Published in final edited form as:

Ann Hum Genet. 2010 September 1; 74(5): 406–415. doi:10.1111/j.1469-1809.2010.00594.x.

Identify Interaction Genes in Genome-Wide Association Studies Using a Model-Based Two-Stage Approach

Zhaogong Zhang^{1,2}, Adan Niu¹, and Qiuying Sha¹

Zhaogong Zhang: zhaogong@mtu.edu; Adan Niu: aniu@mtu.edu; Qiuying Sha: qsha@mtu.edu

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931

²School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China

Abstract

In this paper, we propose a two-stage approach based on seventeen biological plausible models to search for two-locus combinations that have significant joint effects on the disease status in genome-wide association (GWA) studies. In the two-stage analyses, we only test two-locus joint effects of SNPs that show modest marginal effects. We use simulation studies to compare the power of our two-stage analysis with a single-marker analysis and a two-stage analysis by using a full model. We find that for most plausible interaction effects, our two-stage analysis can dramatically increase the power to identify two-locus joint effects compared to a single-marker analysis and a two-stage analysis based on the full model. We also compare two-stage methods with one-stage methods. Our simulation results indicate that two-stage methods are more powerful than one-stage methods. We applied our two-stage approach to a GWA study for identifying genetic factors that might be relevant in the pathogenesis of sporadic Amyotrophic Lateral Sclerosis (ALS). Our proposed two-stage approach found that two SNPs have significant joint effect on sporadic ALS while the single-marker analysis and the two-stage analysis based on the full model did not find any significant results.

Introduction

Genome-wide association (GWA) studies have identified loci for cardiac repolarization (QT interval) (Arking et al. 2006), type 2 diabetes (Sladek et al. 2007), and BMI (Herbert et al. 2006). Currently, single marker tests that test marginal effects are still the most commonly used methods in GWA studies, although there is increasing evidence suggesting that complex diseases are the results of interactions of many genes and environmental factors (Risch 2000; Aston et al. 2005; Dong et al. 2005; Roldan et al. 2005; Millstein et al. 2006). Single marker tests ignore the possibility that effects of multi-locus functional genetic units play a larger role than the single-locus effect in determining the trait variability (Templeton 2000, Nelson et al. 2001, Sha et al. 2006).

Recently, several methods have been proposed to search for a set of interacting loci that jointly have significant effects on the disease. This group of methods includes the Multifactor Dimensionality Reduction (MDR) method proposed by Ritchie et al. (2001, 2003) and reviewed recently by Moore (2004), the Combinatorial Partitioning Method (CPM) proposed by Nelson et al. (2001), the Restrict Partitioning Method (RPM) proposed by Culverhouse et al. (2004), the Combinatorial Searching Method (CSM) proposed by Sha et al. (2006), the Focused Interaction Testing Framework (FITF) proposed by Millstein et al. (2006), and the Ensemble Learning Approach (ELA) proposed by Zhang et al. (2008)

among others. These methods use exhaustive searching approach in which each of the 1-locus, 2-locus, ..., and L-locus combinations is considered, and thus are computationally too intensive for applying to GWA studies.

More recently, Marchini et al. (2005) and Evans et al. (2006) investigated whether a two-stage analysis was a viable approach to improve the power to identify SNPs that have epistatic effects and modest marginal effects. Possibly because they used a very conservative correction for multiple comparisons and used a full model to test two-locus joint effects (used a test with eight degrees of freedom (df)), they found that their two-stage analysis did not improve the power for identifying SNPs that jointly have an epistatic effect in GWAs. One of the main goals of this paper is to investigate the power of an two-stage analysis by using a better correction for multiple comparisons (permutation tests) and using seventeen biologically plausible one df models instead of using the eight df full model. In our two-stage analyses, we only investigate joint effect of SNPs that show modest marginal effect. In the first stage, we select SNPs that show some marginal effect. In the second stage, each of the two-locus combinations of the selected SNPs is tested for joint effects under each of the seventeen models. We use simulation studies to compare the power of our two-stage analysis with a single-marker analysis and a two-stage analysis by using the full model. Simulation results show that our two-stage analysis is consistently more powerful than the single-marker analysis and the two-stage analysis using the full model in all of the cases considered in our simulation studies. We also evaluate the performance of the proposed two-stage approach by applying it to a GWA study for identifying genetic factors that might be relevant in the pathogenesis of sporadic ALS.

