Efficient Algorithms for Genome-Wide Association Study

XIANG ZHANG, FEI ZOU, and WEI WANG
University of North Carolina at Chapel Hill

Studying the association between quantitative phenotype (such as height or weight) and single nucleotide polymorphisms (SNPs) is an important problem in biology. To understand underlying mechanisms of complex phenotypes, it is often necessary to consider joint genetic effects across multiple SNPs. ANOVA (analysis of variance) test is routinely used in association study. Important findings from studying gene-gene (SNP-pair) interactions are appearing in the literature. However, the number of SNPs can be up to millions. Evaluating joint effects of SNPs is a challenging task even for SNP-pairs. Moreover, with large number of SNPs correlated, permutation procedure is preferred over simple Bonferroni correction for properly controlling family-wise error rate and retaining mapping power, which dramatically increases the computational cost of association study.

In this article, we study the problem of finding SNP-pairs that have significant associations with a given quantitative phenotype. We propose an efficient algorithm, FastANOVA, for performing ANOVA tests on SNP-pairs in a batch mode, which also supports large permutation test. We derive an upper bound of SNP-pair ANOVA test, which can be expressed as the sum of two terms. The first term is based on single-SNP ANOVA test. The second term is based on the SNPs and independent of any phenotype permutation. Furthermore, SNP-pairs can be organized into groups, each of which shares a common upper bound. This allows for maximum reuse of intermediate computation, efficient upper bound estimation, and effective SNP-pair pruning. Consequently, FastANOVA only needs to perform the ANOVA test on a small number of candidate SNP-pairs without the risk of missing any significant ones. Extensive experiments demonstrate that FastANOVA is orders of magnitude faster than the brute-force implementation of ANOVA tests on all SNP pairs. The principles used in FastANOVA can be applied to categorical phenotypes and other statistics such as Chi-square test.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*data mining*; J.3 [**Life and Medical Sciences**]—*biology and genetics*

General Terms: Algorithm, Performance

Additional Key Words and Phrases: Association study, ANOVA test, permutation test

This research was partially supported by NSF grant IIS-0448392, NSF grant CCF-0523875, NSF grant IIS-0812464, and a Micosoft New Faculty Fellowship.
Authors' addresses: University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; email: {xiang,weiwang}@cs.unc.edu; fzou@bios.unc.edu.

DOI 10.1145/1631162.1631167 http://doi.acm.org/10.1145/1631162.1631167

ACM Reference Format:
Zhang, X., Zou, F., and Wang, W. 2009. Efficient algorithm for genome-wide association study. ACM Trans. Knowl. Discov. Data. 3, 4, Article 19 (November 2009), 28 pages.
DOI = 10.1145/1631162.1631167 http://doi.acm.org/10.1145/1631162.1631167

1. INTRODUCTION

Quantitative phenotype association study analyzes genetic variation across a population in order to find the genetic factors underlying continuous phenotypes (such as height or weight). These phenotypes are often complex in the sense that they are likely due to the effects of multiple genes [Carlson et al. 2004; Segr et al. 2005]. The most abundant source of genetic variation is represented by single nucleotide polymorphisms (SNPs). A SNP is a DNA sequence variation occurring when a single nucleotide (A, T, G, or C) in the genome differs between individuals of a species. For inbred species, a SNP usually shows variation between only two of the four possible nucleotide types [Ideraabdullah et al. 2004], which allows us to represent it by a binary variable. The binary representation of a SNP is also referred to as the *genotype* of the SNP. Table I shows an example dataset consisting of 1000 SNPs $\{X_1, X_2, \ldots, X_{1000}\}$ and a quantitative phenotype Y for 12 individuals.

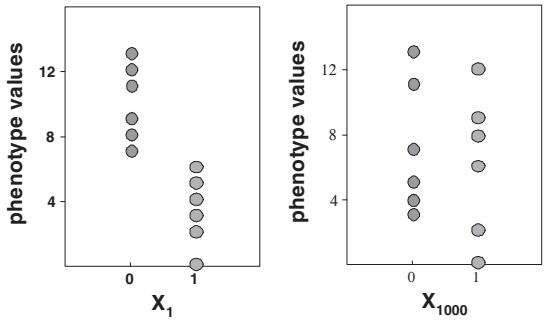
Various statistics can be applied to measure the association between SNPs and the phenotypes of interest, among which ANOVA (analysis of variance) test is one of the standard statistic methods and has been routinely used in quantitative phenotype association study [Pagano and Gauvreau 2000]. The goal of ANOVA test is to determine whether the group means are significantly different after accounting for the variances within groups. It accomplishes the comparison by decomposing the total variance in the data into within-group variance and between-group variance. If the between-group variance is sufficiently larger than the within-group variance, then the test concludes that there is significant (phenotypic) difference between the groups.

In the application of phenotype-SNP association study, the individuals' phenotype values are grouped by the genotype of a SNP or a subset of SNPs. Using the dataset showing in Table I, Figure 1(a) shows an example of strong association between the phenotype and SNP X_1. 0 and 1 on the x-axis represent the binary SNP genotype and the y-axis represents the phenotype. Each point in the figure represents an individual. It is clear from the figure that the phenotype values are partitioned into two groups with distinct means, hence indicating a strong association between the phenotype and the SNP. On the other hand, if the genotype of a SNP partitions the phenotype values into groups as shown in Figure 1(b), the phenotype and the SNP are not associated with each other.

Recent advances in high-throughput techniques enable genotyping SNPs in genome-wide scale, resulting in large datasets containing thousands to millions of SNPs, for example, the genotype datasets available in the Broad Institute (http://www.broad.mit.edu/) and the Jackson Laboratory (http://www.jax.org/). The vast number of SNPs has posed great computational challenge to genome-wide association study. In order to understand the underlying biological mechanisms of complex phenotype, one needs to consider the joint effect of multiple

Table I. An Example Dataset for Phenotype-SNP Association Study

SNPs							Phenotype
X_1	X_2	X_3	X_4	X_5	$\cdots$	X_{1000}	Y
0	0	0	1	0		1	8
0	0	0	0	0		0	7
0	1	1	0	0	$\cdots$	1	12
0	1	0	0	1		0	11
0	1	0	1	0		1	9
0	1	0	0	0	$\cdots$	0	13
1	0	1	1	1		1	6
1	0	0	0	1		0	4
1	1	1	1	1	$\cdots$	1	2
1	0	0	1	0		0	5
1	0	0	1	0		1	0
1	0	1	1	0	$\cdots$	0	3



(a) Strong association (b) No association

Fig. 1. Examples of associations between a phenotype and two different SNPs.

SNPs simultaneously. Although the idea of studying the association between phenotype and multiple SNPs is straightforward, the implementation is nontrivial. For a study with total N SNPs, in order to find the association between n SNPs and the phenotype, a brute-force approach is to exhaustively enumerate all $\binom{N}{n}$ possible SNP combinations and evaluate their associations with the phenotype. The computational burden imposed by this enormous search space often makes the complete genome-wide association study intractable.

The computational challenge of genome-wide association study is further compounded by another well-known statistical problem – the multiple testing problem [Miller 1981]. The multiple testing problem can be described as the potential increase in Type I error (false positive, the error of rejecting a null hypothesis when it is true) when statistical tests are performed multiple times. Let α be the Type I error for each independent test. If n independent comparisons are performed, the experimental-wise error α' is given by

$$\alpha' = 1 - (1 - \alpha)^n.$$

For example, when $\alpha = 0.05$ and $n = 20$, $\alpha' = 1 - 0.95^{20} = 0.64$. We have 64% probability to get at least one spurious result. Determining the statistical

significance of the association between the phenotype and SNPs is crucial. Bonferroni correction based on the assumption that all n tests are independent is too conservative for the genome-wise association studies since SNPs are often correlated. Alternatively, permutation procedure can be used and much preferred in association studies which automatically takes the correlation structure of SNPs into consideration.

The null hypothesis is that there is no association between the genotype and the phenotype. Permutation test is used to estimate the null distribution. The idea is to randomly permute the phenotype K times, where K can be hundreds to thousands. The association analysis will be repeated in order to find the maximum test value for each permutated phenotype. Then the distribution of the K maximum test values is used as the approximated null distribution to assess the statistical significance of the findings from the original phenotype. Permutation test is usually very time-consuming since the test procedure needs to be performed in all permutations in order to find the maximum values.

Algorithm development to support these large scale analysis is still in its infancy stage. Most existing work focuses on studying associations between the phenotype and SNP-pairs and can only handle a small number of SNPs. Given a pair of SNPs, the phenotype values can be partitioned into at most four groups by the genotype of the SNP-pair, that is, 00, 01, 10, and 11. Since each SNP has a distinct location on the genome, the association study of a phenotype and SNP-pairs is also called ***two-locus association mapping***. Important findings are appearing in the literature from studying the association between phenotypes and SNP-pairs [Saxena et al. 2007; Scuteri et al. 2007; Weedon et al. 2007].

Although the standard ANOVA test has been a valuable tool to find association between SNP-pairs and phenotype, it is usually not performed in genome-wide scale. This is due to the fact that the search space of two-locus association mapping in genome-wide scale prohibits an exhaustive search. Suppose that the dataset consists of N SNPs and the number of permutations is K. The total number of ANOVA tests is $KN(N-1)/2$. Given a moderate number of SNPs $N = 10,000$ and number of permutations $K = 1,000$, the number of ANOVA tests is around 5×10^{10}. Therefore, ANOVA test is often reserved for validating a small set of candidates identified by other methods [Ohno et al. 2000; Shimomura et al. 2001].

In this article, we examine the *computational aspect* of ANOVA test. We present an efficient algorithm, FastANOVA, and show that the standard ANOVA test can be applied in genome-wide scale for two-locus association mapping even when the permutation procedure is needed. Unlike algorithms applying heuristics, FastANOVA is a *complete* algorithm, that is, it guarantees to find the optimal solution, though it does not explicitly examine all possible SNP-pairs. In fact, a large portion of the SNP-pairs are pruned without the need of performing the tests. FastANOVA establishes an upper bound on the two-locus ANOVA test. The upper bound is the sum of two terms: one based on the ANOVA test between phenotype and a single SNP, and the other based on the pair-wise SNP genotype and the ordered phenotype values. This formulation of the upper bound allows the algorithm to calculate the bound for a large number of SNPs together, which enables fast candidate retrieval. Moreover,

the intermediate results for calculating the second term of the upper bound is independent of phenotype permutations. Hence, they only need to be computed once and can be reused in all permutations. Applying this bound, FastANOVA is able to identify SNP-pairs with significant ANOVA test values using only a small fraction of the time required by performing ANOVA test on all SNP-pairs.

