different definitions and interpretations of epistasis is provided in [17]. From the statistical point of view, the epistatic effect is the statistical deviation from the joined effects of two loci on the phenotype [18]. There is a series of interesting methods which use the statistical tests at their first step in order to reduce the set of SNPs. These are FastANOVA [19], FastChi [20], COE [21], TEAM [22]. We can also point out methods which differs from the filter methods, for example, the Bayesian epistasis association mapping method (BEAM) proposed by Zhang and Liu [23], tree-based methods like the random forests [24], the multifactor dimensionality reduction [25], modifications of the Lasso techniques [26], the ant colony optimization [27]. Comparative analyses of methods devoted to the epistatic interaction effect were provided by several authors [28, 29]. Analyzing these methods, we have to conclude that most of them have two steps (except for the methods with exhaustive consideration of all SNP pairs) such that the first step is for reducing the set of all SNPs to the most important ones, and the second step solves the SNP-SNP interaction problem.

From many approaches for solving the GWAS taking into account the epistatic effect, we would like to mark out a very interesting and efficient algorithm [30] that is subquadratic in the number of SNPs $\{0,1,2\}$. The authors [30] propose an algorithm for efficiently retrieving some predefined number of top scoring pairs among all pairs of SNPs, assuming binary phenotypes and the difference-in-correlation as the association criterion. Some implicit ideas of the algorithm will be used below.

In the present study, we propose a computationally extremely simple GWAS algorithm. It is based on the intuitive assumption that changes of alleles corresponding to important SNPs in a pair of individuals lead to large difference of phenotype values of these individuals. The main advantage of the algorithm is that it weakly depends on the number of SNPs in a genotype matrix. It mainly depends on the number of individuals, which is typically very small in comparison with the number of SNPs. We called the algorithm FAPI-GWAS (Fast Analysis of Pairs of Individuals for GWAS).

A preprint of the paper is given in https://arxiv.org/abs/1708.01746.

2. The proposed algorithm. We start with the following general definition of the association mapping problem. Let $\mathbf{X} = [X_1,..,X_p]$ be a genotype matrix for n individuals and p SNPs. From a statistical point of view, the genotype matrix can be treated as a predictor matrix and the marker genotypes as qualitative explanatory variables, i.e., $X_j = (x_{1j},..,x_{nj})^{\mathbf{T}}$ is a predictor representing the j-th SNP, $j = 1,...,p$. For bi-allelic SNPs, every x_{ij} is an allele of the i-th individual at the j-th SNP locus. It can be represented by the set $\{0,1\}$, where 0 and 1 stand for majority and minority alleles, respectively. A genotype may also be represented with numbers $\{0,1,2\}$ to represent the homozygous major allele $AA = 0$, heterozygous allele $Aa / aA = 1$, and homozygous minor allele $aa = 2$, respectively. A vector of alleles corresponding to the i-th individual will be denoted as $\mathbf{x}_i^{\mathbf{T}} = (x_{i1},..,x_{ip})$, $i = 1,..,n$. A quantitative trait of interest or a set of the phenotype values $y_i \in \mathbf{R}$, $i = 1,...,n$, can be viewed as a response vector $Y = (y_1,..,y_n)^{\mathbf{T}}$. A goal of GWAS is to find SNPs in $\mathbf{X}$, that are highly associated with Y, which will be called as important or significant SNPs.

In order to explain the introduced notation by means of an example, we provide Figure 1, where the genotype matrix $\mathbf{X}$ and the phenotype vector Y are illustrated. It can be seen from Figure 1 that $n = 12$ individuals with different plant height values $Y = (50,15,...,10,60)^{\mathbf{T}}$ (phenotype values) are defined by $p = 10$ SNPs with alleles x_{ij} taking the values 0 and 1. We have to develop an algorithm which selects the most important SNPs or their combinations from the point of view of their impact on the plant height.

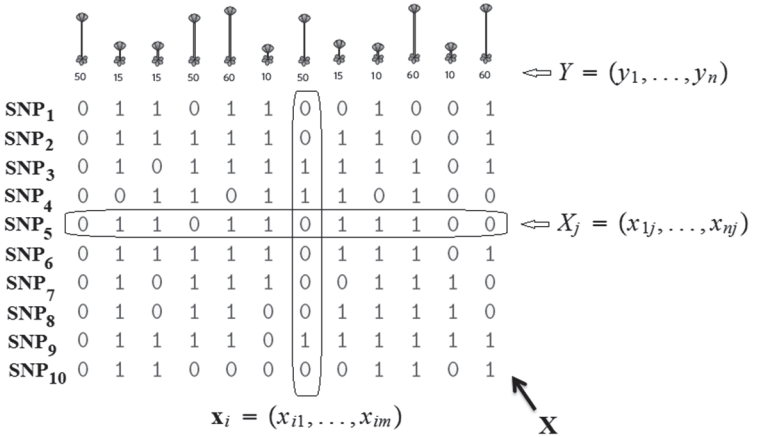

Fig. 1. An example of the genotype matrix $\mathbf{X}$ and the phenotype vector Y

The main idea underlying the FAPI-GWAS is based on comparison of genotypes of pairs of individuals and comparison of the corresponding phenotype values. At that, we use the following intuitive assumption. If genotypes of two individuals are close to each other and the corresponding phenotype values of these two individuals are far from each other, then the SNP-markers which correspond to different elements of the considered two genotypes *might be* important or contribute to the phenotype values. Indeed, if two individuals differ by some small number of genotype elements, then it is naturally to expect that their phenotypes are similar. However, if the corresponding phenotypes are substantially different, then it is naturally to suppose that this small number of distinguishing genotype elements define this large difference of phenotypes values. Of course, the large difference of the phenotype values may be caused by the noise or other random factors. Therefore, we cannot make any conclusions only on the basis of one pair of individuals. That is why the word combination *might be* used above means that this assumption may be wrong due to random character of the phenotype values. But we can make the conclusion by analyzing all pairs of individuals or a part of all pairs.

Informally, the FAPI-GWAS can be written as follows. First of all, we find all pairs $(\mathbf{x}_i, \mathbf{x}_j)$ of vectors of alleles. Then, we select some predefined number of the pairs which have largest differences of phenotype values and smallest distances between the vectors of alleles for every pair in accordance with some combined measure jointly characterizing the differences and the distances. The next step is to make a decision which SNPs contribute to the difference between the vectors of alleles for the best pairs. The use of the predefined number of pairs allows us to smooth possible outliers of the phenotype values due to random factors.