Methods

Two-locus Analysis Based on Seventeen Two-marker Models

In this paper, we propose a two-stage approach to analyze GWA data under seventeen two-locus models. We call the proposed approach Two-Stage approach based on Seventeen two-locus Models (TSSM). For a causal SNP D_1 with alleles A, a and another casual SNP D_2 with alleles B, b , Figure 1 and 2 give eight epistatic two-locus models and nine multiplicative two-locus models with high-risk alleles A and B , respectively. These models have been used and discussed by Xiong et al. (2003), Zhao et al. (2006), Hodge (1981), and Risch (1990).

We constructed a two-locus association test based on the seventeen two-locus models as follows. Let $P(D | g)$ denote the penetrance of two-locus genotype g , that is, the probability of a disease given two-locus genotype g . Let β_0 denote the logarithm of the penetrance of genotypes with a relative risk of 1 in the seventeen models and $\beta_1 = \log \theta$, where θ is the relative risk given in Figure 1 and 2. The seventeen two-locus models can be described by the following log linear model

$$\log P(D | g) = \beta_0 + \beta_1 X,$$

where g is the two-locus genotype. In each of the eight epistatic models given in Figure 1, the nine two-locus genotypes can be divided into two groups: high-risk group G_H and low-risk group G_L . For example, under the $\text{Dom} \cap \text{Dom}$ model, $G_H = \{bBaA, bBAA, BBaA, BBAA\}$ and $G_L = \{bbaa, bbaA, bbAA, bBaa, BBaa\}$. Under the eight epistatic models, we define $X = 1$ if the genotype g belongs to the high-risk group and $X = 0$ otherwise. For the nine multiplicative models given in Figure 2, we define $X = x_1 + x_2$, where x_1 is the numerical code of g_1 (the genotype in the first marker) and is given by

$$x_1 = \begin{cases} 0, & \text{if } g_1 = aa \\ 1, & \text{if } g_1 = aA \text{ or } AA \end{cases}, x_1 = \begin{cases} 0, & \text{if } g_1 = aa \text{ or } aA \\ 1, & \text{if } g_1 = AA \end{cases}, \text{ or } x_1 = \begin{cases} 0, & \text{if } g_1 = aa \\ 0.5, & \text{if } g_1 = aA \\ 1, & \text{if } g_1 = AA \end{cases}$$

for a dominant, recessive or multiplicative model, respectively; x_2 is similarly defined as a numerical code of the genotype in the second marker. Under this log linear model, test of association between the disease and these two loci is equivalent to test of the null hypothesis $H_0: \beta_1 = 0$. For the i^{th} individual, let y_i denote the trait value (1 for diseased individuals and 0 for normal individuals) and X_i denote the numerical code of the genotype (X in the log linear model). The score test statistic is given by

$$T_{score} = \frac{(\sum_{i=1}^N (X_i - \bar{X})(y_i - \bar{y}))^2}{\bar{y}(1 - \bar{y}) \sum_{i=1}^N (X_i - \bar{X})^2}, \quad (1)$$

where N is the sample size, $\bar{X}$ is the average of $X_1, \dots, X_N$, and $\bar{y}$ is the average of $y_1, \dots, y_N$. Under the null hypothesis of no association, T_{score} asymptotically follows a χ^2 distribution with one df.

The method to search for significant two-locus combinations for each of the seventeen models has the following two steps.

Step 1: For each SNP, let n and m denote the number of individuals in cases and controls (different SNPs may have different number of cases and controls due to missing genotypes). Let n_0, n_1, n_2 and m_0, m_1, m_2 denote the number of three genotypes in cases and controls, respectively. The genotypic chi-square test with two df is given by

$$T_G = \sum_{i=0}^2 \frac{(\hat{p}_i - \hat{q}_i)^2}{\frac{\hat{p}_i}{m} + \frac{\hat{q}_i}{n}}, \quad (2)$$

where $\hat{p}_i = \frac{n_i}{n}$ and $\hat{q}_i = \frac{m_i}{m}$. This test is applied to each SNP to calculate a p-value. We then select M SNPs with the smallest p-values for further analysis.

Step 2: Under each of the seventeen two-locus models, we apply the two-locus association test T_{score} to each of the L two-locus combinations among the M selected SNPs, where $L = M(M-1)/2$. In this step, we get a p-value (called raw p-value) for each of the L two-locus combinations and each of the seventeen two-locus models.

