JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 6, 2015

Pp. 563–576
DOI: 10.1089/cmb.2014.0163

Gene–Gene Interactions Detection Using a Two-stage Model

ZHANYONG WANG,[1] JAE HOON SUL,[2] SAGI SNIR,[3] JOSE A. LOZANO,[4] and ELEAZAR ESKIN[1]

ABSTRACT

Genome-wide association studies (GWAS) have discovered numerous loci involved in genetic traits. Virtually all studies have reported associations between individual single nucleotide polymorphisms (SNPs) and traits. However, it is likely that complex traits are influenced by interaction of multiple SNPs. One approach to detect interactions of SNPs is the brute force approach which performs a pairwise association test between a trait and each pair of SNPs. The brute force approach is often computationally infeasible because of the large number of SNPs collected in current GWAS studies. We propose a two-stage model, Threshold-based Efficient Pairwise Association Approach (TEPAA), to reduce the number of tests needed while maintaining almost identical power to the brute force approach. In the first stage, our method performs the single marker test on all SNPs and selects a subset of SNPs that achieve a certain significance threshold. In the second stage, we perform a pairwise association test between traits and pairs of the SNPs selected from the first stage. The key insight of our approach is that we derive the joint distribution between the association statistics of a single SNP and the association statistics of pairs of SNPs. This joint distribution allows us to provide guarantees that the statistical power of our approach will closely approximate the brute force approach. We applied our approach to the Northern Finland Birth Cohort data and achieved 63 times speedup while maintaining 99% of the power of the brute force approach.

Key words: epistasis, gene–gene interaction, GWAS.

1. INTRODUCTION

Genome-wide association studies (GWAS) attempt to discover genetic variation associated with disease traits. To perform GWAS, studies collect genetic variation of individuals and their disease status or disease related traits. GWAS studies typically collect single nucleotide polymorphisms (SNPs) because technologies allow for very cost efficient collections of SNPs. Current GWAS studies collect about a

[1]Computer Science Department, University of California Los Angeles, Los Angeles, California.
[2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.
[3]Institute of Evolution, Department of Evolutionary and Environmental Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel.
[4]Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country, Donostia Spain.

million SNPs in thousands of individuals. The standard approach for identifying associations between SNPs and traits is that for each SNP, we compare the average trait value of individuals who have one allele of an SNP and that of individuals who have the other allele of the SNP. If the difference between the two average trait values is above a certain threshold, we declare that the SNP is significantly associated with the trait. We refer to computing the difference in the average trait values for each SNP as the ''single marker test,'' and it has successfully identified many individual SNPs associated with several complex diseases (Corder et al., 1993; Bertina et al., 1994; Altshuler et al., 2000; Saxena et al., 2007; Consortium, 2007). Since SNPs are so prevalent in the genome, they are likely to be correlated with other genetic variations.

Current studies on certain complex diseases have also suggested that some SNPs influence diseases through interactions (Williams et al., 2000; Brem et al., 2005; Yanchina et al., 2004). In an extreme scenario, two SNPs may not have any effect on a disease independently, but they may affect the disease when both are present. To detect an interaction of SNPs, one needs to consider the association between a trait and a pair of SNPs. One approach to finding such associations is to divide individuals into two groups: one group of individuals who have a certain combination of alleles for a pair of SNPs and the other group of individuals who have different combinations of alleles for the pair of SNPs. We then compute the difference in the average trait value between the two groups to determine whether the pair of SNPs is significantly associated with the trait. Finding an association between a trait and a pair of SNPs is called the ''pairwise association test,'' and recently, several different methods have been proposed for pairwise association tests (Evans et al., 2006; Zhang et al., 2010; Prabhu and Pe'er, 2012; Yang et al., 2009; Millstein et al., 2006; Ljungberg et al., 2004).

One major challenge in discovering pairs of SNPs associated with a trait is that it requires enormous computation. One needs to compute associations between a trait and $4 \times \binom{M}{2}$ pairs of SNPs, where M is the number of SNPs available for testing. When M is close to one million as in current GWAS, an exhaustive pairwise search that goes through all pairs of SNPs considers 2,000 billion pairs of SNPs, which is a computationally challenging task. As the number of SNPs in GWAS keeps increasing with the improvement of technologies to collect SNPs, the exhaustive search becomes even more computationally infeasible.

In this article, we present a threshold-based efficient pairwise association approach (TEPAA) for detecting associations between traits and pairs of SNPs using a two-stage model. In the first stage, our method performs the single marker test on all individual SNPs and selects a subset of SNPs that exceed a certain SNP-specific predetermined significance threshold for further consideration. In the second stage, individual SNPs that are selected in the first stage are paired with each other, and we perform the pairwise association test on those pairs. In this method, there exists a trade-off between the probability of detecting a pair of SNPs associated with a trait and the computational burden. Intuitively, statistical power increases as we include more SNPs in the second stage, which means a higher computational burden. The first stage thresholds determine this trade-off. We derive the analytical power of our method, which allows us to control this trade-off by choosing appropriate thresholds. The key insight of our approach is that we derive the joint distribution between the association statistics of single SNP and the association statistics of pairs of SNPs. This joint distribution allows us to provide guarantees that the statistical power of our approach will closely approximate the brute force approach. We can accurately compute the analytical power of our two-stage model and compare it to the power of the brute force approach. Hence, we are able to choose as few SNPs as possible in the first stage while achieving almost the same power as the brute force approach. We demonstrate the utility of TEPAA applied to the Northern Finland Birth Cohort (Rantakallio, 1969; Jarvelin et al., 2004).

Several methods have recently been developed to increase the efficiency of detecting gene–gene interactions. Most of these methods are only applicable to case/control data. These methods include TEAM (Zhang et al., 2010), which uses a dynamic programming approach to identify significant pairs, SIXPAC (Prabhu and Pe'er, 2012), which utilizes a novel randomization technique, BOOST (Wan et al., 2010), which utilizes a Boolean representation of data and a screening stage to filter out most nonsignificant SNP interactions, and RAPID (Brinza et al., 2010), which utilizes geometric properties of the χ^2 statistic to identify significant pairs. However, none of these methods are applicable to quantitative traits.

The only existing method that is feasible on a GWAS dataset to detect SNP pairs associated with quantitative traits is FastEpistasis (Schpbach et al., 2010). FastEpistasis is a brute-force approach that conducts pairwise associations for all pairs of SNPs, or SNP pairs specified by users. The advantage of FastEpistasis is that their method utilizes architectures with multiple cores. In essence, FastEpistasis is an extremely efficient implementation of the brute force approach.

