COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study

Xiang Zhang [1], Feng Pan [1], Yuying Xie [2], Fei Zou [3], and Wei Wang [1]
Departments of [1]Computer Science, [2] Genetics, and [3]Biostatistics
University of North Carolina at Chapel Hill
[1]{xiang, panfeng,weiwang}@cs.unc.edu, [2]xyy@email.unc.edu, [3]fzou@bios.unc.edu

Abstract. The availability of high density single nucleotide polymorphisms (SNPs) data has made genome-wide association study computationally challenging. Two-locus epistasis (gene-gene interaction) detection has attracted great research interest as a promising method for genetic analysis of complex diseases. In this paper, we propose a general approach, COE, for efficient large scale gene-gene interaction analysis, which supports a wide range of tests. In particular, we show that many commonly used statistics are convex functions. From the observed values of the events in two-locus association test, we can develop an upper bound of the test value. Such an upper bound only depends on single-locus test and the genotype of the SNP-pair. We thus group and index SNP-pairs by their genotypes. This indexing structure can benefit the computation of all convex statistics. Utilizing the upper bound and the indexing structure, we can prune most of the SNP-pairs without compromising the optimality of the result. Our approach is especially efficient for large permutation test. Extensive experiments demonstrate that our approach provides orders of magnitude performance improvement over the brute force approach.

1 Introduction

High throughput genotyping technologies produce vast amounts of genetic polymorphism data which empowers genome-wide association study, and at the same time, makes it a computationally challenging task [16, 20, 24]. As the most abundant source of genetic variations, the number of single nucleotide polymorphisms (SNPs) in public datasets is up to millions[1, 3]. Through analyzing genetic variation across a population consisting of disease (case) and healthy (control) individuals, the goal of disease association study is to find the genetic factors underlying the disease phenotypes. Growing evidence suggests that many diseases are likely caused by the joint effect of multiple genes [8, 29]. The interaction between genes is also referred to as epistasis [10]. In an epistatic interaction, each gene may only have weak association with the disease. But when combined, they have strong effect on the disease. A large amount of research has been devoted to find epistatic interactions between genes [4, 11, 17], among which the *two-locus association mapping* has attracted most attention. The goal is to find SNP-pairs having strong association with the phenotype. Important findings are appearing in the literature from studying the association between phenotypes and SNP-pairs [26, 27, 33].

Two critical issues need to be addressed in epistasis detection – one from the *statistical* side, and one from the *computational* side. The statistic issue is to develop statistical tests that have strong power in capturing epistatic interactions. Commonly used statistics in disease association study include: chi-square test, G-test, information-theoretic association measurements, and trend test [4, 21, 31]. Different tests are good at detecting different epistatic interactions, and there is no single winner. Another thorny challenge in epistasis detection is the computational burden posed by the huge amount of SNPs genotyped in the whole genome. The enormous search space often makes the complete genome-wide epistasis detection intractable.

The computational issue is further compounded by the well-known multiple test problem, which can be described as the potential increase in Type I error when tests are performed multiple times. Let α be the significant level for each independent test. If n independent comparisons are performed, the family-wise error α' is given by $\alpha' = 1 - (1 - \alpha)^n$. For example, if $\alpha = 0.05$ and $n = 20$, then $\alpha' = 1 - 0.95^{20} = 0.64$. We have probability 0.64 to get at least one spurious result. Permutation test is a standard procedure for family-wise error rate controlling. By repeating the test many times with randomly permuted phenotype, a critical threshold can be established to assess the statistical significance of the findings. Ideally, permutation test should be performed in the genome-wide scale. In practice, however, permutation test is usually reserved for a small number of candidate SNPs. This is because large permutation test usually entails prohibitively long computation time. For example, if the number of SNPs is 10,000, and the number of permutations is 1,000. The number of SNP-pairs need to be tested in a two-locus epistasis detection is about 5×10^{10}. In this paper, we focus on addressing the computational challenges of two-locus epistatic detection when large permutation test is needed. In the following discussion, we briefly review the related work from a computational point of view.

Exhaustive algorithms [19, 22] have been developed for small datasets consisting of tens to hundreds of SNPs. Since they explicitly enumerate all possible SNP combinations, they are not well adapted to genome-wide association studies. Genetic algorithm [7] has been proposed. However, this heuristic approach does not guarantee to find the optimal solution. A two-step approach [14, 30] is commonly used to reduce the computational burden. The idea is to first select a subset of important SNPs according to some criteria, which is also known as SNP tagging [9, 15, 28]. Then in the second step, an exhaustive search is performed to find the interactions among the selected SNPs. This approach is incomplete since it ignores the interactions among the SNPs that individually have weak association with the phenotype. A case study on colon cancer demonstrates that the two-step approach may miss important interacting SNPs.

1.1 A Case Study of Colon Cancer

We perform two-locus chi-square test on genome-wide mouse SNP data. The dataset is extracted from [32] and [2]. There are 14 cases out of 32 individuals. The number of SNPs is 132,896. The top-100 SNP-pairs having highest chi-square test values are recorded. We show that a two-step approach will fail to identify most of the signficant pairs. We compute the single-locus chi-square test value for every SNP in these pairs. In Figure 1, the yellow bars show the histogram of their single-locus test values. More

than half of the SNPs have very low test values. However, when combined with other SNPs, the test values dramatically increases. The green bar in the figure represents the histogram of the two-locus test values of the pairs. Since the two-step approach ignores the interactions between SNPs that individually have weak association with the phenotype, a majority of the top-100 SNP-pairs will not be identified.