Let p_{lj} denote the raw p-value of the l^{th} two-locus combination under the j^{th} two-locus model. A permutation procedure is used to adjust for multiple tests and multiple models. Based on the permuted data, we redo the single-marker test for each SNP, select M SNPs with the smallest p-values, and then test each of the two-locus combinations among the M selected SNPs under each of the seventeen models. For the k^{th} permutation, let p_{lj}^k denote the p-value of the two-locus test for the l^{th} two-locus combinations under the j^{th} model and let $p_{\min}^k = \min_{1 \leq l \leq L, 1 \leq j \leq 17} \{p_{lj}^k\}$. Suppose that we perform the permutation K times (we use $K = 1000$

in this study). Then, the overall p-value of the two-locus test for the l^{th} two-locus

combination under the j^{th} model is given by $P_{lj} = \frac{\#\{k: p_{\min}^k < p_{lj}\}}{K}$. In this way, we can adjust for multiple testing for L two-locus tests and seventeen models and we also can adjust for the fact that the M SNPs to do two-locus analysis were chosen as the “top” signals from a much larger set. A small p-value (e.g., $P_{lj} \leq 0.05$) indicates a significant genetic association between the two-locus combination and the disease.

Comparison of Methods

We compare our proposed method TSSM with the following four methods. The first is a Two-Stage method based on the Full Model (TSFM). The TSFM is the same as our TSSM in the first stage. In the second stage, however, instead of using 17 tests corresponding to 17 models in the TSSM, the TSFM uses one genotypic chi-square test with eight df. In fact, the genotypic chi-square test is the score test under a full model in which we assume that the nine two-locus genotypes follow a multinomial distribution. Let n_i and m_i denote the number of the i^{th} two-locus genotype in cases and controls, respectively, for $i = 1, \dots, 9$. Let n and m denote the number of cases and controls, respectively. The genotypic chi-square test with eight degrees of freedom is given by

$$\chi_8^2 = \sum_{i=1}^9 \frac{(\widehat{p}_i - \widehat{q}_i)^2}{\frac{\widehat{p}_i}{m} + \frac{\widehat{q}_i}{n}},$$

where $\widehat{p}_i = \frac{n_i}{n}$ and $\widehat{q}_i = \frac{m_i}{m}$. If the frequency of one or more genotypes is zero, the df of χ_8^2 will decrease accordingly. Again, a permutation procedure is used to adjust for multiple testing.

The second method that we compare with is a Single-Marker Test (SMT). In this method, we use the genotypic chi-square test with two df as given in equation (2) to test each marker and use a permutation procedure to adjust for multiple testing.

The third and fourth methods are the one-stage version of TSSM and TSFM and denoted by TSSM1 and TSFM1, respectively. In fact, TSSM1 is TSSM and TSFM1 is TSFM when we select all the markers in the first stage.

Simulation Studies

We simulated samples based on the haplotype data of 120 European chromosomes (CEU) released by the HapMap project (International HapMap Consortium 2007). However, we used only the haplotype data on chromosome 21 with 48,635 SNPs. To generate the genotype of an individual, we generated two haplotypes that are the recombinants of the 120 HapMap chromosomes. To generate a recombinant of the 120 chromosomes, we first generated a number of crossovers across the chromosome by a Poisson process with an average of 6 crossovers per Morgan. The crossover locations were generated according to a uniform distribution. The crossover locations divided the chromosome into segments. Each segment of the recombinant was a random chosen haplotype from the HapMap chromosomes in the same segment.

Under the null hypothesis, we randomly signed an individual as a case or control independent of the genotype. Under alternative hypothesis, we generated the disease status according to four two-locus disease models as described in Table 1. These four models were also used by Pickrell et al. (2007). In these models, we assumed that the disease prevalence was 10%. For power comparisons, we considered different scenarios which included

different values of LD between the two disease markers: 0, 0.2, 0.4, 0.6, different minor allele frequency (MAF) pairs at the two disease markers: (0.2, 0.2), (0.3, 0.4), (0.1, 0.3), and different sample sizes: 1,000, 1,500, 2,000 (half cases and half controls).

