Li *et al. BMC Genetics* 2010, **11**:79
http://www.biomedcentral.com/1471-2156/11/79

BMC Genetics

METHODOLOGY ARTICLE Open Access

Mapping Haplotype-haplotype Interactions with Adaptive LASSO

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

Abstract

Background: The genetic etiology of complex diseases in human 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. single nucleotide polymorphisms or SNPs) have been extensively studied. Recently, haplotype-based analysis has gained its popularity among genetic association studies. 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, largely 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 propose a new parameter estimation method which estimates and selects parameters by the modified Gauss-Seidel method nested within the EM algorithm. Simulation studies show that it has low false positive rate and reasonable power in detecting haplotype interactions. The method is 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 are detected.

Conclusions: As demonstrated by the simulation studies and real data analysis, the approach developed provides an efficient tool for the modeling and testing of haplotype interactions. The implementation of the method in R codes can be freely downloaded from http://www.stt.msu.edu/~cui/software.html.

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

* Correspondence: fuw@epi.msu.edu; cui@stt.msu.edu
[1]Department of Epidemiology, Michigan State University, East Lansing, Michigan 48824, USA
[2]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, USA
Full list of author information is available at the end of the article

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 opportunities 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 as the 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 are detected between mother and offspring genomes that might be 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 have a study sample of n unrelated subjects with n_1 cases and n_2 controls. A number of SNPs are 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 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, 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 as H and all other non-risk haplotype as $\overline{H}$. In doing so, we can map the observed genotypes to three possible composite diplotypes, i.e., HH, $H\overline{H}$ and $\overline{H}\overline{H}$. Except for the double heterozygote 12/12 which is phase ambiguous and could be from two possible composite diplotypes, all other genotypes can 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 $(H_1, \bar{H}_1)$ and $(H_2, \bar{H}_2)$ be the risk and non-risk haplotypes at blocks s and t, respectively. Expressed in terms of composite diplotypes, the four haplotypes can form nine distinct composite diplotypes expressed as $H_1H_1H_2H_2$, $H_1\bar{H}_1H_2H_2$, $H_1H_1H_2\bar{H}_2$, $H_1H_1H_2\bar{H}_2$, $H_1\bar{H}_1H_2\bar{H}_2$, $\bar{H}_1\bar{H}_1H_2\bar{H}_2$, $H_1\bar{H}_1\bar{H}_2\bar{H}_2$, $H_1\bar{H}_1\bar{H}_2\bar{H}_2$ 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.

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]	2 $p_{11}p_{12}$	1	$H\overline{H}$
11/22	[12][12]	p_{12}^2	1	$\overline{H}\overline{H}$
12/11	[11][21]	2 $p_{11}p_{21}$	1	$\overline{H}\overline{H}$
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]	2 $p_{12}p_{22}$	1	$\overline{H}\overline{H}$
22/11	[21][21]	p_{21}^2	1	$\overline{H}\overline{H}$
22/12	[21][22]	2 $p_{21}p_{22}$	1	$\overline{H}\overline{H}$
22/22	[22][22]	p_{22}^2	1	$\overline{H}\overline{H}$

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

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

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

where

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

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_i denote a measured disease trait for subject i, 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. Let x_{ig} and x_{ie} denote the i^{th} row of X_g and X_e. 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. To simplify the notations, we also use $\beta = [\beta_1, \beta_2, ...,\beta_8]^T$ for the genetic effects in the equations below. Given a binary disease response, we can apply a conditional logistic model with the form

$$\log \frac{p(y_i = 1 \mid x_{ig}, x_{ie})}{p(y_i = 0 \mid x_{ig}, x_{ie})} = x_{ig}\beta + x_{ie}\gamma \quad (2)$$