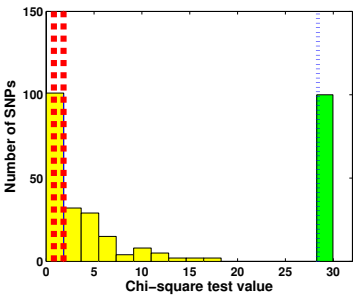

Fig. 1: Histogram test values

Among the SNP pairs identified above, the interaction between the SNP located at 86,627,952 base pair on Chromosome 2 and the SNP located at 94,546,781 base pair on Chromosome 6 was reported previously [25]. They correspond to two candidate genes: Ptprj, located on Chromosome 2 from 90,269,911 to 90,319,327, and Lrig1, located on Chromosome 6 from 95,067, to95,079,646. Numerous studies have emphasized the crucial importance of Ptprj to colon cancer susceptibility. It is well known that immune system may play a protective role on colon-cancer, indicating the potential importance of Ptprj to cancer susceptibility [18]. Ptprj knock-out mice display an impaired immune system with decreased B cell number, abnormal B cell differentiation and shortened life-span. Previous evidence also suggests that Lrig1 acts as a feedback negative regulator of signaling by receptor tyrosine kinases through a mechanism that involves enhancement of receptor ubiquitination and accelerated intracellular degradation. Rceptor tyrosine kinases including EGFR/ERBB1 are believed to be a main player on colon cancer-genesis, and Egfr expression has correlated with poor prognosis of colon cancer. Therefore, Lrig1 is a good candidate for colon cancer susceptibility [13]. Each of the two genes individually shows very weak association signal. Their single-locus chi-square test values are 1.81 and 0.79 respectively, depicted by the two red dotted lines to the left in Figure 1. However, this pair of genes jointly show much stronger association. The two-locus test value is 28.4, depicted by the blue dotted line to the right in Figure 1. This implies a strong epistatic interaction between the two genes. Using a two-step approach, however, these two SNPs will not be selected for interaction study since they both have very low single-locus test values.

Some recent work [34, 35] has taken the initial steps to develop complete algorithms for genome-wide two-locus epistasis detection: FastANOVA [34] for two-locus ANOVA (analysis of variance) test on quantitative traits and FastChi [35] for two-locus chi-square test on case-control phenotypes. Both methods rework the formula of ANOVA test and Chi-square test to estimate an upper bound of the test value for SNP pairs. These upper bounds are used to identify candidate SNP pairs that may have strong epistatic effect. Repetitive computation in a permutation test is also identified and performed once whose results are stored for use by all permutations. These two strategies lead to substantial speedup, especially for large permutation test, without compromising the accuracy of the test. These approaches guarantee to find the optimal solutions. However, a common drawback of these methods is that they are designed for specific tests, i.e., chi-square test and ANOVA test. The upper bounds used in these methods do not work for other statistical tests, which are also routinely used by researchers. In

addition, new statistics for epistasis detection are continually emerging in the literature [5, 12, 36]. Therefore, it is desirable to develop a general model that supports a variety of statistical tests.

In this paper, we propose a general approach, COE[1], to scale-up the process of genome-wide two-locus epistasis detection. Our method guarantees to find the optimal solution. A significant improvement over previous methods is that our approach can be applied to a wide range of commonly used statistical tests. We show that a key property of these statistics is that they are all convex functions of the observed values of certain events in two-locus tests. This allows us to apply the convex optimization techniques [6]. Specifically, by examining the contingency tables, we can derive constraints on these observed values. Utilizing these constraints, an upper bound can be derived for the two-locus test value. Similar to the approaches in [34, 35], this upper bound only depends on single-locus test and the genotype of the SNP-pairs. It avoids redundant computation in permutation test by grouping and indexing the SNP-pairs by their genotypes. An important difference, however, is that the upper bound presented in this paper is general and much tighter than those in previous methods such as FastChi. It supports all tests using convex statistics and can prune the search space more efficiently. As a result, our method is orders of magnitude faster than the brute force approach, in which all SNP-pairs need to be evaluated for their test values, and is an order of magnitude faster than the pruning strategies used in previous methods such as FastChi. In this paper, we focus on the case where SNPs are binary variables which can be encoded by $\{0, 1\}$. The principle introduced here is also applicable to heterozygous case where SNPs are encoded using $\{0, 1, 2\}$.

2 Problem Formalization

Let $\{X_1, X_2, \cdots, X_N\}$ be the set of all biallelic SNPs for M individuals, and Y be the binary phenotype of interest (e.g., disease or non-disease). We adopt the convention of using 0 to represent majority allele and 1 to represent minority allele, and use 0 for non-disease and 1 for disease. We use $\mathscr{T}$ to denote the statistical test. 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 1 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 $\mathscr{T}(X_iX_j, Y)$.

The goal of permutation test is to find a critical threshold value. A two-locus epistasis detection with permutation test is typically conducted as follows [21, 34, 35]. A permutation Y_k of Y represents a random reshuffling of the phenotype Y. In each permutation, the phenotype values are randomly reassigned to individuals with no replacement. Let $Y' = \{Y_1, Y_2, \cdots, Y_K\}$ be the set of K permutations of Y. For each permutation $Y_k \in Y'$, let $\mathscr{T}_{Y_k}$ represent the maximum test value among all SNP-pairs, i.e., $\mathscr{T}_{Y_k} = \max\{\mathscr{T}(X_iX_j, Y_k) | 1 \leq i < j \leq N\}$. The distribution of $\{\mathscr{T}_{Y_k} | Y_k \in Y'\}$ is used as the null distribution. Given a Type I error threshold α, the *critical value* $\mathscr{T}_\alpha$ is

[1] COE stands for Convex Optimization-based Epistasis detection algorithm.

(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

Table 1. Contingency Tables

the αK-th largest value in $\{\mathscr{T}_{Y_k}|Y_k \in Y'\}$. After determining the critical value $\mathscr{T}_\alpha$, a SNP-pair (X_iX_j) is considered significant if its test value with the original phenotype Y exceeds the critical value, i.e., $\mathscr{T}_(X_iX_j, Y) \geq \mathscr{T}_\alpha$.

Determining the critical value is computationally more demanding than finding significant SNP-pairs, since the test procedure needs to be repeated for every permutation in order to find the maximum values. These two problems can be formalized as follows.