In Section 7, we discuss further extensions of the FastANOVA algorithm to case-control study whose phenotypes can be represented as binary variables. We first show that the principle of FastANOVA can be applied to Chi-square test [Zhang et al. 2009b]. Then we briefly describe a more general approach that can be applied to a variety of statistics used in case-control study [Zhang et al. 2009a].

2. RELATED WORK

The problem of phenotype-SNP association study has attracted extensive research interests and is an ongoing research area in biology and statistic communities. In this section, we review the related work from a computational point of view. Please refer to Doerge [2002], Hoh and Ott [2003], and Balding [2006] for excellent surveys of existing work.

Different machine learning models have been adopted in multilocus association study. In Curtis et al. [2001] and Sherriff and Ott [2001], the authors investigate using neural networks to study the relationship between complex traits and multilocus genotypes. These models are theoretically well suited for analyzing high-order interactions. However, the results of these methods are usually expressed as weights associated with SNPs. They are difficult to interpret and do not clearly identify the interacting SNPs. Recursive partitioning methods [Zhang and Bonney 2000; Province et al. 2001] utilize classification and regression tree (CART) [Breiman et al. 1984] to pick the SNP that minimizes some pre-specified measure of impurity in each iteration. These methods are not effective in detecting SNP combinations if there is little or no marginal effect.

Under the assumption that the number of SNPs is limited, for example, from tens to hundreds, exhaustive algorithms that explicitly enumerate all possible SNP combinations have been developed. Combinatorial partitioning method (CPM) [Nelson et al. 2001] is designed to identify multilocus genotypic partitions that predict quantitative trait variation. Given a small set of SNPs, CPM searches for the partitions of multilocus genotypes that are the most predictive in terms of phenotypic variability. Motivated by CPM, multifactorial dimension reduction (MDR) [Ritchie et al. 2001; Moore et al. 2006] is designed for case/control studies. By pooling genotypes of multilocus into two groups at high disease risk and low disease risk, MDR reduces the genotype of multiple SNPs into one dimension. Among all possible combinations, MDR selects the one that maximizes the case/control ratio of the high risk group. Since these methods explicitly enumerate all possible SNP combinations, they are not well adapted to genome-wide association studies.

To avoid exhaustively enumerating the search space, a common approach is to break the problem into two steps [Hoh et al. 2000; Evans et al. 2006].

First, a subset of important SNPs are selected. Second, within the selected subset, the association between SNPs and the phenotypes are searched. These methods are not complete since the SNPs with weak marginal effects may not be selected in the first step. Genetic algorithm [Carlborg et al. 2000; Nakamichi et al. 2001] has been applied in finding SNP-pairs for quantitative phenotypes. These methods cannot guarantee to find the optimal solution.

Feature selection methods [Liu and Motoda 1998] have been proposed to address the problem of finding important SNPs. In feature selection, the selected feature subset usually contains features that have low correlation with each other but have strong correlation with the target feature. In the application of selecting SNPs, the goal is to select a subset of SNPs that can be used as proxies for all SNPs in the genome [Sebastiani et al. 2003; Chi et al. 2006; Halperin et al. 2005]. The selected SNPs can then be used as the input SNPs in the association study. These methods are also not complete since some important SNPs may not be tagged.

3. TWO-LOCUS ANOVA TEST

In this section, we formalize the problem of two-locus ANOVA test with permutation procedure. Let $\{X_1, X_2, \ldots, X_N\}$ be the set of SNPs of M individuals. Each SNP X_i $(1 \leq i \leq N)$ is a binary variable coded by $\{0, 1\}$. Let $Y = \{y_1, y_2, \ldots, y_M\}$ be the quantitative phenotype of interest, where y_m $(1 \leq m \leq M)$ is the phenotype value of individual m. For any SNP X_i $(1 \leq i \leq N)$, we represent the F-statistic from the ANOVA test of X_i and Y as $F(X_i, Y)$. For any SNP-pair $(X_i X_j)$, we represent the F-statistic from the ANOVA test of $(X_i X_j)$ and Y as $F(X_i X_j, Y)$.

The basic idea of ANOVA test is to partition the total sum of squared deviations SS_T into between-group sum of squared deviations SS_B and within-group sum of squared deviations SS_W

$$SS_T = SS_B + SS_W.$$

Suppose that phenotype values are partitioned into k groups, with m_i individuals in group i $(1 \leq i \leq k)$. Let y_{ij} be the jth observation in group i. Let $\bar{y}$ be the mean of all the observed phenotype values, and $\bar{y}_i$ be the mean of the observed phenotype values in group i. The terms used in an ANOVA test are defined as follows.

$$SS_B = \sum_{i=1}^{k} m_i(\bar{y}_i - \bar{y})^2;$$

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2; SS_T = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y})^2.$$

In the application of two-locus association study, Table II(a) and Table II(b) show the possible groupings of phenotype values by the genotypes of X_i and $(X_i X_j)$ respectively. Let A, B, a_1, a_2, b_1, b_2 represent the groups as indicated in Table II(a) and Table II(b). We use $SS_B(X_i, Y)$ and $SS_B(X_i X_j, Y)$ to distinguish the one locus (i.e., single-SNP) and two locus (i.e., SNP-pair) analysis.

Table II. Possible Groupings of Phenotype Values by the Genotypes of X_i and (X_iX_j)

(a) Grouping of Y by X_i

$X_i = 1$	$X_i = 0$
group A	group B

(b) Grouping of Y by X_iX_j

	$X_i = 1$	$X_i = 0$
$X_j = 1$	group a_1	group b_1
$X_j = 0$	group a_2	group b_2

Specifically, we have

$$SS_T(X_i, Y) = SS_B(X_i, Y) + SS_W(X_i, Y),$$

$$SS_T(X_iX_j, Y) = SS_B(X_iX_j, Y) + SS_W(X_iX_j, Y).$$

The F-statistics for ANOVA tests on X_i and (X_iX_j) are:

$$F(X_i, Y) = \frac{M-2}{2-1} \times \frac{SS_B(X_i, Y)}{SS_T(X_i, Y) - SS_B(X_i, Y)}, \tag{1}$$

$$F(X_iX_j, Y) = \frac{M-g}{g-1} \times \frac{SS_B(X_iX_j, Y)}{SS_T(X_iX_j, Y) - SS_B(X_iX_j, Y)}, \tag{2}$$

where g in Eq. (2) is the number of groups that the genotype of (X_iX_j) partitions the individuals into. Possible values of g are 3 or 4, assuming all SNPs are distinct: If none of groups A, B, a_1, a_2, b_1, b_2 is empty, then $g = 4$. If one of them is empty, then $g = 3$.

Let $T = \sum_{y_m \in Y} y_m$ be the sum of all phenotype values. The total sum of squared deviations does not depend on the groupings of individuals:

$$SS_T(X_i, Y) = SS_T(X_iX_j, Y) = \sum_{y_m \in Y} y_m^2 - \frac{T^2}{M}.$$

Let $T_{group} = \sum_{y_m \in group} y_m$ be the sum of phenotype values in a specific group, and n_{group} be the number of individuals in that group. $SS_B(X_i, Y)$ and $SS_B(X_iX_j, Y)$ can be calculated as follows:

$$SS_B(X_i, Y) = \frac{T_A^2}{n_A} + \frac{T_B^2}{n_B} - \frac{T^2}{M},$$

$$SS_B(X_iX_j, Y) = \frac{T_{a_1}^2}{n_{a_1}} + \frac{T_{a_2}^2}{n_{a_2}} + \frac{T_{b_1}^2}{n_{b_1}} + \frac{T_{b_2}^2}{n_{b_2}} - \frac{T^2}{M}.$$

Note that for any group of A, B, a_1, a_2, b_1, b_2, if $n_{group} = 0$, then T_{group}^2/n_{group} is defined to be 0.

The two-locus association mapping with permutation test is typically conducted in the following way [Westfall and Young 1993; Dudoit and van der Laan 2008]:

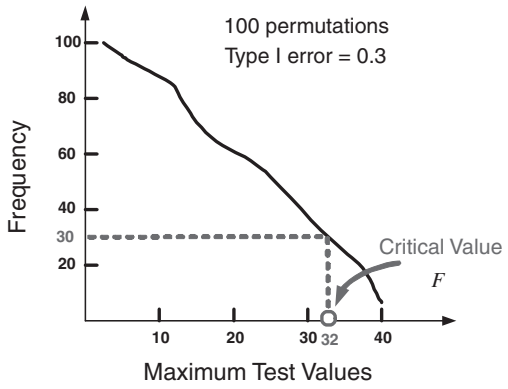


Fig. 2. An example of determining the critical value using permutation test.

First, for every SNP-pair (X_iX_j) $(1 \leq i < j \leq N)$, the ANOVA test is performed and $F(X_iX_j, Y)$ is recorded.

Second, a permutation test is performed to get a reference distribution in order to assess the statistical significance of previous findings. More specifically, a permutation Y_k of Y is generated by sampling the phenotype Y without replacement. In other words, phenotype values are randomly assigned to individuals in the dataset with no single phenotype value being assigned to more than one individual. Let $Y' = \{Y_1, Y_2, \ldots, Y_K\}$ be the set of K permutations of Y. For each permutation $Y_k \in Y'$, let F_{Y_k} represent the maximum F-statistic value of all SNP-pairs, that is,

$$F_{Y_k} = \max\{F(X_iX_j, Y_k) | 1 \leq i < j \leq N\}.$$

The distribution of $\{F_{Y_k} | Y_k \in Y'\}$ is then used as the reference distribution for assessing the statistical significance of $F(X_iX_j, Y)$ values found using the original phenotype Y: Given a Type I error threshold α, the *critical value* F_α is the αK-th largest value in $\{F_{Y_k} | Y_k \in Y'\}$. The SNP-pair (X_iX_j) whose F-statistic value $F(X_iX_j, Y) \geq F_\alpha$ is considered as significant at α.

For example, Figure 2 shows the cumulative distribution of the maximum values for $K = 100$ permutations. Suppose that $\alpha = 0.3$, then F_α is the 30th largest value among the 100 maximum test values, which is 32 as shown in this example.

Two computational problems need to be solved in this procedure. The first one is to find the critical value F_α for a given Type I error threshold α. The second one is to find all SNP-pairs (X_iX_j) whose F-statistics are greater than F_α. We formalize these two problems as follows:

Problem (1). Given the Type I error threshold α, find the critical value F_α, which is the αK-th largest value in $\{F_{Y_k} | Y_k \in Y'\}$.