The above is illustrated in Figure 2, where three pairs of individuals are analyzed. The first pair does not show a large difference between the phenotype values. It is 5. Therefore, this pair is not interesting for us. The second and the third pairs have the difference 50 between the phenotype values. However, this difference for the third pair is caused by many (5) transitions between genotype values, which are underlined. Therefore, the third pair is also not interesting for us. At the same time, the second pair has only one transition. This implies that the large difference between phenotypes is caused by the 7-th SNP. Hence, we can conclude that this SNP is important.

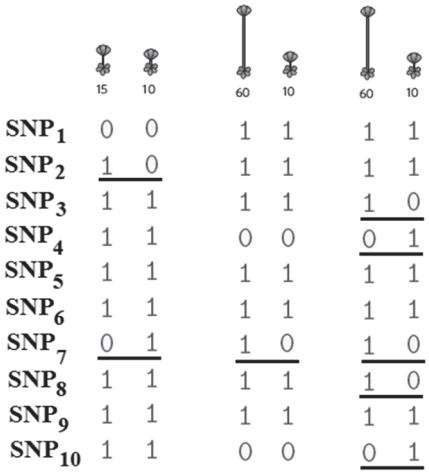

Fig. 2. An example of three different pairs of individuals

Formally, the proposed algorithm FAPI-GWAS can be represented as follows.

Step 1. All vectors of alleles $\mathbf{x}_1,...,\mathbf{x}_n$ are sorted in descending order of the corresponding phenotypes, i.e., $y_1 \geq ... \geq y_n$. This step simplifies comparison of phenotypes because the condition $y_i - y_j \geq 0$ for all $y_i \neq y_j$ is valid.

Step 2. All different pairs of individuals are composed. The number of pairs is $n(n-1)/2$. Only pairs $(\mathbf{x}_i, \mathbf{x}_j)$ such that $i < j$ are studied.

Step 3. For every pair $(\mathbf{x}_i, \mathbf{x}_j)$, the distance $\rho(\mathbf{x}_i, \mathbf{x}_j)$ between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, $i, j = 1,...,n$, $i < j$, is computed. A type of the distance depends on data. It can be the standard Hamming distance for binary variables x_{ij}. The standard Euclidean distance metric can be also used here.

Step 4. For every pair (i, j), the difference $d(y_i, y_j)$ between phenotype values y_i and $a_k(t) = \sum_{(i,j)\in J^*} r^*_{\text{norm}}(i,j)\mathbf{1}(\mathbf{z}_{ij}(k) = t)$, $t = -1, 0, 1$, $i, j = 1,...,n$, $i < j$, is computed. The condition $d(y_i, y_j) \geq 0$ is valid because phenotypes are sorted in descending order (see Step 1).

Step 5. For every pair (i, j), the ratio

$$r(i,j) = d(y_i, y_j) / \rho(\mathbf{x}_i, \mathbf{x}_j)$$

is computed. The larger the difference d and the smaller the distance ρ are, the larger ratio r is. The ratio r is a measure of target pairs.

Step 6. N largest values of $r(i,j)$ are selected. Denote these values as $r^*(i,j)$ and the set of their indices (i,j) as J^*. The value N can be regarded as a tuned parameter later. Another way is to compute the value N by constructing a cumulative probability distribution of the random variable r whose sample values are $r(i,j)$. It was observed by many numerical experiments that values $r(i,j)$ have a unimodal distribution. Moreover, if we assume that random variables taking values $d(y_i, y_j)$ and $\rho(\mathbf{x}_i, \mathbf{x}_j)$ have some distributions, for example, normal distributions, then r has one of the so-called ratio distributions, for example, the Cauchy distribution, the t-distribution, the F-distribution. Therefore, we take a predefined value of $q\%$ quantile of the random variable r and find all values of the ratio such that their empirical distribution function is larger than $q/100$. In this case, we derive some value of N from the above procedure, and q can be viewed as a tuned parameter of the algorithm.

Step 7. For every pair (i,j) from J^*, we find a subset of elements of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ which differentiate these vectors. In particular, if $x_{ij} \in \{0,1\}$, then $\mathbf{z}_{ij} = \mathbf{x}_i - \mathbf{x}_j$. The vector $\mathbf{z}_{ij}$ has element -1 at the k-th position if there is the transition from 0 in $\mathbf{x}_i$ to 1 in $\mathbf{x}_j$ at the k-th position, element 1 if there is the transition from 1 in $\mathbf{x}_i$ to 0 in $\mathbf{x}_j$ at the same position, and element 0 by transitions from 0 to 0 or from 1 to 1 at the same position, i.e., there holds

$$\mathbf{z}_{ij}(k) = \begin{cases} -1, & \text{if } \mathbf{x}_i(k) = 0,\ \mathbf{x}_j(k) = 1, \\ 1, & \text{if } \mathbf{x}_i(k) = 1,\ \mathbf{x}_j(k) = 0, \\ 0, & \text{if } \mathbf{x}_i(k) = \mathbf{x}_j(k). \end{cases}$$

Only elements of $\mathbf{z}_{ij}$ with values -1 and 1 are interesting for us because they indicate positions where vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are different, which, in turn, indicate *possible* important SNPs. In the case $x_{ij} \in \{0,1,2\}$, we have six transitions $0 \to 1$, $0 \to 2$, $1 \to 0$, $1 \to 2$, $2 \to 0$, $\mathbf{x}_j$ enumerated as $-3, -2, -1, 1, 2, 3$, and three transitions $0 \to 0, 1 \to 1, 2 \to 2$ denoted as 0.

Step 8. For the k-th SNP under condition $x_{ij} \in \{0,1\}$, we use the ratio $r^*(i,j)$ for computing summed weights of elements -1, 0, 1 at the k-th position in $\mathbf{z}(i,j)$ denoted as $a_k(-1)$, $a_k(0)$, $a_k(1)$, i.e., we compute

$$a_k(t) = \sum_{(i,j) \in J^*} r^*_{\text{norm}}(i,j)\mathbf{1}(\mathbf{z}_{ij}(k) = t),\ t = -1, 0, 1.$$

Here $\mathbf{1}(\mathbf{z}_{ij}(k) = t)$ is the indicator function taking the value 1 if $\mathbf{z}_{ij}(k) = t$, and the value 0 otherwise; r^*_{norm} is the normalized ratio. We can also take $r^*_{\text{norm}}(i, j) = 1$ for all values $(i, j) \in J^*$. In this simplified case, we find the numbers of elements -1, 0, 1 at the k-th position in vectors $\mathbf{z}_{ij}$, $(i, j) \in J^*$. If $x_{ij} \in \{0,1,2\}$, then t takes values from $T = \{-3,-2,-1,0,1,2,3\}$.