Determining Critical Value: For a given Type I error threshold α, find the critical value $\mathscr{T}_\alpha$, which is the αK-th largest value in $\{\mathscr{T}_{Y_k}|Y_k \in Y'\}$.

Finding Significant SNP-pairs: For a given critical value $\mathscr{T}_\alpha$, find the significant SNP-pairs (X_iX_j) such that $\mathscr{T}(X_iX_j, Y) \geq F_\alpha$.

In the reminder of the paper, we first show the convexity of common statistics. Then we discuss how to establish an upper bound of two-locus test and use it in the algorithm to efficiently solve the two problems.

3 Convexity of Common Test Statistics

In this section, we show that many commonly used statistics are convex functions. Since there are many statistics in the literature, it is impossible to exhaustively enumerate all of them. We focus on four widely used statistics: chi-square test, G-test, entropy-based statistic, and Cochran-Armitage trend test.

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 1. 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:

$$\chi^2(X_iX_j, Y) = \underbrace{\sum_{event \in \mathbb{E}_1} \frac{(O_{event} - E_{event})^2}{E_{event}}}_{\chi^2_1(X_iX_jY)} + \underbrace{\sum_{event \in \mathbb{E}_2} \frac{(O_{event} - E_{event})^2}{E_{event}}}_{\chi^2_2(X_iX_jY)}. \quad (1)$$

Note that we intentionally break the calculation into two components: one for the events in $\mathbb{E}_1$, denoted as $\chi^2_1(X_iX_jY)$, 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 1).

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

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

Information-theoretic measurements have been proposed for association study [12, 36]. 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

$$-H(X_iX_jY) = \sum_{event \in \mathbb{E}_1} \frac{O_{event}}{M} \cdot \log \frac{O_{event}}{M} + \sum_{event \in \mathbb{E}_2} \frac{O_{event}}{M} \cdot \log \frac{O_{event}}{M}. \quad (3)$$

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 1. *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}.*

Proof. See Appendix.

The Cochran-Armitage test for trend is another widely used statistic in genetic association study. Let $Z = (O_{c_1} - pO_S)(s_1 - \bar{s}) + (O_{c_2} - pO_P)(s_2 - \bar{s}) + (O_{d_1} - pO_T)(s_3 - \bar{s}) + (O_{d_2} - pO_Q)(s_4 - \bar{s})$. The Cochran-Armitage two-locus test can be calculated as

$$z^2 = Z^2/[p(1-p)(O_S(s_1 - \bar{s})^2 + O_P(s_2 - \bar{s})^2 + O_T(s_3 - \bar{s})^2 + O_Q(s_4 - \bar{s})^2)],$$

where p is the percentage of cases in the case-control population, s_i ($i \in \{1, 2, 3, 4\}$) are user specified scores for the four possible genotype combinations of (X_iX_j): $\{00, 01, 10, 11\}$, and $\bar{s} = (O_S s_1 + O_P s_2 + O_T s_3 + O_Q s_4)/M$ is the weighted average score. The following theorem shows the convexity of the trend test.

Lemma 2. *Given the values of $O_A, O_B, O_C, O_D, O_P, O_Q$, the Cochran-Armitage test for trend z^2 is a convex function of (O_{c_2}, O_{d_2}).*

Proof. See Appendix.

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 [6]. Thus, from Lemmas 1 and 2, we have the following theorem.

Theorem 1. *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* $O_{d_2} = u_{d_2}$*. The maximum value of Cochran-Armitage test* z^2 *is attained when* (O_{c_2}, O_{d_2}) *takes one of the four values in* $\{(l_{c_2}, l_{d_2}), (l_{c_2}, u_{d_2}), (u_{c_2}, l_{d_2}), (u_{c_2}, u_{d_2})\}$.

Therefore, we can develop an upper bound of the two-locus test if we identify the range of O_{c_2} and O_{d_2}. For example, suppose that the value of vector $(O_A, O_B, O_C, O_D, O_P, O_Q)$ is $(6, 10, 10, 6, 7, 6)$. In Figure 2, we plot function $\chi^2_1(X_iX_j, Y)$. The blue stars represent the values of $\chi^2_1(X_iX_j, Y)$ when O_{c_2} takes different values. Clearly, $\chi^2_1(X_iX_j, Y)$ is a convex function of O_{c_2} , and its upper bound is determined by the two end points of the range of O_{c_2}. Since O_{c_2} is always less than O_C, in this example, the default range of O_{c_2} is $[0, O_C] = [0, 10]$. Typically, the actual range of O_{c_2} is tighter, as indicated by the red dotted lines, which leads to a tighter upper bound of the test value. In the next section, by examining the contingency tables, we derive a set of constraints that determine the range of O_{c_2} and O_{d_2}.

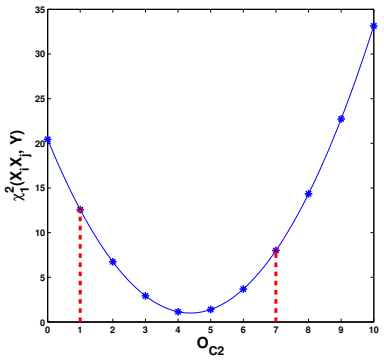

Fig. 2: Convexity Example

4 Constraints on Observed Values

$$\begin{cases} O_{a_1} + O_{a_2} = O_A \\ O_{b_1} + O_{b_2} = O_B \\ O_{c_1} + O_{c_2} = O_C \\ O_{d_1} + O_{d_2} = O_D \\ O_{a_1} + O_{c_1} = O_S \\ O_{a_2} + O_{c_2} = O_P \\ O_{b_1} + O_{d_1} = O_T \\ O_{b_2} + O_{d_2} = O_Q \end{cases} \Longrightarrow \begin{pmatrix} 1&1&0&0&0&0&0&0 \\ 0&0&0&0&1&1&0&0 \\ 0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&1&1 \\ 0&1&0&0&0&1&0&0 \\ 0&0&0&1&0&0&0&1 \end{pmatrix} \begin{pmatrix} O_{a_1} \\ O_{a_2} \\ O_{b_1} \\ O_{b_2} \\ O_{c_1} \\ O_{c_2} \\ O_{d_1} \\ O_{d_2} \end{pmatrix} = \begin{pmatrix} O_A \\ O_C \\ O_B \\ O_D \\ O_P \\ O_Q \end{pmatrix}$$

