A novel strategy for detecting multiple loci in Genome-Wide Association Studies of complex diseases

Jing Li

Electrical Engineering and Computer Science Department,
Case Western Reserve University,
Cleveland, OH 44106, USA
E-mail: jingli@eecs.case.edu

Abstract: Large-scale Genome-Wide Association Studies (GWAS) for complex diseases are increasingly common, due to recent advances in genotyping technology. Gene-gene interactions play an important role in the etiology of complex diseases and have to be addressed in GWAS. In this paper, an efficient strategy based on two-stage analysis is proposed. It combines a single-locus approach with a Goodness-Of-Fit (GOF) test in stage one, and selects a promising subset of SNPs to be modelled using a full interaction model in stage two. Extensive simulations using different disease models with different levels of epistasis demonstrate that it achieves higher power than existing approaches.

Keywords: Genome-Wide Association Studies; GWAS; gene-gene interactions; two-stage analysis; case-control studies; bioinformatics.

Reference to this paper should be made as follows: Li, J. (2008) 'A novel strategy for detecting multiple loci in Genome-Wide Association Studies of complex diseases', *Int. J. Bioinformatics Research and Applications*, Vol. 4, No. 2, pp.150–163.

Biographical notes: Jing Li received his PhD in Computer Science from the University of California at Riverside in 2004. He joined Case Western Reserve University as an Assistant Professor of Electrical Engineering and Computer Science in August 2004. His research interest is in the area of computational biology and bioinformatics. More specifically, he mainly focuses on the design and development of efficient computational and statistical algorithms for the characterisation of DNA variations in human populations, and for the identification of the correlations of DNA variations and phenotypic differences such as diseases.

1 Introduction

With the completion of the human genome project and the HapMap project and with advances of Single Nucleotide Polymorphism (SNP) genotyping technology, GWAS for complex diseases are finally feasible (Hirschhorn and Daly, 2005; Marchini et al., 2005) and results from several large scale GWAS have been reported in the

literature (Hampe et al., 2006; Hu et al., 2005; Ozaki and Tanaka, 2006). Most existing methods for GWAS are single-locus based approaches, which examine one SNP at a time. However, empirical data have shown that many complex diseases may actually involve multiple genes, and gene-gene interactions play an important role in the etiology of complex diseases. Single-locus based methods usually are unable to recover all involved loci, especially when individual loci have little or no marginal effects, which is not uncommon for many gene-gene interaction models (Culverhouse et al., 2002; Hoh and Ott, 2003; Millstein et al., 2006).

Traditionally, estimating epistatic effects has been investigated mainly in the context of quantitative trait mapping of plants or animals (e.g., Zhang and Xu, 2005; Xu and Jia, 2007). In recent years, data mining approaches have also been adopted in case-control studies. Hoh and Ott (2003) provided an excellent review. In the context of genome-wide association studies with hundreds of thousands of markers, Marchini et al. (2005) have shown that explicitly modelling of interactions between loci is computationally feasible. And it can actually achieve reasonably high power with realistic sample sizes under three interaction models with some marginal effects, even after adjustments of multiple testing using a very conservative approach (Bonferroni correction). However, directly modelling of interactions is still computationally demanding and it can hardly be extended to include more than two loci. On the other hand, a two-stage analysis strategy, for which a small subset of promising loci is identified in the first stage and multi-locus methods are used in the second stage to model interactions based on the selection in the first stage, is promising. In the same paper (Marchini et al., 2005), results have also shown that one particular two-stage strategy called simultaneous approach, in which only loci showing moderate associations under a single-locus model will be subsequently tested using a full two-locus model, has almost identical power as the full model using all SNP pairs. Another commonly used two-stage approach in practice (called conditional two-stage approach), where combinations between SNPs selected from a single-locus model and original SNPs will be tested using a full model, was not evaluated in their paper. In a more recent study, Lonita and Man (2006) have compared the two two-stage approaches using a different set of disease models, and have concluded that the conditional two-stage approach is more powerful and more robust than the simultaneous approach. Unfortunately, no direct comparisons between the conditional two-stage approach and the full model on all pairs under the same set of disease models have been made in either of the two papers (Lonita and Man, 2006; Marchini et al., 2005). Furthermore, one underlying assumption for both studies is that some marginal effects must exist at either or both loci. Otherwise, none of the two two-stage strategies will work.

In this paper, I extend the conditional two-stage approach by incorporating a GOF test in stage one. The modified strategy, called Combined Conditional Two-Stage Test (CCTST) can effectively identify a promising subset of SNPs for subsequent tests, regardless of the magnitude of marginal effects. The GOF test examines deviations of observed multi-locus genotype counts from the expected counts based on single locus genotypes without separating cases from controls, and it implicitly assumes such deviations are due to deviations in cases because of different penetrances (probabilities of being affected) from different genotypes. The framework can easily incorporate any single-locus based approaches in stage one and any tests for interactions in stage two, and it can be extended to multiple stages and/or multiple loci. To evaluate the

performance of CCTST, three two-locus disease models have been chosen ranging from a pure additive model without epistasis to a model with little or no marginal effects. Simulations show that CCTST outperforms other four commonly used or recently proposed searching strategyies, and achieves highest power in detecting both loci.