Because data were simulated based on the 120 Hapmap chromosomes, we may not find two markers with exactly the given LD and MAF pair. In our simulations, we randomly chose two markers with LD within ± 0.1 of the given value and MAF at each marker within ± 0.05 of the given value as the two disease markers. We assumed that the minor allele was the high risk allele in all of the four models.

We used 1000 replicated samples to evaluate the type I error and 100 replicated samples to evaluate power. With 1,000 replicated samples, the standard deviation for the type I error rates is $\sqrt{0.05 \times 0.95/1000} \approx 0.007$ for the nominal level of 0.05. Thus, the 95% confidence interval (CI) is (0.036, 0.064) for the nominal level of 0.05. Table 2 displays the estimated type I error rates of TSSM method. From this table, we can see that all the type I error rates are within the CI which indicates that the estimated type I error rates are not significantly different from the nominal level 0.05.

Power Comparisons

In each simulation scenario, we use 100 replications to evaluate the power. Let S_i denote all significant markers in the i th replication (single-marker method or two-marker method).

Then, the power is calculated by $power = \frac{1}{200} \sum_{i=1}^{100} I_i$, where I_i is the number of disease markers contained in S_i .

For power comparisons, we generated data under the four disease models. Among the four disease models, only one model (model 1: Heterogeneity) belongs to the seventeen models we used to create our method. We first compare the power of our proposed TSSM with that of the TSFM and SMT. The results of power comparisons are summarized in Figures 3 to 6. Figure 3 gives the results of the power comparison of the three methods (TSSM, TSFM, and SMT) for different sample sizes under four disease models and four LD patterns when the MAF pair at the two disease loci is (0.2, 0.2). Figure 3 shows that our proposed method, TSSM, is consistently more power than TSFM and SMT and TSFM is consistently more power than SMT regardless of different disease models and LD patterns. Figure 3 also shows that the power of all the three methods will increase with the increases of LD between the two disease loci. Figures 4 and 5 give the results of power comparisons when the MAF pairs at the two disease loci are (0.3, 0.4) and (0.1, 0.3), respectively. The patterns of power comparisons given in Figures 4 and 5 are very similar to that given in Figure 3. Summarizing the results of Figures 3 to 5, we conclude that our proposed method TSSM is consistently more power than the other two methods in all the cases we considered.

Figure 6 shows the results of the power comparisons of the three methods for different values of M (the number of markers selected in the first stage). In this set of simulations, we aim to investigate the impact of M on the power of TSSM and TSFM and to better choose the value of M . Figure 6 shows that although M has impact on the power of both TSSM and TSFM, the impact is not large. When the value of M varies from 100 to 600 (about 0.2% to 1% of the total number of markers), the power of both TSSM and TSFM change only slightly. In addition, Figure 6 shows again that our proposed method TSSM is consistently more powerful than the other two methods.

In order to compare the power of two-stage methods with that of the one-stage methods, we also perform another set of simulations to compare the power of the five methods (TSSM,

TSFM, TSSM1, TSFM1, and SMT) for different values of M under the four disease models. Since TSSM1 and TSFM1 need to evaluate every two-locus combination, and thus, are computationally intensive, we use a small number of markers (5000 markers) in this set of simulations. The results of the power comparison in this set of simulation are summarized in Figure 7. Figure 7 shows that the two-stage methods TSSM and TSFM are consistently more powerful than the corresponding one-stage methods TSSM1 and TSFM1 when M varies from 10 to 100. The results also show that the two-locus analysis (TSSM, TSFM, TSSM1, and TSFM1) is more powerful than the single-marker analysis (SMT), which is consistent with the results of Marchini et al. (2005) and Evans et al. (2006).

Analysis of ALS Data Set

Amyotrophic lateral sclerosis (ALS) is a fatal progressive neurodegenerative disease that attacks nerve cells in the brain and spinal cord resulting in muscle weakness and atrophy. Recently, Schymick et al. (2007) did a GWA study attempting to identify genetic factors that might be relevant in the pathogenesis of sporadic ALS and made their data set publicly available through the website of the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics Resource Center at the Coriell Institute (<http://ccr.coriell.org/ninds>). The data set contains 555,352 unique SNPs across the genome in 276 patients with sporadic ALS and 271 neurologically normal controls. The 555,352 SNPs are carefully chosen tagging SNPs from phase I and II of the HapMap Project. The sampled individuals are all from non-Hispanic white Americans.

