Mapping Haplotype-haplotype Interactions with Adaptive LASSO

Ming Li[1], Roberto Romero[3], Wenjiang J. Fu[1*], and Yuehua Cui[2*]

[1]Department of Epidemiology, and [2]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824

[3]The Perinatology Research Branch, NICHD, NIH, DHHS, Bethesda, MD, and Detroit, 48201

* Corresponding author

Email addresses:

ML: liming@msu.edu

RR: prbchiefstaff@med.wayne.edu

WJF: fuw@epi.msu.edu

YC: cui@stt.msu.edu

Abstract

Background: The genetic etiology of complex human disease has been commonly viewed as a complex process involving both genetic and environmental factors functioning in a complicated manner. Quite often the interactions among genetic variants play major roles in determining the susceptibility of an individual to a particular disease. Statistical methods for modeling interactions underlying complex diseases between single genetic variants (e.g. SNPs) have been extensively studied. Recently, haplotype-based association analysis gains its popularity in genetic association study. When multiple sequence or haplotype interactions are involved in determining an individual's susceptibility to a disease, it presents daunting challenges in statistical modeling and testing of the interaction effects, due to the complicated higher order epistatic complexity.

Results: In this article, we propose a new strategy in modeling haplotype-haplotype interactions under the penalized logistic regression framework with adaptive L_1-penalty. We consider interactions of sequence variants between haplotype blocks. The adaptive L_1-penalty allows simultaneous effect estimation and variable selection in a single model. We proposed a new parameter estimation method which estimates and shrinks parameters by the modified Gauss-Seidel method nested within the EM algorithm. Simulation studies showed low false positive rate and reasonable power in detecting haplotype interactions. The method was applied to test haplotype interactions involved in mother and offspring genome in a small for gestational age (SGA) neonates data set, and significant interactions between different genomes were detected.

Conclusions: Demonstrated by the simulation studies and real data analysis, the developed approach provides an efficient tool for the modeling and testing of haplotype interactions.

Availability: The implementation of the method in R codes can be freely downloaded from http://www.stt.msu.edu/~cui/software.html

Key words: Gauss-Seidel algorithm, Penalized logistic regression, Risk haplotype, Single nucleotide polymorphism

Background

It has been commonly recognized that most human diseases are complex involving joint effort of multiple genes, complicated gene–gene as well as gene–environment interactions [1]. The identification of disease risk factors for monogenic diseases has been quite successful in the past. Due to the small effect of many single genetic variants on the risk of a disease, the identification of disease variants for complex multigenic diseases has not been very successful [2]. There are multiple reasons for this. First, most complex diseases involve multiple genetic variants each conferring a small or moderate effect on a disease risk. Second, the complexity relies on the complicated interactions among disease variants, on a single-single variants or multiple-multiple variants basis. Third, but not the last, gene-environment interaction also plays pivotal roles in determining the underlying complexity of disease etiology. Studies on testing gene-gene interactions have been commonly pursued in the past, but little has been achieved, despite its importance in determining a disease risk (see [3] for a comprehensive review).

Mapping genetic interactions has been traditionally pursued in model organisms to identify functional relationships among genes [4]-[6]. With the seminal work in quantitative trait loci (QTL) mapping by Lander and Botstein [7], extensive work has been focused on experimental crosses to study the genetic architecture of complex traits. Along the line, methods for mapping QTL interactions have also been developed [8][9]. The recent development of human HapMap and radical breakthrough in genotyping technology have enabled us to generate high throughput single nucleotide polymorphisms (SNPs) data which are dense enough to cover the whole genome [10]. This advancement allows us to characterize variants at a sequence level that encode a complex disease phenotype, and opens a prospective future for disease variants identification [11][12].

Genetic interaction, or termed epistasis, occurs when the effect of one genetic variant is suppressed or enhanced by the existence of other genetic variants [13]. In align with this definition, Mani et al. [14] recently defined two distinct genetic interactions, namely the *synergistic interaction* in which extreme phenotype is expected whenever double mutations are present, and the *alleviating interaction* where one mutation in one gene masks the effect of another mutation by impairing the function of relative pathways. As an important component of the genetic architecture of many biological traits, the role of epistasis in shaping an organism's development has been unanimously recognized [15][16]. An increasing number of empirical studies have also revealed the role of epistasis in the pathogenesis of most common human diseases, such as cancer or cardiovascular disease [17][18].