2 Methods

2.1 Disease models

The most general form of two-locus models for diallelic loci consists of nine parameters, one for each genotype combination. There are also many restricted models with less than nine free parameters that are of great interests. By focusing on fully penetrant models (the probability of an individual being affected is either 1 or 0 for any given two-locus genotype), Li and Reich (2000) enumerated all the 512 possible combinations and summarised 50 unique ones. With incomplete penetrances, the possible number of models becomes infinite. Three submodels are selected in this study based on two criteria, the level of epistasis and evidences from empirical studies. To test different approaches under a variety of epistatic effects, the three models represent low/none, medium and high level of epistasis. More detailed discussions of these models and others, as well as references to some diseases that confer these theoretical models can be found in Li and Reich (2000). Tables 1–3 present the three models by specifying the penetrance of each genotype combination at two unlinked loci A and B. Based on their definitions, the marginal penetrance for each genotype at each individual locus and the population prevalence are also included in the tables. They are calculated based on individual penetrances and allele frequencies under the assumption of Hardy-Weinberg equilibrium. Let A/a (B/b) denote the two allele at locus *A* (*B*). Let p_1 and p_2 denote the frequencies of alleles A and B, respectively. Let $g_{i.}$ ($i = 1, 2, 3$) denote the three genotypes (AA, Aa, aa) at locus A and $g_{.j}$ ($j = 1, 2, 3$) denote the three genotypes at locus B. Let $g_{i,j}$ denote the two-locus genotype combinations. Similarly, let $\lambda_{i.}, \lambda_{.j}$ and $\lambda_{i,j}$ denote the marginal penetrance for $g_{i.}$ at locus *A*, the marginal penetrance for $g_{.j}$ at locus *B*, and the penetrance for the two-locus genotype $g_{i,j}$, respectively. By definition,

$$\lambda_{i.} = \sum_j \lambda_{i,j} Pr(g_{.j}), \quad \lambda_{.j} = \sum_i \lambda_{i,j} Pr(g_{i.}), \tag{1}$$

$$p = \sum_{i,j} \lambda_{i,j} Pr(g_{i,j}), \tag{2}$$

where p is the population prevalence, and genotype frequencies $Pr(g_{i.})$ and $Pr(g_{.j})$ can be calculated based on allele frequencies under Hardy-Weinberg equilibrium. The frequency of $Pr(g_{i,j})$ can be calculated from genotype frequencies of individual loci.

Model one (M1) is a pure additive model without epistasis. An individual with a genotype $aabb$ has a baseline probability η of being affected. The parameter θ represents the level of increasing risk of having one or more disease associated alleles A and/or B in a genotype. More specifically, for each of all other genotype combinations in M1, the probability of being affected increases by $\eta\theta$ with each additional disease-associated allele, from either locus A or locus B. Based on marginal

penetrances, it is easy to see that each individual locus is also an additive single-locus model. Model two (M2) is a threshold model that has been widely studied by different groups (Lonita and Man, 2006; Marchini et al., 2005). It requires at least one copy of a disease-associated allele from each locus for the corresponding genotype to have a higher penetrance. On the other hand, having more disease-associated alleles does not increase the chance of being affected. Each individual locus is a dominant single-locus model and the marginal effect actually depends on the allele frequencies at the other locus. The third model (M3) is an epistatic model and has been investigated by Culverhouse et al. (2002) and some other researchers. Based on this model, an individual has a higher risk only if it has one of the three genotypes ($AAbb, AaBb, aaBB$). This model is of particular interest because when the allele frequency at one locus is 0.5, no marginal effects exist at the other locus, regardless of the size of interaction effect θ. The three models are chosen so that algorithms can be tested across various degrees of epistasis.

Table 1 Penetrance table for an additive model (M1), where each entry is the value of penetrance (λ) of a particular genotype, characterised using a function of the baseline penetrance η (for genotype $aabb$) and a parameter θ (representing the level of increasing risk of having one or more disease associated alleles A or B in a genotype). More specifically, the middle entries are penetrances for two-locus genotypes, and the last column and the last row are penetrances of one locus genotypes with the last entry the population prevalence. p_1 and p_2 are frequencies of allele A and B, respectively

	BB	*Bb*	*bb*	
AA	$\eta(1+4\theta)$	$\eta(1+3\theta)$	$\eta(1+2\theta)$	$\eta(1+2p_2\theta+2\theta))$
Aa	$\eta(1+3\theta)$	$\eta(1+2\theta)$	$\eta(1+\theta)$	$\eta(1+2p_2\theta+\theta))$
aa	$\eta(1+2\theta)$	$\eta(1+\theta)$	η	$\eta(1+2p_2\theta)$
	$\eta(1+2p_1\theta+2\theta)$	$\eta(1+2p_1\theta+\theta)$	$\eta(1+2p_1\theta)$	$\eta(1+2\theta(p_1+p_2))$

Table 2 Penetrance table for a threshold model (M2)

	BB	*Bb*	*bb*	
AA	$\eta(1+\theta)$	$\eta(1+\theta)$	η	$\eta(1+\theta(1-(1-p_2)^2))$
Aa	$\eta(1+\theta)$	$\eta(1+\theta)$	η	$\eta(1+\theta(1-(1-p_2)^2))$
aa	η	η	η	η
	$\eta(1+$ $\theta(1-(1-p_1)^2))$	$\eta(1+$ $\theta(1-(1-p_1)^2))$	η	$\eta(1+\theta(1-(1-p_1)^2)$ $(1-(1-p_2)^2)))$