Fig. 3. Linear equation system derived from contingency tables

From the contingency tables shown in Table 1, we can develop a set of equations, as shown in Figure 3 at the left side of the arrow sign. Although there are 8 equations, the rank of the linear equation system is 6. We choose 6 linear equations to form a full rank system. The matrix multiplication form of these 6 equations is shown in Figure 3 at the right side of the arrow sign. The reason for choosing the 6 equations is two-fold. First, these 6 equations can be used to derive the range of O_{c_2} and O_{d_2}. Second, the values of

$$\begin{pmatrix} O_{a_1} \\ O_{a_2} \\ O_{c_1} \\ O_{c_2} \end{pmatrix} = \begin{pmatrix} O_A - O_P \\ O_P \\ O_C \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix} O_{c_2} \text{ , and } \begin{pmatrix} O_{b_1} \\ O_{b_2} \\ O_{d_1} \\ O_{d_2} \end{pmatrix} = \begin{pmatrix} O_B - O_Q \\ O_Q \\ O_D \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix} O_{d_2}.$$

Fig. 4. Relations between observed values in the contingency table of two-locus test

O_A, O_B, O_C, O_D are determined by the single-locus contingency table in Table 1(a). The remaining two values, O_P and O_Q, only depend on the SNP-pair's genotype. It enables us to index the SNP-pairs by their (O_P, O_Q) values to effectively apply the upper bound. This will become clear when we present the algorithm in Section 5.

From these 6 equations, we obtain the relationships between the observed values shown in Figure 4. Since all observed values in the contingency table must be greater or equal to 0, the ranges of O_{c_2} and O_{d_2} are stated in Theorem 2.

Theorem 2. *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*

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

Given $O_A, O_B, O_C, O_D, O_P, O_Q$, the values of $O_{a_1}, O_{a_2}, O_{c_1}$ are determined by O_{c_2}, the values of $O_{b_1}, O_{b_2}, O_{d_1}$ are determined by O_{d_2}. So all values in the contingency table for two-locus test in Table 1(c) depend only on O_{c_2} and O_{d_2}. The maximum value, $ub(\mathscr{T}(X_iX_j, Y))$, is attained when O_{c_2} and O_{d_2} take the boundary values shown in Theorems 1 and 2[2]. Continuing with the example in Figure 2, the value of $(O_A, O_B, O_C, O_D, O_P, O_Q)$ is $(6, 10, 10, 6, 7, 6)$. From Theorem 2, the range of O_{c_2} is $[1, 7]$, as indicated by the red lines. The upper bound of $\chi^2_1(X_iX_j, Y)$ is reached when $O_{c_2} = 1$.

Note that the upper bound value only depends on $O_A, O_B, O_C, O_D, O_P, O_Q$. This property allows us to group and index SNP-pairs by their genotypes so that the upper bound can effectively estimated and applied to prune the search space.

5 Applying the Upper Bound

Theorems 1 and 2 show that the upper bound value of the two-locus test $\mathscr{T}(X_iX_j, Y)$ (for any one of the four tests discussed in Section 3) is determined by the values of $O_A, O_B, O_C, O_D, O_P, O_Q$. As shown in Table 1, these values only depend on the contingency table for the single-locus test $\mathscr{T}(X_i, Y)$ and the contingency table for the SNP-pair (X_iX_j)'s genotype. This allows us to group the SNP-pairs and index them by their genotypes. The idea of building such indexing structure has also been explored in [34, 35]. For self-containment, in this section, we first discuss how to apply the upper

[2] For entropy-based statistic, so far we have focused on the joint entropy $-H(X_iX_jY)$. Note that, given the values of $O_A, O_B, O_C, O_D, O_P, O_Q$, the upper bound for the mutual information $I(X_iX_j, Y)$ can also be easily derived.

bound to find the significant SNP-pairs. Then we show that a similar idea can be used to find the critical values $\mathscr{T}_\alpha$ using permutation test.

For every X_i $(1 \leq i \leq N)$, let $AP(X_i) = \{(X_iX_j)|i+1 \leq j \leq N\}$ be the SNP-pairs with X_i being the SNP of lower index value. We can index the SNP-pairs in $AP(X_i)$ by their (O_P, O_Q) values in a 2D array, referred to as $Array(X_i)$. Note that O_P is the number of 1's in X_j when X_i takes value 0. O_Q is the number of 1's in X_j when X_i takes value 1.

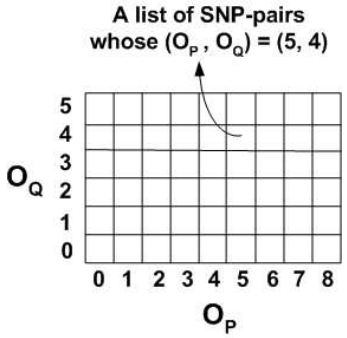

Fig. 5: Indexing SNP-pairs

For example, suppose that there are 13 individuals in the dataset. SNP X_i consists of 8 0's and 5 1's. Thus for the SNP-pairs in $AP(X_i)$, the possible values of O_P are $\{0, 1, 2, \cdots, 8\}$. The possible values of O_Q are $\{0, 1, 2, \cdots, 5\}$. Figure 5 shows the 6×9 array, $Array(X_i)$, whose entries represent the possible values of (O_P, O_Q) for the SNP-pairs (X_iX_j) in $AP(X_i)$. Each entry of the array is a pointer to the SNP-pairs (X_iX_j) having the corresponding (O_P, O_Q) values. For example, all SNP-pairs in $AP(X_i)$ whose (O_P, O_Q) value is (5,4) are indexed by the entry (5,4) in Figure 5.