2. RESULTS

2.1. Overview of the two-stage model TEPAA

We present a two-stage model, TEPAA, for detecting associations between traits and pairs of SNPs. In the first stage, the association statistics for all SNPs are computed. Any SNP that has a statistic higher than a predetermined SNP-specific threshold advances to the second stage in which all pairs of these SNPs are evaluated. The specific first-stage thresholds are important in determining the tradeoff between statistical power and computational cost; they control the number of SNPs to be selected in the first stage. For a truly associated pair of SNPs to be identified using our approach, both SNPs must advance to the second round and thus must have association statistics higher than the thresholds. Clearly, the more stringent the thresholds are, the smaller the number of SNPs in the second stage and the smaller the number of pairs of SNPs that must be evaluated. More stringent thresholds speed up this method. On the other hand, more stringent thresholds increase the chance that at least one of the pair of truly associated SNPs will not be more significant than its first-stage threshold, and the pair will not be identified by the method. Hence, there is a trade-off between power and cost, which is determined by the first-stage thresholds. The thresholds themselves are influenced by the desired statistical power versus computational time tradeoff and also the minor allele frequency as we show below.

Our method chooses the first-stage thresholds T_A and T_B such that the two-stage model loses only a small amount of power but increases computational efficiency dramatically compared to the exhaustive search. To find such thresholds, we first derive the analytical power and cost of both the brute force approach and the two-stage model. This analysis allows us to choose the threshold that yields the desired power and cost, and hence it allows us to control the trade-off between the two. To derive the analytical power of our two-stage model, we use the framework of multivariate normal distribution (MVN) to model the association statistics (Han et al., 2009; Kostem and Eskin, 2013; Kostem et al., 2011), where the statistics follow standard normal distributions under the null hypothesis. We use an MVN to approximate the joint distribution between the association statistics of single SNP S_A and S_B, and the association statistic of pairs of SNPs S_{AB}. The noncentrality parameters (NCPs) of statistics are considered to be the mean vector in the MVN, and correlations among statistics are considered as the covariance matrix in the MVN. The NCPs and correlations can be calculated from the data, and thus we obtained all the parameters of the MVN.

In order to demonstrate the power loss of TEPAA, we consider the scenario in which we have two SNPs, A and B, and are also considering their interaction resulting in three statistics (S_A, S_B, S_{AB}). For the simplicity of the example, we show how to compute the power loss when there are two dimensions (S_A, S_{AB}), under the assumption that $S_A \geq 0$ and $S_{AB} \geq 0$. We project the MVN surface to the (S_A, S_{AB}) plane as shown in Figure 1. The probability density function of the MVN is represented by the contour lines in the plane. The center of the contour lines corresponds to the mean vector of the MVN. For the brute force approach, we declare an SNP pair AB to be significant whenever we see the statistic S_{AB} is greater than a threshold T_2. So the area on the right of the vertical line $S_{AB} = T_2$ in Figure 1 is the significant area of the brute force approach. In our method TEPAA, we first examine the statistic S_A. Only when $S_A \geq T_A$ will we further compute S_{AB} to see whether it is greater than T_2. T_2 is obtained using the desired overall significance level for discovering gene-by-gene interactions, which in our case is $T_2 = -\Phi^{-1}(\alpha/2)$, where $\alpha = 10^{-12}$ (Prabhu and Pe'er, 2012). So, the Area 1 in Figure 1 corresponds to the power of TEPAA. Compared to

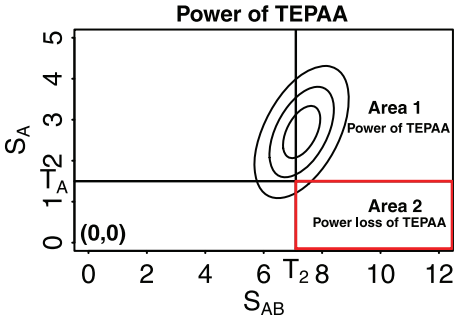

FIG. 1. The power loss region of the threshold-based efficient pairwise association approach (TEPAA). The contour lines represent the probability density function of the multivariate normal distribution (MVN). The area surrounded by the red rectangle corresponds to the power loss region.

brute force approach, if a pair of SNP AB has $S_{AB} \geq T_2$ but $S_A < T_A$, our method will not be able to detect it. So Area 2 in Figure 1 corresponds to the power loss of TEPAA compared to the brute force approach. The details of the analysis are discussed in sections 3.3 and 3.4.

From our analysis, we observe that the thresholds that control the power loss of the two-stage approach depend on the minor allele frequency (MAF) of the SNPs. In particular, more common SNPs can be filtered out with less significant thresholds than rare SNPs. In order to efficiently implement TEPAA using MAF-dependent thresholds for each pair, we group the SNPs into bins based on their MAFs to apply the correct thresholds to each possible pair. After disregarding rare variants with $MAF < 0.05$, we categorize all common SNPs into nine bins according to their MAF, with step size 0.05. Each pair of SNPs would have two thresholds, one for each SNP in the first stage. In total, we have $\binom{9}{2} + 9$ categories of SNP pairs. We precompute the first-stage thresholds for each combination of two MAFs in order to achieve 1% power loss, while achieving high cost savings. We sort the SNPs within each bin by their association statistics and use binary search to rapidly obtain the set of SNPs above a single threshold to efficiently implement the first stage of our method.

2.2. *Application of TEPAA to the NFBC data*

We applied TEPAA to the Northern Finland Birth Cohort (NFBC) data to demonstrate the utility of our two-stage model and the cost savings on real data. The Northern Finland Birth Cohort data contains 5,326 individuals, and 331,476 SNPs are genotyped. The histogram of all SNPs' MAFs is shown in Figure 2. As described in detail in section 3.5, we categorize all common SNPs into nine bins according to their MAFs. The number of SNP pairs in each category is shown in Table 1. The first-stage thresholds of TEPAA are precomputed for each category in order to have the power loss at 1% using the methods described in section 3.5. The cost saving for each category is summarized in Tables 2. Based on Tables 1 and 2, the estimated overall cost savings is 63.2 times, which is the ratio between total number of pairwise association tests in brute force approach and that of TEPAA.