The high-dimensional SNP data present unprecedented opportunity as well as daunting challenges in statistical modeling and testing in identifying genetic interactions. However, for most complex diseases, it remains largely unknown which combination of genetic variants is causal to the disease. Given that most traits or diseases are multifactorial and genetically complex, it is very unlikely that the function of a single variant can induce an overt disease signal without modeling the gene networks or pathways. Lin and Wu [19] proposed a sequence interaction model in a linear regression framework for a quantitative phenotype. Zhang et al. [20] proposed an entropy-based method for searching haplotype-haplotype interactions using unphased genotype data with applications in type I diabetes. Musani et al. [21] and Cordell [3] recently gave

a comprehensive review of statistical methods developed for detecting gene-gene interactions. While most methods are nonparametric in nature such as the popular multifactor dimensionality reduction (MDR) method [22], they do not provide effect estimates for gene-gene interactions. Thus methods focusing on data reduction ignore the biological interpretation of the interaction. For instance, if two SNPs are identified to have interaction, how do they interact in genetics? What are the modes of gene action?

In Cui et al. [12], a novel approach was proposed to group haplotypes to detect risk haplotypes associated with a disease. In an extension to this work, we proposed a new statistical method to model haplotype-haplotype interactions responsible for a binary disease phenotype. We assume a population-based case-control design where a disease phenotype is assumed dichotomous. Due to high-order interactions, we propose a penalized logistic regression framework with adaptive L_1-penalty, commonly termed adaptive LASSO [23]. The adaptive L_1-penalty allows effect estimation and variable selection simultaneously in a single model. Moreover, it preserves the oracle property of variable selection [23]. Due to the binary nature of the response, we proposed a modified Gauss-Seidel method nested within the EM algorithm to estimate parameters. The model is applied to a real data set in which significant haplotype interactions were detected between mother and offspring genomes in responsible for disease risks in pregnancy.

Methods

We first explain our method for a model involving interactions of haplotypes in 2 different haplotype blocks containing 2 SNPs in each. More complex models could be easily extended. Assume we are studying a population of n samples with n_1 cases and n_2 controls. All the n individuals are unrelated. A number of SNPs can be genotyped either in a genome-wide scale or in a candidate gene-based scale. Following the notation given in Liu et al. [11] and Cui et al. [12], we can construct composite diplotypes by defining a distinct haplotype termed as "risk" haplotype for each haplotype block. Assuming two SNPs in each block, there could be nine possible genotypes observed numerically denoted as 11/11, 11/12, 11/22, 12/11, 12/12, 12/22, 22/11, 22/12, 22/22. Without loss of generality, we assume [11] to be the "risk" haplotype. We denote the risk haplotype [11] as H and all other non-risk haplotype as $\overline{H}$. By doing this, we can map the observed genotypes to three possible composite diplotypes, i.e., HH, $H\overline{H}$ and $\overline{HH}$. Except for the double heterozygote 12/12 which is phase ambiguous and could be from two possible composite diplotypes, all other genotypes could be mapped to unique composite diplotypes. A detailed list of the configuration is given in Table 1.

The epistasis model

We consider two haplotype blocks *s* and *t*, each with two SNPs. There are total 81 possible genotype combinations. In each block, only the double heterozygote has ambiguous linkage phase, thus 64 genotypes could be mapped to unique composite diplotypes. Let H_1, $\overline{H}_1$ and

H_2, $\bar{H}_2$ be the risk and non-risk haplotypes at the two blocks, respectively. Expressed in terms of composite diplotypes, the four haplotypes can form nine distinct composite diplotypes expressed as $H_1H_1H_2H_2$, $H_1\bar{H}_1H_2H_2$, $\bar{H}_1\bar{H}_1H_2H_2$, $H_1H_1H_2\bar{H}_2$, $H_1\bar{H}_1H_2\bar{H}_2$, $\bar{H}_1\bar{H}_1H_2\bar{H}_2$, $H_1H_1\bar{H}_2\bar{H}_2$, $H_1\bar{H}_1\bar{H}_2\bar{H}_2$ and $\bar{H}_1\bar{H}_1\bar{H}_2\bar{H}_2$. The effects of the nine distinct composite diplotypes can be modeled through the traditional quantitative genetics model. Specifically, we use the Cockerham's orthogonal partition method [24] in which the genetic mean of an interaction model between blocks *s* and *t* can be expressed as

$$\mu_{st} = \mu + a_s x_s + a_t x_t + d_s z_s + d_t z_t + i_{aa} x_s x_t + i_{ad} x_s z_t + i_{da} z_s x_t + i_{dd} z_s z_t \quad (1)$$

where

$$x_s = \begin{cases} 1 & \text{for } H_1H_1 \\ 0 & \text{for } H_1\bar{H}_1 \\ -1 & \text{for } \bar{H}_1\bar{H}_1 \end{cases} \qquad z_s = \begin{cases} -1/2 & \text{for } H_1H_1 \\ 1/2 & \text{for } H_1\bar{H}_1 \\ -1/2 & \text{for } \bar{H}_1\bar{H}_1 \end{cases}$$

x_t and z_t can be defined similarly. With the above definition, $a_{s(t)}$ and $d_{s(t)}$ can be interpreted as the additive and dominance effects for the risk haplotype at block *s*(*t*); i_{aa}, i_{ad}, i_{da}, i_{dd} can be interpreted as the additive×additive, additive×dominance, dominance×additive, and dominance×dominance interaction effects between the two blocks, respectively.