It is obvious that for any SNP-pair $(X_iX_j) \in AP(X_i)$, if the upper bound value of the two-locus test is less than the critical value, i.e., $ub(\mathscr{T}(X_iX_j, Y)) < \mathscr{T}_\alpha$, then this SNP-pair cannot be significant since its actual test value will also be less than the threshold. Only the SNP-pairs whose upper bound values are greater than the threshold need to be evaluated for their test values. We refer to such SNP-pairs as *candidates*.

Recall that from Theorems 1 and 2, the upper bound of two-locus test value is a constant for given $O_A, O_B, O_C, O_D, O_P, O_Q$. Given SNP X_i and phenotype Y, the values of O_A, O_B, O_C, O_D are fixed. For SNP-pairs $(X_iX_j) \in AP(X_i)$, once we index them by their (O_P, O_Q) values as shown in Figure 5, we can identify the candidate SNP-pairs by accessing the indexing structure: For each entry of the indexing structure, we calculate the upper bound value. If the upper bound value is greater than or equal to the critical value $\mathscr{T}_\alpha$, then all SNP-pairs indexed by this entry are candidates and subject to two-locus tests. The SNP-pairs whose upper bound values are less than the critical value are pruned without any additional test.

Suppose that there are m 1's and $(M-m)$ 0's in SNP X_i. The maximum size of the indexing structure $Array(X_i)$ is $m(M-m)$. Usually, the number of individuals M is much smaller than the number of SNPs N. Therefore, the number of entries in the indexing structure is also much smaller than N. Thus there must be a group of SNP-pairs indexed by the same entry. Since all SNP-pairs indexed by the same entry have the same upper bound value, the indexing structure enables us to calculate the upper bound value for this group of SNP-pairs together.

So far, we have discussed how to use the indexing structure and the upper bound to prune the search space to find significant SNP-pairs for a given critical value $\mathscr{T}_\alpha$. The problem of finding this critical value $\mathscr{T}_\alpha$ is much more time consuming than finding the significant SNP-pairs since it involves large scale permutation test. The indexing struc-

ture $Array(X_i)$ can be easily incorporated in the algorithm for permutation test. The key property is that the indexing structure $Array(X_i)$ is independent of the phenotype. Once $Array(X_i)$ is built, it can be reused in all permutations. Therefore, building the indexing structure $Array(X_i)$ is only a one time cost. The permutation procedure is similar to that of finding significant SNP-pairs. The only difference is that the threshold used to prune the search space is a dynamically updated critical value found by the algorithm so far. The overall procedure of our algorithm COE is similar to that in [34, 35]. An important difference is that COE utilizes the convexity of statistical tests and is applicable to all four statistics. We omit the pseudo code of the algorithm in the main body of the paper. Please refer to the Appendix for further details.

Property 1. The indexing structure $Array(X_i)$ can be applied in computing the upper bound value for all four statistical tests, i.e., chi-square test, G-test, mutual information, and trend test.

The correctness of Property 1 relies on the fact that the upper bound is always a function of $O_A, O_B, O_C, O_D, O_P, O_Q$, regardless of the choice of test. All SNP-pairs having the same (O_P, O_Q) value will always share a common upper bound. This property shows that there is no need to rebuild the indexing structure if the users want to switch between different tests. It only needs to be built once and retrieved for later use.

The time complexity of COE for permutation test is $O(N^2M + KNM^2 + CM)$, where N is the number of SNPs, M is the number of individuals, K is the number of permutations, and C is the number of candidates reported by the algorithm. Experimental results show that C is only a very small portion of all SNP-pairs. A brute force approach has time complexity $O(KN^2M)$. Note that N is the dominant factor, since $M \ll N$. The space complexity of COE is linear to the size of the dataset. The derivation of the complexity is similar to that in [34, 35] and can be found in the Appendix.

6 Experimental Results

In this section, we present extensive experimental results on evaluating the performance of the COE algorithm. COE is implemented in C++. We use COE_Chi, COE_G, COE_MI, COE_T to represent the COE implementation for the chi-square test, G-test, mutual information, and trend test respectively. The experiments are performed on a 2.4 GHz PC with 1G memory running WindowsXP system.

Dataset and Experimental Settings: The SNP dataset is extracted from a set of combined SNPs from the 140k Broad/MIT mouse dataset [32] and 10k GNF [2] mouse dataset. This merged dataset has 156,525 SNPs for 71 mouse strains. The missing values in the dataset are imputed using NPUTE [23]. The phenotypes used in the experiments are simulated binary variables which contain half cases and half controls. This is common in practice, where the numbers of cases and controls tend to be balanced. If not otherwise specified, the default settings of the experiments are as follows: #individuals = 32, #SNPs=10,000, #permutations=100. There are 62,876 unique SNPs for these 32 strains.

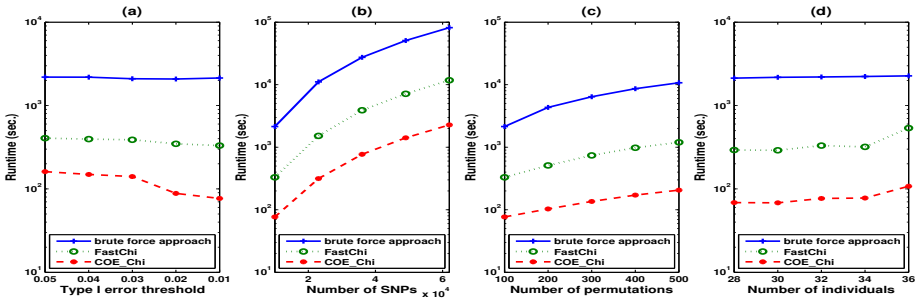

Fig. 6. Performance comparison of the brute force approach, FastChi, and COE_Chi

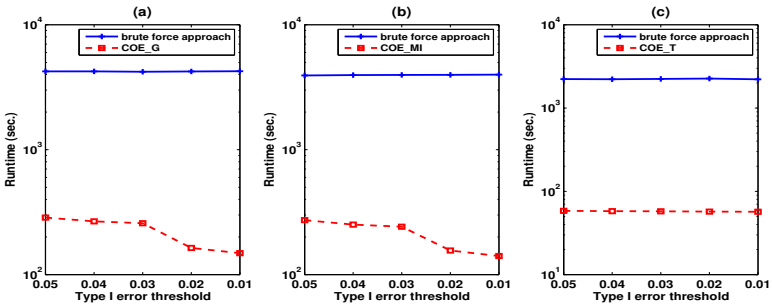

Fig. 7. Performance comparison of the brute force approach, COE_G, COE_MI, and COE_T