Step 9. For the k-th SNP, we compare two numbers $a_k(t=0)$ and $\sum_{t\neq 0} a_k(t)$ with a decision threshold h, i.e., we compare the summed weights of transitions which do not contribute to the difference of vectors of alleles and which correspond to transitions $0 \to 0$, $1 \to 1$, and the weights of transitions with different values of alleles. If the inequality

$$\sum_{t\neq 0} a_k(t) / a_k(t=0) \geq h$$

is valid, then the corresponding k-th SNP is important, otherwise it does not belong to the subset of important SNPs. The decision threshold is typically equal to 1.

Step 10. For every target SNP, we write the value

$$S_k = \arg \max_{t\in\{-1,1\}} a_k(t).$$

Values $S_k = -1$ or 1 mean that the allele, corresponding to the k-th SNP and having values 1 or 0, respectively, contributes to decreasing of the phenotype.

Let us illustrate the above algorithm by means of a toy example. Suppose we have $n = 3$ individuals whose genotype matrix for 5 bi-allelic SNPs is represented by symbols 0 and 1 which stand for major and minor alleles, respectively. The sorted phenotype values are 45, 15, 10. The initial data are shown in Table 1. We have three pairs of vectors of alleles such that the phenotype differences $d(y_i, y_j)$, the genotype transitions, the corresponding Hamming distances between vectors of alleles in every pair and the ratios $r(i, j)$ are given in Table 2.

Table 1. Genotypes and phenotypes of three individuals

Phenotypes		45	15	10
SNPs	1	0	0	1
	2	0	0	1
	3	1	0	0
	4	1	1	0
	5	1	1	1

Table 2. The genotype transitions and the values ρ and r

$d(y_i, y_j)$		30	35	5
SNPs	1	$0 \to 0$	$0 \to 1$	$0 \to 1$
	2	$0 \to 0$	$0 \to 1$	$0 \to 1$
	3	$1 \to 0$	$1 \to 0$	$0 \to 0$
	4	$1 \to 1$	$1 \to 0$	$1 \to 0$
	5	$1 \to 1$	$1 \to 1$	$1 \to 1$
$\rho(\mathbf{x}_i, \mathbf{x}_j)$		1	4	3
$r(i, j)$		30	8.75	1.667

Suppose that the threshold N for selecting the largest values of $r(i, j)$ is 2. Table 3 shows individuals satisfying this condition and the values $\mathbf{z}(i, j)$ of transitions taking the values $-1, 0, 1$ (see Step 7). It can be seen from Table 3 that only the third SNP has two non-zero elements $\mathbf{z}(i, j)$. This implies that only the third SNP is important. Indeed, it is obviously from Table 3 that the largest difference is observed between phenotypes of the first and the second individuals. Moreover, only the third SNP separates the first and the second vectors of alleles. Intuitively, we can conclude that this SNP is a reason for the large difference between phenotypes of the first and the second individuals.

Table 3. Vectors $\mathbf{z}_{ij}$ and decision making about the important SNP

$r^*(i, j)$		30	8.75
SNPs	1	0	-1
	2	0	-1
	3	*1*	*1*
	4	0	1
	5	0	1

The FAPI-GWAS for determining important SNPs is given as Algorithm 1.

Algorithm 1. A simple FAPI-GWAS algorithm

Require: $\mathbf{X}_{n \times p} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ (binary genotype matrix), Y (phenotype vector), N, h (parameters)

Ensure: S_k (imported SNPs)

1. Order $(\mathbf{x}_1, ..., \mathbf{x}_n)$ such that $y_1 \geq ... \geq y_n$

2. *for each* $i \leq n$ и $j > i$ *do*

3. Build a pair $(\mathbf{x}_i, \mathbf{x}_j)$

4. Compute $\rho(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2$

5. Compute $r(i,j) = (y_i - y_j) / \rho(\mathbf{x}_i, \mathbf{x}_j)$

6. Compute $\mathbf{z}_{ij} = \mathbf{x}_i - \mathbf{x}_j$

7. *end for*

8. $J^* = \{(i,j):\ N$ largest values of $r(i,j)\}$

9. *for each* $k \le p$ *do*

10. Compute $a_k(t) = \sum_{(i,j)\in J^*} r^*_{\text{norm}}(i,j)\mathbf{1}(\mathbf{z}_{ij}(k) = t),\ t = -1, 0, 1.$

11. *If* $\sum_{t\neq 0} a_k(t) \ge h \cdot a_k(t=0)$ *then* the *k*-th SNP is important,

12. Compute $S_k = \arg\max_{t\in\{-1,1\}} a_k(t)$

13. *end if*

14. *end for*

The Euclidean distance $\rho(\mathbf{x}_i, \mathbf{x}_j)$ is taken in Algorithm 1. However, other distance metrics depending on the analyzed dataset can be used. These distance metrics can be regarded as tuning elements of the model.

3. Properties of the algorithm. Let us point out some properties and advantages of the FAPI-GWAS.

1) The epistatic effect which is viewed as gene-gene interaction should not be separately analyzed. It is implicitly included into the proposed algorithm. Indeed, we do not consider single SNPs. For every pair of vectors of alleles, the difference of the vectors is computed for all SNPs simultaneously. So, if there is a combination of alleles which significantly impact on the phenotype, it produces a large difference between the corresponding phenotype values. This is a very important property which allows us to significantly reduce the computational burden needed for consideration of many SNP pairs.

2) The FAPI-GWAS is very simple. Its computational complexity is $O(p + n^2)$, i.e. the complexity is linear with the number of SNPs p. This is a very important property of the algorithm because the number of SNPs p is typically 10-100 times the number of individuals n in the training sample for many problems. Moreover, the algorithm does not require special procedures like Lasso, etc. For comparison purposes, a very interesting algorithm for 2-locus genome-wide association studies [30] has the complexity $O(p^{3/2})$. Algorithms FastANOVA [19] and TEAM [22] have the complexity $O(p^2 n)$.

3) The FAPI-GWAS does not depend on the set of allele values. For example, a few trivial changes are needed to consider the case $x_{ij} \in \{0,1,2\}$. Moreover, the important feature of the algorithm is that the values $\{0,1,2\}$ or $\{0,1\}$ are viewed as categorical numbers without order, for example, $0<1<2$. The FAPI-GWAS can be modified for the case $x_{ij} \in \mathbf{R}$ which takes place in the microarray gene expression data analysis.

4) Another advantage of the FAPI-GWAS is handling missing data in the genotype matrix. We do not need to apply special procedures for preprocessing missing data and their imputation. The missing data just extend the set of values of every x_{ij}. We use the conservative strategy. For example, suppose $x_{ij} \in \{0,1\}$ and the missing value is denoted as 2. If we have two missing values at the same *k*-th position in vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, then $\mathbf{z}_{ij}(k) = 0$. This value means that we do not consider the *k*-th position in vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ as a candidate for getting an important SNP. At the same time, when we have a single missing value at the *k*-th position in vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, then $\mathbf{z}_{ij}(k) \neq 0$ in accordance with the strategy that a larger number of important SNPs is preferable because the second selection from a small subset of important SNPs should be carried out by means of the well-known standard procedures.