Let *y* denote a measured disease trait which is dichotomous taking value 1 or 0, corresponding to affected or unaffected individual, respectively. Let X_g denote a matrix of numerical codes corresponding to the two composite diplotypes as well as their interactions, and let X_e denote a matrix of measured covariates, including the intercept as the first column. Assuming that these factors influence the mean of a trait, so that their effects can be summarized by a function of linear predictors $\eta = X_g\beta + X_e\gamma$, where $\beta = [a_s, a_t, d_s, d_t, i_{aa}, i_{ad}, i_{da}, i_{dd}]^T$ contain regression parameters for the genetic effects of composite diplotypes on a disease trait; γ contain the effects of overall mean and the covariates. Given a binary disease response, we can apply a conditional logistic model with the form

$$\log \frac{p(y=1 \mid x_g, x_e)}{p(y=0 \mid x_g, x_e)} = X_g\beta + X_e\gamma \quad (2)$$

Compared to most non-parametric methods in detecting gene-gene interactions, such as the multifactor dimensionality reduction (MDR) method which only provides an interaction test [19], the above interaction model allows one to identify which ones are the risk haplotypes in two haplotype blocks, and to further quantify the specific structure and effect size of epistatic

interactions between the two haplotype blocks. We argue that this model-based epistatic test provides biologically more meaningful results than a non-parametric method such as MDR.

Likelihood function

Assuming independence between individuals, we can construct the joint likelihood function. Define

$$c_{1i} = \begin{cases} 1 & \text{if the composite diplotype in block } s \text{ is } H_1\bar{H}_1 \\ 0 & \text{if the composite diplotype in block } s \text{ is } \bar{H}_1\bar{H}_1 \end{cases}$$

and

$$c_{2i} = \begin{cases} 1 & \text{if the composite diplotype in block } t \text{ is } H_2\bar{H}_2 \\ 0 & \text{if the composite diplotype in block } t \text{ is } \bar{H}_2\bar{H}_2 \end{cases}$$

and let $\pi_i = p(y_i = 1 | X_g, X_e)$, then the log likelihood function can be expressed as

$$\begin{aligned} L = & \sum_{i=1}^{n_{00}} \log\left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] + \sum_{i=1}^{n_{10}} \left\{ c_{1i} \log\left[\pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i}\right] + (1-c_{1i}) \log\left[\pi_{0i}^{y_i}(1-\pi_{0i})^{1-y_i}\right] \right\} \\ & + \sum_{i=1}^{n_{01}} \left\{ c_{2i} \log\left[\pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i}\right] + (1-c_{2i}) \log\left[\pi_{0i}^{y_i}(1-\pi_{0i})^{1-y_i}\right] \right\} \\ & + \sum_{i=1}^{n_{11}} \left\{ c_{1i}c_{2i} \log\left[\pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i}\right] + c_{1i}(1-c_{2i}) \log\left[\pi_{2i}^{y_i}(1-\pi_{2i})^{1-y_i}\right] + \right. \\ & \left. (1-c_{1i})c_{2i} \log\left[\pi_{3i}^{y_i}(1-\pi_{3i})^{1-y_i}\right] + (1-c_{1i})(1-c_{2i}) \log\left[\pi_{4i}^{y_i}(1-\pi_{4i})^{1-y_i}\right] \right\} \end{aligned}$$

where n_{00} is the number of individuals without phase ambiguity in both blocks; n_{10} is the number of individuals with only phase ambiguity in block *s*; n_{01} is the number of individuals with only phase ambiguity in block *t*; and n_{11} is the number of individuals with phase ambiguity in both blocks. The total data set can be grouped as four distinct groups according to the above definition. Except the n_{00} group, all other groups involve phase ambiguity genotypes, hence are modeled with mixture distributions.

Variable selection methods such as LASSO [25] or adaptive LASSO [23] have been commonly applied when the number of predictors is large. These methods can achieve parameter estimation and variable selection simultaneously and have gained large popularity in genetic and genomic data analysis. Considering the large number of genetic parameters to be estimated in the model, we applied the adaptive LASSO to our model since it has been shown that the adaptive LASSO preserves the oracle property and is consistent for variable selection [23]. Instead of maximizing the above log likelihood, we estimate the parameters by maximizing the log likelihood with the adaptive lasso penalty.

$$L' = -2L + \lambda \sum_i w_i |\beta_i| \quad (3)$$

where λ is a balance parameter for the likelihood and penalty term, and is chosen by the minimum BIC criterion. When $w_i = 1$ for every i, this leads to a general LASSO penalty. Previous study showed that when $w_i = 1/|\beta_{OLS}|$, the adaptive LASSO estimate enjoys the oracle property, which is much more attractive than the general LASSO estimate [23].