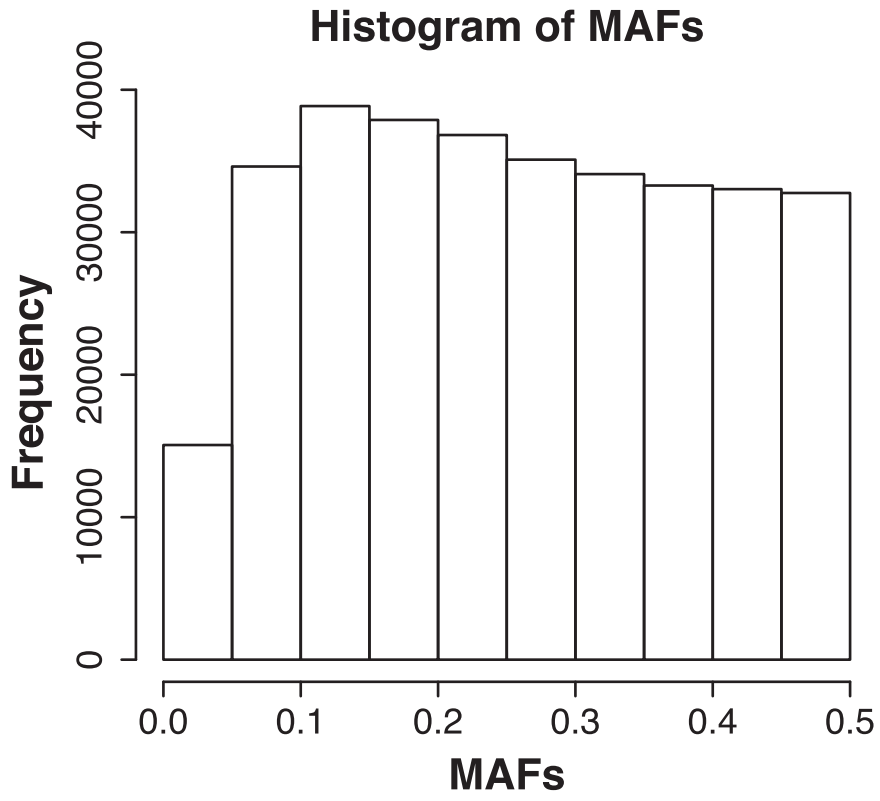

FIG. 2. The distribution of all single nucleotide polymorphisms (SNPs'), minor allele frequencies (MAFs).

Table 1. The Number of SNP Pairs in Each Category

		MAF of SNP B									
		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
MAF of SNP A	0.05	1.13	5.2	5.86	5.70	5.54	5.30	5.13	5.00	4.97	4.94
	0.1	—	5.97	13.45	13.07	12.72	12.16	11.78	11.48	11.40	11.34
	0.15	—	—	7.57	14.71	14.32	13.69	13.26	12.93	12.84	12.77
	0.2	—	—	—	7.15	13.91	13.31	12.89	12.56	12.48	12.41
	0.25	—	—	—	—	6.77	12.95	12.54	12.22	12.14	12.07
	0.3	—	—	—	—	—	6.19	11.99	11.69	11.61	11.55
	0.35	—	—	—	—	—	—	5.81	11.32	11.25	11.18
	0.4	—	—	—	—	—	—	—	5.52	10.96	10.90
	0.45	—	—	—	—	—	—	—	—	5.44	10.83
	0.5	—	—	—	—	—	—	—	—	—	5.38

Numbers are shown in factor of 100 million. MAF, minor allele frequency; SNP, single nucleotide polymorphism.

For all SNPs in each bin, we calculate the association statistics and sort the SNPs in descending order of their statistics. We perform our analysis using the dominant model, which is standard for analysis of epistatic interactions. We note that the basic approach of TEPAA can be extended to other models, such as the recessive model or additive model as well.

We compare the performance of the brute force approach and TEPAA when detecting the SNP pairs associated with the phenotype ''CRP'' (C-reactive protein) on a machine with 2.3 GHz AMD Opteron Processor. Since it is impractical to run the brute force approach on the whole chromosome, the CPU time of the brute force approach is estimated from one single chromosome by scaling, which is estimated to be 1,542 hours for phenotype ''CRP.'' The CPU time of TEPAA is 24.5 hours for the same phenotype. We achieved 62.9 times of cost saving, which verifies our analysis of the cost savings of TEPAA when achieving 1% of power loss. However, both the brute-force approach and two-stage model report no significant SNP interactions under the significance threshold 10^{-12}. This is understandable since this data set contains only 5,326 individuals. In the next section, we show that the brute force approach and TEPAA have similar power when there exist significant SNP interactions.

2.3. TEPAA controls power loss in simulated data

To demonstrate that TEPAA has only 1% power loss using the precomputed first stage thresholds, we perform simulations where we implant a significant SNP–SNP interaction to the NFBC data and then detect the SNP pair using TEPAA.

Table 2. The Threshold for SNP A/SNP B and Cost Savings in Various Combinations of MAFs to Achieve Power Loss of 1%

		MAF of SNP B								
		0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
MAF of SNP A	0.1	34/34/8	8/50/25	7/58/25	5/62/32	2/76/66	0.82/84/145	0.26/79/487	0.10/84/1190	0.02/90/5555
	0.15	—	14/14/51	3/24/139	3/31/107	2/46/108	1/58/172	0.35/54/529	0.13/62/1241	0.03/69/4830
	0.2	—	—	5/5/400	2/9/556	2/16/312	1/21/476	0.47/31/686	0.19/58/907	0.05/69/2899
	0.25	—	—	—	3/3/1100	2/5/1000	1/7/1429	1/16/625	0.26/21/1831	0.10/42/2380
	0.3	—	—	—	—	1/1/1e5	1/3/3333	1/4/2500	0.62/12/1344	0.13/16/4807
	0.35	—	—	—	—	—	0.6/0.6/2.7e4	0.5/1/2e4	0.1/2/5e4	0.03/8/4e4
	0.4	—	—	—	—	—	—	0.3/0.3/1.1e5	0.1/0.6/1.6e5	0.1/1/1e5
	0.45	—	—	—	—	—	—	—	0.2/0.2/2.5e5	0.1/0.5/2e5
	0.5	—	—	—	—	—	—	—	—	0.1/0.1/1e6

Here we assume that the MAF of SNP A is smaller than that of SNP B in each pair. The first and second number in each cell is the threshold for SNP A (α_A) and SNP B (α_B), respectively. These two thresholds are scaled by 10^{-2}. The third number in each cell is the cost saving, which is the ratio between cost of brute-force method and that of the two-stage model.