5) The FAPI-GWAS can be used when the phenotype takes only two values (the case-control study). It is obvious in this case that only a set composed from pairs of individuals taken from the case and control groups, respectively, is analyzed. Indeed, $d(y_i, y_j) = 0$, $r(i,j) = 0$ if $y_i = y_j$, and $d(y_i, y_j) = 1$, $r(i,j) = 1/\rho(\mathbf{x}_i, \mathbf{x}_j)$ if $y_i \neq y_j$ (we assume that the vectors of alleles are sorted in descending order of the corresponding phenotypes).

6) For many available algorithms of GWAS using filter methods for selection of the most important SNPs like the Fisher exact test, the one-way ANOVA, etc. we have to predefine a limit number of the important SNPs. The FAPI-GWAS determines this number itself.

7) The FAPI-GWAS can be tuned by means of the parameter N (the number of largest values of the rate r) or parameter q. On the one hand, too small values of the parameter N may lead to a large number of target SNPs. As a result, we have to use some additional procedures for restricting the number of SNPs. On the other hand, large values of N may lead to possible missing SNPs which actually may be very important. There is a compromise choice

of N which can be carried out by considering all possible values of N in a predefined grid. Another parameter for tuning is the decision threshold h.

8) The FAPI-GWAS is flexible. This means that many its elements can be changed. For example, there are many metrics for computing distances between vectors of alleles such that the choice of an appropriate metric might improve the algorithm. Similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ and dissimilarity $\rho(\mathbf{x}_i, \mathbf{x}_j)$ measures of two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ can be applied. If we use similarity measures, then $r(i, j) = d(y_i, y_j) \cdot S(\mathbf{x}_i, \mathbf{x}_j)$. Another element which could be changed is the choice of the ratio r. The proposed ratio is one of the possible measures for the target pair localization. It is just a simplest way for defining the measure. Perhaps, other measures might also improve the algorithm.

4. Numerical experiments. Numerical experiments are carried out on three populations of double haploid (DH) lines of barley:

1) The first dataset consists of 175 DH lines of barley [31, 32]. The data are available at Oregon Wolfe Barley Data (OWBD) and GrainGenes Tools (http://wheat.pw.usda.gov/ggpages/maps/OWB/). The lines are analyzed with respect to the heading date trait. The linkage map consists of 1328 SNPs.

2) The second dataset consists of 92 DH lines of barley from the Dicktoo x Morex cross and described in [33, 34, 35]. The data are available at http://wheat.pw.usda.gov/ggpages/DxM/. We analyze the lines with respect to two phenotypic traits: heading date with and without vernalization with an 8-h light/16-h dark photoperiod regime. The linkage map consists of 117 SNPs.

3) The third population dataset includes 150 DH lines of barley from the Steptoe x Morex cross [36]. The corresponing data are available at http://wheat.pw.usda.gov/ggpages/SxM. The linkage map consists of 223 SNPs. The lines are analyzed with respect to the heading date trait measured in 16 environments and grain yield trait measured in 6 environments.

The missing data are handled by means of extending the set of values of every x_{ij}, i.e., the set of values $\{0,1\}$ is extended on the set $\{0,1,2\}$.

First, we investigate DH lines of barley from OWBD. The parameter q is 97%. In order to compare the proposed algorithm, we apply the standard tool ANOVA to testing the association between a single marker and a continuous outcome. The F-test is used to assess whether the expected values of a quantitative variable within several predefined groups differ from each other. From this, we can retrieve a p-value for the significance of association between each SNP and the phenotype. Then we correct for multiple testing using the Holm-Bonferroni method. The Manhattan plot generated from the obtained p-values is shown in Figure 3 (the left plot). One can see from the Figure 3 that the significant SNPs have numbers close to 139, 725, 1100. SNPs with these numbers have the smallest p-values.

Let us look at Figure 3 (the right plot) now. It shows a similar Manhattan plot, but significant SNPs are obtained by using the FAPI-GWAS, and p-values are computed for this set again using the Holm-Bonferroni correction. However, the first step of the FAPI-GWAS provides not only the significant SNPs which coincide with the SNPs derived by the standard tool ANOVA. It provides SNPs with numbers 1169 and 1302, which do not belong to the set of significant SNPs obtained by means of the ANOVA. It turns out that the p-values of these single SNPs are larger than 0.05, i.e., they cannot be viewed as significant ones. In contrast to the single-locus approach applied before, we perform the ANOVA test in order to identify interacting SNP-pairs that have strong association with the phenotype. It is important to note that the two-locus ANOVA test is performed on a small number of candidate SNP-pairs which have been obtained by means of the FAPI-GWAS. It turns out that SNPs with numbers 1169 and 1302 interact with SNPs 729 and 725, respectively, such that the corresponding p-values (0.021 and 0.047) after the Holm-Bonferroni correction are smaller than 0.05. In other words, the FAPI-GWAS allows us to implement the efficient epistasis detection.

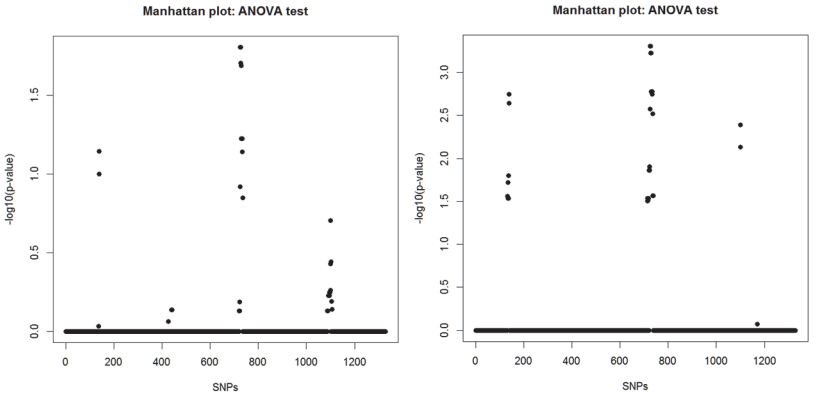

Fig. 3. The Manhattan plot for the OWBD using standard method (left) and the FAPI-GWAS (right)

Let us study the dataset obtained from the Dicktoo x Morex cross. According to Pan et al. [35] (Page 905), top ranked SNPs for heading date with and without vernalization are ABC170-CD064 and Dhn1-BCD265b which correspond to the following numbers of SNPs 22-24 and 111-113, respectively. The ANOVA is applied here again. We get two SNPs with numbers 22 and 112 having the smallest p-values 1.32×10^{-5} and 2.66×10^{-9}, respectively. The corresponding Manhattan plot is shown in Figure 4 (the left plot). Numerical experiments with using the FAPI-GWAS provide quite the same

results. They are shown in Figure 4 (the right plot). However, the FAPI-GWAS indicates that there is the 49-th SNP (saflp35) which has a large p-values, but its interaction with SNPs 112 and 22 gives the p-values 0.0135 and 0.0144, respectively. All p-values are computed by using the Holm-Bonferroni correction.