Problem (2). Given the threshold F_α, find all significant SNP-pairs (X_iX_j) such that $F(X_iX_j, Y) \geq F_\alpha$.

A brute force approach to these two problems is to enumerate all SNP-pairs and find their F-statistics. In Problem (1), for each permutation $Y_k \in Y$, all

SNP-pairs need to be enumerated in order to find the maximum value F_{Y_k}. In Problem (2), all SNP-pairs need to be enumerated to see if their test values are above the threshold F_α. Computationally, Problem (1) is more challenging, since the permutation number K can range form hundreds to thousands, which means the running time of finding the critical value F_α can be hundreds to thousands times longer than the running time of finding the significant SNP-pairs in Problem (2) using a brute-force search.

In the reminder of this article, we first derive an upper bound on two-locus ANOVA test value and discuss how this upper bound enables an efficient ANOVA testing for a single phenotype. Then, we show how this approach can be easily extended to handle the permutation procedure.

4. THE UPPER BOUND

4.1 Updating F-Statistic

Since the total sum of squared deviations does not change, from the calculation of $F(X_i, Y)$ and $F(X_iX_j, Y)$ (Eqs. (1) and (2)), we know that the relationship between these two tests only depends on the relationship between $SS_B(X_i, Y)$ and $SS_B(X_iX_j, Y)$. Next, we show that $SS_B(X_iX_j, Y)$ can be updated from $SS_B(X_i, Y)$.

For groups A, a_1 and a_2, let

$$\begin{aligned}\Delta A &= \frac{T_{a_1}^2}{n_{a_1}} + \frac{T_{a_2}^2}{n_{a_2}} - \frac{T_A^2}{n_A} \\ &= \frac{n_{a_2}T_{a_1}^2 + n_{a_1}T_{a_2}^2}{n_{a_1}n_{a_2}} - \frac{(T_{a_1}+T_{a_2})^2}{n_{a_1}+n_{a_2}} \\ &= \frac{(n_{a_2}T_{a_1} - n_{a_1}T_{a_2})^2}{n_{a_1}n_{a_2}n_A} \\ &= \frac{(n_AT_{a_1} - n_{a_1}T_A)^2}{n_{a_1}(n_A - n_{a_1})n_A}.\end{aligned}$$

Similarly, we have

$$\Delta B = \frac{T_{b_1}^2}{n_{b_1}} + \frac{T_{b_2}^2}{n_{b_2}} - \frac{T_B^2}{n_B} = \frac{(n_BT_{b_1} - n_{b_1}T_B)^2}{n_{b_1}(n_B - n_{b_1})n_B}.$$

Thus, $SS_B(X_iX_j, Y)$ can be updated using $SS_B(X_i, Y)$:

$$SS_B(X_iX_j, Y) = SS_B(X_i, Y) + \Delta A + \Delta B. \tag{3}$$

Note that if any one of $\{n_{a_1}, n_{a_2}, n_A\}$ is 0, then $\Delta A = 0$. Similarly, if any one of $\{n_{b_1}, n_{b_2}, n_B\}$ is 0, then $\Delta B = 0$.

Next, we develop an upper bound of $SS_B(X_iX_j, Y)$. We first show the derivation of an upper bound of Δ_A. A similar idea can be applied to find an upper bound of Δ_B.

4.2 Bounds of ΔA and ΔB

Let $\{y_m | y_m \in A\} = \{y_{A_1}, y_{A_2}, \ldots, y_{A_{n_A}}\}$ be the phenotype values in group A. Without loss of generality, assume that these phenotype values are arranged in ascending order, that is,

$$y_{A_1} \leq y_{A_2} \leq \cdots \leq y_{A_{n_A}}.$$

The derivative of ΔA with respect to T_{a_1} is:

$$\frac{d\,\Delta A}{d\,T_{a_1}} = \frac{2n_A(n_A T_{a_1} - n_{a_1} T_A)}{n_{a_1}(n_A - n_{a_1})n_A}.$$

Thus we have

$$\Delta A \text{ monotonically} \begin{cases} \text{increases} & \text{if } T_{a_1} \geq \frac{n_{a_1} T_A}{n_A}; \\ \text{decreases} & \text{if } T_{a_1} \leq \frac{n_{a_1} T_A}{n_A}. \end{cases}$$

We have the range of T_{a_1}:

$$T_{a_1} \in [l_{a_1}, u_{a_1}] = \left[\sum_{i=1}^{n_{a_1}} y_{A_i}, \sum_{i=n_A - n_{a_1}+1}^{n_A} y_{A_i} \right].$$

The maximum value of ΔA is attained when $T_{a_1} = l_{a_1}$ or $T_{a_1} = u_{a_1}$, i.e.,

$$\Delta A \leq \frac{\max\{(n_A l_{a_1} - n_{a_1} T_A)^2, (n_A u_{a_1} - n_{a_1} T_A)^2\}}{n_{a_1}(n_A - n_{a_1})n_A}. \tag{4}$$

We use $R_1(X_i X_j Y)$ to denote this upper bound.

Let $\{y_m | y_m \in B\} = \{y_{B_1}, y_{B_2}, \ldots, y_{B_{n_B}}\}$ be the phenotype values in group B. Without loss of generality, assume that these phenotype values are arranged in ascending order, that is,

$$y_{B_1} \leq y_{B_2} \leq \cdots \leq y_{B_{n_B}}.$$

Similarly, we can derive the bound on ΔB:

$$\Delta B \leq \frac{\max\{(n_B l_{b_1} - n_{b_1} T_B)^2, (n_B u_{b_1} - n_{b_1} T_B)^2\}}{n_{b_1}(n_B - n_{b_1})n_B}. \tag{5}$$

We use $R_2(X_i X_j Y)$ to denote this upper bound. The symbols used in Inequalities (4) and (5) are summarized in Table III.

From Eq. (3), Inequalities (4) and (5), we have the overall upper bound on $SS_B(X_i X_j, Y)$:

THEOREM 4.1 (UPPER BOUND OF $SS_B(X_i X_j, Y)$).

$$SS_B(X_i X_j, Y) \leq SS_B(X_i, Y) + R_1(X_i X_j Y) + R_2(X_i X_j Y).$$

PROPERTY 4.2. *The upper bound in Theorem* 4.1 *is tight.*

The tightness of the bound is obvious from the derivation of the upper bound, since there exists some genotype of SNP-pair $(X_i X_j)$ that makes the equality hold. For the same reason, we have the following property.

Table III. Notations for the Bounds on ΔA and ΔB

Symbols	Formulas
l_{a_1}	$\sum_{i=1}^{n_{a_1}} y_{A_i}$
u_{a_1}	$\sum_{i=n_A-n_{a_1}+1}^{n_A} y_{A_i}$
$R_1(X_iX_jY)$	$\frac{\max\{(n_A l_{a_1}-n_{a_1}T_A)^2,\ (n_A u_{a_1}-n_{a_1}T_A)^2\}}{n_{a_1}(n_A-n_{a_1})n_A}$
l_{b_1}	$\sum_{i=1}^{n_{b_1}} y_{B_i}$
u_{b_1}	$\sum_{i=n_B-n_{b_1}+1}^{n_B} y_{B_i}$
$R_2(X_iX_jY)$	$\frac{\max\{(n_B l_{b_1}-n_{b_1}T_B)^2,\ (n_B u_{b_1}-n_{b_1}T_B)^2\}}{n_{b_1}(n_B-n_{b_1})n_B}$

PROPERTY 4.3. *The upper bound in Theorem* 4.1 *does not exceed the total sum of squared deviations, that is,*

$$SS_B(X_i, Y) + R_1(X_iX_jY) + R_2(X_iX_jY) \leq SS_T(X_iX_j, Y).$$

5. THE FASTANOVA ALGORITHM

In this section, we show how our algorithm FastANOVA utilizes the upper bound in Theorem 4.1 to achieve efficient two-locus ANOVA testing. In Section 5.1, we describe the method for Problem (2) discussed in Section 3; that is, given a threshold F_α, we want to find all SNP-pairs whose F-statistics are greater than F_α. Then, in Section 5.2, we discuss how FastANOVA performs in permutation procedure, that is, the scenario of Problem (1) in Section 3.

5.1 A Single Phenotype

Given the threshold F_α, to find all SNP-pairs whose F-statistics are greater than F_α, a brute-force approach is to enumerate all SNP-pairs. To expedite this process, we employ the inequality in Theorem 4.1 to prune SNP pairs that will have no chance to pass the significance threshold F_α. From Eq. (2), we know that finding SNP-pairs (X_iX_j) whose F-statistics $F(X_iX_j, Y) \geq F_\alpha$ is equivalent to finding SNP-pairs satisfying

$$SS_B(X_iX_j, Y) \geq \frac{SS_T(X_i, Y)}{\frac{M-g}{(g-1)F_\alpha} + 1} = \theta.$$

Theorem 4.1 suggests that we only need to compute the F-statistics for the SNP-pairs that satisfy:

$$SS_B(X_i, Y) + R_1(X_iX_jY) + R_2(X_iX_jY) \geq \theta.$$

We refer to these SNP-pairs as *candidate* SNP-pairs.

We now discuss how to apply the upper bound in Theorem 4.1 in detail. The set of all SNP-pairs is partitioned into nonoverlapping groups such that each group has a common upper bound. For every X_i $(1 \leq i \leq N)$, let $AP(X_i)$ be the set of SNP-pairs

$$AP(X_i) = \{(X_iX_j) | i+1 \leq j \leq N\}.$$

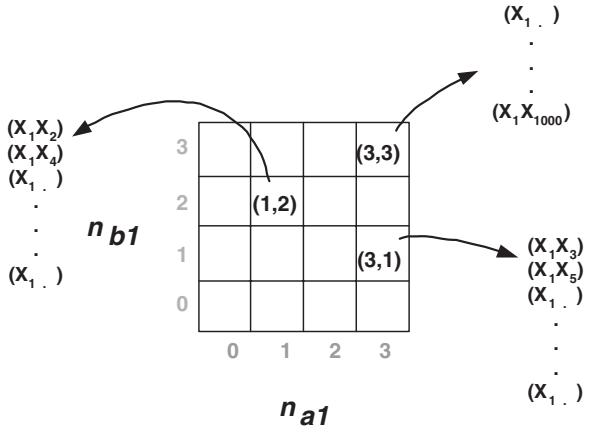


Fig. 3. The index array $Array(X_1)$ for efficient retrieval of the candidate SNP-pairs.