2.2 Simulations with realistic marginal effects at both loci

It is a well accepted fact that gene-gene interactions account, at least partially, for many unsuccessful stories in mapping and replicating susceptibility genes for complex traits. However, little information is available about the nature and mechanisms of gene-gene interactions, not to mention the magnitudes of joint effects (Marchini et al., 2005), partially due to the lack of efficient and effective algorithms in dealing with the multiple testing problem arising from testing interactions directly. On the other hand,

increasing information on sizes of marginal effects is available. Marchini et al. (2005) fixed the marginal effect at one locus and derived the parameters for the joint effect. But the marginal effect at the other locus was not controlled. In reality, it is more likely that marginal effects at both loci are small. To closely mimic real data, the magnitudes of marginal effects at both loci are bounded using realistic empirical values for all three models in this study. In general, there are two genotype relative risk parameters for each locus, namely, $\lambda_{1.}/\lambda_{2.}$ and $\lambda_{2.}/\lambda_{3.}$ at locus A, and $\lambda_{.1}/\lambda_{.2}$ and $\lambda_{.2}/\lambda_{.3}$ at locus *B*. Let γ denote the maximum allowed marginal genotype relative risk for both loci. Based on empirical data about complex diseases, it takes a value in the range of 1.2–1.5 in this study. For each possible value of γ, and for any fixed allele frequencies p_1 and p_2, the parameter θ is determined by its possible maximum value that satisfies the following conditions:

$$\lambda_{1.}/\lambda_{2.} \leq \gamma, \qquad \lambda_{2.}/\lambda_{3.} \leq \gamma, \qquad \lambda_{.1}/\lambda_{.2} \leq \gamma, \qquad \lambda_{.2}/\lambda_{.3} \leq \gamma.$$

It is interesting to notice that in certain areas of the allele frequency spectrum of model 3, the size of marginal effects can not reach γ regardless of the value of θ. For example, when $p_1 = p_2 = 0.4$, the marginal effects $(\lambda_{1.}/\lambda_{2.}, \lambda_{2.}/\lambda_{3.}, \lambda_{.1}/\lambda_{.2}, \lambda_{.2}/\lambda_{.3})$ based on Table 3 are always less than 1.5. In such cases, a small fixed value $\theta = 0.25$ is used in the simulation. Therefore it is possible that the final assignment of the joint effect θ with a greater γ may be less then the assignment of θ with a smaller γ. The baseline value η is then determined using equation (2) for any fixed population prevalence p. I take model 2 as an example to illustrate the relationship of the joint effect, marginal effects, and allele frequencies. Based on Table 2, each individual locus itself is a dominant single-locus model and only one genotype relative risk needs to be considered at each locus. Once γ, p_1, p_2 and p are fixed, θ and η can be calculated via

$$\theta = argmax((1 + \theta(1 - (1 - p_1)^2)) \leq \gamma, \qquad (1 + \theta(1 - (1 - p_2)^2)) \leq \gamma)$$

and

$$\eta = p/(1 + \theta(1 - (1 - p_1)^2)(1 - (1 - p_2)^2))).$$

Table 3 Penetrance table for a model with strong epistatic effects (M3)

	BB	*Bb*	*bb*	
AA	η	η	$\eta(1 + 4\theta)$	$\eta(1 + 4\theta(1 - p_2)^2)$
Aa	η	$\eta(1 + 2\theta)$	η	$\eta(1 + 4\theta(1 - p_2)p_2)$
aa	$\eta(1 + 4\theta)$	η	η	$\eta(1 + 4\theta p_2^2)$
	$\eta(1 + 4\theta(1 - p_1)^2)$	$\eta(1 + 4\theta(1 - p_1)p_1)$	$\eta(1 + 4\theta p_1^2)$	$\eta(1 + 4\theta(p_1 + p_2 - 2p_1p_2))$

Table 4 presents corresponding values of θ and the actual marginal effect at loci *A* and *B* given allele frequencies at loci *A* and *B* for a fixed $\gamma = 1.5$. One can easily see that allele frequencies can greatly influence the magnitude of joint effect θ for a fixed maximum marginal effect. A much higher joint effect is required to accommodate low disease-associated allele frequencies in order to keep a fixed maximum marginal effect. This is equivalent to say that, when θ is fixed, the ability of detecting each individual locus greatly depends on the frequency of the disease-associated allele at the other locus (see Table 2). The parameters of the other two models can be determined similarly.