Missing data and the EM algorithm

The phase ambiguous genotypes lead to missing data. The currently developed algorithms LASSO or adaptive LASSO estimation can not be directly applied to maximize the penalized likelihood (3). However, this could be solved by applying an EM algorithm detailed as follows:

E-step:

- Initialize β, γ, and calculate $\pi_i = p(y_i = 1 | X_i) = \dfrac{\exp(X_g\beta + X_e\gamma)}{1 + \exp(X_g\beta + X_e\gamma)}$
- Estimate c_{1i}, c_{2j} by $u_j = E(c_{ji}) = \dfrac{\phi_j \pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i}}{\phi_j \pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i} + (1-\phi_j)\pi_{1i}^{y_i}(1-\pi_{1i})^{1-y_i}}$

M-step:

- Update β, γ by maximize the penalized log likelihood function (3).

Repeat until convergence.

Computational algorithm for maximizing the penalized log likelihood

In the M step, parameters β, γ are updated by calculating LASSO estimate. The LASSO regression with continuous response has been well studied. Some very efficient algorithms have been proposed, such as the shooting algorithm and the LARS [26][27]. The estimation has been a challenge for the generalized linear model due to the non-linearity of the likelihood function, especially with an adaptive penalty term. No exact solution exists for parameter estimation in this setting. Here we propose a computational algorithm using a Gauss-Seidel method [28] to solve an unconstrained optimization problem. More detail about this method can be found in Shevade [29].

We first derive the first order optimality conditions for likelihood (3) which is defined by

$$F_j = \sum_i \frac{1}{1 + e^{-y_i \sum_k x_{ik}\beta_k}} y_i x_{ij}$$

The optimality conditions are

$$
\begin{aligned}
F_j &= 0 \qquad if \;\; j = 0 \\
F_j &= w_j \lambda \qquad if \;\; \beta_j > 0, j > 0 \\
F_j &= -w_j \lambda \qquad if \;\; \beta_j < 0, j > 0 \\
-w_j \lambda \le F_j &\le w_j \lambda \qquad if \;\; \beta_j < 0, j > 0
\end{aligned}
$$

Based on the above conditions, we define

$$
\begin{aligned}
viol_j &= | F_j | \qquad if \;\; j = 0 \\
&= | w_j \lambda - F_j | \quad if \;\; \beta_j > 0, j > 0 \\
&= | w_j \lambda + F_j | \quad if \;\; \beta_j < 0, j > 0 \\
&= \max(F_j - w_j \lambda, -F_j - w_j \lambda, 0) \qquad if \;\; \beta_j < 0, j > 0
\end{aligned}
$$

For a given λ and $w_j, j = 1.....p$, we further define $I_z = \{ j : \beta_j = 0, j > 0 \}$; and $I_{nz} = \{0\} \cup \{ j : \beta_j \neq 0, j > 0 \}$. The detailed estimation procedure is given as:

1. Initialize $\beta_j = 0, j = 0, 1.....p$
2. While any $Vilo_j > 0$ in I_z

   Find the maximum violator V_k

   Update β_k by optimize L'.

   While any $Vilo_j > 0$ in I_{nz}

   Find the maximum violator V_l

   Update β_l by optimize L'.

   Until no violator exists in I_{nz}

   Until no violator exists in I_z

For computation precision reasons, the $Vilo_j > 0$ condition is relaxed to $Vilo_j > 10^{-5}$ in our computation.

This method is based on the convexity of the likelihood function. The computation procedure tries to update one β_j by the violation of the optimality conditions. The algorithm is relatively efficient because it does not involve matrix inverse. The convexity condition warrants

one and only one solution for each update.

Risk haplotype selection

We treat each possible haplotype as a potential "risk" haplotype. The one with minimum BIC information defined below is chosen as the "risk" haplotype.

$$BIC = -2L + d\log(n)$$

where d is number of non-zero parameters in the model and n is the total sample size.

Results

Simulation study

We conducted a series of simulation scenarios to evaluate the statistical property of the proposed method. Within each block, the minor allele frequencies of the two SNPs are assumed to be 0.3 and 0.4 with a linkage disequilibrium D=0.02. The simulation is conducted under different sample sizes (i.e., n=200, 500, 1000)

Data were simulated by assuming one haplotype is distinct from the other ones for each block. Haplotypes were simulated assuming Hardy-Weinberg equilibrium. A disease status was simulated from a Bernoulli distribution with given genetic effects under different scenarios (Table 2). The intercept was adjusted to make the sample size ratio between cases and controls at approximately 1. Scenario S0 assumes no genetic effect at all. Other scenarios assume different structure of genetic effects. Scenario S1 is an extreme case where all parameters are significant. The purpose of this simulation is to compare the selection power of different genetic parameters. Scenario S2 assumes that only one haplotype block has effects; Scenario S3 assumes both blocks having a genetic contribution to a disease phenotype without interaction between them; and Scenario S4 assumes both main and interaction effects between the two blocks. Data simulated with these configurations were subject to analysis with the proposed method. Results from 200 Monte Carlo repetitions were recorded.