6.1 Performance Comparison

Figure 6 shows the runtime comparison of the brute force two-locus chi-square test, the FastChi algorithm [35], and the COE implementation of chi-square test, COE_Chi, in permutation test under various settings. Note that the runtime reported in this section are based on the complete executions of all methods including the one time cost for building the indexing structures. Figure 6(a) shows the comparison when the Type I error threshold varies. The y-axis is in logarithm scale. COE_Chi improves the efficiency of two-locus epistasis detection by one order of magnitude over FastChi (which was specifically designed for two-locus chi-square test), and two orders of magnitude over the brute force approach. Figure 6 (b), (c), and (d) demonstrate similar performance improvements of COE_Chi over the other two approaches when varying number of SNPs, number of permutations, and number of individuals respectively. This is consistent with the pruning effect of the upper bounds which will be presented later.

Figure 7(a) shows the runtime comparison between the brute force two-locus G-test and COE_G when varying the type I error threshold. The runtime of COE_G dramatically reduces as the type I error threshold decreases. COE_G is one to two orders magnitudes faster than the brute force approach. Similar performance improvement can also be observed for COE_MI and COE_T in Figures 7(b) and 7(c). Note that for these three

		FastChi	COE_Chi	COE_G	COE_MI	COE_T
α	0.05	87.59%	95.70%	95.84%	95.80%	99.90%
	0.04	87.98%	96.11%	96.23%	96.23%	99.92%
	0.03	88.12%	96.32%	96.40%	96.43%	99.93%
	0.02	89.43%	98.18%	98.31%	98.28%	99.96%
	0.01	90.03%	98.59%	98.65%	98.62%	99.98%
# SNPs	10k	90.03%	98.59%	98.65%	98.62%	99.98%
	23k	91.52%	99.08%	99.50%	99.13%	99.99%
	36k	91.39%	99.03%	99.43%	99.09%	99.99%
	49k	91.39%	99.04%	99.43%	99.09%	99.99%
	62k	91.22%	99.04%	99.43%	99.09%	99.99%
# Perm.	100	90.03%	98.59%	98.65%	98.62%	99.98%
	200	91.79%	99.03%	99.42%	99.08%	99.99%
	300	91.90%	99.04%	99.43%	99.09%	99.99%
	400	91.91%	99.04%	99.43%	99.09%	99.99%
	500	91.99%	99.04%	99.43%	99.09%	99.99%
# Indiv.	28	91.05%	98.77%	99.83%	99.06%	99.99%
	30	91.23%	98.83%	98.94%	99.06%	99.98%
	32	90.03%	98.59%	99.65%	98.62%	99.98%
	34	91.54%	98.80%	99.74%	98.84%	99.97%
	36	89.08%	97.94%	95.74%	93.55%	99.94%

Table 2. Pruning effects of FastChi and COE on four different statistics

tests, we also have similar results when varying other settings. Due to space limitation, we omit these results here.

6.2 Pruning Power of the Upper Bound

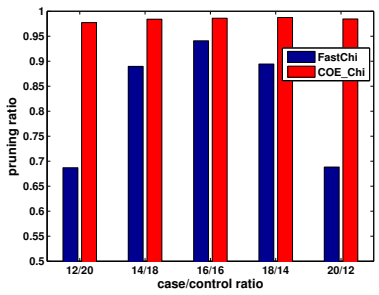

Fig. 8: FastChi v.s. COE_Chi

Table 2 shows the percentage of SNP-pairs pruned under different experimental settings for the four statistical tests. We also include the pruning ratio of FastChi in the table for comparison. From the table, most of the SNP-pairs are pruned by COE. Note that COE_Chi has more pruning power than FastChi. The upper bound used in FastChi is derived by loosening the observed values for the events in two-locus test without using the convexity property. The tighter upper bound of COE_Chi demonstrates the strength of convex optimization in finding the maximum values. In addition, the upper bound derived by applying convex optimization is not only more effective, but also more robust for unbalanced datasets.

Figure 8 shows the pruning effectiveness of FastChi and COE_Chi when the ratio of case/control varies. It is clear that the pruning power of FastChi is weakened when the case/control ratio becomes unbalanced. Therefore, FastChi is not very effective for

unbalanced case-control datasets. In contrast, COE_Chi maintains a steady pruning percentage under different case/control ratios. Thus it remains effective for the unbalanced datasets. Similar behaviors of COE are also observed in the other three statistical tests.

7 Discussion

Genome-wide epistasis detection is computationally demanding due to the large number of SNPs. As a golden standard for proper family-wise error controlling, the permutation test dramatically increases the computation burden. In this paper, we present a general approach COE that support genome-wide disease association study with a wide range of statistics composing of convex terms. We use four commonly used statistics as prototypes: chi-square test, G-test, entropy-based test, and Cochran-Armitage trend test. COE guarantees optimal solution and performs two orders of magnitude faster than brute force approaches.