Table 4 Joint effect θ, marginal effects at loci A and B when γ (= 1.5) and p (=0.1) are fixed for various allele frequencies for the threshold model (M2)

$A \diagdown B$	*0.1*	*0.3*	*0.5*
0.1	(2.63, 1.50, 1.50)	(0.98, 1.50, 1.19)	(0.67, 1.50, 1.13)
0.2	(1.39, 1.26, 1.50)	(0.98, 1.50, 1.35)	(0.67, 1.50, 1.24)
0.3	(0.98, 1.19, 1.50)	(0.98, 1.50, 1.50)	(0.67, 1.50, 1.34)
0.4	(0.78, 1.15, 1.50)	(0.78, 1.40, 1.50)	(0.67, 1.50, 1.43)
0.5	(0.67, 1.13, 1.50)	(0.67, 1.34, 1.50)	(0.67, 1.50, 1.50)

2.3 Commonly used testing strategies

The most commonly used searching strategies for susceptibility genes of complex diseases for GWAS are still single-locus based tests. At each locus, a logistic regression model or Pearson's χ^2 test for independence can be performed, and Bonferroni correction is commonly used to adjust the overall significance level. Parallel to the standard quantitative genetics model (Cordell, 2002), the full single-locus model under the logistic regression framework is

$$\log(\lambda/(1-\lambda)) = \mu_0 + ax + dz,$$

where μ_0, a, d are genetic parameters representing mean, additive and dominance effects, and x and z are dummy variables with $x = 1$ for genotype AA, 0 for Aa, and -1 for aa, and $z = -0.5$ for AA and aa, and 0.5 for Aa. The log likelihood ratio test comparing the full single-locus model and the null model ($a = d = 0$) is used to test the significance of the model. For a χ^2 independence test of case-control samples, a 3×2 contingency table can be constructed, and the expected counts and the observed counts from each category of genotype and phenotype combination are compared (Collett, 1999).

Both approaches can be readily extended to the two-locus case to construct a fully saturated model. In addition to the mean, additive and dominance effects at both loci, the logistic regression model has four interaction terms (i_{aa} for additive $\times$ additive, i_{ad} for additive $\times$ dominance, i_{da} for dominance $\times$ additive, and i_{dd} for dominance $\times$ dominance),

$$\begin{aligned} log(\lambda/(1-\lambda)) = \mu_0 &+ a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 \\ &+ i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2. \end{aligned}$$

The χ^2 test for two loci is calculated based on a 9×2 contingency table.

A third commonly used strategy is based on a two-stage analysis and has two variants. For both variants, a small subset of promising loci is identified based on a single-locus method in the first stage. In the second stage, a two-locus model will be applied either on each pair of loci in the selected subset only (this variant is named the simultaneous approach), or on pairwise combinations between the selected loci and the original set of loci (this variant is named the conditional approach). The conditional approach is frequently used in genetic studies and has been shown more robust and powerful than the simultaneous approach (Lonita and Man, 2006). Therefore the conditional approach has been chosen in this study. Notice that the two-stage analysis approach is different from a two-stage design for which additional samples will be

recruited in the second stage. Because of the sequential testing nature of any two-stage analysis, obtaining the correct significance level is not a trivial task. When the logistic regression framework is applied in both stages, a conservative approach to account for the 'selection bias' has been proposed by Marchini et al. (2005) and has been used in Lonita and Man (2006). More specifically, let α_1 be the significance level in stage one and let k_{α_1} be the critical value that corresponds to α_1 for the log likelihood ratio test of a single locus. For the conditional approach, let R be the statistic of the log likelihood ratio test of the two-locus model for a pair of loci i and j, where locus i has been selected from stage one and j ($j \neq i$) is any of other SNPs. The new defined statistic for the overall approach is just $R - k_{\alpha_1}$ and its significance is assessed against a χ^2 distribution (Lonita and Man, 2006; Marchini et al., 2005). When Pearson's χ^2 test is applied in both stages, a similar correction can be adopted here. Let t_{α_1} be the critical value that corresponds to α_1 for the χ^2 test using a single-locus model and let T be the statistic of the χ^2 test using a two-locus model, the new statistic is defined as $T - t_{\alpha_1}$ and its significance is evaluated against the χ^2 distribution. Simulations (data not shown) suggest that it has the correct type one error rate under the null hypothesis of no associations.

Each of the above three strategies has its own advantages and disadvantages. Single-locus based methods are computationally efficient and easy to perform. The power to detect all involved loci is low because some of them may have little marginal effects. The full two-locus model is powerful when the model is correct and when only epistatic effects exist. But the computational cost is high. Furthermore, because of the problem of multiple testing, the overall significance level across the genome has to be adjusted and Bonferroni correction is commonly used in practice. The total number of tests for a full two-locus model is much greater than the number of tests from single-locus based methods. For example, if the number of SNPs is m, the number of tests is $\binom{m}{2}$ for a two-locus model, vs. m for a single-locus model. The conditional two-stage approach lies in between. It has the potential to detect all involved loci but with much reduced computational costs. With a liberal threshold in stage one, it is possible that loci with small marginal effects can pass the screen. The number of tests greatly reduces to $l \times (m - l)$, where l is the number of SNPs identified in stage one with its expectation $E(l) = m \times \alpha_1$. Furthermore, this is the only strategy that can be extended to directly model complex diseases involving more than two loci, for any realistic sample sizes.

2.4 The proposed strategy

One of the drawbacks of the conditional two-stage approach is that when marginal effects are extremely weak or do not exist, the approach is unlikely to work, even for a liberal threshold α_1 in the first stage. I propose a new strategy called CCTST, which is an improved version of the original conditional approach. In addition to the screen based on each individual locus, it directly assesses pairs of loci in hope that some promising pairs will be prompted into stage two even when marginal effect sizes from individual loci are weak.