We get similar results for the unvernalized treatment (the second phenotypic trait). In addition, we obtain SNPs with numbers 36, 59, 76, which are called as saflp219, SOLPRO, HorB, respectively, and which are located on different chromosomes. These SNPs interact with the SNP 22 with the corresponding p-values 0.0034, 0.038, 0.045, respectively.

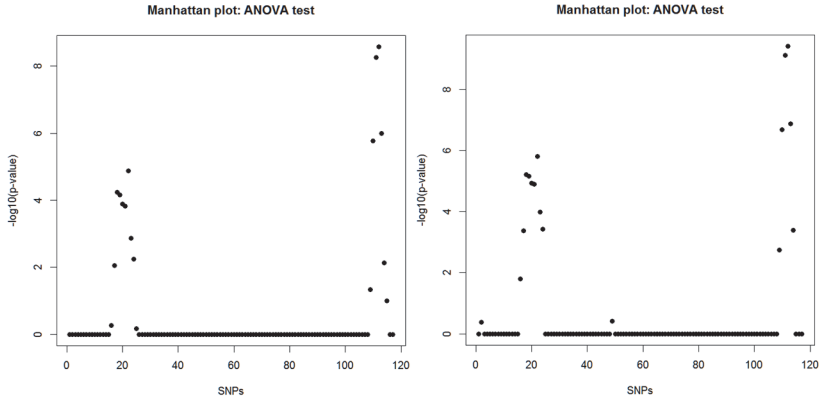

Fig. 4. The Manhattan plots for the Dicktoo x Morex data set using standard method (left) and the FAPI-GWAS (right)

The third dataset obtained from the Steptoe x Morex cross. First, we analyze lines with respect to the heading date trait. According to the standard ANOVA test, the 47-th SNP has the smallest p-value which is 8.5×10^{-19}. Other significant SNPs have numbers 68, 82, 205. However, they have larger p-values, namely, 1.48×10^{-3}, 1.37×10^{-5}, 8.12×10^{-3}. The Manhattan plot generated from the obtained p-values is shown in Figure 5 (the left plot). By using the FAPI-GWAS, we get quite the same results. The Manhattan plot generated from the p-values obtained by means of the FAPI-GWAS is shown in Figure 5 (the right plot). Moreover, we obtain the strong interactions of SNPs 47×82 (p-value is 1.7×10^{-26}), 47×205 (p-value is 7.66×10^{-24}), 47×68 (p-value is 2.07×10^{-21}), 47×165 (p-value is 4.65×10^{-13}), 47×102 (p-value is 1.07×10^{-12}), 47×134 (p-value is 2.26×10^{-12}).

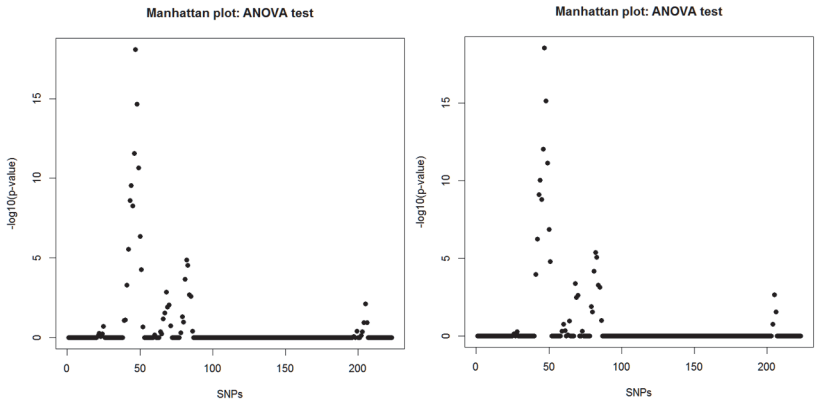

Fig. 5. The Manhattan plots for the Steptoe x Morex data set (the heading date trait) using standard method (left) and the FAPI-GWAS (right)

The standard analysis with respect to the grain yield trait gives the following significant SNPs and their p-values in parentheses: 82 (2.69×10^{-11}), 20 (1.03×10^{-2}), 68 (3.02×10^{-2}), 129 (4.04×10^{-1}). The FAPI-GWAS provides the same significant SNPs. Additionally, we get the following interacting SNPs: 82×112 (1.8×10^{-11}), 82×151 (5.56×10^{-8}), 82×135 (5.12×10^{-6}), 82×195 (1.22×10^{-5}). The corresponding Manhattan plots generated from the p-values for the grain yield trait are shown in Figure 6.

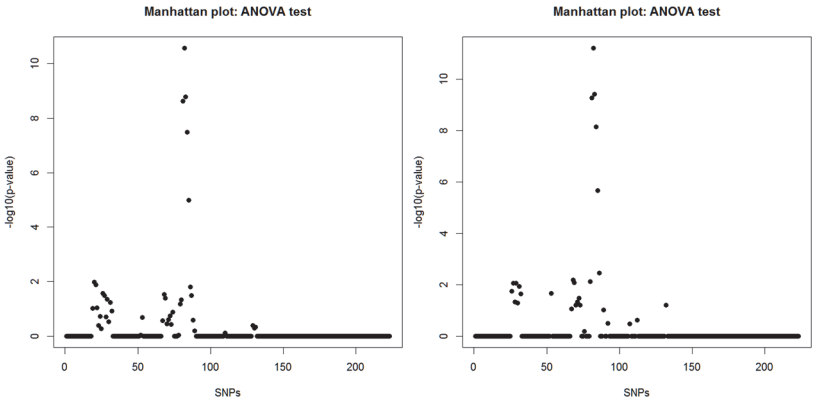

Fig. 6. The Manhattan plots for the Steptoe x Morex data set (the grain yield trait) using standard method (left) and the FAPI-GWAS (right)

5. Discussion how to improve the algorithm. Let us point out shortcomings of the FAPI-GWAS and discuss possible ways to overcome them and to improve the algorithm.