We created phenotype data using the phenotype ''CRP'' (C-reactive protein) in the NFBC data as a starting point. To simulate the significant SNP pairs, we randomly sample the MAF of each SNP from [0.05, 0.5). The alleles of each of the individuals at these two simulated SNPs are then sampled from the MAF. The phenotypes of the individuals with causal alleles at the SNP pairs are increased by a selected effect size so that the pairs have 50% power in the brute-force approach. Then we apply both the brute-force approach and the two-stage approach to the simulated dataset. The first stage significance thresholds in the two-stage approach are selected in order to obtain 1% power loss.

We generated 10,000 simulated SNP pairs and applied both approaches. The power for each approach is calculated as the proportion of experiments that the approach detected the implanted SNP pairs among all 10,000 experiments. The power of the brute-force approach is 51% while the power of TEPPA is 50.8%. The practical power loss is 0.4%. We note that the power loss is lower than we expected because the thresholds are chosen for MAF frequency bins to be conservative and valid for all members of that bin.

3. METHODS

3.1. Association test between traits and one SNP

We first illustrate the method to detect associations between traits and one SNP. A traditional approach to identify the association is that for each SNP, we compare the average trait value of individuals who carry the causal allele at the SNP and that of the individuals who do not have the causal allele at the SNP of interest. If the difference between those two values is above a certain threshold, we declare that the investigated SNP has a significant correlation with the trait. This approach is referred to as ''single marker test'' and has been successful in many association studies. We analyze the power of the ''single marker test'' as follows.

Assume we are investigating SNP A, with minor allele frequency to be p_A and the causal allele is the minor allele (for the case where the causal allele is the major allele, we have a similar analysis). Let N be the number of individuals and y_i be the trait value of individual i. Then the number of individuals with the minor allele at SNP A can be denoted as $N_A = N \cdot p_A$, and the number of individuals without the minor allele at SNP A can be denoted as $N_{\neg A} = N \cdot p_{\neg A} = N \cdot (1 - p_A)$. We use x_i^A to denote the allele of individual i at SNP A; y_i is any real number and $x_i^A \in \{0, 1\}$. We set $x_i^A = 1$ when the allele of individual i at SNP A is the minor allele and $x_i^A = 0$ otherwise.

We assume that a trait value of individual i follows the normal distribution with a certain mean μ and a variance σ^2. If the minor allele affects the trait, the mean trait value (μ) of individuals with the minor allele will increase by an effect size β_A. Now, we can obtain the distribution of y_i as

$$y_i \sim N(\mu + x_i^A \beta_A, \sigma^2) \tag{1}$$

Let $\bar{Y}_A$ be the average trait value of individuals who have the causal allele at SNP A and $\bar{Y}_{\neg A}$ be the average trait value of individuals who do not carry the causal allele at SNP A. Then we can derive the distributions of $\bar{Y}_A$ and $\bar{Y}_{\neg A}$ as follows:

$$\bar{Y}_A = \frac{\sum_{i:x_i^A=1} y_i}{N_A} \sim N\left(\mu + \beta_A, \frac{\sigma^2}{N \cdot p_A}\right), \quad \bar{Y}_{\neg A} = \frac{\sum_{i:x_i^A=0} y_i}{N_{\neg A}} \sim N\left(\mu, \frac{\sigma^2}{N \cdot p_{\neg A}}\right) \tag{2}$$

We normalize the difference between $\bar{Y}_A$ and $\bar{Y}_{\neg A}$ to obtain the following statistic S_A, which is normally distributed with mean $\lambda_A \sqrt{N}$ (the noncentrality parameter) and unit variance.

$$S_A = \frac{\bar{Y}_A - \bar{Y}_{\neg A}}{\sqrt{\frac{\sigma^2}{N \cdot p_A \cdot (1-p_A)}}} \sim N(\lambda_A \sqrt{N}, 1), \text{ where } \lambda_A = \frac{\beta_A \sqrt{p_A(1-p_A)}}{\sigma} \tag{3}$$

Given the significance level α and the observed value of the test statistic S_A, the SNP is deemed as significant, or statistically associated with the trait, if $|S_A| \geq \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the quantile function of the standard normal distribution. For simplicity, we use the notation $T = \Phi^{-1}(1 - \alpha/2)$ as the per-SNP threshold.

We declare all those SNPs with statistic $|S_A| > T$ to be associated with the trait. So the per-causal-SNP power of a putative causal SNP A, which is the probability of $|S_A| > T$, can be calculated as

$$P_1(\mathrm{A}) = P(|S_A| > T) = \Phi(-T + \lambda_A\sqrt{N}) + 1 - \Phi(T + \lambda_A\sqrt{N}) \tag{4}$$

The average power $\overline{P_1}$ is obtained by averaging per-causal-SNP powers over all putative causal SNPs.

3.2. The brute-force approach for pairwise association test

Current studies on complex diseases have also suggested that some SNPs influence traits in pairs. Only when both causal alleles appear on a pair of SNPs is the trait value increased. To detect the interaction of SNPs that influence the trait, we need to consider the association between a trait and a pair of SNPs (pairwise association test). We analyze the power of the brute force approach, which calculates the association between a trait and all pairs of SNPs as follows.