For all SNP-pairs in $AP(X_i)$, n_A, T_A, n_B, T_B and $SS_B(X_i, Y)$ are constants. Moreover, l_{a_1}, u_{a_1} are determined by n_{a_1}, and l_{b_1}, u_{b_1} are determined by n_{b_1}. Therefore, in the upper bound, n_{a_1} and n_{b_1} are the only variables that depend on X_j and may vary for different SNP-pairs (X_iX_j) in $AP(X_i)$.

Note that n_{a_1} is the number of 1's in X_j when X_i takes value 1, and n_{b_1} is the number of 1's in X_j when X_i takes value 0. In Section 4.2, we have shown the upper bound of ΔA (ΔB) using the phenotype values in group a_1 (b_1). We can also develop a similar bound based on group a_2 (b_1). Therefore, without loss of generality, we always assume that n_{a_1} is the smaller one between the number of 1's and number of 0's in X_j when X_i takes value 1, and n_{b_1} is the smaller one between the number of 1's and number of 0's in X_j when X_i takes value 0.

For example, using the dataset showing in Table I, for SNP-pair (X_iX_2), $n_a = 1$ since the minimum of number of 1's and 0's in X_2 when $X_1 = 1$ is 1 (the number of 1's), and $n_b = 2$ since the minimum of number of 1's and 0's in X_2 when $X_1 = 0$ is 2 (the number of 0's).

The following property specifies the values that n_{a_1} and n_{b_1} can take. The proof is straightforward and omitted here.

PROPERTY 5.1. *If there are m 1's and $(M-m)$ 0's in X_i, then for any $(X_iX_j) \in AP(X_i)$, the possible values that n_{a_1} can take are $\{0, 1, 2, \ldots, \lfloor m/2 \rfloor\}$. The possible values that n_{b_1} can take are $\{0, 1, 2, \ldots, \lfloor (M-m)/2 \rfloor\}$.*

To efficiently retrieve the candidates, the SNP-pairs (X_iX_j) in $AP(X_i)$ are grouped by their (n_{a_1}, n_{b_1}) values and indexed in a 2D array, referred to as $Array(X_i)$.

Example 5.2. Using the example dataset shown in Table I, we consider the SNP-pairs in $AP(X_1)$, that is, $\{(X_1X_2), (X_1X_3), (X_1X_4), (X_1X_5), \ldots, (X_1X_{1000})\}$. There are 12 individuals in the dataset, and the genotype of X_1 contains 6 0's and 6 1's. Therefore, the possible values of n_{a_1} and n_{b_1} are $\{0, 1, 2, 3\}$. Figure 3 shows the 4×4 array, $Array(X_1)$, whose entries represent the possible

values of (n_{a_1}, n_{b_1}) for the SNP-pairs in $AP(X_i)$. The entries in the same column have the same n_{a_1} value. The entries in the same row have the same n_{b_1} value. The n_{a_1} value of each column is noted beneath each column. The n_{b_1} value of each row is noted left to each row. Each entry of the array is a pointer to the SNP-pairs having the corresponding (n_{a_1}, n_{b_1}) values. For example, for SNP-pair (X_1X_3), its $(n_{a_1}, n_{b_1}) = (3, 1)$. Thus, it is indexed by entry (3,1).

Note that for a SNP-pair $(X_iX_j) \in AP(X_i)$, n_{a_1} and n_{a_2} can be calculated faster than performing the two-locus ANOVA test. To obtain n_{a_1} and n_{a_2}, we only need to count the numbers of 0's and 1's of X_j when X_i is equal to 0 and 1 respectively, which can be done by a linear scan of the $M \times 2$ binary matrix consisting of the genotypes of X_i and X_j. In contrast, to calculate the F-statistic, we first need to scan the $M \times 3$ binary matrix consisting of X_i, X_j and Y in order to find out how the phenotype values are grouped by the genotype of (X_iX_j). Then a constant time $O(t)$ is required to compute the F-statistic.

PROPERTY 5.3. *For any SNP X_i, the maximum number of the entries in $Array(X_i)$ is*

$$\left(\left\lceil \frac{M}{4} \right\rceil + 1\right)^2.$$

The proof of Property 5.3 is straightforward and omitted here. In order to find candidate SNP-pairs, we scan all entries in $Array(X_i)$ to calculate their upper bounds. Since the SNP-pairs indexed by the same entry share the same (n_{a_1}, n_{b_1}) value, they have the same upper bound.

PROPERTY 5.4. *Given phenotype Y, for any SNP X_i, the SNP-pairs indexed by the same entry in $AP(X_i)$ have the same upper bound value.*

For typical genome-wide association studies, the number of individuals M is much smaller than the number of SNPs N. From Property 5.3, there must be a group of SNP-pairs indexed by the same entry of $AP(X_i)$. In Example 5.2, there are in total 16 entries in $Array(X_1)$, and 999 SNP-pairs in $AP(X_1)$. Thus many SNP-pairs share the same (n_{a_1}, n_{b_1}) value and hence indexed by the same entry in $Array(X_1)$. Moreover, from Property 5.4, we can calculate the upper bound for the group of SNP-pairs indexed by the same entry together. It is these two key properties of the index structure that help to reduce the complexity of the algorithm. The additional cost for accessing $Array(X_i)$ is minimal compared to performing ANOVA tests for all pairs $(X_iX_j) \in AP(X_i)$ since $M \ll N$.

Algorithm 1 describes the FastANOVA algorithm for finding the SNP-pairs whose F-statistics are greater than the threshold F_α. The inputs of FastANOVA include the N SNPs, the phenotype Y and the critical value F_α. For each X_i, FastANOVA first indexes $(X_iX_j) \in AP(X_i)$ using $Array(X_i)$. Then it retrieves the candidate SNP-pairs by accessing $Array(X_i)$ and records them in $Cand(X_i, Y)$. The candidates in $Cand(X_i, Y)$ are then evaluated for their F-statistics. The candidates whose F-statistics are greater than or equal to F_α are reported by the algorithm.

Algorigthm 1: FastANOVA (no phenotype permutation)

Input: SNPs $X' = \{X_1, X_2, \ldots, X_N\}$, phenotype Y, and threshold F_α
Output: find the set of SNP-pairs
$Result(Y) = \{(X_iX_j)|F(X_iX_j, Y) \geq F_\alpha, 1 \leq i < j \leq N\}$

```
1 for every X_i ∈ X', do
2 |   index (X_iX_j) ∈ AP(X_i) by Array(X_i);
3 |   access Array(X_i) to find the candidate SNP-pairs and store them in Cand(X_i, Y);
4 |   for every (X_iX_j) ∈ Cand(X_i, Y) do
5 |   |   if F(X_iX_j, Y) ≥ F_α then
6 |   |   |   Result(Y) ← (X_iX_j);
7 |   |   end
8 |   end
9 end
10 return Result(Y).
```

5.2 Permutation Procedure

For multiple tests, permutation procedure is often used in genetic analysis for controlling family-wise error rate. For genome-wide association study, permutation is less commonly used because it often entails prohibitively long computation time. Our FastANOVA algorithm makes permutation procedure feasible in genome-wide association study.

Let $Y' = \{Y_1, Y_2, \ldots, Y_K\}$ be K permutations of the phenotype Y. Following the idea discussed in Section 5.1, the upper bound in Theorem 4.1 can be easily incorporated in the algorithm to handle the permutations.

PROPERTY 5.5. *For every SNP X_i, the indexing structure $Array(X_i)$ is independent of the permuted phenotypes in Y'.*

The correctness of this property relies on the fact that, for any $(X_iX_j) \in AP(X_i)$, n_{a_1} and n_{b_1} only depend on the genotype of the SNP-pair and thus remain constant for different phenotype permutations. Therefore, for each X_i, once we build $Array(X_i)$, it can be reused in all permutations.

The FastANOVA algorithm for permutation test is described in Algorithm 2. The inputs include the N SNPs, K phenotype permutations, and the Type I error threshold α. The goal is to find the critical value F_α, which is the αK-th largest value in $\{F_{Y_k}|Y_k \in Y'\}$. Recall that F_{Y_k} is the maximum F-statistic value for phenotype Y_k. We use *Tlist* to keep the αK phenotype permutations having the largest F-statistics found by the algorithm so far. Initially, *Tlist* contains αK dummy phenotype permutations with test values 0. The smallest F-statistic value in *Tlist*, initially 0, is used as the threshold to prune the SNP-pairs. For each X_i, FastANOVA first indexes $(X_iX_j) \in AP(X_i)$ using $Array(X_i)$. Then it finds the set of candidate SNP-pairs $Cand(X_i, Y_k)$ by accessing $Array(X_i)$ for every phenotype permutation Y_k. The candidates in $Cand(X_i, Y_k)$ are then evaluated for their F-statistics. If a candidate's F-statistic value is greater than the current threshold, then *Tlist* is updated accordingly: If the candidate's phenotype Y_k is not in the *Tlist*, then the phenotype in *Tlist* having the smallest

Algorigthm 2: FastANOVA (for permutation test)

Input: SNPs $X' = \{X_1, X_2, \ldots, X_N\}$, phenotype permutations $Y' = \{Y_1, Y_2, \ldots, Y_K\}$, and the Type I error α
Output: find the critical value F_α

```
1  Tlist ← αK dummy phenotype permutations with F-statistics 0;
2  F_α = 0;
3  for every X_i ∈ X', do
4  |   index (X_i X_j) ∈ AP(X_i) by Array(X_i);
5  |   for every Y_k ∈ Y', do
6  |   |   access Array(X_i) to find the candidate SNP-pairs and store them in
   |   |   Cand(X_i, Y_k);
7  |   |   for every (X_i X_j) ∈ Cand(X_i, Y_k) do
8  |   |   |   if F(X_i X_j, Y_k) ≥ F_α then
9  |   |   |   |   update Tlist;
10 |   |   |   |   F_α = the smallest test value in Tlist;
11 |   |   |   end
12 |   |   end
13 |   end
14 end
15 return F_α.
```

F-statistic value is replaced by Y_k. If the candidate's phenotype Y_k is already in *Tlist*, we only need to update its corresponding F-statistic value to be the maximum value found for the phenotype so far. The threshold is also updated to be the smallest F-statistic value in *Tlist*.

5.3 Complexity Analysis

In this section, we study the time and space complexities of the FastANOVA algorithm for permutation test. The complexity for a single phenotype can be analyzed in a similar way.