First, numerous experiments with real data illustrate that the FAPI-GWAS selects groups of adjacent strongly correlated SNPs in the same chromosomal region which are not inherited randomly. This effect is similar to those taking a place in the ridge regression algorithm which tends to select all of the correlated SNPs and make their importance coefficients to be equal. In contrast to the ridge regression, the Lasso method tends to select only one SNP from the group of correlated ones. Therefore, the problem of correlated SNPs can be solved by using a two-step procedure. The first step is based on the FAPI-GWAS. The result of this step is a small set of important SNPs. The second step uses the Lasso method or its modification, for example, the adaptive Lasso, in order to remove the correlated SNPs from the available small set. Moreover, we can use a modification of the Lasso which takes into account the epistatic effect because the number of possible pairs of SNPs after the first step is rather small.

Another way to treat with the correlated SNPs is to use the standard tools for testing the association between single SNPs and a continuous phenotype, including for example, one-way ANOVA. In order to identify two-locus epistatic effect or interacting SNP-pairs that have strong association with the phenotype, an algorithm for the two-locus ANOVA test can be used. There are many approximated methods for reducing the computational burden. They are reviewed in detail for a case-control study when the phenotype can be represented as a binary variable with 0 representing controls and 1 representing cases as well as for the quantitative trait locus analysis when the phenotype is quantitative [4]. Most methods are reduced to two steps. The first step is reduction of a set of SNPs in order to apply standard statistical procedures to this reduced set of SNPs. The standard statistical procedures make up the second step. The reduction of the set of correlated SNPs can be successfully implemented by means of the FAPI-GWAS as the first step. As a result, we get a small subset of important SNPs which can be processed by statistical tests, for instance, ANOVA test, in order to remove the correlated SNPs located on the same chromosome.

We point out another shortcoming which has been observed in numerical experiments. Since the number of SNPs is much larger than the number of individuals, then we observe only a very small number of vectors $\mathbf{x}_i$ among all possible vectors. This implies that contributions of some important SNPs in a pair of vectors of alleles $(\mathbf{x}_i, \mathbf{x}_j)$ may be hidden when there are many transitions in this pair, for example, from 0 to 1 and from 1 to 0. In this case, the distance between vectors is large, and this

pair does not get to a set of N best pairs with the largest ratios $r(i, j)$. One of the ways to overcome the difficulty is to apply the combination of the bagging method [37] and the random subspace method [38]. The FAPI-GWAS can be improved by using a combination of the bagging method for individuals and the random subspace method for SNPs. The random sampling of individuals in the proposed method allows us to smooth some outliers of the phenotype caused by random factors. By means of the random sampling of SNPs, we try to reduce the effect of SNPs which mask the effect of subsets of important SNPs.

6. Conclusion. In this paper, a very fast and simple algorithm for GWAS, including SNP interaction detection, has been presented. In spite of its simplicity, the FAPI-GWAS can be applied to various GWAS problems and cases from the analysis of binary genotype matrices to the microarray gene expression data analysis. Moreover, the algorithm can be simply extended, for example, on the bagging method.

At the same time, it is important to note that the algorithm should be used jointly with another algorithm, for example, with the ANOVA tests to identify the association between a single marker or interacting SNP-pairs and a continuous outcome. At that, the second stage uses a set of significant SNPs which is obtained at the first stage by means of the FAPI-GWAS.

The results of numerical experiments and the logic underlying the FAPI-GWAS have demonstrated that it outperforms the standard algorithms from the computational point of view for many real data sets. Moreover, it takes into account the epistatic effect or the SNP-SNP interaction. We have analyzed DH populations of barley for purposes of numerical experiments. The experiments have illustrated the FAPI-GWAS efficiency. The obtained sets of significant SNPs have coincided with similar sets obtained by means of standard algorithms. Moreover, we could see that SNP-SNP interactions detected by means of the FAPI-GWAS were successfully validated by performing the two-locus ANOVA test. However, we have investigated only rather small data sets and only a simplest implementation of the FAPI-GWAS. It has been done because we aimed to compare results of the FAPI-GWAS with the well-known standard technique. We aimed to get added evidence that the algorithm copes with tasks of the GWAS. It should be noted that a lot of experiments have to be performed in order to evaluate how the FAPI-GWAS handles various types of data set, large data sets, how its modifications and extensions outperform the available algorithms. These questions are directions for further research.

References

1. Mieth B. et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific reports*. 2016. vol. 6. Article number: 36671. pp. 1–14.

2. Koo C.L., Liew M.J., Mohamad M.S., Salleh A.H.M. Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. *BioMed Research International.* 2013. vol. 2013. Article ID 432375. 13 p.
3. Yang J., Lee S.H., Goddard M.E., Visscher P.M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Genome-wide association studies and genomic prediction.* 2013. pp. 215–236.
4. Korte A., Farlow A. The advantagees and limitations of trait analysis with GWAS: a review. *Plant Methods*. 2013. vol. 9(29). pp. 1–9.
5. Chandrashekar G., Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014. vol. 40(1). pp. 16–28.
6. Kim T.K. Understanding one-way ANOVA using conceptual figures. *Korean Journal of Anesthesiology.* 2017. vol. 70(1). pp. 22–26.
7. Dimou N.L., Tsirigos K.D., Elofsson A., Bagos P.G. GWAR: robust analysis and meta-analysis of genome-wide association studies. *Bioinformatics*. 2017. vol. 33(10). pp. 1521–1527.
8. Duan W. et. al. A fast algorithm for Bayesian multi-locus model in genome-wide association studies. *Molecular Genetics and Genomics*. 2017. vol. 292(4). pp. 923–934.
9. Franberg M. et. al. Fast and general tests of genetic interaction for genome-wide association studies. *PLOS Computational Biology*. 2017. vol. 13(6). pp. e1005556.
10. Wang S., He S., Yuan F., Zhu X. Tagging SNP-set selection with maximum information based on linkage disequilibrium structure in genome-wide association studies. *Bioinformatics.* 2017. vol. 33(14). pp. 2078–2081.
11. Lander E.S., Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. 1989. vol. 121(1). pp. 185–199.
12. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning. Springer. 2013. 426 p.
13. Wray N.R. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*. 2013. vol. 14(7). pp. 507–515.
14. Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*. 2013. vol. 1019. pp. 149–169.
15. Visscher P.M. et. al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*. 2017. vol. 101(6). pp. 5–22.
16. Bühlmann P. High-dimensional statistics, with applications to genome-wide association studies. *EMS Surveys in Mathematical Sciences*. 2017. vol. 4(1). pp. 45–75.
17. Uppu S., Krishna A., Gopalan R. A review of machine learning and statistical approaches for detecting SNP interactions in high-dimensional genomic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016. vol. PP(99). pp. 1545–5963.
18. Li J., Zhong W., Li R., Wu R. Boost: A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *The Annals of Applied Statistics*. 2014. vol. 8(4). pp. 2292–2318.
19. Zhang X., Zou F., Wang W. FastANOVA: an efficient algorithm for genome-wide association study. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008. pp. 821–829.
20. Zhang X., Zou F., Wang W. FastChi: an effcient algorithm for analyzing gene-gene interactions. Proceedings of the Pacific Symposium on Biocomputing PMC. 2009. vol. 14. pp. 528–539.
21. Zhang X. et. al. COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *Research in Computational Molecular Biology*. 2009. pp. 253–269.
22. Zhang X., Huang S., Zou F., Wang W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*. 2010. vol. 26(12). pp. i217–i227.