More specifically, the preprocessing step based on marginal effects of single loci is performed first, as it was in stage one for the original conditional approach. In addition, another procedure is proposed to preprocess all possible SNP pairs in stage one in order to identify promising pairs that might have little marginal effects.

Under the assumption that the joint effect of two disease susceptibility genes/SNPs is mainly from epistasis, the multilocus genotype distribution among cases will be different from the expected distribution that can be estimated based on genotype frequencies of each individual locus. The magnitude of such differences will depend on the differences of penetrances from different genotypes. However, any test with explicit use of disease status is almost the same as performing the test on all pairs of SNPs thus must be avoided. I argue that in the case that the epistatic effect is great enough, say the penetrance λ_{ij} of genotype G_{ij} is very different from the population prevalence λ, the total number of individuals with genotype G_{ij} from the pooled case and control samples might also deviate from its expected value greatly, because of excessive sampling of cases. Certainly, the magnitude of the deviation will depend on the size of the epistatic effect, as well as the genotype/allele frequencies. Based on this observation, a χ^2 GOF test that compares the observed genotype count $n_{G_{i,j}}$ for $G_{i,j}$ and its expected count $E(n_{G_{i,j}})$, which is calculated based on genotype counts at each individual locus in the combined case and control samples, can be used to pre-screen all pairs of SNPs:

$$W = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(n_{G_{i,j}} - E(n_{G_{i,j}}))^2}{E(n_{G_{i,j}})},$$

where $E(n_{G_{i,j}}) = n_{i.}n_{.j}/n$, $n_{i.}$ and $n_{.j}$ are the numbers of individuals with genotype $g_{i.}$ and $g_{.j}$ respectively, and n is the total number of samples. Only pairs with a statistic that exceeds a predefined threshold (for example, a point-wise significance level of α_2) will be further examined in stage two. Notice that unlike the single SNP screen, the χ^2 GOF test does not introduce biases for subsequent tests in stage two. Under the null hypothesis of no associations between genotypes and the disease, the above GOF test is independent of disease status because it only uses information about genotypes. Therefore, no adjustments are needed in the second stage in testing the selected subset from the GOF test. The same argument and a similar test have been proposed in a recent paper Millstein et al. (2006). The new proposed method differs from the method in Millstein et al. (2006) because it does not assume an underlying model of risk, which is usually unknown in practice.

Under the assumption of strong pairwise interactions, CCTST can potentially identify both loci regardless of the magnitude of marginal effects. And the framework can be easily extended to include interactions from more than two loci and to include more than two stages. The number of pairs needs to be examined by a full model in stage two is much smaller than the number of all possible pairs. Assume a significance level α_1 for the single SNP screen and α_2 for the GOF screen, the expected number of pairs that will be prompted to stage two will be bounded from above by $m\alpha_1 m(1-\alpha_1) + \binom{m}{2}\alpha_2$ because some pairs will be selected by both screens.

3 Results

3.1 Simulation details

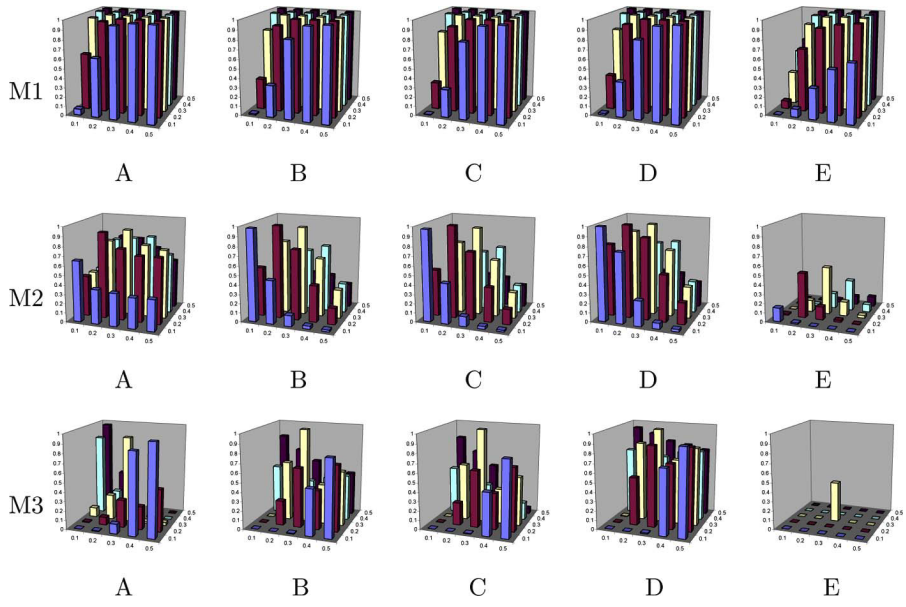
For each of the three disease models, simulated genotype data at two unlinked marker loci have been generated using the program *gs* (Li and Chen, 2008) under a variety of parameter values, assuming a population-based genome wide association study design

with $m = 300,000$ markers. The maximum value of marginal effects γ takes realistic values ($\gamma = 1.2, 1.3, 1.4, 1.5$). The disease allele frequencies at both loci vary from 0.1 to 0.5. The population prevalence p is fixed as 0.1. The penetrance table can then be calculated based on previous discussions. The linkage disequilibrium r^2 between marker loci and unobservable disease loci varies from 0.6 to 1.0. Equal numbers of cases and controls have been generated with $n = 1000, 1500, 2000$. For each parameter combination, 1000 random replicates are utilised to compare the power of five strategies, namely,

- the single SNP test, detecting either of the two loci
- the full interaction model
- the conditional two-stage test
- the new proposed method CCTST
- the single SNP test, detecting both loci.