Time Complexity. For each X_i, FastANOVA needs to index $(X_i X_j)$ in $AP(X_i)$. The complexity to build the indexing structure for all SNPs is $O(N(N-1)M/2)$. The worst case for accessing all $Array(X_i)$ for all permutations is $O(N \times K \times (\lceil \frac{M}{4} \rceil + 1)^2) = O(NKM^2)$. Let $C = \sum_{i,k} |Cand(X_i, Y_k)|$ represent the total number of candidates. The overall time complexity of FastANOVA is thus $O(N(N-1)M/2) + O(NK \times (\lceil \frac{M}{4} \rceil + 1)^2) + O(\sum_{i,k} |Cand(X_i, Y_k)|M) = O(N^2M + NKM^2 + CM)$. The experimental results show that the overhead of building the indexing structures and accessing them for candidate retrieval are negligible when large permutation tests are needed. The time complexity of the brute-force approach is $O(KN(N-1)M/2) = O(KN^2M)$. Note that in a typical genotype-phenotype association study, the number of SNPs N is much lager than the number of individuals M. Therefore, when the number of permutations K is large, e.g. thousands, the complexity of FastANOVA is much less than the complexity of the brute force approach.

Space Complexity. The total number of variables in the dataset, including the SNPs and the phenotype permutations, is $N + K$. The maximum space of

Table IV. Statistics of the Genotype Datasets

	Cardiovascular	Metabolism	Neurosensory
# individuals	19	26	34
# SNPs	14,513	43,856	66,006

the indexing structure $Array(X_i)$ is $O((\lceil \frac{M}{4} \rceil + 1)^2 + N)$. Note that for each SNP X_i, FastANOVA only needs to access one indexing structure, $Array(X_i)$, for all permutations. Once the evaluation process for X_i is done for all permutations, $Array(X_i)$ can be cleared from the memory. Therefore, the space complexity of FastANOVA is $O((N + K)M) + O((\lceil \frac{M}{4} \rceil + 1)^2 + N) = O((N + K)M)$ since $M \ll N$. The space complexity is linear to the dataset size.

6. EXPERIMENTAL RESULTS

In this section, we present extensive experimental results on evaluating the performance of the FastANOVA algorithm. We show (1) the runtime comparison between FastANOVA and the brute-force approach under various experimental settings, (2) the punning effect of the upper bound, and (3) the relative computational cost of each component of FastANOVA. FastANOVA is implemented in C++. The experiments are performed on a 2.4 GHz PC with 1G memory running WindowsXP system.

Dataset. The SNP dataset used for the experiments is extracted from a set of combined SNPs from the 140k Broad/MIT mouse dataset (*http://www.broad.mit.edu/*) and 10k GNF mouse dataset (*http://www.gnf.org/*). This merged dataset has 156,525 SNPs for 71 individuals. The missing values in the dataset are imputed using NPUTE [Roberts et al. 2007]. We use both real phenotypes and synthetic phenotypes in our experiments. The real phenotype data is available from the Jackson Laboratory (*http://www.jax.org/*).

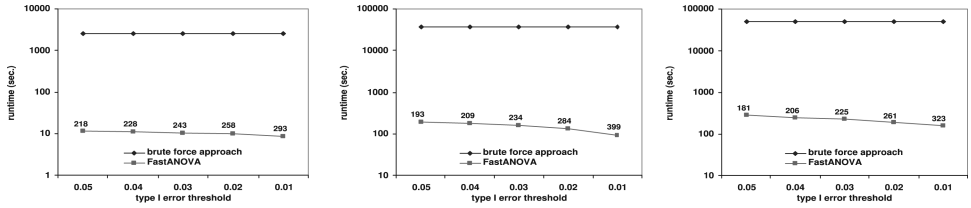
6.1 Real Phenotypes

We use three real phenotypes in our experiments: cardiovascular (blood pressure), metabolism (water intake), and neurosensory (acoustic startle response). Table IV shows the statistics of the genotype datasets corresponding to the three phenotypes. The number of SNPs in the table indicates the number of unique SNPs in each genotype dataset.

We first show the results on finding the critical value F_α, which is more time-consuming than finding the significance SNP-pairs given the critical value F_α for a single phenotype.

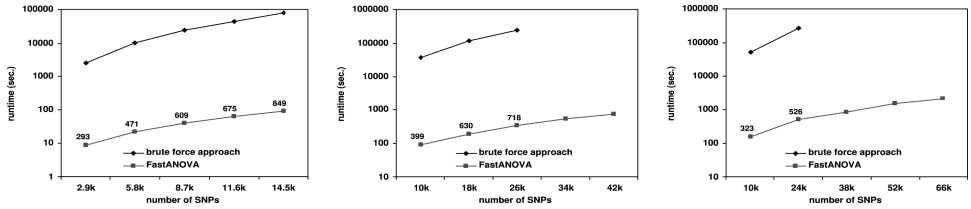
6.1.1 *Finding Critical Value F_α*

6.1.1.1 *FastANOVA vs. the Brute-Force Approach.* We compare FastANOVA with the brute-force approach under various experimental settings. Since the brute-force approach is very time-consuming, we use a moderate number of SNPs and permutations in the default setting in order to show the performance comparisons. The default setting is as follows: The Type I error threshold $\alpha = 0.01$. The number of permutations is 100. The number of SNP is 10,000 for the two larger datasets of metabolism and neurosensory, and 2,900 for the



(a) cardiovascular (b) metabolism (c) neurosensory

Fig. 4. Performance comparison between FastANOVA and the brute-force approach when varying Type I error thresholds.



(a) cardiovascular (b) metabolism (c) neurosensory

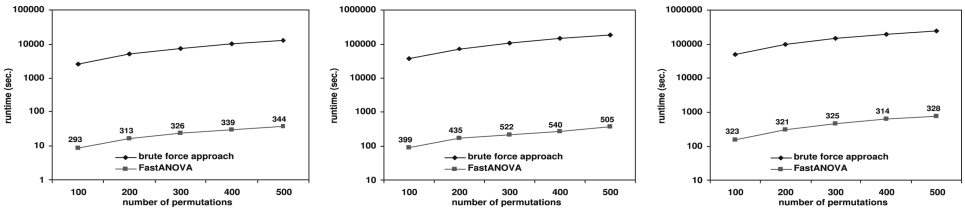
Fig. 5. Performance comparison between FastANOVA and the brute-force approach when varying the number of SNPs.

cardiovascular SNP dataset. These experimental settings are chosen to demonstrate the performance gain and enhanced scalability offered by FastANOVA over the brute-force implementation. FastANOVA can handle much larger SNP panels and larger number of permutation tests. The performance of FastANOVA is expected to follow the same trends presented in the remainder of this section.

Figures 4, 5, and 6 show the running time comparison of FastANOVA and the brute-force approach on the three genotype phenotype datasets using different settings. The y-axis is in logarithm scale. The numbers above the runtime line of FastANOVA indicate the ratio of the runtimes of the brute-force approach over FastANOVA. We terminate the programs that have run over 72 hours without completion.

Figure 4 shows the runtime comparison when varying the Type I error thresholds. For each dataset, the runtime of the brute-force approach does not change over different Type I error thresholds. The runtime of FastANOVA decreases as the threshold decreases. FastANOVA offers 218-fold speedup when $\alpha = 0.05$ and 293 fold speedup when $\alpha = 0.01$ on cardiovascular dataset. We can also observe a similar two-orders-of-magnitude speedup in the metabolism and neurosensory datasets. This is consistent with the pruning effect of the upper bound, which will be presented later in this section. In general, the lower the Type I error threshold, the more powerful the pruning effect, hence the faster the algorithm.

Figure 5 depicts the comparison of these two approaches when the number of SNPs changes. From these figures, it is clear that FastANOVA is about two orders of magnitude faster than the brute-force approach. The brute-force approach cannot finish in 72 hours when the number of unique SNPs is greater than 26k in the metabolism dataset and greater than 24k in the neurosensory



(a) cardiovascular (b) metabolism (c) neurosensory

Fig. 6. Performance comparison between FastANOVA and the brute-force approach when varying the number of permutations.

dataset. We observe that the runtime ratio tends to increase (approaching three-orders-of-magnitude speedup) as the number of SNPs increases. This indicates that the performance gain of FastANOVA is even higher for larger SNP datasets.

Figure 6 shows the runtime comparison when the number of phenotype permutations changes. The runtime of the brute-force approach is linear with respect to the number of permutations. FastANOVA is consistently two orders of magnitude faster than the brute-force approach. The performance gap increases as the number of permutations increases.

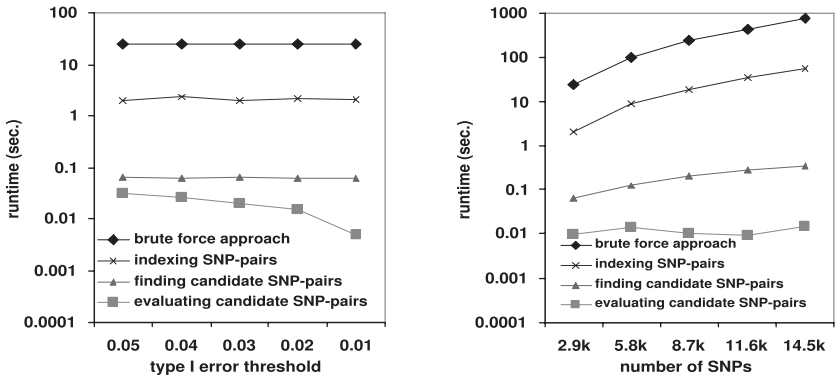
6.1.1.2 *Pruning Effect of the Upper Bound.* Table V shows the percentage of SNP-pairs pruned under different experimental settings. Since the three datasets have different numbers of SNPs, the 1st to 5th rows in the column of "# SNPs" correspond to the settings from left to right on x-axis in each plot in Figure 5. Most SNP-pairs are pruned under all settings. Moreover, as the Type I error threshold α decreases, the pruning ratio increases, which is consistent with runtime comparison shown in Figure 4. As the number of SNPs increases, the pruning ratio also increases. This is because, with more SNPs, the dynamic threshold used to prune the search space becomes higher. Hence, a larger portion of SNPs are pruned. This is consistent with results shown in Figure 5. Note that from Table V, we observe that the pruning ratio tends to remain steady when the number of permutations changes. However, we observe that the runtime ratio increases as the number of permutations increases. The reason for these two different trends will become clear after we show the results on the computational cost of each component of FastANOVA in the next subsection.

6.1.2 *Finding Significant SNP-Pairs.* In this section, we study the comparison between FastANOVA and the brute-force approach in finding significant SNP-pairs given a critical value F_α. Only the original phenotype (without permutations) is used in this procedure. We examine the detailed computation cost of each component of the FastANOVA algorithm. FastANOVA has three major components: building the indexing structure $Array(X_i)$ for every SNP X_i, accessing $Array(X_i)$ to find the candidate SNP-pairs, and performing ANOVA tests on these candidates.