We applied our TSSM to this data set. In the first step, we returned 1,000 SNPs with the smallest p-values which corresponded to use a p-value cut-off of 0.0023. Then we tested all of the $L = 499,500$ two-locus combinations under each of the seventeen models and used 1,000 permutations to evaluate the adjusted p-value for each of the two-locus combinations. Table 3 gives the best two-locus combination and its p-value. The best two-locus combination consists of rs4363506 (SNP1) on chromosome 10 and rs3733242 (SNP2) on chromosome 4. The details of the two SNPs are given in Table 4. The combination of SNP1 and SNP2 followed the Dom $\cap$ Dom model with a raw p-value of 2.766×10^{-13} and an adjusted p-value of 0.032. Table 5 gives the number of cases and controls in each of the nine genotypes for the two-locus combination. This table shows that there are more cases than controls for genotypes with at least one C allele at SNP1 and at least one G allele at SNP2 and there are more controls than cases for the other genotypes, which indicates that SNP1 and SNP2 follow the Dom $\cap$ Dom model. Figures 8 and 9 give the p-values of SNPs around the two SNPs and the overall p-values of all two-locus combinations at the second stage, respectively.

SNP1 lies within 24Kb of the dedicator of cytokinesis 1 (*DOCK1*) gene, which is important in neural outgrowth regulated by nerve growth factor (Katoh et al. 2003). SNP2 lies in Shroom3 gene that is associated with regulation and metabolism of the actin cytoskeleton. The mouse homologue of human Shroom3 is also involved at several levels in regulating cytoarchitecture (Hildebrand and Soriano 1999). The importance of the actin cytoskeleton and axonal transport in motor neuron disease is increasingly recognized (Schymick et al. 2007).

We also applied TSFM and SMT to this dataset. The results of analyzing ALS data set by these two methods are summarized in Table 3. The best SNP found by SMT is SNP1 with adjusted p-value of 0.15. Although the best two-locus combination found by TSFM is the same as that found by TSSM, the adjusted p-value of this two-locus combination using TSFM is 0.1. In summary, none of TSFM and SMT found significant results.

Discussion

In this study, we proposed a two-stage approach for GWA studies to detect two-locus combinations that have significant joint effects. In the proposed two-stage approach, we test joint effects under each of the seventeen biologically plausible models for the SNPs that show modest marginal effects. Simulation studies and application to a real GWA study show advantages of the proposed two-stage approach over a single-marker analysis and a two-stage approach based on the full model.

Our proposed two-stage approach has two simplifications. One is the “two-stage”, which means that we only search two-locus combinations that show some marginal effects. The other simplification is that we use seventeen one df models to model two-locus interactions, although the patterns of two-locus interaction are much more than seventeen. TSSM uses seventeen one df tests based on the seventeen interaction models while TSFM uses an eight df test. It is well known that the larger of df, the less power of the test will be. If the model is correct, the one df test should be much more powerful than the eight df test. Thus, we would expect that TSSM will be more powerful than TSFM if the disease model is similar to one of the seventeen interaction models. In our simulation studies, three out of four disease models used to generate data are not included in the seventeen models. However, in all the cases considered, TSSM is consistently more powerful than the single-marker analysis and TSFM. The simulation results indicate that our proposed approach has its advantages under a range of interaction patterns that may be well beyond the seventeen models. Of course, the simplifications have their disadvantages. If a two-locus combination has a strong joint effect but very small or no marginal effects, the two-stage approach will lose power in detecting the two-locus combination. Furthermore, if an interaction pattern departs a lot from the seventeen models, the approach based on the seventeen models may also lose power.

The number of markers selected at the first stage, M , is an important parameter. If M is too small, we cannot detect the two-locus combinations that show modest marginal effects. If M is too large, multiple comparisons for a large number of two-locus tests at the second stage may lose power. According to our experience, the values around 1% of the total number of markers are good choices for M . However, further investigations might be needed on finding the optimal value for M . Another thing needed to be mentioned is the computational time. The computational time heavily depends on M . For $M = 1,000$, the computational time of analyzing ALS data set using TSSM (1,000 permutations) in a PC computer (2.0 GHz CPU and 2 G memory) is about one day. Using parallel computing in our cluster computer with 64 CPUs, the analysis of ALS data set only takes about half an hour.