The logistic regression method is used in testing of single and two-locus models and log-likelihood ratio tests are performed comparing full models vs. the null model.

Figure 1 Power comparison of the five strategies across the three models. Each row is an interaction model and each column is one strategy. The five strategies are: (A) single-locus (either); (B) full two-locus model; (C) conditional approach; (D) CCTST and (E) single-locus (both). Within each panel, the x and y axes are risk allele frequencies at the two interaction loci, and the z axis is the power for each strategy. The bound on the size of marginal effects $\lambda = 1.5$. For model one, the samples consist of 1000 cases and 1000 controls. For the other two models, there are 1500 cases and 1500 controls. The genome wise significance threshold is 0.05 after Bonferroni corrections. For the two-stage tests, the initial threshold $\alpha_1 = 0.05$ and the threshold for the goodness of fit test is $\alpha_2 = 0.05$



3.2 Results

Under the null hypothesis of no associations, simulations indicate that the new proposed approach has the correct type one error rate (data not shown). A subset of representative results is presented in Figures 1–3 to illustrate some general features of power of the five strategies across different interaction models. The subset is selected so that not all the five strategies have extremely high or extremely low power across the parameter space. A notable result from Figure 1 is that the interaction model has great influence on power of any method, even when the maximum value of marginal effect size is fixed. For example, for the additive model, all five methods achieve high power when the frequency of at least one of the risk alleles is high ($\geq$0.3). But with increase of epistatic effects from model 1 to model 3, the power of all methods decreases in most of the spectrum of allele frequencies, even with increase of sample sizes from 1000 in M1 to 1500 in M2 and M3. Not surprisingly, the decrease in power is more serious in the two single-locus based strategies (methods A and E). For a fixed disease model, risk allele frequencies can also greatly affect power in a specific way that depends on the model. For example, for model 2, all the methods achieve higher power when the frequencies of risk alleles from two loci are the same.

Among the five strategies, the method to detect both loci simultaneously based on single-locus screens (method E) has least power across models and allele frequencies. The problem is more serious when marginal effects are small. Strategy A (the single-locus test to detect either SNP) has high power in many cases, but with the limitation of detecting one of the two sites. Furthermore, its power decreases with increase of epistatic effects and essentially becomes no power when there only epistatic effects exist. In those cases, a full interaction model (strategy B) may actually achieve higher power than strategy A. For the other three strategies (B–D), the conditional method (C) have slightly lower power than the full interaction model (B) in many settings (Figures 1 and 2). This result is consistent with results obtained by other researchers using different models (Lonita and Man, 2006; Marchini et al., 2005). For example, results in Marchini et al. (2005) have shown that the full model has equivalent power comparing with the simultaneous two-stage approach, while results in Lonita and Man (2006) have shown the conditional two-stage approach is more powerful and more robust than the simultaneous method. However, the power of strategy (C) is much lower than that of strategy (B) when marginal effects are small, for example, in the cases of Model 3 when allele frequencies at both loci are great than 0.4 (Figure 2). On the other hand, the new proposed CCTST (strategy (D)), is consistently better than both strategy B and strategy (C) across all parameters tested in this study. All three methods have similar power when the signal is extremely strong or extremely weak, otherwise, CCTST achieves much higher power than the other two approaches (Figure 2). For example, when allele frequencies at both loci are 0.5 in model 3, the power of CCTST (73.4%) represents a 10-fold increase comparing with the original conditional approach (6.2%). It also represents a 25% increase in power comparing with the full two-locus model (46.7%). The gains in power by CCTST reflect that it can balance well the magnitude of marginal effects and the number of tests for the full model in stage two. When there are some marginal effects, which is probably true for many models, the pairs being prompted to stage two by CCTST are mostly from the single-locus screen. When there are little marginal effects, the pairs being prompted to stage two by CCTST are mostly from the GOF screen. In these cases, CCTST

and strategy C have similar performances and both of them have higher power than strategy B. The results suggest that not only is CCTST computationally appealing, it is also more robust than those commonly used strategies.