We assume that there exists an SNP pair AB, composed of SNP A and SNP B, that influences a trait. Assume the causal alleles are minor alleles at both SNPs. Our statistic is the difference between the average trait value of individuals who have minor alleles on both SNPs and that of individuals who do not have minor allele on at least one of the two SNPs A and B. Here we assume the two SNPs have the same (positive) direction of effect. We use the same notation as in section 3.1. The expected number of individuals who have minor alleles at both SNPs can be computed as $N_{AB} = N \cdot p_A \cdot p_B$, and the expected number of individuals who do not have minor alleles at both SNPs can be computed as $N_{\neg AB} = N \cdot (1 - p_A \cdot p_B)$. If an individual carries the causal alleles at both SNPs A and B, the mean of trait value is increased or decreased by the effect size of the SNP pairs, which is denoted as β_{AB}. Then we can write the distribution of y_i as

$$y_i \sim N(\mu + x_i^A x_i^B \beta_{AB}, \sigma^2) \tag{5}$$

Let $\bar{Y}_{AB}$ be the average trait value of individuals with causal alleles at both SNPs and, let $\bar{Y}_{\neg AB}$ be the average trait value of individuals without causal alleles at both SNPs. For simplicity, let $\sum_{11}$ denote $\sum_{i:x_i^A=1 \wedge x_i^B=1}$, and similarly for $\sum_{10}$, $\sum_{01}$, $\sum_{00}$ for different alleles of SNPs A and B. We can calculate $\bar{Y}_{AB}$ and $\bar{Y}_{\neg AB}$ as

$$\begin{aligned}
\bar{Y}_{AB} &= \frac{1}{N_{AB}} \sum_{11} y_i \sim N\left(\mu + \beta_{AB}, \frac{\sigma^2}{Np_Ap_B}\right), \\
\bar{Y}_{\neg AB} &= \frac{1}{N_{\neg AB}} \sum_{00,\,01,\,10} y_i \sim N\left(\mu, \frac{\sigma^2}{N(1-p_Ap_B)}\right)
\end{aligned} \tag{6}$$

We normalize the difference between $\bar{Y}_{AB}$ and $\bar{Y}_{\neg AB}$ to obtain the following statistic S_{AB}, which is normally distributed with mean $\lambda_{AB}\sqrt{N}$ (the noncentrality parameter) and unit variance.

$$S_{AB} = \frac{\bar{Y}_{AB} - \bar{Y}_{\neg AB}}{\sqrt{\frac{\sigma^2}{Np_Ap_B(1-p_Ap_B)}}} \sim N(\lambda_{AB}\sqrt{N}, 1), \text{ where } \lambda_{AB} = \frac{\beta_{AB}\sqrt{p_Ap_B(1-p_Ap_B)}}{\sigma} \tag{7}$$

According to Prabhu and Pe'er (2012), we set the per-SNP-pair significance level $\alpha = 10^{-12}$. The per-SNP-pair statistic threshold is then $T_2 = -\Phi^{-1}(\alpha/2) = 7.13$. The per-causal-SNP-pair power of a putative causal SNP pair AB can be estimated as

$$P_{BF}(\mathrm{AB}) = \Phi(-T_2 + \lambda_{AB}\sqrt{N}) + 1 - \Phi(T_2 + \lambda_{AB}\sqrt{N}) \tag{8}$$

The average power $\overline{P_{BF}}$ is obtained by averaging per-causal-SNP-pair powers over all putative causal SNP pairs.

Assuming the total number of SNPs is M, we define the cost of brute-force method to be the total number of SNP pairs needed for association analysis, that is, $C_{BF}(M) = \binom{M}{2}$.

3.3. Two-stage model

In the brute force approach, the total number of SNP pairs to be considered is $\binom{M}{2}$, and we need to compute the statistic S_{AB} for all these pairs. Considering the number of SNPs involved in current GWAS, the computational burden makes this strategy infeasible.

We propose a two-stage model to reduce the number of tests needed while maintaining similar power with the brute force approach. In the first stage, we propose two thresholds T_A and T_B and perform the single marker test on all SNPs. In the second stage, we pair all SNPs that are significant under threshold T_A with those significant SNPs under threshold T_B. Then we perform a pairwise association test between traits and all those pairs. The SNP pairs that pass the per-SNP-pair statistic threshold T_2 are considered to be statistically associated with the trait.

The analysis of single marker tests in the first stage is quite similar to that of the single SNP association test in Section 3.1. We derive the similar equations with (1), (2) and (3) except that the effect size of SNP A becomes $p_B\beta_{AB}$, when the pair of SNP A and SNP B is the causal SNP pair. So the statistic S_A of SNP A is approximately distributed as

$$S_A = \frac{\bar{Y}_A - \bar{Y}_{\neg A}}{\sqrt{\frac{\sigma^2}{N \cdot p_A \cdot (1-p_A)}}} \sim N(\lambda_A \sqrt{N}, 1), \text{ where } \lambda_A = \frac{p_B \beta_{AB} \sqrt{p_A(1-p_A)}}{\sigma} \tag{9}$$

The analysis of SNP B is the same except that we switch p_A and p_B in the equations. We note that this is an approximation, because for the individuals that have SNP A, the distribution of the phenotypic mean is actually a mixture of two normal distributions where one component is of the individuals who have SNP A and also have SNP B and the other component is of the individuals who have SNP A but do not have SNP B. We approximate this mixture as a single normal distribution, which in principal may underestimate the variance for phenotypic mean. However, for the plausible range of effect sizes, samples sizes, and minor allele frequencies in gene–gene interaction studies, the effect of this approximation is negligible. We describe the empirical verification of this observation in our experiments in the Appendix.

Assume a pair of SNPs A and B are putatively associated with a trait. The underlying effect size β_{AB} could either be positive or negative. Here we first analyze the case where the true effect size is positive. To find such positive pairwise association in our model, S_A must be no less than T_A, S_B must be no less than T_B (or vice versa, but here we only analyze one case since we will show in section 3.5 that the other case is not necessary), and S_{AB} must be at least T_2. Hence, we need to consider three statistics and three thresholds to compute the analytical power of the two-stage model. Under the assumption that we are aware the effect size is positive, the per-causal-SNP-pair power of a putative causal SNP pair AB can be denoted as

$$P_2+(\mathrm{AB}) = P(S_A \geq T_A, S_B \geq T_B \text{ and } S_{AB} \geq T_2) \tag{10}$$

However, considering the fact that whether the effect of the SNP pair is positive or negative is unknown, we also need to calculate the probability where S_{AB} is less than $-T_2$, that is,

$$P_2-(\mathrm{AB}) = P(S_A \leq -T_A, S_B \leq -T_B \text{ and } S_{AB} \leq -T_2) \tag{11}$$

So, the per-causal-SNP-pair power of a putative causal SNP pair AB is

$$P_2(\mathrm{AB}) = P_2+(\mathrm{AB}) + P_2-(\mathrm{AB}) \tag{12}$$

The analysis for the case where the true effect size is negative is exactly the same except that the noncentrality parameters for S_A, S_B, and S_{AB} are negative.