Figure 1 shows the results for variable selection under different simulation scenarios. For each genetic parameter, the three bars in color correspond to different sample sizes (see figure legend). The top figure corresponds to Scenario S0, in which the proportion of selection is equivalent to the false positive (or selection) rate. It can be seen that the false selection rates for all parameters are all under the nominal level of 0.05, indicating a good false positive control. For the other scenarios (S1-S4), a clear pattern is that the selection power increases as the sample size increases. The selection rates for true negatives are also under reasonable control. Also as we expected, the selection power for the main effects is generally larger than the interaction effect (S1). Among the four interaction effects, the dominance×dominance effect performs the worst (S1 and S4). The simulation results also indicate that small sample size (n=200) generally performs badly given the large number of genetic parameters to be estimated. Generally, at least 500 samples are required to achieve reasonable power to detect interactions.

A case study

We applied our model to a perinatal case-control study on small for gestational age (SGA) neonates as part of a large-scale candidate gene-based genetic association studies of pregnancy

complication conducted in Chile. A total of 991 mother-offspring pairs (406 SGA cases and 585 controls) were genotyped for 1331 SNPs involving 200 genes. Maternal and fetal genome interaction is a primary genetic resource for SGA neonates. So we focus our analysis on identifying haplotype interactions between the maternal and fetal genome.

We first excluded SNPs that had a minor allele frequency of less than 5% or that did not satisfy Hardy-Weinberg equilibrium (HWE) in the combined mother and offspring control population by a Chi-squares test with a cut-off p-value of 0.001. We further used the computer software Haploview [30] to identify haplotype blocks for SNPs within each gene. Two tag SNPs were used to represent each block. A sliding window approach was applied to search for interactions between two blocks.

We picked two SNPs within each block and applied our model to study the main effects as well as the haplotype interaction effects between a mother and her offspring genome. By fitting our model as described in previous section and controlling other variables including maternal age and BMI, we successfully identified several SNP haplotypes with interaction effects through the adaptive LASSO logistic regression model. To ensure the significance, permutation tests of 1000 runs were further conducted to assess the significance. In each permutation test, the phenotypes were permuted and the model was fitted with different parameter estimate. An empirical p-value for effect *j* was calculated which is defined by

$$p-value\_j = \frac{\sum I_{|\beta_{perm}|>0}}{1000}$$

Results of the real data analysis were summarized in Table 3. Among the identified pairs, genes HPGD and MMP9 only show main block effects. All the other five show significant interaction effect. Permutation p-values confirm the statistical significance of the detected effects. We used the maternal-fetal pairs to show the utility of our method. We could also do the analysis focusing on the fetal genome only. We thought an interaction between the maternal and fetal genome is more interesting, thus used this as an example.

Model extension

Our method is illustrated with two SNPs only. The model can be easily extended to more than two SNPs. When three or more SNPs are involved in each haplotype block, Cui et al. [12] gave an explicit derivation for possible "risk" haplotype structure. In fact no matter how may SNPs are involved, three possible composite diplotypes can be constructed as illustrated by Cui et al. [12]. The only challenge for this extension is to deal with the number of heterozygous loci. For example, when three SNPs are considered in a block, there are a total of seven possible phase-ambiguous genotypes. In a single block haplotype analysis, there could be four mixture distributions when constructing the likelihood function. When we consider interactions between two blocks, there are a total of 16 possible mixture distributions in the likelihood function. This, however, definitely will increase the programming challenge and the computing burden. Fortunately, the increasing of the mixture components will not affect the number of parameters to be estimated. We still have four main effects and four interactions, as these parameters are defined based on the "risk" haplotype structure.

Another possible solution to the challenges mentioned above is to do a sliding window search with each window covering two SNPs at a time. This is similar to the sliding window haplotype analysis commonly applied in some software such as PLINK.

Discussion

Although it has been reported that gene-gene interaction plays a major role in genetic studies of complex diseases, the detection of gene-gene interaction has been traditionally pursued on a single SNP level, i.e., focusing on single SNP interaction. Intuitively, SNP-SNP interaction can not represent gene-gene interaction as single SNPs can not capture the total variation of a gene. Thus, extending the idea of single SNP interaction to haplotype interaction could potentially gain much in terms of capturing variations in genes. The proposed method defines gene-gene interaction through haplotype block interactions and offers an alternative strategy in finding potential interactions between two genes. We argue that the definition of haplotype block interaction could provide additional biological insights into a disease etiology, compared to a single SNP-based interaction analysis.