Figures 7 to 9 show the performance comparison on the three datasets. The default experimental setting is the same as before. We examine the performance on metabolism dataset in detail. Similar behaviors can be observed on the other

Table V. Pruning Effects on Cardiovascular, Metabolism and Neurosensory Datasets when Finding Critical Value F_α

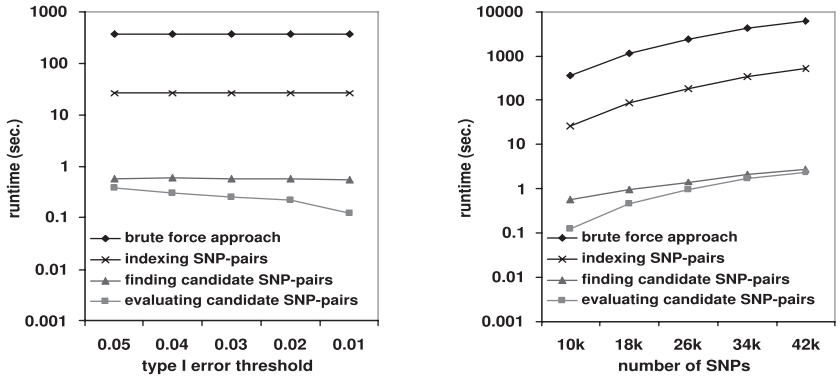
		Cardiovascular	Metabolism	Neurosensory
α	0.05	99.881%	99.724%	99.701%
	0.04	99.907%	99.758%	99.751%
	0.03	99.928%	99.797%	99.792%
	0.02	99.949%	99.877%	99.853%
	0.01	99.974%	99.929%	99.911%
# SNPs	1st	99.974%	99.929%	99.911%
	2nd	99.991%	99.985%	99.979%
	3rd	99.996%	99.996%	99.997%
	4th	99.998%	99.996%	99.997%
	5th	99.998%	99.993%	99.998%
# Perm.	100	99.974%	99.929%	99.911%
	200	99.966%	99.935%	99.917%
	300	99.977%	99.962%	99.919%
	400	99.977%	99.961%	99.914%
	500	99.974%	99.953%	99.907%



(a) Varying threshold values (b) Varying number of SNPs

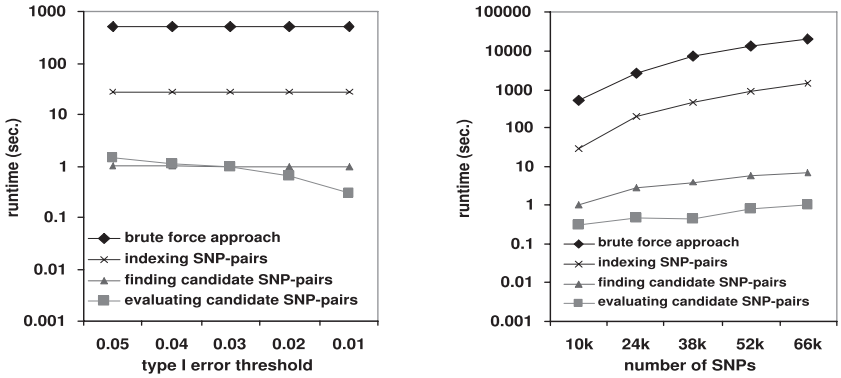
Fig. 7. Finding significant SNP-pairs (cardiovascular dataset).

two datasets. Figure 8(a) and Figure 8(b) show the runtime of these three components when varying the Type I error threshold and number of SNPs in the metabolism dataset respectively. Since F_α is a function of α, in Figure 8(a), we plot the runtime with respect to α. In both figures, the three lines from the bottom show the runtime of these three components. The runtime of the brute-force approach is the top line. As we can see from these two figures, performing two-locus ANOVA tests on candidate SNP pairs is two to three orders of magnitude faster than performing such tests on all SNP-pairs. This is the benefit of the upper bound pruning since most SNP-pairs have been pruned and only a very small portion of candidates need to be evaluated for their F-statistics. The cost for accessing the indexing structures is also small, which demonstrates the efficiency of the method introduced in Section 5.1 for candidate retrieval. Among the three components of FastANOVA, the most time-consuming one is



(a) Varying threshold values (b) Varyingn umber of SNPs

Fig. 8. Finding significant SNP-pairs (metabolism dataset).



(a) Varying threshold values (b) Varying number of SNPs

Fig. 9. Finding significant SNP-pairs (neurosensory dataset).

building the index structures. Yet, its runtime is only a small fraction of the runtime of performing the two-locus ANOVA tests on all SNP pairs. Note that, in permutation test, building the index structures is a one time cost. Once the index structures are built, they can be reused in all permutations. Therefore, the amortized overhead per permutation decreases when the number of permutations increases. This is why the pruning ratio remains steady in Table V while the runtime ratio increases in Figure 6 when the number of permutations increases.

Figure 10 shows the histogram of the sizes of the indexing structures for the three datasets. From Property 5.3, the maximum sizes of the indexing structures are 36 for the cardiovascular dataset, 64 for the metabolism dataset, and 100 for the neurosensory dataset. It is clear from the figure that the actual sizes of the indexing structures are much smaller than the maximum sizes.

6.1.3 *Finding F_{Y_k} for All Permutations.* Sometimes the users may be interested in finding F_{Y_k} values of all phenotype permutations. In this way, the

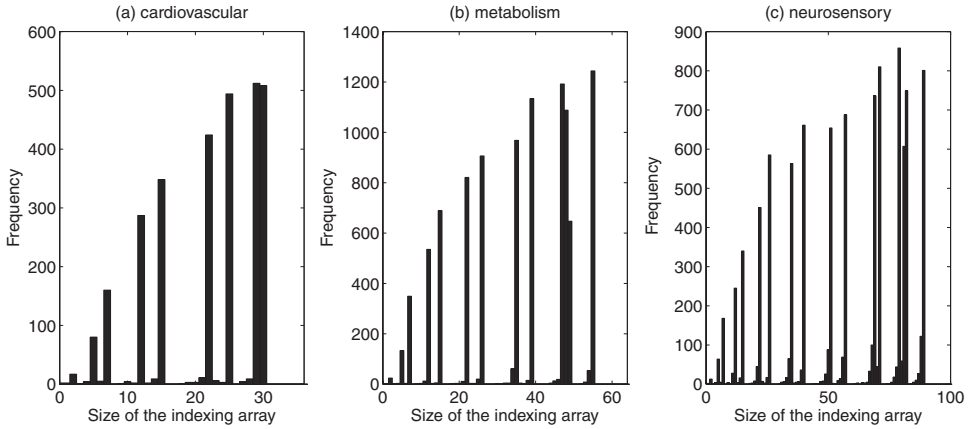


Fig. 10. Histogram of the sizes of the indexing structures.

Table VI. Pruning Effect on Cardiovascular, Metabolism and Neurosensory Datasets when Finding F_{Y_k} for all Permutations

Cardiovascular	Metabolism	Neurosensory
97.865%	97.844%	98.061%

users can get the critical value F_α for any Type I error threshold α ranging from 0 to 1, without re-running the permutation tests for different thresholds. Recall that, given a set of phenotype permutations $Y' = \{Y_1, Y_2, \ldots, Y_K\}$, $F_{Y_k} = \max\{F(X_i X_j, Y_k) | 1 \leq i < j \leq N\}$ is the maximum F-statistic value for permutation Y_k. F_α is the αKth largest value in $\{F_{Y_k} | Y_k \in Y'\}$. In this section, we show the pruning effect of the upper bound when it is applied to determine F_{Y_k} for every Y_k $(1 \leq k \leq K)$. Note that in this case, for each permutation Y_k, the dynamic threshold used to prune the search space is the largest F-statistic value of Y_k identified by the algorithm so far.

Table VI shows the pruning ratio of applying the upper bound to the three real phenotype datasets. The experimental setting is the same as the default setting before. As expected, the pruning ratios are slightly lower than those in Table V, where smaller Type I error thresholds are used to prune the search space. However, the pruning ratios on all three datasets are still above 97%. Moreover, finding all F_{Y_k} provides the advantage that we can get the F_α values for all possible α values instead of just for a specific one.

6.2 Synthetic Phenotypes

To further study the performance of FastANOVA, we generate three synthetic phenotypes whose values follow three different distributions: uniform, standard normal (with mean 0 and variance 1), and standard exponential distribution (with the probability density function $f(x) = e^{-x}$). Our purpose is to study the pruning effect of the upper bound under different phenotype distributions. The default setting of the experiments in this subsection is as follows:

Table VII. Pruning Effect when Finding Critical Value F_α using three Synthetic Phenotypes

		Uniform	Normal	Exponential
α	0.05	96.469%	97.793%	99.335%
	0.04	96.888%	98.222%	99.401%
	0.03	97.695%	98.631%	99.502%
	0.02	98.712%	99.072%	99.617%
	0.01	99.605%	99.506%	99.737%
# SNPs	10k	99.605%	99.506%	99.737%
	22k	99.864%	99.814%	99.924%
	34k	99.907%	99.905%	99.967%
	46k	99.928%	99.889%	99.965%
	58k	99.941%	99.942%	99.963%
# Perm.	100	99.605%	99.506%	99.737%
	200	98.891%	99.398%	99.726%
	300	98.897%	99.072%	99.780%
	400	98.623%	99.315%	99.762%
	500	98.709%	99.199%	99.759%
# indiv.	28	99.756%	99.695%	99.893%
	30	99.422%	99.577%	99.880%
	32	99.605%	99.506%	99.737%
	34	99.073%	99.289%	99.773%
	36	98.736%	98.832%	99.745%

#individuals = 32, #SNPs = 10,000, #permutations = 100, $\alpha = 0.01$. There are 60,970 unique SNPs for these 32 individuals.

Table VII shows the pruning ratio of FastANOVA under different settings using permutation test. In this table, we also include the pruning ratio when the number of individuals varies. We observe that the pruning effects are similar to that of real phenotypes, which indicates that the upper bound pruning is effective and insensitive to different phenotype distributions.

7. DISCUSSION

The large number of available SNPs poses great computational challenge to the genome-wide association study. To assess the significance of the findings, permutation test is usually required. These factors make the association study a very time-consuming process. Thus tools that can improve the efficiency of the association study are in demand.

In this article, we present an efficient algorithm, FastANOVA, for genome-wide two-locus ANOVA test. FastANOVA is a complete algorithm which guarantees to find the optimal solution. Experimental results demonstrate that FastANOVA is two to three orders of magnitude faster than the brute-force alternative. The efficiency of FastANOVA is gained from two sources. First, it utilizes an upper bound of the two-locus ANOVA test value to prune a majority of the SNP-pairs. Second, it identifies and reuses computation units that are independent of the phenotype and hence are invariant in permutation test. By eliminating redundant computation of these invariant units, FastANOVA is much more efficient than the brute-force method.