To calculate the value of $P_2(\mathrm{AB})$, we need to take into account correlations between statistics. The two statistics S_A and S_{AB} are correlated because both involve SNP A. Similarly, we have a correlation between S_B and S_{AB}. We assume SNPs are independent, and hence there is no correlation between S_A and S_B. The average power $\overline{P_2}$ is obtained by averaging per-causal-SNP-pair powers over all putative causal SNP pairs. Computing the analytical power of the two-stage model is complicated because of the correlations between statistics. We estimate the power using a multivariate normal distribution framework as in section 3.4.

Denote the per-SNP significance level corresponding to the statistic thresholds T_A and T_B in the first stage to be α_A and α_B, respectively. Then we have $\alpha_A = 2\Phi(-T_a)$ and $\alpha_B = 2\Phi(-T_b)$. The cost of the two-stage model can be computed as $C_{TS}(M, \alpha_A, \alpha_B) \approx M^2\alpha_A\alpha_B/2$.

We measure the cost saving by the ratio between cost of brute-force method (C_{BF}) and that of the two-stage model (C_{TS}):

$$\frac{C_{BF}(M)}{C_{TS}(M, \alpha_A, \alpha_B)} = \frac{\binom{M}{2}}{M^2\alpha_A\alpha_B/2} \approx \frac{1}{\alpha_A\alpha_B} \tag{13}$$

And we define the power loss to be

$$1 - \frac{\overline{P_2}}{\overline{P_{BF}}} \tag{14}$$

For a given dataset, there exists a trade-off between the power loss and cost saving. The trade-off is controlled by the two thresholds T_A and T_B. We carefully design the thresholds to achieve high cost saving while maintaining low power loss. The details of the algorithm are summarized in section 3.5.

3.4. Estimating the two-stage power using the MVN

In this section, we provide an approach to compute the power of the two-stage model in Equation (12). The distribution of association statistics S_A, S_B, and S_{AB} has been derived in sections 3.2 and 3.3. We aim to compute the power in Equation (12) for any given thresholds T_A, T_B, and T_2.

For many widely used statistical tests, the statistics over multiple markers asymptotically follow an MVN (Seaman and Muller-Myhsok, 2005; Lin, 2005). To derive the analytical power of our two-stage model, we use the framework of MVN proposed by Han et al. (2009). The statistics (S_A,S_B,S_{AB}) follow an MVN. The mean of each statistic of the MVN is the noncentrality parameters (NCPs) of the statistics. The NCPs of S_A, S_B, and S_{AB} are already derived in Equations (7) and (9), so the mean vector of the joint random vector (S_A, S_B, S_{AB}) is ($\lambda_A\sqrt{N}$, $\lambda_B\sqrt{N}$, $\lambda_{AB}\sqrt{N}$). The covariance and the correlation between any pair of statistics are equivalent in this case because variances of the statistics are 1. So, the covariance matrix in the MVN will be the correlations among statistics. We assume the SNPs are independent of each other, so the correlation between S_A and S_A is 1, and the correlation between S_A and S_B is 0. The covariance matrix is as follows:

$$\begin{pmatrix} 1 & 0 & Cor(S_A, S_{AB}) \\ 0 & 1 & Cor(S_B, S_{AB}) \\ Cor(S_A, S_{AB}) & Cor(S_B, S_{AB}) & 1 \end{pmatrix}$$

We only need to compute the correlation between S_A (or S_B) and S_{AB}, which is denoted as $Cor(S_A, S_{AB})$, to derive the complete MVN. We provide a simple but accurate way to compute the correlation. First we create a virtual SNP C, where the allele of SNP C is exactly the same with the value of the SNP pair AB. The minor allele frequency of SNP C is denoted as $p_C = p_{AB} = p_A p_B$. The statistic S_C will be equivalent to statistic S_{AB}. Instead of computing $Cor(S_A, S_{AB})$, now we can compute $Cor(S_A, S_C)$. Since we adopt the dominant model, the genotypes of SNP A and SNP B are binary values {0, 1}. The genotype of SNP C is obtained by multiplying the genotypes of SNP A and SNP B, so it is also a binary value. The Pearson correlation r_{AC} between the genotypes of SNP A and SNP C is then

$$r_{AC} = \frac{p_C(1-p_A)}{\sqrt{p_C(1-p_C)p_A(1-p_A)}} = \frac{p_A p_B(1-p_A)}{\sqrt{p_A p_B(1-p_A p_B)p_A(1-p_A)}} \tag{15}$$

It is well known that the correlation $Cor(S_A, S_C)$ between the test statistic S_A and S_C is equal to r_{AC} (Han et al., 2009). Similarly, we can compute the correlation $Cor(S_B, S_{AB})$.

Up to now we obtained all parameters for the MVN framework. Then, we can compute the power as the volume outside of the significance threshold under the MVN we created. Figure 3 helps to illustrate the idea. We can see that in the three-dimension space of the MVN framework for statistics S_A, S_B, and S_{AB}, the two cubes on the corners correspond to the significance region. Using the MVN, we can compute the power of our two-stage model for any given thresholds T_A, T_B, and T_2 by summing up the volume of these two cubes under the MVN. This method yields a very accurate estimate of power when there exist correlations among statistics, and hence it provides an appropriate framework to compute the analytical power of our model.

3.5. Efficient pairwise association test using TEPAA

In previous sections, we have illustrated how to calculate the power and cost savings of our two-stage model for any given threshold. In this section, we provide a framework, TEPAA, to

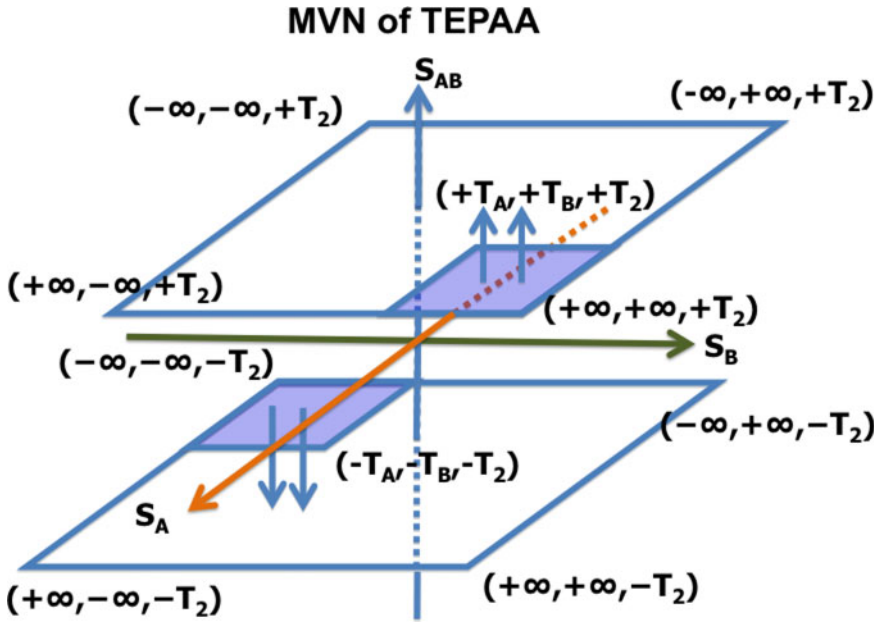

FIG. 3. The volume of the two cubes under the MVN on the two corners is the power of our two-stage model.