Figure 2 Power comparison of the three strategies (B, C, D) across the three models. The bound on the size of marginal effects $\lambda = 1.4$ and all other parameters are the same as those in Figure 1

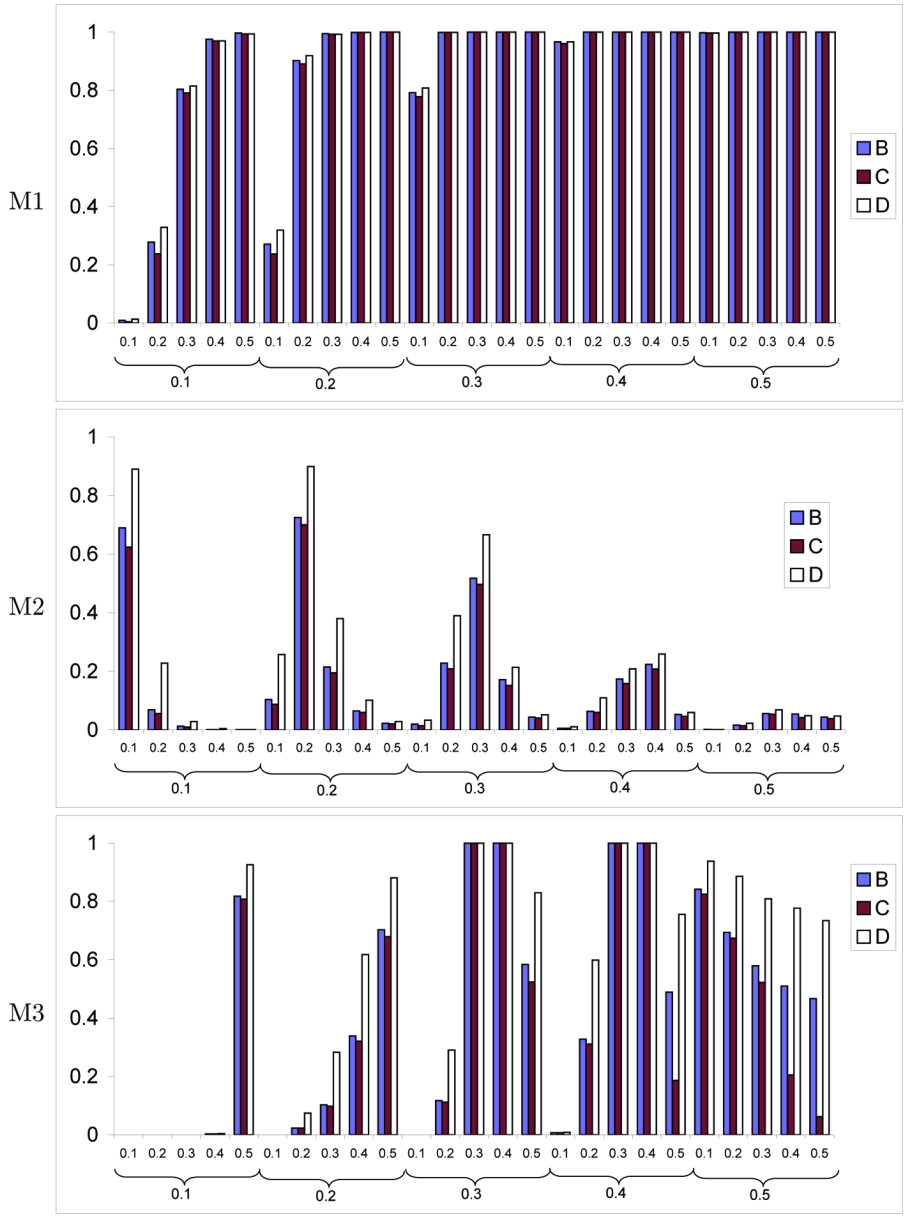
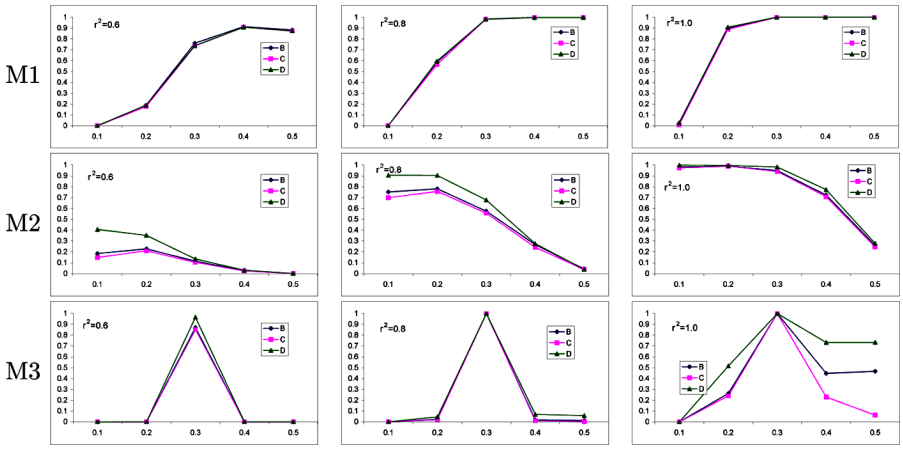


Figure 3 illustrates the power of the three methods when the LD (r^2) between markers and disease loci varying from 0.6 to 1. Not surprisingly, the power of all three methods

decreases with the decay of LD. But the magnitude of decrease not only depends on r^2, but also depends on genetic models and allele frequencies. Models with high epistatic effects are more easily being adversely affected by dropping of LD. For the three methods, (strategies B–D), they have similar power regardless of LD values under the additive model. For both the threshold model (M2) and the epistatic model (M3), the new proposed strategy (D) has greater power than both strategies (B) and (C). However, the advantage of strategy (D) is most noticeable when LD are small for M2 and when LD $r^2 = 1$ for M3. A possible explanation is that all three methods have relatively high power for M2 when $r^2 = 1$ and have extremely low power for M3 when $r^2 \neq 1$ for most allele frequencies.