The performance gain is attributed to two main contributions of COE. The first is a tight upper bound estimated using convex optimization. It has much higher pruning power than any upper bounds used in previous methods such as FastChi. As a result, COE_Chi is an order of magnitude faster than FastChi. Moreover, COE serves as a general platform for two-locus epistasis detection, which eliminates the need of designing specific pruning methods for different statistical tests. Recall that any observed value in a two-locus test is a function of O_{c_2} and O_{d_2} for given $O_A, O_B, O_C, O_D, O_P, O_Q$. Let $x = O_{c_2}$ and $y = O_{d_2}$. A wide spectrum of functions of x and y are convex [6], which include all linear and affine functions on x and/or y, exponential terms e^{ax} ($a \in \mathbb{R}$), powers x^a ($a \geq 1$ or $a \leq 0$), negative logarithm $-\log x$, maximum $\max\{x, y\}$. In addition, many operations preserve convexity. For example, if $f(x, y)$ is a convex function, and $g(x, y)$ is an affine mapping, then $f(g(x, y))$ is also a convex function. Please refer to [6] for further details.

The second source of performance improvement is from indexing SNP-pairs by their genotypes. Applying this indexing structure, we can compute a common upper bound value for each group. The indexing structure is independent of the phenotype permutations and the choice of statistical test . We can eliminate redundant computation in permutation test and provide the flexibility of supporting multiple statistical tests on the fly.

In this paper, we focus on binary SNPs and case-control phenotypes. The principle is also applicable to the heterozygous case, where SNPs are encoded using $\{0, 1, 2\}$, and to evaluate quantitative phenotypes, where phenotypes are continuous variables. We will investigate these two cases in our future work.

References

1. *http://www.fnih.org/GAIN2/home_new.shtml*.
2. *http://www.gnf.org/*.
3. *http://www.jax.org/*.
4. D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

5. S. Bohringer, C. Hardt, B. Miterski, A. Steland, and J. T. Epplen. Multilocus statistics to uncover epistasis and heterogeneity in complex diseases: revisiting a set of multiple sclerosis data. *European Journal of Human Genetics*, 11:573–584, 2003.
6. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
7. O. Carlborg, L. Andersson, and B. Kinghom. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, 155:2003–2010, 2000.
8. C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429:446–452, 2004.
9. P. B. Chi and et al. 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, 2006.
10. H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.
11. R. W. Doerge. Multifactorial genetics: Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3:43–52, 2002.
12. C. Dong and et al. Exploration of genegene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16:229–235, 2008.
13. C. Erlichman and D. J. Sargent. New treatment options for colorectal cancer. *N. Engl. J. Med.*, 351:391–392, 2004.
14. D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet.*, 2: e157, 2006.
15. E. Halperin, G. Kimmel, and R. Shamir. Tag snp selection in genotype data for maximizing snp prediction accuracy. In *Proc. ISMB*, 2005.
16. A. Herbert and et al. A common genetic variant is associated with adult and childhood obesity. *Science*, 312:279–284, 2006.
17. J. Hoh and J. Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4:701–709, 2003.
18. I. Kirman, E. H. Huang, and R. L. Whelan. B cell response to tumor antigens is associated with depletion of b progenitors in murine colocarcinoma. *Surgery*, 135:313–318, 2004.
19. M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, 11:458–470, 2001.
20. K. Ozaki and et al. Functional snps in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.*, 32:650–654, 2002.
21. M. Pagano and K. Gauvreau. *Principles of Biostatistics*. Pacific Grove, CA: Duxbury Press, 2000.
22. M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, 69:138–147, 2001.
23. A. Roberts, L. McMillan, W. Wang, J. Parker, I. Rusyn, and D. Threadgill. Inferring missing genotypes in large snp panels using fast nearest-neighbor searches over sliding windows. In *Proc. ISMB*, 2007.
24. A. Roses. The genome era begins. *Nat. Genet.*, 33(Supp2):217, 2003.
25. C. A. Ruivenkamp, T. Csikos, A. M. Klous, T. van Wezel, and P. Demant. Five new mouse susceptibility to colon cancer loci, scc11-scc15. *Oncogene*, 22:7258–7260, 2003.
26. R. Saxena and et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316:1331–1336, 2007.
27. A. Scuteri and et al. Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS Genet.*, 3(7), 2007.

28. P. Sebastiani, R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane, and M. F. Ramoni. Minimal haplotype tagging. *Proc. Natl. Acad. Sci. USA*, 100(17):9900–9905, 2003.
29. D. Segr, A. DeLuna, G. M. Church, and R. Kishony. Modular epistasis in yeast metabolism. *Nat. Genet.*, 37:77–83, 2005.
30. J. Storey, J. Akey, and L. Kruglyak. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 8: e267, 2005.
31. D. C. Thomas. *Statistical methods in genetic epidemiology*. Oxford Univeristy Press, Oxford, 2004.
32. C. M. Wade and M. J. Daly. Genetic variation in laboratory mice. *Nat. Genet.*, 37:1175–1180, 2005.
33. M. N. Weedon and et al. A common variant of hmga2 is associated with adult and childhood height in the general population. *Nat. Genet.*, 39:1245–1250, 2007.
34. X. Zhang, F. Zou, and W. Wang. Fastanova: an efficient algorithm for genome-wide association study. In *KDD*, 2008.
35. X. Zhang, F. Zou, and W. Wang. FastChi: an efficient algorithm for analyzing gene-gene interactions. In *PSB*, 2009.
36. J. Zhao, E. Boerwinkle, and M. Xiong. An entropy-based statistic for genomewide association studies. *Am J Hum Genet*, 77:27–40, 2005.