determine the first thresholds, which generate a relatively small number of SNP pairs for pairwise association test in the second stage while losing a small amount of power compared to the brute force approach.

From Equation (12) and section 3.4, we can see that the joint distribution between the association statistics of single SNPs and the association statistic of a pair of SNPs depends on the MAFs of the pair of SNPs. MAFs are observable values, so we can categorize all SNP pairs based on the combination of their MAFs. Since MAFs are continuous values, we can discretize the MAFs into bins to have a small number of combinations. After removing rare variants, we can categorize all SNPs into nine bins, with step size 0.05. In order to detect the pairwise association for all SNP pairs, we break all combinations of SNP pairs into two cases. First we pair SNPs within different bins and this results in $\binom{9}{2}$ categories. The second case is to combine SNPs within one bin. So totally we have $\binom{9}{2}+9$ categories of SNP pairs.

Assuming the power of brute force approach is 50%, we can calculate the effect size β_{AB} from Equation (8). Then for each category of SNP pairs, we can compute the power loss and cost savings from Equations (13) and (14) with the MVN, given two first-stage significance levels α_A and α_B. We do an exhaustive search over the space [0, 1) with a small step size to find the optimal values of α_A and α_B to achieve best cost savings while maintaining power loss of 1%. The values of α_A and α_B are shown in Table 2 when there are 5,326 samples in the dataset.

For SNPs in each bin, we carry out the single marker test and sort the association statistics of the SNPs. Then for each category of SNP pairs, we do a binary search in each involved bin to find all significant SNPs under the precomputed significance level. The selected SNPs are then paired for the second-stage pairwise association test. Based on the precomputed values of α_A and α_B, we can estimate the cost savings for each category of SNP pairs as in Table 2. We propose a threshold for each bin in each category of SNP pairs, and the bins are disjoint. So, in the calculation of Equation (10), we only need to consider the case where $S_A > T_A$ and $S_B > T_B$, and it is not necessary to consider the case $S_A > T_B$ and $S_B > T_A$. We have the same conclusion in the calculation of Equation (11). We summarize the framework of *TEPAA* in Algorithm 1.

Although the calculation is based on the assumption that the brute force approach has a power of 50%, our approach is robust to the effect size. We did simulations for different effect sizes, which generate different power for the brute force approach. The cost saving of TEPAA is stable when achieving 1% power loss under various effect sizes.

Algorithm 1: Framework of TEPAA

Input: A GWAS data set with genotype and phenotype for each individual.
Output: SNP pairs associated with the phenotype.
1 Remove rare variants, categorize rest SNPs into 9 bins according to MAFs, with step size 0.05.
2 Precompute the thresholds for each combination of bins as in Table 2, which only depend on the number of samples in the dataset and second-stage threshold.
3 For SNPs in each bin, we carry out the single marker test and sort the association statistics of each SNP in the bin.
4 For each pair of bins, we perform binary searches within the corresponding bin to determine which SNPs have association statistics more significant than the predetermined thresholds.
5 For each of these identified pairs of SNPs, we perform the pairwise association test and report any significant pairs.

4. CONCLUSIONS

In this article, we proposed a two-stage model to detect SNP pairs associated with traits. The key idea behind our method is that we model the joint distribution between association statistics at single SNPs and association statistics at pairs of SNPs to allow us to apply a two-stage model that provides guarantees that we detect associations of pairs of SNPs with small numbers of tests while losing very little power. We rapidly eliminate from consideration pairs of SNPs in which high probability is not associated with the trait. Using extensive simulations, we show that our approach can reduce the computational time by a factor of 63 while only losing approximately 1% of the power compared to the brute-force approach.

We note that in this article, we are only considering pairs of SNPs that are far apart from each other. There is another class of methods that consider multiple SNPs close to each other (Wu et al., 2010, 2011; Listgarten et al., 2013). These problems are completely different and characterized by very different challenges. For example, the computational burden that is the focus of our article is different because the number of pairs of SNPs near each other is significantly smaller than the total number of pairs of SNPs. In addition, neighboring SNPs are typically correlated with each other, referred to as in linkage disequilibrium (LD). Pairs of SNPs far from each other are typically independent or unlinked, which is an observation that we leverage in our approach.

In our case, we assume that the effect of the interaction is in the same direction as the independent effects of the SNPs. Under this assumption, the worst case scenario in terms of statistical power of our two-stage approach is when the marginal effects are zero, which is how we determine our thresholds. If the independent effects are nonzero, the statistical power will be higher. If the independent effects are in a direction inconsistent with the interaction effect, then our model is not appropriate and may not discover the interactions.

5. APPENDIX

In order to justify the assumption that we can approximate the phenotypic variance in Equation (9) with a single uniform variance, we did extensive simulations to demonstrate the negligible effect of gene–gene interaction on the phenotypic variance of different subgroups determined by the genotype. We first use the phenotype CRP(C-reactive protein) of the NFBC data as a starting point, which has mean value -0.0256. The phenotype is collected from 5,326 individuals and the variance of phenotype (σ^2) is 2.65.

We simulate the genotype of one causal SNP pair in each iteration. The MAF of the SNPs are randomly chosen from the [0.05, 0.5). In order to simulate the phenotype, the effect of gene–gene interaction β is computed to guarantee that the brute force approach have power 50% at the 10^{12} significance level. Thus β depends on the MAFs of the causal SNP pairs. After the genotype of the causal SNP is simulated, the phenotype is simulated according to Equation (5). The phenotype of individuals with causal haplotype is elevated with the effect size β. We then check the variance of phenotype within each group determined by the genotype to see whether there is significant difference. The process is repeated 10,000 iterations. The average value of phenotypic mean and variance within each group throughout the 10,000 iterations is summarized in Table 3.