Table VIII. Notations Used in the Derivation of the Upper Bound for Two-Locus Chi-Square Test

Symbols	Formulas
T_1	$\frac{M^2}{(O_A+O_B)(O_A+O_C)(O_C+O_D)}$
S_1	$\max\{O_A^2, O_C^2\}$
$\mathscr{R}_1$	$\min\left\{\left[\frac{O_{X_j=1}}{O_{X_j=0}}\middle\vert X_i=0\right], \left[\frac{O_{X_j=0}}{O_{X_j=1}}\middle\vert X_i=0\right]\right\}$
T_2	$\frac{M^2}{(O_A+O_B)(O_B+O_D)(O_C+O_D)}$
S_2	$\max\{O_B^2, O_D^2\}$
$\mathscr{R}_2$	$\min\left\{\left[\frac{O_{X_j=1}}{O_{X_j=0}}\middle\vert X_i=1\right], \left[\frac{O_{X_j=0}}{O_{X_j=1}}\middle\vert X_i=1\right]\right\}$

7.1 Extension to Chi-Square Test

As our initial attempt to develop scalable algorithms for genome-wide association study, FastANOVA is specifically designed for the ANOVA test on quantitative phenotypes. Another category of phenotypes is generated in case-control study, where the phenotypes are binary variables representing disease/non-disease individuals. Chi-square test is one of the most commonly used statistics in binary phenotype association study. We can extend the principles in FastANOVA for efficient two-locus chi-square test [Zhang et al. 2009b]. The general idea of FastChi is similar to that of FastANOVA, that is, re-formulating the chi-square test statistic to establish an upper bound of two-locus chi-square test, and indexing the SNP-pairs according to their genotypes in order to effectively prune the search space and reuse redundant computations. Here we briefly introduce the FastChi algorithm.

For SNP X_i, we represent the chi-square test value of X_i and the binary phenotype Y as $\chi^2(X_i, Y)$. For any SNP-pair X_i and X_j, we use $\chi^2(X_iX_j, Y)$ to represent the chi-square test value for the combined effect of (X_iX_j) with Y. Let A, B, C, D represent the following events respectively: $Y = 0 \wedge X_i = 0$; $Y = 0 \wedge X_i = 1$; $Y = 1 \wedge X_i = 0$; $Y = 1 \wedge X_i = 1$. Let O_{event} denote the observed value of an event. T_1, T_2, S_1, S_2, $\mathscr{R}_1$, and $\mathscr{R}_2$ represent the formulas shown in Table VIII. We have the upper bound of $\chi^2(X_iX_j, Y)$ stated in Theorem 7.1.

THEOREM 7.1 (UPPER BOUND OF $\chi^2(X_iX_j, Y)$).

$$\chi^2(X_iX_j, Y) \leq \chi^2(X_i, Y) + T_1S_1\mathscr{R}_1 + T_2S_2\mathscr{R}_2.$$

For given phenotype Y and SNPX_i, $\chi^2(X_i, Y)$, T_1, S_1, T_2, and S_2 are constants. $\mathscr{R}_1$ and $\mathscr{R}_2$ are the only variables that depend on X_j and may vary for different SNP-pairs $(X_iX_j) \in AP(X_i)$. (Recall that $AP(X_i) = \{(X_iX_j)|i + 1 \leq j \leq N\}$.) Thus, for a given X_i, we can treat equation $\chi^2(X_i, Y) + T_1S_1\mathscr{R}_1 + T_2S_2\mathscr{R}_2 = \theta$ as a *straight line* in the 2-D space of $\mathscr{R}_1$ and $\mathscr{R}_2$. The ones whose $(\mathscr{R}_1(X_iX_j), \mathscr{R}_2(X_iX_j))$ values fall below the line can be pruned without any further test.

Suppose that there are 32 individuals, X_i contains half 0's, and half 1's. For the SNP-pairs in $AP(X_i)$, the possible values of $\mathscr{R}_1$ (and $\mathscr{R}_2$) are

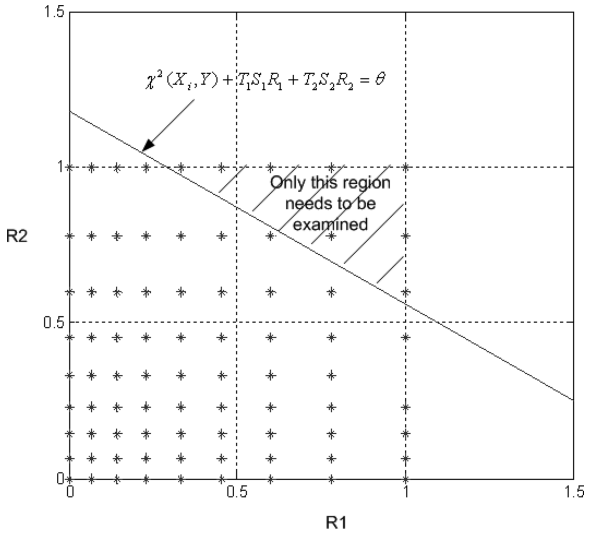


Fig. 11. Pruning SNP-Pairs in $AP(X_i)$ using the Upper Bound.

$\{\frac{0}{16}, \frac{1}{15}, \frac{2}{14}, \frac{3}{13}, \frac{4}{12}, \frac{5}{11}, \frac{6}{10}, \frac{7}{9}, \frac{8}{8}\}$. Figure 11 shows the 2-D space of $\mathscr{R}_1$ and $\mathscr{R}_2$. The blue stars represent the values that ($\mathscr{R}_1$, $\mathscr{R}_2$) can take. The line $\chi^2(X_i, Y) + T_1S_1\mathscr{R}_1 + T_2S_2\mathscr{R}_2 = \theta$ is plotted in the figure. Only the SNP-pairs whose ($\mathscr{R}_1$, $\mathscr{R}_2$) values are in the shaded region are subject to two-locus Chi-square test.

Similar to FastANOVA, in FastChi, we can index the SNP-pairs in $AP(X_i)$ according to their genotype relationships, i.e., by the values of ($\mathscr{R}_1$, $\mathscr{R}_2$). Experimental results demonstrate that FastChi is an order of magnitude faster than the brute force alternative. For further details of FastChi, please refer to Zhang et al. [2009b].

7.2 A General Approach for Binary Phenotypes

The common drawback of FastANOVA and FastChi is that they are specifically designed for ANOVA and chi-square tests, and cannot be applied to other statistics. In Zhang et al. [2009a], a generalized approach, COE, for case-control study is proposed. The major contribution is that COE can be applied to a wide range of statistics. One key observation is that many commonly used statistics are convex functions. This property allows to use convex optimization techniques to find tight upper bound for tow-locus statistical tests.

We use $\mathscr{T}$ to denote the statistical test that will be used for two-locus association study. Specifically, we represent the test value of SNP X_i and phenotype Y as $\mathscr{T}(X_i, Y)$, and represent the test value of SNP-pair (X_iX_j) and Y as $\mathscr{T}(X_iX_j, Y)$. A contingency table, which records the observed values of all events, is the basis for many statistical tests. Table IX shows contingency tables for the single-locus test $\mathscr{T}(X_i, Y)$, genotype relationship between SNPs X_i and X_j, and two-locus test $\mathscr{T}(X_iX_j, Y)$. Next, we use chi-square test,

Table IX. Contingency Tables

(a) X_i and Y

	$X_i = 0$	$X_i = 1$	Total
$Y = 0$	event A	event B	
$Y = 1$	event C	event D	
Total			M

(b) X_i and X_j

	$X_i = 0$	$X_i = 1$	Total
$X_j = 0$	event S	event T	
$X_j = 1$	event P	event Q	
Total			M

(c) X_iX_j and Y

	$X_i = 0$		$X_i = 1$		Total
	$X_j = 0$	$X_j = 1$	$X_j = 0$	$X_j = 1$	
$Y = 0$	event a_1	event a_2	event b_1	event b_2	
$Y = 1$	event c_1	event c_2	event d_1	event d_2	
Total					M

G-test (likely ratio test), and entropy based test as concrete examples to show that they are convex statistics.

Let $A, B, C, D, S, T, P, Q, a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2$ represent the events as shown in Table IX. Let E_{event} and O_{event} denote the expected value and observed value of an event. Suppose that $\mathbb{E}_0 = \{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2\}$, $\mathbb{E}_1 = \{a_1, a_2, c_1, c_2\}$, and $\mathbb{E}_2 = \{b_1, b_2, d_1, d_2\}$. The two-locus chi-square tests can be calculated as follows:

$$\chi^2(X_iX_j, Y) = \underbrace{\sum_{event \in \mathbb{E}_1} \frac{(O_{event} - E_{event})^2}{E_{event}}}_{\chi_1^2(X_iX_jY)} + \underbrace{\sum_{event \in \mathbb{E}_2} \frac{(O_{event} - E_{event})^2}{E_{event}}}_{\chi_2^2(X_iX_jY)}. \quad (6)$$

Note that we intentionally break the calculation into two components: one for the events in $\mathbb{E}_1$, denoted as $\chi_1^2(X_iX_jY)$, and one for the events in $\mathbb{E}_2$, denoted as $\chi_2^2(X_iX_jY)$. The reason for separating these two components is that each of these two components is a convex function (See Lemma 7.2).

The G-test, also known as a likelihood ratio test for goodness of fit, is an alternative to the chi-square test. The formula for two-locus G-test is

$$G(X_iX_j, Y) = 2 \sum_{event \in \mathbb{E}_1} O_{event} \cdot \ln\left(\frac{O_{event}}{E_{event}}\right) + 2 \sum_{event \in \mathbb{E}_2} O_{event} \cdot \ln\left(\frac{O_{event}}{E_{event}}\right). \quad (7)$$

Information-theoretic measurements have been proposed for association study. We examine the mutual information measure, which is the basic form of many other measurements. The mutual information between SNP-pair (X_iX_j) and phenotype Y is $I(Y; X_iX_j) = H(Y) + H(X_iX_j) - H(X_iX_jY)$, in which the joint entropy $-H(X_iX_jY)$ is calculated as

$$-H(X_iX_jY) = \sum_{event \in \mathbb{E}_1} \frac{O_{event}}{M} \cdot \log \frac{O_{event}}{M} + \sum_{event \in \mathbb{E}_2} \frac{O_{event}}{M} \cdot \log \frac{O_{event}}{M}. \quad (8)$$

Let $\mathscr{T}(X_iX_j, Y)$ represent any one of $\chi^2(X_iX_j, Y)$, $G(X_iX_j, Y)$, and $-H(X_iX_jY)$. Let $\mathscr{T}_1(X_iX_jY)$ denote the component for events in $\mathbb{E}_1$, and $\mathscr{T}_2(X_iX_jY)$ denote the component for events in $\mathbb{E}_2$. The following lemma shows the convexity of $\mathscr{T}_1(X_iX_jY)$ and $\mathscr{T}_2(X_iX_jY)$.