One of the advantages of our method is in grouping, hence reducing data dimension. By mapping genotypes to composite diplotypes, the data dimension is significantly reduced. Then we can use Bayesian information criterion to select potential "risk" haplotypes (Cui et al. 2007). The selection of "risk" haplotype renders another advantage of the method. We can identify significant haplotype structures and further quantify its main and interaction effects. This greatly enhances our model interpretability and biological relevance.

Our simulation study showed the reasonable false positive control and selection power for the genetic parameters. As we expected, the interaction effects have lower selection power compared to the main effects. As sample size increases, we are able to achieve an optimal power for the interaction effects. Another novelty of the method is through the modeling of the "risk" haplotype, which leads to the partition of composite diplotypes. No matter how many SNPs are modeled, it always ends up with three types of composite diplotypes. Thus, the number of genetic parameters is always fixed regardless of the number of SNPs modeled. The only cost is that we need to search for possible "risk" haplotypes through a larger parameter space.

We applied our method to a SGA study data set. Several SNP pairs were selected with either main or interaction effects. The permutation test confirmed the statistical significance of the selected effect. Our findings confirmed other findings in gene selection in the literature. Gene PON1 was previously reported to be associated with preterm birth, which is one of the potential genetic resources leading to SGA [31]. Gene FLT4 has been found to be association with the growth of human fetal endothelia cells and early human development [32][33]. Gene HPGD was also reported being involved in human intrauterine growth restriction [34]. Gene MMP9 has been suggested to be related with placenta function [35]. These evidences strongly indicate the biological relevance of our method.

Authors' contributions

ML performed the analysis and wrote the manuscript; RR collected the data; WF participated in the design and manuscript writing; YC conceived the idea, designed and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported in part by NSF grant DMS-0707031 and by the Perinatology Research Branch, Division of Intramural Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, DHHS. The authors thank Dr. Kelian Sun for helping data processing.

Reference

1. Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. Am. J Hum Genet. 2006 79(5):831-45
2. Drysdale C.M, McGraw D.W, Stack C.B, Stephens J.C, Judson R.S, Nandabalan K, Arnold K, Ruano G, Liggett SB. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci 2000 97(19):10483-8*
3. Cordell, H.J Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet. 2009 10:392-404*
4. Phillips PC, Otto SP, Whitelock MC. In Epistasis and the Evolutionary Process eds Wold JB, Brodie ED,Wade MJ *(Oxford Univ Press, New York) 2000*
5. Hartman JL, Garvik B, Hartwell L. Principles for the buffering of genetic variation. *Science 2001 291:1001-1004.*
6. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. *Nat Rev Genet. 2007 8:437-449*
7. Lander E.S and Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 1989 121(1):185-99*
8. Kao C.H, Zeng Z.B, and Teasdale R.D. Multiple interval mapping for quantitative trait loci. *Genetics 1999 152(3):1203-16*
9. Cui Y and Wu R. Mapping genome-genome epistasis: a high-dimensional model. *Bioinformatics. 2005 21(10):2447-55*
10. The international HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature 2007 449, 851-861*
11. Liu T, Johnson JA, Casella G, Wu R. Sequencing complex diseases with HapMap. *Genetics 2004, 168:503-511*
12. Cui Y, Fu W, Sun K, Romero R and Wu R. Mapping Nucleoide sequences that encode complex binary disease traits with Hapmap. *Current Genomics 2007 5:307-22*
13. Bateson W. Mendel's Principles of Heredity. *Cambridge University Press, 1909 Cambridge*
14. Mani R, St Onge RP, Hartman JL 4th, Giaever G, and Roth FP. Defining genetic interaction. *Proc Natl Acad Sci. 2008 105(9):3461-6.*
15. Wolf JB, Frankino WA, Agrawal AF, Brodie ED 3rd, Moore AJ. Developmental interactions