Table 3. The Average Value of Phenotypic Mean and Variance Within Each Group Throughout 10,000 Iterations

Phenotype group	*Average mean*	*Average variance*
Y_A	0.209	2.689
$Y_{\neg A}$	−0.026	2.653
Y_B	0.138	2.680
$Y_{\neg B}$	−0.026	2.658
Y_{AB}	0.378	2.652
Y_{Ab}	−0.024	2.650

Y_A represents the phenotype of individuals with causal allele at SNP A. $Y_{\neg A}$ represents the phenotype of individuals without causal allele at SNP A. Y_B represents the phenotype of individuals with causal allele at SNP B. $Y_{\neg B}$ represents the phenotype of individuals without causal allele at SNP B. Y_{AB} represents the phenotype of individuals with causal allele at both SNP A and SNP B. Y_{Ab} represents the phenotype of individuals with causal allele at SNP A but not SNP B.

Table 4. The Mean and Variance of Different Phenotypes of the NFBC Data

Phenotype	*Mean*	*Variance*
TG	−0.430	3.910
INS	−0.908	7.461
DBP	−0.076	126.5
BMI	−0.483	4.141
GLU	−0.817	6.70
HDL	−0.431	3.809
SBP	−0.124	163
LDL	−0.451	4.593
Height	−0.011	39.9

TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low-density lipoprotein.

We can see that the average variance in each group does not differ much, the largest difference is within 2%. We also compared the variance of two subgroups within group A, where one subgroup has causal SNP pair AB and the other subgroup only has SNP A but not SNP B. We found that the variance of these two subgroups are also close and within 2% difference of the variance of group A. So, although the distribution of group A is a mixture of two normal distributions, these two normal distributions share similar variance with the distribution of group A and the effect of mixture is negligible.

We also did simulations on the other 9 phenotypes of the NFBC data, which have various phenotypic mean and variance, to see the effect of different phenotypic mean and variance on the approximation in Equation (9). The phenotypic mean and variance is summarized in Table 4. Similar with the phenotype CRP, our simulation shows that the average variance of different groups differ at most 2% for all nine phenotypes. So we conclude that under various phenotypic mean and variance, we can approximate the mixture of two normal distributions as a single normal distribution in Equation (9).

ACKNOWLEDGMENTS

Z.W., J.H.S., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, and 1320589, and the National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-MH101782, and R01-ES022282. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691). S.S. is supported by the USA-Israel Binational Science Foundation and

Israel Science Foundation. J.A.L. has been partially supported by the Saiotek and IT-609-13 programs (Basque government), TIN2013-41272-P (Spanish Ministry of Science and Innovation), and the COMBIOMED network in computational biomedicine (Carlos III Health Institute).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Altshuler, D., Hirschhorn, J.N., Klannemark, M., et al. 2000. The common pparγ pro12ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26, 76–80.

Bertina, R.M., Koeleman, B.P.C., Koster, T., et al. 1994. Mutation in blood coagulation factor v associated with resistance to activated protein c. *Nature* 369, 64–67.

Brem, R.B., Storey, J.D., Whittle, J., et al. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436, 701–703.

Brinza, D., Schultz, M., Tesler, G., et al. 2010. Rapid detection of gene-gene interactions in genome-wide association studies. *Bioinformatics* 26, 2856–2862.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., et al. 1993. Gene dose of apolipoprotein e type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 921–923.

Evans, D.M., Marchini, J., Morris, A.P., et al. 2006. Two-stage two-locus models in genome-wide association. *PLoS Genet.* 2, e157.

Han, B., Kang, H.M., and Eskin, E. 2009. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* 5, e1000456.

Jarvelin, M.-R., Sovio, U., King, V., et al. 2004. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension* 44, 838–846.

Kostem, E., and Eskin, E. 2013. Efficiently identifying significant associations in genomewide association studies. *J. Comput. Biol.* 20, 817–830.

Kostem, E., Lozano, J.A., and Eskin, E. 2011. Increasing power of genome-wide association studies by collecting additional snps. *Genetics* 188, 449–460.

Lin, D.Y. 2005. An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21, 781–787.

Listgarten, J., Lippert, C., Kang, E.Y., et al. 2013. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 29, 152633.

Ljungberg, K., Holmgren, S., and Carlborg, O. 2004. Simultaneous search for multiple qtl using the global optimization algorithm direct. *Bioinformatics* 20, 1887–1895.

Millstein, J., Conti, D.V., Gilliland, F.D., et al. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.* 78, 15–27.

Prabhu, S., and Pe'er, I. 2012. Ultrafast genome-wide scan for snp-snp interactions in common complex disease. *Genome Res.* 22, 2230–2240.

Rantakallio, P. 1969. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr. Scand.* Suppl 193, 43.

Saxena, R., Voight, B.F., Lyssenko, V., et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331–1336.

Schpbach, T., Xenarios, I., Bergmann, S., et al. 2010. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 26, 1468–1469.

Seaman, S., and Muller-Myhsok, B. 2005. Rapid simulation of *p*-values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* 76, 399–408.

Wan, X., Yang, C., Yang, Q., et al. 2010. Boost: a fast approach to detecting genegene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* 87, 325–340.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678.

Williams, S.M., Addy, J.H., Phillips, J.A., et al. 2000. Combinations of variations in multiple genes are associated with hypertension. *Hypertension* 36, 2–6.

Wu, M.C., Kraft, P., Epstein, M.P., et al. 2010. Powerful snp-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* 86, 929–942.

Wu, M.C., Lee, S., Cai, T., et al. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93.

Yanchina, E.D., Ivchik, T.V., Shvarts, E.I., et al. 2004. Gene-gene interactions between glutathione-s transferase m1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease. *Bull. Exp. Biol. Med.* 137, 64–66.

Yang, C., He, Z., Wan, X., et al. 2009. Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 25, 504–511.

Zhang, X., Huang, S., Zou, F., et al. 2010. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* 26, i217–i227.

Address correspondence to:
Prof. Eleazar Eskin
Computer Science Department
University of California Los Angeles
Mail Code: 1596
3532-J Boelter Hall
Los Angeles, CA 90095-1596

E-mail: eeskin@cs.ucla.edu