Currently, it has been recognized that many SNPs identified in GWA studies do not show any biological evidence that they might be causal. One reason for this phenomenon may be the false positive results due to population stratification, genotyping error, or other unknown factors. If false positive is the reason, the more markers we test, the more false positive results we can find. Thus, the problem is likely to get worse in identifying gene-gene interaction, because we usually perform more tests for identifying gene-gene interactions. For a typical GWA study with 5×10^5 markers, a single-marker analysis will perform 5×10^5 tests and our two-stage two-locus analysis, TSSM, with $M=1000$ will perform about 10^6 tests while a one-stage two-locus analysis will perform about 1.25×10^{11} tests. We can see that the chance of false positive results from TSSM is in the same order as that from a single-marker analysis and is much smaller than that from a one-stage two-locus analysis. We may have several ways to reduce the chance of false positive and make the results easy to interpret biologically. (1) Reduce the number of tests, such as the proposed two-stage two-locus analysis that reduces the number of tests of a one-stage two-locus analysis greatly. (2) Develop association tests that are robust to population stratification and genotyping

errors. Currently, we are developing a two-stage two-locus analysis that is robust to population stratification. (3) Use biological information as prior. For example, we can perform pathway-wise association tests instead of SNP-wise association tests. We can also test interactions within pathways.

Acknowledgments

This work was supported by NIH grants R01 GM069940 and the Overseas-Returned Scholars Foundation of Department of Education of Heilongjiang Province (1152HZ01).

References

- Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, et al. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat Genet.* 2006; 38:644–651. [PubMed: 16648850]
- Aston CE, Ralph DA, Lalo DP, Manjeshwar S, Gramling BA, Defreese DC, West AD, Branam DE, Thompson LF, Craft MA, Mitchell DS, Shimasaki CD, Mulvihill JJ, Jupe ER. Oligogenic combinations associated with breast cancer risk in women under 53 years of age. *Hum Genet.* 2005; 116:208–221. [PubMed: 15611867]
- Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol.* 2004; 27:141–152. [PubMed: 15305330]
- Dong C, Li WD, Li D, Price RA. Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21. *Eur J Hum Genet.* 2005; 13:102–108. [PubMed: 15470360]
- Evans DM, Marchini J, Morris AP, Cardon LR. Two-Stage Two-Locus Models in Genome-Wide Association. *PLoS. Genetics.* 2006; 2(9):1424–1432.
- Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al. A common genetic variant is associated with adult and childhood obesity. *Science.* 2006; 312:279–283. [PubMed: 16614226]
- Hildebrand JD, Soriano P. Shroom, a PDZ domain-containing actin-binding protein, is required for neural tube morphogenesis in mice. *Cell.* 1999; 99:485–497. [PubMed: 10589677]
- Hodge SE. Some epistatic two-locus models of disease.I. Relative risks and identity-by-descent distributions in affected sib pairs. *Am J Hum Genet.* 1981; 33:381–395. [PubMed: 6941695]
- International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007; 449(7164):851–861. [PubMed: 17943122]
- Katoh H, Negishi M. RhoG activates Rac1 by direct interaction with the Dock180-binding protein Elmo. *Nature.* 2003; 424:461–464. [PubMed: 12879077]
- Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet.* 2005; 37:413–417. [PubMed: 15793588]
- Millstein J, Conti DV, Gilliland FD, W. James Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet.* 2006; 78:15–27. [PubMed: 16385446]
- Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Review of Molecular Diagnostics.* 2004; 4(6):795–803. [PubMed: 15525222]
- Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 2001; 11:458–470. [PubMed: 11230170]
- Pickrell J, Clerget-Darpoux F, Bourgain C. Power of Genome-Wide Association Studies in the Presence of Interacting Loci. *Genetic Epidemiology.* 2007; 31(7):748–762. [PubMed: 17508359]
- Risch N. Linkage strategies for genetically complex traits.I. Multilocus models. *Am J Hum Genet.* 1990; 46:222–228. [PubMed: 2301392]
- Risch NJ. Search for genetic determinations in the new millenium. *Nature.* 2000; 405:847–856. [PubMed: 10866211]

- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-Dimensionality reduction reveals high-order interactions among Estrogen- Metabolism genes in sporadic breast cancer. *Am J Hum Genet.* 2001; 69:138–147. [PubMed: 11404819]
- Ritchie MD, Hahn LW, Moore JH. Power of Multifactor-Dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol.* 2003; 24:150–157. [PubMed: 12548676]
- Roldan V, Gonzalez-Conejero R, Marin F, Pineda J, Vicente V, Corral J. Five pro-thrombotic polymorphisms and the prevalence of premature myocardial infarction. *Haematologica.* 2005; 90:421–423. [PubMed: 15749685]
- Schymick JC, Scholz SW, Fung H-C, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* 2007; 6:322–328. [PubMed: 17362836]
- Sha Q, Zhu X, Zuo Y, Cooper R, Zhang S. A combinatorial searching method for detecting a set of interacting loci associated with complex traits. *Ann Hum Genet.* 2006; 70:677–692. [PubMed: 16907712]
- Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature.* 2007; 445:881–885. [PubMed: 17293876]
- Templeton, AR. Epistasis and complex trait. In: Wade, M.; Brodie, B., III; Wolf, J., editors. *Epistasis and the evolutionary process.* Oxford: Oxford University Press; 2000.
- Xiong M, Zhao J, Boerwinkle E. Generalized T^2 test for genome association studies. *Am J Hum Gene.* 2003; 70:1257–1286.
- Zhang Z, Zhang S, Wong MY, Wareham NJ, Sha Q. An ensemble learning approach jointly modeling main and interaction effects in genetic association studies. *Genetic Epidemiol.* 2008; 32:285–300.
- Zhao J, Jin L, Xiong M. Test for Interaction between Two Unlinked Loci. *Am J Hum Genet.* 2006; 79:831–845. [PubMed: 17033960]

Dom $\cap$ Dom				Dom $\cup$ Dom				Dom $\cap$ Rec				Dom $\cup$ Rec			
bb	1	1	1	bb	1	0	0	bb	1	1	1	bb	1	0	0
bB	1	0	0	bB	0	0	0	bB	1	1	1	bB	1	0	0
BB	1	0	0	BB	0	0	0	BB	1	0	0	BB	0	0	0
	aa	aA	AA		aa	aA	AA		aa	aA	AA		aa	aA	AA
Rec $\cap$ Dom				Rec $\cup$ Dom				Rec $\cap$ Rec				Rec $\cup$ Rec			
bb	1	1	1	bb	1	1	0	bb	1	1	1	bb	1	1	0
bB	1	1	0	bB	0	0	0	bB	1	1	1	bB	1	1	0
BB	1	1	0	BB	0	0	0	BB	1	1	0	BB	0	0	0
	aa	aA	AA		aa	aA	AA		aa	aA	AA		aa	aA	AA

Figure 1.

Eight two-locus epistatic models. A and B are the high-risk alleles in the two markers. The symbol in each cell denotes the relative risk.

Dom x Dom				Dom x Rec				Dom x Mul			
bb	1	θ	θ	bb	1	θ	θ	bb	1	θ	θ
bB	θ	φ	φ	bB	1	θ	θ	bB	θ	φ	φ
BB	θ	φ	φ	BB	θ	φ	φ	BB	φ	ρ	ρ
	aa	aA	AA		aa	aA	AA		aa	aA	AA
Rec x Dom				Rec x Rec				Rec x Mul			
bb	1	1	θ	bb	1	1	θ	bb	1	1	θ
bB	θ	θ	φ	bB	1	1	θ	bB	θ	θ	φ
BB	θ	θ	φ	BB	θ	θ	φ	BB	φ	φ	ρ
	aa	aA	AA		aa	aA	AA		aa	aA	AA
Mul x Dom				Mul x Rec				Mul x Mul			
bb	1	θ	φ	bb	1	θ	φ	bb	1	θ	φ
bB	θ	φ	ρ	bB	1	θ	φ	bB	θ	φ	ρ
BB	θ	φ	ρ	BB	θ	φ	ρ	BB	φ	ρ	γ
	aa	aA	AA		aa	aA	AA		aa	aA	AA

Figure 2.

Nine two-locus multiplicative models. A and B are high-risk alleles in the two markers. The symbol in each cell denotes the relative risk. $\varphi = \theta^2$, $\rho = \theta^3$ and $\gamma = \theta^4$.

Figure 3. Power comparison of the three methods (TSSM, TSFM and SMT) under four disease models and four LD patterns. The x-axis is the sample size and y-axis is power. The MAF pair is (0.2, 0.2) and the number of markers selected in the first stage is 600.

Figure 4.