and the constituents of quantitative variation. *Evolution. 2001 55(2):232-45*

16. Segrè, D., DeLuna, A., Church, G.M. and Kishony, R. Modular epistasis in yeast metabolism. *Nat. Genet. 2005 37, 77–83.*
17. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. *Hered. 2003 56: 73-82.*
18. Nagel R.L. Epistasis and the genetics of human diseases. *C. R. Biol. 2005 328(7): 606-615.*
19. Lin M, and Wu R.L. Detecting sequence-sequence interactions for complex diseases. *Current Genomics 2006 7: 59-72.*
20. Zhang J, Liang F, Dassen WR, Veldman BA, Doevendans PA, and DeGunst M. Search for haplotype interactions that influence susceptibility to type 1 diabetes through use of unphased genotype data. *Am J Hum Genet. 2003 73(6):1385-401*
21. Musani S.K, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari H.K, and Allison D.B. Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum Hered. 2007 63(2): 67-84.*
22. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Plummer, W.D., Parl, F.F. and Moore, J.H. Multifactor Dimensionality Reduction Reveals High-Order Interactions among Estrogen Metabolism Genes in Sporadic Breast Cancer. *American Journal of Human Genetics, 2001 69:138-147.*
23. Zou H. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association. 2006 101:1418-1429*
24. Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistatis is present. *Genetics 1954 39:859-882*
25. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B., 1996 58(1):267-288*
26. Fu W. Penalized regressions: the Bridge versus the Lasso, J. *Computational and Graphical Statistics, 1998; 7,3: 397-416.*
27. Efron B, Hastie T, Johnstone I and Tibshirani R. Least Angle Regression. *Annals of Statistics 2004 32(2), 407-499*
28. Bertsekas, DT. and Tsitsiklis, JN. Parallel and Distributed Computation: Numerical Methods. *Prentice Hall, Englewood Cliffs, NJ, USA. 1989*
29. Shevade S.K and Keerthi S.S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 2003 19(17):2246-53*
30. Barrett J.C, Fry B, Maller J, Daly M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics. 2005 21(2):263-5*
31. Lawlor D.A, Gaunt T.R, Hinks L.J, Davey S.G, Timpson N, Day I.N, Ebrahim S. The association of the PON1 Q192R polymorphism with complications and outcomes of pregnancy: findings from the British Women's Heart and Health cohort study. *Paediatr Perinat Epidemiol. 2006 May;20(3):244-50*
32. Kaipainen A, Korhonen J, Pajusola K, Aprelikova O, Persico MG, Terman BI, Alitalo K. The related FLT4, FLT1, and KDR receptor tyrosine kinases show distinct expression patterns in human fetal endothelial cells. *J Exp Med. 1993 Dec 1;178(6):2077-88*

33. Boutsikou T, Malamitsi-Puchner A, Economou E, Boutsikou M, Puchner KP, Hassiakos D. Soluble vascular endothelial growth factor receptor-1 in intrauterine growth restricted fetuses and neonates. *Early Hum Dev. 2006 Apr;82(4):235-9*.
34. Nevo O, Many A, Xu J, Kingdom J, Piccoli E, Zamudio S, Post M, Bocking A, Todros T, Caniggia I. Placental expression of soluble fms-like tyrosine kinase 1 is increased in singletons and twin pregnancies with intrauterine growth restriction. *J Clin Endocrinol Metab. 2008 Jan;93(1):285-92*.
35. Kiess W, Chernausek S.D, Hokken-Koelega ACS (eds): Small for Gestational Age. Causes and Consequences. *Pediatr Adolesc Med. Basel, Karger, 2009, vol 13, pp 11-25*