Figure 3 Statistical power of the three methods when markers and disease-associated loci are in different levels of LD. All the markers have the same allele frequencies as those unobserved disease loci, and markers at the two loci also have the same frequencies. The bound on the size of marginal effects $\lambda = 1.5$. There are 1500 cases and 1500 controls for model one, and 2000 cases and 2000 controls simulated for models 2 and 3



4 Conclusion

Identification of genetic risks underlying complex diseases is a great challenge, mainly because many involved genes have small individual effects, among other reasons. Computational approaches to model gene-gene interactions are greatly needed. In this paper, I demonstrate that a new proposed strategy based on a two-stage analysis is quite robust and more powerful than several existing strategies for three models from pure additive to pure epistatic. It is not my intention to claim that this is the best strategy, because the properties of each method greatly depend on the form of interactions and there are many different types of interactions. But the new proposed framework is appealing not only because it has great performance in the experiments, but it also is easy to extend to models with multiple loci.

There are several ways in which the proposed framework can be extended. As pointed out in Marchini et al. (2005), obtaining the significance level in such a sequential test framework is not trivial. And the problem is coupled with the multiple testing problem in the context of genome wide association studies. The corrections

for both problems in this study are conservative, and more sophisticated approaches can further improve the power of the proposed framework. The extension to multiple stages is straightforward and haplotype effects can be considered at some point within a multi-stage analysis framework. Given the large number of SNPs for GWAS, some datamining approaches (Hoh and Ott, 2003) can be adopted here to preprocess data so that formal statistical tests can be performed to detect high order interactions. Knowledge on biological pathways and genomic data should also be incorporated in a multi-stage analysis.

Acknowledgements

The author gratefully acknowledges the support of K.C. Wong Education Foundation, Hong Kong. Research supported by NIH/NLM grant LM008991 and a US Public Health Service Resource grant (RR03655) from the National Center for Research Resources.

References

Collett, D. (1999) *Modelling Binary Data*, Chapman & Hall/CRC.

Cordell, H.J. (2002) 'Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans', *Hum. Mol. Genet.*, Vol. 11, pp.2463–2468.

Culverhouse, R., Suarez, B.K., Lin, J. and Reich, T. (2002) 'A perspective on epistasis: limits of models displaying no main effect', *Am. J. Hum. Genet.*, Vol. 70, pp.461–471.

Evans, D.M., Marchini, J., Morris, A.P. and Cardon, L.R. (2006) 'Two-stage two-locus models in genome-wide association', *PLoS Genet*., p.2.

Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., Gunther, S., Prescott, N.J., Onnie, C.M., Hasler, R., Sipos, B., Folsch, U.R., Lengauer, T., Platzer, M., Mathew, C.G., Krawczak, M. and Schreiber, S. (2006) 'A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1', *Nat. Genet.*, Vol. 39, pp.207–211.

Hirschhorn, J.N. and Daly, M.J. (2005) 'Genome-wide association studies for common diseases and complex traits', *Nat. Rev. Genet.*, Vol. 6, pp.95–108.

Hoh, J. and Ott, J. (2003) 'Mathematical multi-locus approaches to localizing complex human trait genes', *Nat. Rev. Genet*., Vol. 4, pp.701–709.

Hu, N., Wang, C., Hu, Y., Yang, H.H., Giffen, C., Tang, Z.Z., Han, X.Y., Goldstein, A.M., Emmert-Buck, M.R., Buetow, K.H., Taylor, P.R. and Lee, M.P. (2005) 'Genome-wide association study in esophageal cancer using GeneChip mapping 10K array', *Cancer Res*., Vol. 65, pp.2542–2546.

Li, J. and Chen, C. (2008) 'Generating samples for association studies based on HapMap data', *BMC Bioinformatics*, Vol. 9, p.44.

Li, W. and Reich, J. (2000) 'A complete enumeration and classification of two-locus disease models', *Hum. Hered.*, Vol. 50, pp.334–349.

Lonita, L. and Man, M. (2006) 'Optimal two-stage strategy for detecting interacting genes in complex diseases', *BMC Genet*., Vol. 7, p.39.

Marchini, J., Donnelly, P. and Cardon, L.R. (2005) 'Genome-wide strategies for detecting multiple loci that influence complex diseases', *Nat. Genet.*, Vol. 37, pp.413–417.

Millstein, J., Conti, D.V., Gilliland, F.D. and Gauderman, W.J. (2006) 'A testing framework for identifying susceptibility genes in the presence of epistasis', *Am. J. Hum. Genet.*, Vol. 78, pp.15–27.

Ozaki, K. and Tanaka, T. (2006) 'Genome-wide association study to identify single-nucleotide polymorphisms conferring risk of myocardial infarction', *Methods Mol. Med.*, Vol. 128, pp.173–180.

Xu, S. and Jia, Z. (2007) 'Genomewide analysis of epistatic effects for quantitative traits in barley', *Genetics*, Vol. 175, pp.1955–1963.

Zhang, Y-M. and Xu, S. (2005) 'A penalized maximum likelihood method for estimating epistatic effects of QTL', *Heredity*, Vol. 95, pp.96–104.