Power comparison of the three methods (TSSM, TSFM and SMT) for four disease models and four LD patterns. The x-axis is the sample size and y-axis is the power. The MAF pair is (0.3, 0.4) and the number of markers selected in the first stage is 600.

Figure 5.

Power comparison of the three methods (TSSM, TSFM and SMT) for four disease models and four LD patterns. The x-axis is the sample size and y-axis is the power. The MAF pair is (0.1, 0.3) and the number of markers selected in the first stage is 600.

Figure 6.

Power comparison of the three methods (TSSM, TSFM and SMT) under different number of markers selected in the first stage. The x-axis is the number of markers selected in the first stage and y-axis is the power. The MAF pair is (0.3, 0.4), the LD between the two disease markers is 0.6, and the sample size is 1000.

Figure 7.

Power comparison of the five methods (TSSM, TSFM, TSSM1, TSFM1 and SMT) under four disease models. The x-axis is the number of markers selected in the first stage for two-stage methods (TSSM and TSFM) and y-axis is the power. The MAF pair is (0.3, 0.4), the LD between the two disease markers is 0.6, and the sample size is 500.

Figure 8.
The $-\log(p\text{-value})$ of SNPs around (1000 SNPs in front and behind) the two SNPs:
rs4363506 (SNP1) and rs3733242 (SNP2).

Figure 9.
The $-\log(\text{overall p-value})$ of all two-locus combinations (with an overall p-value less than 1) at the second stage.

Table 1

Two-locus genotype relative risk (GRR) matrices used in power comparisons.

Model	Model Name	Genotype	AA	Aa	aa
1	Heterogeneity	BB	1+a ₁	1+a ₁	1+a ₁
		Bb	1+a ₁	1	1
		bb	1+a ₁	1	1
2	Crossover	BB	1+a ₂	1+a ₂	1
		Bb	1+a ₂	1+a ₂	1
		bb	a ₁₂	a ₁₂	1
3	Additive	BB	1+4a ₃	1+3a ₃	1+2a ₃
		Bb	1+3a ₃	1+2a ₃	1+a ₃
		bb	1+2a ₃	1+a ₃	1
4	Multiplicative	BB	(1+2a ₄)(1+2a ₄)	(1+a ₄)(1+2a ₄)	1+2a ₄
		Bb	(1+a ₄)(1+2a ₄)	(1+a ₄) ²	1+a ₄
		bb	1+2a ₄	1+a ₄	1

In our simulation study, a₁=1 for Heterogeneity, a₂=1 and a₁₂=0.8 for Crossover, and a₃=a₄=0.5 for Additive and Multiplicative.

Table 2

Type I error rates of the proposed method TSSM with different values of M (the number of markers selected at the first stage) and different sample sizes.

Number of Markers	M	Sample Size		
		500	1000	2000
100	10	0.051	0.042	0.064
	20	0.048	0.041	0.06
	50	0.047	0.045	0.062
1000	10	0.042	0.056	0.05
	20	0.042	0.056	0.048
	50	0.041	0.055	0.051
	100	0.04	0.053	0.056
10000	100	0.058	0.042	0.038
	200	0.053	0.04	0.038
	400	0.053	0.04	0.042
	600	0.053	0.041	0.041

Table 3

The results of analyzing ALS data using three methods

Methods	Best marker or marker combination	Raw p-value	Adjusted p-value
TSSM	rs4363506 (SNP1), rs3733242 (SNP2)	2.766×10^{-13}	0.032
TSFM	rs4363506 (SNP1), rs3733242 (SNP2)	2.53×10^{-10}	0.1
SMT	rs4363506 (SNP1)	6.8×10^{-7}	0.15

Table 4

Information of the two SNPs

SNP	dbSNP ID	Chromosome Location	Gene	Two alleles	Allele frequency	
					Controls	Cases
SNP1	rs4363506	10q26.13	Intergenic	T	0.656	0.505
				C	0.344	0.495
SNP2	rs3733242	4q21.1	SHROOM3	A	0.467	0.341
				G	0.533	0.659

Table 5

(number of cases) / (number of controls) in each of the nine two-locus genotypes. Three individuals (one case and two controls) are deleted because of missing genotypes at these two SNPs.

SNP1	TT	TC	CC
SNP2			
AA	11/23	29/50	23/45
AG	14/37	73/56	65/24
GG	3/7	29/11	28/16