Figures

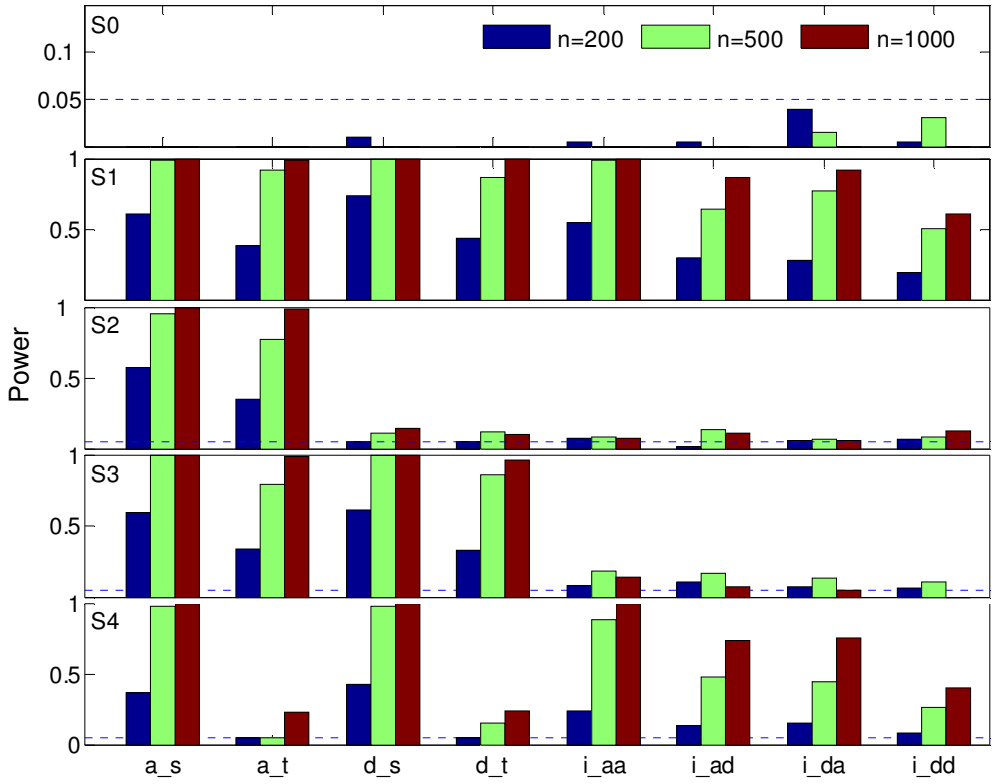

Figure 1: The bar plot of variable selection results under different simulation scenarios (Parameter values are listed in Table 2). The three sets of colored bars correspond to different sample sizes (Blue:200; Green:500; Red:1000). The horizontal dashed line indicates the nominal level of 0.05.

Tables

Table 1: The configuration of two SNP combinations

Observed Genotype	Diplotype			Composite Diplotype
	Configuration	Frequency	Relative Freq.	
11/11	[11][11]	p_{11}^2	1	HH
11/12	[11][12]	$2p_{11}p_{12}$	1	$H\bar{H}$
11/22	[12][12]	p_{12}^2	1	$\bar{H}\bar{H}$
12/11	[11][21]	$2p_{11}p_{21}$	1	$H\bar{H}$
12/12	$\begin{cases}[11][22]\\ [12][21]\end{cases}$	$\begin{cases}p_{11}p_{22}\\ p_{12}p_{21}\end{cases}$	$\begin{cases}\phi\\ 1-\phi\end{cases}$	$\begin{cases}H\bar{H}\\ \bar{H}\bar{H}\end{cases}$
12/22	[12][22]	$2p_{12}p_{22}$	1	$\bar{H}\bar{H}$
22/11	[21][21]	p_{21}^2	1	$\bar{H}\bar{H}$
22/12	[21][22]	$2p_{21}p_{22}$	1	$\bar{H}\bar{H}$
22/22	[22][22]	p_{22}^2	1	$\bar{H}\bar{H}$

Where $\phi = \dfrac{p_{11}p_{22}}{p_{11}p_{22} + p_{12}p_{21}}$

Table 2: List of parameter values under different simulation designs

Scenario	a_s	a_t	d_s	d_t	i_{aa}	i_{ad}	i_{da}	i_{dd}
S0	0	0	0	0	0	0	0	0
S1	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
S2	0.8	0.8	0	0	0	0	0	0
S3	0.8	0.8	0.8	0.8	0	0	0	0
S4	0.8	0	0.8	0	0.8	0.8	0.8	0.8

Table 3: List of selected genes, corresponding "risk" haplotype structure, effect estimates and permutation p-values

SNP ID (allele)	Gene (region)	"Risk" haplotype	a_s	d_s	a_t	d_t	i_{aa}	i_{ad}	i_{da}	i_{dd}
9508994 (C/T)^M	PON1 (intron 1)^M	[TC]^M	0	0	0	0	0	-0.45	0	0
20209376 (C/T)^O	PON1 (intron 5)^O	[CC]^O						p*=0.001		
659435566 (C/T)^M	NFKB1 (exon 12)^M	[CC]^M	0	0	0	0	-0.33	0	0	0
659435702 (C/G)^O	NFKB1 (intron 22)^O	[TC]^O					p*=0.001			
22767327 (A/T)^M	FLT4 (intron 7)^M	[AT]^M	0	0	0	0	0	-0.30	0	0
22175087 (C/T)^O	FLT4 (intron 8)^O	[TC]^O						p*<0.001		
1125300 (G/T)^M	SPARC (intron 3)^M	[TT]^M	0	-0.38	0	0	0	0	0	0.245
1125290 (G/T)^O	SPARC (intron 5)^O	[TT]^O		p*=0.001						p*<0.001
634841108 (A/C)^M	TIMP2 (intron 2)^M	[AG]^M	0	0	0	0	0	0	0	0.68
634841123 (A/G)^O	TIMP2 (exon 3)^O	[CG]^O								p*<0.001
634018768 (A/G)^M	HPGD (promoter)^M	[AG]^M	0	0	0.44	0	0	0	0	0
636105057 (A/G)^O	HPGD (promoter)^O	[GA]^O			p*<0.001					
17252653 (G/T)^M	MMP9 (intron)^M	[GC]^M	0	0	0.53	0	0	0	0	0
17254821 (C/G)^O	MMP9 (exon 10)^O	[TC]^O			p*<0.001					

^{M} mother's SNP, gene and "risk" haplotype information; ^{O} offspring's SNP, gene and "risk" haplotype information. p* is the permutation p-value.