Appendix

Proof of Lemma 1 and Lemma 2

Proof. We first show that $\chi_1^2(X_iX_j, Y)$ is a convex function of O_{c_2}. Recall that

$$\chi_1^2(X_iX_j, Y) = \sum_{event \in \{a_1, a_2, c_1, c_2\}} \frac{(O_{event} - E_{event})^2}{E_{event}}.$$

For fixed $O_A, O_B, O_C, O_D, O_P, O_Q$, we know that the expected values of the four events are constants:

$$\begin{cases} E_{a_1} = \dfrac{O_S(O_A + O_B)}{M} = \dfrac{(O_A + O_C - O_P)(O_A + O_B)}{M} \\ E_{a_2} = \dfrac{O_P(O_A + O_B)}{M} \\ E_{c_1} = \dfrac{O_S(O_C + O_D)}{M} = \dfrac{(O_A + O_C - O_P)(O_C + O_D)}{M} \\ E_{c_2} = \dfrac{O_P(O_C + O_D)}{M} \end{cases}$$

From the relations between the observed values of the events in two-locus test (as shown in Figure 4), we have that $O_{a_1}, O_{a_2}, O_{c_1}$ are linear functions of O_{c_2}[3]. So $\chi_1^2(X_iX_j, Y)$ is a positive quadratic function of O_{c_2}. Thus $\chi_1^2(X_iX_j, Y)$ is a convex function of O_{c_2}.

Next, we show that

$$G_1(X_iX_j, Y) = \sum_{event \in \{a_1, a_2, c_1, c_2\}} O_{event} \cdot \ln \frac{O_{event}}{E_{event}}$$

is a convex function of O_{c_2}. From previous result, for fixed $O_A, O_B, O_C, O_D, O_P, O_Q$, the expected values of the four events $\{a_1, a_2, c_1, c_2\}$ are constants, and $O_{a_1}, O_{a_2}, O_{c_1}$ are linear functions of O_{c_2}. Thus $G_1(X_iX_j, Y)$ is a function of O_{c_2}. To prove the convexity of $G_1(X_iX_j, Y)$, it suffices to show that the second derivative $\nabla^2 G_1(X_iX_j, Y) = \dfrac{\partial^2 G_1(X_iX_j, Y)}{\partial O_{c_2}^2}$ is nonnegative. We show this is the case for the component of event a_2:

$$\nabla^2(O_{a_2} \cdot \ln \frac{O_{a_2}}{E_{a_2}}) = \nabla^2((O_P - O_{c_2}) \cdot \ln \frac{O_P - O_{c_2}}{E_{a_2}}) = \frac{1}{O_P - O_{c_2}} \geq 0.$$

Similarly, we can prove that the second derivative of other components are nonnegative. Therefore, $G_1(X_iX_j, Y)$ is a convex function of O_{c_2}.

Following the similar idea, i.e., by showing the second derivative of $-H(X_iX_jY)$ is nonnegative, we can prove that $-H_1(X_iX_jY)$ is a convex function of O_{c_2}.

Thus we have shown the $\mathscr{T}_1(X_iX_jY)$ is a convex function of O_{c_2}. The convexity $\mathscr{T}_2(X_iX_jY)$ can be proven in a similar way.

[3] Note that, although these relations are presented after shown the convexity of the statistics, it is easy to see that the derived relations are independent of whether the statistics are convex.

We now prove that the Cochran-Armitage trend test is a convex function of (O_{c_2}, O_{d_2}). Observe that the O_{c_1} is a linear function of O_{c_2}, and O_{d_1} is a linear function of O_{d_2}. The values of p, s_i ($i \in \{1, 2, 3, 4\}$), and $\bar{s}$ are fixed. Thus the trend statistic z^2 is a quadratic function of the two variables (O_{c_2}, O_{d_2}). This completes the proof. □

Pseudo code of COE for permutation test

Algorithm 1: COE for permutation test

Input: SNPs $X' = \{X_1, X_2, \cdots, X_N\}$, phenotype permutations $Y' = \{Y_1, Y_2, \cdots, Y_K\}$, and the Type I error α.
Output: the critical value $\mathscr{T}_\alpha$.

```
1  Tlist ← αK dummy phenotype permutations with test value 0 ;
2  T_α = 0;
3  for every X_i ∈ X', do
4      index (X_iX_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_iX_j) ∈ Cand(X_i, Y_k) do
8              if T(X_iX_j, Y_k) ≥ T_α then
9                  update Tlist;
10                 T_α = the smallest test value in Tlist;
11             end
12         end
13     end
14 end
15 return T_α.
```

Algorithm 1 describes our COE algorithm for finding critical values in two-locus epistasis detection using permutation test. The algorithm for finding significant SNP-pairs is similar. The overall process is similar to that in [34, 35]. The goal is to find the critical value $\mathscr{T}_\alpha$, which is the αK-th largest value in $\{\mathscr{T}_{Y_k} | Y_k \in Y'\}$. We use $Tlist$ to keep the αK phenotype permutations having the largest test values found by the algorithm so far. Initially, $Tlist$ contains αK dummy permutations with test values 0. The smallest test value in $Tlist$, initially 0, is used as the threshold to prune the SNP-pairs. For each X_i, the algorithm first builds the indexing structure $Array(X_i)$ for the SNP-pairs $(X_iX_j) \in AP(X_i)$. Then it accesses $Array(X_i)$ to find the set of candidates $Cand(X_i, Y_k)$ for every phenotype permutation. Two-locus tests are performed on these candidates to get their test values. If a candidate's test value is greater than the current threshold, then $Tlist$ is updated: If the candidate's phenotype Y_k is not in the $Tlist$, then the phenotype in $Tlist$ having the smallest test value is replaced by Y_k. If

the candidate's phenotype Y_k is already in $Tlist$, we only need to update its corresponding test value to be the maximum value found for the phenotype so far. The threshold is also updated to be the smallest test value in $Tlist$.

Time complexity: For each X_i, the algorithm needs to index (X_iX_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 $O(KNM^2)$. Let $C = \sum_{i,k} |Cand(X_i, Y_k)|$ represent the total number of candidates. The overall time complexity of our algorithm is $O(N(N-1)M/2) + O(KNM^2) + O(CM) = O(N^2M + KNM^2 + CM)$.

Space complexity: The total number of variables in the dataset, including the SNPs and the phenotype permutations, is $N + K$. The maximum space of the indexing structure $Array(X_i)$ is $O(M^2 + N)$. For each SNP X_i, our algorithm 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 COE is $O((N + K)M) + O(M^2 + N) = O((N + K + M)M + N)$. Since $M \ll N$, the space complexity is linear to the dataset size.