23. Zhang Y., Liu J.S. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*. 2007. vol. 39(9). pp. 1167–1173.
24. Li J. et. al. Detecting gene-gene interactions using a permutation-based random forest method. *BioData Mining*. 2016. vol. 9(1). pp. 14–30.
25. Moore J.H., Andrews P.C. Epistasis analysis using multifactor dimensionality reduction. *Epistasis.* 2015. pp. 301–314.
26. Bocianowski J. Estimation of epistasis in doubled haploid barley populations considering interactions between all possible marker pairs. *Euphytica*. 2014. vol. 196(1). pp. 105–115.
27. Zhou Z., Liu G., Su L. A new approach to detect epistasis utilizing parallel implementation of ant colony optimization by MapReduce framework. *International Journal of Computer Mathematics.* 2016. vol. 93(3). pp. 511–523.
28. Ma L., Clark A.G., Keinan A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS genetics*. 2013. vol. 9(2). pp. e1003321-12.
29. Niel C., Sinoquet C., Dina C., Rocheleau G. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*. 2015. vol. 6. Article 285. pp. 1–19.
30. Achlioptas P., Scholkopf B., Borgwardt K. Two-locus association mapping in subquadratic time. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011. pp. 726–734.
31. Chutimanitsakun Y. Construction and application for QTL analysis of a restriction site associated DNA (rad) linkage map in barley. *BMC Genomics*. 2011. vol. 12. no. 4. pp. 1–13.
32. Cistue L. et. al. Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics*. 2011. vol. 122(7). pp. 1399–1410.
33. Hayes P.M. et. al. Quantitative trait loci on barley (Hordeum vulgare L.) chromosome 7 associated with components of winterhardiness. *Genome*. 1993. vol. 36(1). pp. 66–71.
34. Hayes P. et. al. The Dicktoo x Morex population. *Plant Cold Hardiness*. 1997. pp. 77–87.
35. Pan A. et. al. Genetic analysis of the components of winterhardiness in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*. 1994. vol. 89(7-8). pp. 900–910.
36. Close T.J. et. al. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*. 2009. vol. 10. no. 1. pp. 582.
37. Breiman L. Bagging predictors. *Machine Learning*. 1996. vol. 24(2). pp. 123–140.
38. Ho T.K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998. vol. 20(8). pp. 832–844.

Utkin Lev Vladimirovich — Ph.D., Dr. Sci., professor, professor of telematics (under Russian state scientific center for robotics and technical cybernetics (RTC)) of Institute of applied mathematics and mechanics, Peter the Great Saint-Petersburg Polytechnic University. Research interests: machine learning, uncertainty representation, decision making under incomplete information, reliability theory, bioinformatics. The number of publications — 344. lev.utkin@gmail.com, http://levvu.narod.ru; 29, Polytechnicheskaya, St. Petersburg, 195251; office phone: +7(812)552-6521.

Utkina Irina Lvovna — Master's degree student, Skolkovo Institute of Science and Technology, research assistant of research and innovation complex “Nanobiotechnologies”, Peter the Great Saint-Petersburg Polytechnic University. Research interests: bioinformatics, biophysics, mathematical statistics. The number of publications — 4. ira.l.utkina@gmail.com; 29, Polytechnicheskaya, St.Petersburg, 195251; office phone: +7(977)821-0425.

.

УДК 006.72 **DOI** 10.15622/sp.58.1

Л.В. Уткин, И.Л. Уткина

БЫСТРОДЕЙСТВУЮЩИЙ АЛГОРИТМ ПОЛНОГЕНОМНОГО ПОИСКА АССОЦИАЦИЙ НА ОСНОВЕ АНАЛИЗА ПАР ОБЪЕКТОВ

Уткин Л.В., Уткина И.Л. **Быстродействующий алгоритм полногеномного поиска ассоциаций на основе анализа пар объектов.**

Аннотация. Предложен простой быстродействующий алгоритм полногеномного поиска ассоциаций для оценки основного и эпистатического эффекта влияния маркеров или единичных нуклеотидных полиморфизмов (SNP). Основная идея, лежащая в основе алгоритма, заключается в сравнении генотипов пар объектов популяции и сравнении соответствующих значений фенотипа. В алгоритме используется интуитивное предположение, что изменения аллелей, соответствующих важным SNP у пары объектов, приводят к большому различию значений фенотипа этих индивидуумов. Алгоритм основан на рассмотрении пар индивидуумов вместо SNP или пар SNP. Основным преимуществом алгоритма является то, что он слабо зависит от количества SNP в матрице генотипов. В основном он зависит от количества объектов, которое, как правило, очень мало по сравнению с количеством SNP. Другое важное преимущество алгоритма заключается в том, что он позволяет без дополнительных вычислений обнаруживать эпистатический эффект, рассматриваемый как взаимодействие генов. Алгоритм также может использоваться в случае, когда фенотип принимает только два значения (схема случай–контроль). Кроме того, алгоритм может быть достаточно просто расширен с анализа двоичной матрицы генотипов на случай количественного анализа экспрессии генов. Численные эксперименты с реальными наборами данных, состоящими из популяций удвоенных гаплоидных линий ячменя, иллюстрируют преимущество предлагаемого алгоритма по сравнению со стандартными алгоритмами полногеномного поиска ассоциаций с вычислительной точки зрения, особенно для обнаружения эпистатического эффекта. Пути для повышения эффективности предлагаемого алгоритма также обсуждаются в статье.

Ключевые слова: полногеномный поиск ассоциаций; ANOVA; машинное обучение; эпистаз; SNP; метрика расстояния.

Уткин Лев Владимирович — д-р техн. наук, профессор, профессор кафедры телематики (при ЦНИИ РТК) института прикладной математики и механики, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: интеллектуальный анализ данных, представление неопределенности, принятие решений при неполной информации, теория надежности, биоинформатика. Число научных публикаций — 344. lev.utkin@gmail.com, http://levvu.narod.ru; Политехническая, 29, Санкт-Петербург, 195251; р.т.: +7(812)552-6521.

Уткина Ирина Львовна — магистрант, Сколковский институт науки и технологий (Сколтех), лаборант-исследователь научно-исследовательского комплекса «Нанобиотехнологии», Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: биофизика, биоинформатика, математическая статистика. Число научных публикаций — 4. ira.l.utkina@gmail.com; Политехническая, 29, Санкт-Петербург, 195251; р.т.: +7(977)821-0425.