LEMMA 7.2. *Given the values of $O_A, O_B, O_C, O_D, O_P, O_Q$, $\mathscr{T}_1(X_iX_jY)$ is a convex function of O_{c_2}, and $\mathscr{T}_2(X_iX_jY)$ is a convex function of O_{d_2}.*

Suppose that the range of O_{c_2} is $[l_{c_2}, u_{c_2}]$, and the range of O_{d_2} is $[l_{d_2}, u_{d_2}]$. For any convex function, its maximum value is attained at one of the vertices of it convex domain [Boyd and Vandenberghe 2004]. Thus, we have the following theorem.

THEOREM 7.3. *Given the values of $O_A, O_B, O_C, O_D, O_P, O_Q$, for chi-square test, G-test, and entropy-based test, the maximum value of $\mathscr{T}_1(X_iX_jY)$ is attained when $O_{c_2} = l_{c_2}$ or $O_{c_2} = u_{c_2}$. The maximum value of $\mathscr{T}_2(X_iX_jY)$ is attained when $O_{d_2} = l_{d_2}$ or $O_{d_2} = u_{d_2}$.*

By further studying the relationships between the observed values shown in Table IX, we can derive the ranges of O_{c_2} and O_{d_2}.

THEOREM 7.4. *Given the values of $O_A, O_B, O_C, O_D, O_P, O_Q$, the ranges of O_{c_2} and O_{d_2} are*

$$\begin{cases} \max\{0, O_P - O_A\} \le O_{c_2} \le \min\{O_P, O_C\}; \\ \max\{0, O_Q - O_B\} \le O_{d_2} \le \min\{O_Q, O_D\}. \end{cases}$$

Experimental results show that the developed upper bound is much tighter than that of the FastChi algorithm. In addition, this approach only requires the test statistic to be a convex function, which is true for a variety of tests. Please refer to Zhang et al. [2009a] for further details about COE.

8. LIMITATIONS AND FUTURE WORK

In general, genome-wide association study is not restricted to two-locus tests. Ideally, one should be able to examine the interactions among any number of SNPs. This dramatically increases the computational burden. For example, suppose that number of SNPs $N = 10,000$ and number of permutations $K = 1,000$, the number of tests needed for two-locus association study is in the order of 10^{10}, and number of tests needed for three-locus association study is in the order of 10^{14}. In practice, this means that if two-locus association testing takes 1 second, the three-locus testing will take about 10^4 seconds. The computational burden increases exponentially when the number of SNPs considered for interaction increases. In our future work, we will investigate scalable algorithms for multi-locus association study involving more than two-locus.

Our work in this article is motivated by the association study for inbred mice, whose genotypes are usually binary. For other subjects, such as human, the genotype are heterozygous, where SNPs are encoded as $\{0, 1, 2\}$. The formulation in this article is for binary SNPs. In the future work, we plan to extend the principles used in this article to the heterozygous case. Another difference between human subjects and inbred mice is that the number of samples of human subjects are usually much larger than that of the mice. This could potentially impair the applicability of the indexing structure which is the key component of FastANOVA. The size of the indexing structure depends on the number of individuals in the dataset: the maximum size of the indexing structure increases

quadratically with respect to the number of individuals. The associated problem is that he number of SNP-pairs indexed by the same entry will decrease and the accessing time of the indexing structure will increase. In the worst case, if the number of entries is larger than the number of SNPs, then there is no advantage to build the indexing structure. We will investigate algorithms for large sample datasets in our future work.

REFERENCES

BALDING, D. J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 10, 781–791.

BOYD, S. AND VANDENBERGHE, L. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, MA.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Inc., Monterey, CA.

CARLBORG, O., ANDERSSON, L., AND KINGHORN, B. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics 155*, 4, 2003–2010.

CARLSON, C. S., EBERLE, M. A., KRUGLYAK, L., AND NICKERSON, D. A. 2004. Mapping complex disease loci in whole-genome association studies. *Nature 429*, 446–452.

CHI, P., DUGGAL, P., KAO, W., MATHIAS, R. A., GRANT, A. V., STOCKTON, M. L., GARCIA, J. G. N., INGERGOLL, R. G., SCOTT, A. F., BENTY, T. H., BARNES, K. C., AND FALLIN, M. D. 2006. Comparison of SNP tagging methods using empirical data: Association study of 713 SNPs on chromosome 12q14.3-12q24.21 for asthma and total serum IgE in an African Caribbean population. *Genet. Epidemiol. 30*, 7, 609–619.

CURTIS, D., NORTH, B. V., AND SHAM, P. C. 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Ann. Hum. Genet. 65*, 95–107.

DOERGE, R. W. 2002. Multifactorial genetics: Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet. 3*, 43–52.

DUDOIT, S. AND VAN DER LAAN, M. J. 2008. *Multiple Testing Procedures with Applications to Genomics*. Springer-Verlag, Berlin, Germany.

EVANS, D. M., MARCHINI, J., MORRIS, A. P., AND CARDON, L. R. 2006. Two-stage two-locus models in genome-wide association. *PLoS Genet. 2*, e157.

HALPERIN, E., KIMMEL, G., AND SHAMIR, R. 2005. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. In *Proceedings of the ISMB*. Oxford University Press.

HOH, J. AND OTT, J. 2003. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet. 4*, 701–709.

HOH, J., WILLE, A., ZEE, R., CHENG, S., REYNOLDS, R., LINDPAINTNER, K., AND OTT, J. 2000. Selecting SNPs in two-stage analysis of disease association data: A model-free approach. *Ann. Hum. Genet. 64*, 413–417.

IDERAABDULLAH, F., DELA CASA-ESPERÓN, E., BELL, T. A., PETWILER, D. A., MAGNUSON, T., SAPIENZA, C., AND PARDO-MANUEL DE VILLENA, F. 2004. Genetic and haplotype diversity among wild-derived mouse inbred strains. *Gen. Res. 14*, 10a, 1880–1887.

LIU, H. AND MOTODA, H. 1998. *Feature selection for knowledge discovery and data mining*. Kluwer Academic, Boston, MA.

MILLER, R. G. 1981. *Simultaneous Statistical Inference*. Springer-Verlag, New York.

MOORE, J. H., GILBERT, J. C., TSAI, C.-T., CHIANG, F.-T., HOLDEN, T., BARNEY, N., AND WHITE, B. C. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theoret. Biol. 241*, 2, 252–261.

NAKAMICHI, R., UKAI, Y., AND KISHINO, H. 2001. Detection of closely linked multiple quantitative trait loci using a genetic algorithm. *Genetics 158*, 1, 463–475.

NELSON, M. R., KARDIA, S. L., FERRELL, R. E., AND SING, C. F. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Gen. Res. 11*, 458–470.

Ohno, Y., Tanase, H., Nabika, T., Otsuka, K., Sasaki, T., Suzawa, T., Morii, T., Yamori, Y., and Saruta, T. 2000. Selective genotyping with epistasis can be utilized for a major quantitative trait locus mapping in hypertension in rats. *Genetics 155*, 785–792.

Pagano, M. and Gauvreau, K. 2000. *Principles of Biostatistics*. Duxbury Press, Pacific Grove, CA.

Province, M. A., Shannon, W. D., and Rao, D. C. 2001. Classification methods for confronting heterogeneity. *Adv. Genet. 42*, 273–286.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Amer. J. Hum. Gen. 69*, 138–147.

Roberts, A., Mcmillan, L., Wang, W., Parker, J., Rusyn, I., and Threadgill, D. 2007. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. In *Proceedings of ISMB*. Oxford University Press.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardle, K., Bengtsson Boström, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rástam, L., Speliotes, E. K., Taskinen, M.-R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurita, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava. M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chim, G.-W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science 316*, 1331–1336.

Scuteri, A., Sanna, S., Chan, W. M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G. R. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLPoS Genet. 3*, 7, e115.

Sebastiani, P., Lazarus, R., Weiss, S. T., Kunkel, L. M., Kohane, I. S., and Ramoni, M. F. 2003. Minimal haplotype tagging. *Proc. Natl. Acad. Sci. USA 100*, 17, 9900–9905.

Segr, D., Deluna, A., Church, G. M., and Kishony, R. 2005. Modular epistasis in yeast metabolism. *Nat. Genet. 37*, 77–83.

Sherriff, A., and Ott, J. 2001. Applications of neural networks for gene finding. *Adv. Genet. 42*, 287–297.

Shimomura, K., Low-Zeddies, S. S., King, D. P., Steeves, T. D., Whiteley, A., Kushla, J., Zemenides, P. D., Lin, A., Vitaterna, M. H., Churchill, G. A., and Takahashi, J. S. 2001. Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Gen. Res. 11*, 6, 959–980.

Weedon, M. N., Lettre, G., Freathy, R. M., Lindgren, C. M., Voight, B. F., Perry, J. R., Elliott, K. S., Hackett, R., Guiducci, C., Shields, B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burtt, N. P., Rayner, N. W., Saxena, R., Ardlie, K., Tobias, J. H., Ness, A. R., Ring, S. M., Palmer, C. N., Morris, A. D., Peltonen, L., Salomaa, V., Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, Davey Smith, G., Groop, L. C., Hattersley, A. T., McCarthy, M. I., Hirschhorn, J. N., and Frayling, T. M. 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet. 39*, 10, 1245–1250.

Westfall, P. H. and Young, S. S. 1993. *Resampling-Based Multiple Testing*. Wiley, New York.

Zhang, H. and Bonney, G. 2000. Use of classification trees for association studies. *Genet. Epidemiol. 19*, 323–332.

Zhang, X., Pan, F., Xie, Y., Zow, F., and Wang, W. 2009a. COE: A general approach for efficient genome-wide two-locus epistatic test in disease association study. In *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Lecture Notes in Computer Science, vol. 5541, Springer-Verlag, Berlin, Germany.

Zhang, X., Zou, F., and Wang, W. 2009b. FastChi: An efficient algorithm for analyzing gene-gene interactions. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific.

Received January 2009; revised May 2009; accepted July 2009