Литература

1. *Mieth B. et al.* Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies // Scientific reports. 2016. vol. 6. Article number: 36671. pp. 1–14.

2. *Koo C.L., Liew M.J., Mohamad M.S., Salleh A.H.M.* Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology // BioMed Research International. 2013. vol. 2013. Article ID 432375. 13 p.
3. *Yang J., Lee S.H., Goddard M.E., Visscher P.M.* Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations // Genome-wide association studies and genomic prediction. 2013. pp. 215–236.
4. *Korte A., Farlow A.* The advantagees and limitations of trait analysis with GWAS: a review // Plant Methods. 2013. vol. 9(29). pp. 1–9.
5. *Chandrashekar G., Sahin F.* A survey on feature selection methods // Computers & Electrical Engineering. 2014. vol. 40(1). pp. 16–28.
6. *Kim T.K.* Understanding one-way ANOVA using conceptual figures // Korean Journal of Anesthesiology. 2017. vol. 70(1). pp. 22–26.
7. *Dimou N.L., Tsirigos K.D., Elofsson A., Bagos P.G.* GWAR: robust analysis and meta-analysis of genome-wide association studies // Bioinformatics. 2017. vol. 33(10). pp. 1521–1527.
8. *Duan W. et. al.* A fast algorithm for Bayesian multi-locus model in genome-wide association studies // Molecular Genetics and Genomics. 2017. vol. 292(4). pp. 923–934.
9. *Franberg M. et. al.* Fast and general tests of genetic interaction for genome-wide association studies // PLOS Computational Biology. 2017. vol. 13(6). pp. e1005556.
10. *Wang S., He S., Yuan F., Zhu X.* Tagging SNP-set selection with maximum information based on linkage disequilibrium structure in genome-wide association studies // Bioinformatics. 2017. vol. 33(14), pp. 2078–2081.
11. *Lander E.S., Botstein D.* Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps // Genetics. 1989. vol. 121(1). pp. 185–199.
12. *James G., Witten D., Hastie T., Tibshirani R.* An introduction to statistical learning // Springer. 2013. 426 p.
13. *Wray N.R.* Pitfalls of predicting complex traits from SNPs // Nature Reviews Genetics. 2013. vol. 14(7). pp. 507–515.
14. *Hayes B.* Overview of statistical methods for genome-wide association studies (GWAS) // Methods in Molecular Biology. 2013. vol. 1019. pp. 149–169.
15. *Visscher P.M. et. al.* 10 Years of GWAS Discovery: Biology, Function, and Translation // The American Journal of Human Genetics. 2017. vol. 101(6). pp. 5–22.
16. *Bühlmann P.* High-dimensional statistics, with applications to genome-wide association studies // EMS Surveys in Mathematical Sciences. 2017. vol. 4(1). pp. 45–75.
17. *Uppu S., Krishna A., Gopalan R.* A review of machine learning and statistical approaches for detecting SNP interactions in high-dimensional genomic data // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2016. vol. PP(99). pp. 1545–5963.
18. *Li J., Zhong W., Li R., Wu R.* Boost: A fast algorithm for detecting gene-gene interactions in genome-wide association studies // The Annals of Applied Statistics. 2014. vol. 8(4). pp. 2292–2318.
19. *Zhang X., Zou F., Wang W.* FastANOVA: an efficient algorithm for genome-wide association study // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008. pp. 821–829.
20. *Zhang X., Zou F., Wang W.* FastChi: an effcient algorithm for analyzing gene-gene interactions // Proceedings of the Pacific Symposium on Biocomputing PMC. 2009. vol. 14. pp. 528–539.
21. *Zhang X. et. al.* COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study // Research in Computational Molecular Biology. 2009. pp. 253–269.
22. *Zhang X., Huang S., Zou F., Wang W.* TEAM: efficient two-locus epistasis tests in human genome-wide association study // Bioinformatics. 2010. vol. 26(12). pp. i217–i227.

23. *Zhang Y., Liu J.S.* Bayesian inference of epistatic interactions in case-control studies // Nature Genetics. 2007. vol. 39(9). pp. 1167–1173.
24. *Li J. et. al.* Detecting gene-gene interactions using a permutation-based random forest method // BioData Mining. 2016. vol. 9(1). pp. 14–30.
25. *Moore J.H., Andrews P.C.* Epistasis analysis using multifactor dimensionality reduction // Epistasis. 2015. pp. 301–314.
26. *Bocianowski J.* Estimation of epistasis in doubled haploid barley populations considering interactions between all possible marker pairs // Euphytica. 2014. vol. 196(1). pp. 105–115.
27. *Zhou Z., Liu G., Su L.* A new approach to detect epistasis utilizing parallel implementation of ant colony optimization by MapReduce framework // International Journal of Computer Mathematics. 2016. vol. 93(3). pp. 511–523.
28. *Ma L., Clark A.G., Keinan A.* Gene-based testing of interactions in association studies of quantitative traits // PLoS genetics. 2013. vol. 9(2). pp. e1003321-12.
29. *Niel C., Sinoquet C., Dina C., Rocheleau G.* A survey about methods dedicated to epistasis detection // Frontiers in Genetics. 2015. vol. 6. Article 285. pp. 1–19.
30. *Achlioptas P., Scholkopf B., Borgwardt K.* Two-locus association mapping in subquadratic time // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011. pp. 726–734.
31. *Chutimanitsakun Y.* Construction and application for QTL analysis of a restriction site associated DNA (rad) linkage map in barley // BMC Genomics. 2011. vol. 12. no. 4. pp. 1–13.
32. *Cistue L. et. al.* Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes // Theoretical and applied genetics. 2011. vol. 122(7). pp. 1399–1410.
33. *Hayes P.M. et. al.* Quantitative trait loci on barley (Hordeum vulgare L.) chromosome 7 associated with components of winterhardiness // Genome. 1993. vol. 36(1). pp. 66–71.
34. *Hayes P. et. al.* The Dicktoo x Morex population // Plant Cold Hardiness. 1997. pp. 77–87.
35. *Pan A. et. al.* Genetic analysis of the components of winterhardiness in barley (Hordeum vulgare L.) // Theoretical and Applied Genetics. 1994. vol. 89(7-8). pp. 900–910.
36. *Close T.J. et. al.* Development and implementation of high-throughput SNP genotyping in barley // BMC Genomics. 2009. vol. 10. no. 1. pp. 582.
37. *Breiman L.* Bagging predictors // Machine Learning. 1996. vol. 24(2). pp. 123–140.
38. *Ho T.K.* The random subspace method for constructing decision forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. vol. 20(8). pp. 832–844.