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

We first introduce notations. Let g_{is} and g_{it} denote the observed genotypes in haolotype block s and t respectively for subject i. With the same numerical notation defined previously, we have g_{is}, $g_{it} \in$ {11/11, 11/12, 11/22, 12/11, 12/12, 12/22, 22/11, 22/12, 22/22}. Let G_{is} and G_{it} be the underlying composite diplotypes for g_{is} and g_{it}, respectively. We have $G_{is} \in \{H_1H_1, H_1\bar{H}_1, \bar{H}_1\bar{H}_1\}$ and $G_{it} \in \{H_2H_2, H_2\bar{H}_2, \bar{H}_2\bar{H}_2\}$. We further define M_1, M_2, M_3 and M_4 as four distinct genotype groups corresponding to the classification of phase (un)ambiguous haplotype blocks:

$$\begin{aligned}
M_1 &= \{i \mid g_{is} \neq 12/12 \ \& \ g_{it} \neq 12/12\},\\
M_2 &= \{i \mid g_{is} = 12/12 \ \& \ g_{it} \neq 12/12\},\\
M_3 &= \{i \mid g_{is} \neq 12/12 \ \& \ g_{it} = 12/12\},\\
M_4 &= \{i \mid g_{is} = 12/12 \ \& \ g_{it} = 12/12\}.
\end{aligned}$$

To construct likelihood function, all three groups, M_2, M_3, M_4, except group M_1, involve phase ambiguity genotypes, hence need to be modeled with mixture distributions.

Define

$$c_{si} = \begin{cases} 1 & G_{is} = H_1\bar{H} \\ 0 & G_{is} = \bar{H}_1\bar{H}_1 \end{cases} \text{ and } c_{ti} = \pi\begin{cases} 1 & G_{it} = H_2\bar{H}_2 \\ 0 & G_{it} = \bar{H}_2\bar{H}_2 \end{cases} \text{ for } i \in M_2, M_3, M_4$$

We further define a set of the logistic regression functions for each genotype group as

$$\begin{aligned}
\pi_{M_1 i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_1) = \frac{\exp(x_{ig}\beta + x_{ie}\gamma)}{1 + \exp(x_{ig}\beta + x_{ie}\gamma)}\\
\pi_{M_2 1i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_2, c_{si} = 1),\\
\pi_{M_2 0i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_2, c_{si} = 0),\\
\pi_{M_3 1i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_3, c_{ti} = 1),\\
\pi_{M_3 0i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_3, c_{ti} = 0),\\
\pi_{M_4 1i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_4, c_{si} = 1, c_{ti} = 1),\\
\pi_{M_3 2i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_4, c_{si} = 1, c_{ti} = 0),\\
\pi_{M_4 3i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_4, c_{si} = 0, c_{ti} = 1),\\
\pi_{M_4 4i} &= p(y_i = 1 \mid x_{ig}, x_{ie}, i \in M_4, c_{si} = 0, c_{ti} = 0).
\end{aligned}$$

Assuming independence between individuals, we construct the joint likelihood function as follows:

$$\begin{aligned}
L = &\sum_{i \in M_1} \log\left[\pi_{M_1 i}^{y_i}(1 - \pi_{M_1 i})^{1-y_i}\right]\\
&+ \sum_{i \in M_2} \left\{ \begin{array}{l} c_{si}\log\left[\pi_{M_2 1i}^{y_i}(1 - \pi_{M_2 1i})^{1-y_i}\right] \\ +(1 - c_{si})\log\left[\pi_{M_2 0i}^{y_i}(1 - \pi_{M_2 0i})^{1-y_i}\right] \end{array} \right\}\\
&+ \sum_{i \in M_3} \left\{ \begin{array}{l} c_{ti}\log\left[\pi_{M_3 1i}^{y_i}(1 - \pi_{M_3 1i})^{1-y_i}\right] \\ +(1 - c_{ti})\log\left[\pi_{M_3 0i}^{y_i}(1 - \pi_{M_3 0i})^{1-y_i}\right] \end{array} \right\}\\
&+ \sum_{i \in M_4} \left\{ \begin{array}{l} c_{si}c_{ti}\log\left[\pi_{M_4 1i}^{y_i}(1 - \pi_{M_4 1i})^{1-y_i}\right] \\ +c_{si}(1 - c_{ti})\log\left[\pi_{M_4 2i}^{y_i}(1 - \pi_{M_4 2i})^{1-y_i}\right] \\ +(1 - c_{si})c_{ti}\log\left[\pi_{M_4 3i}^{y_i}(1 - \pi_{M_4 3i})^{1-y_i}\right] \\ +(1 - c_{si})(1 - c_{ti})\log\left[\pi_{M_4 4i}^{y_i}(1 - \pi_{M_4 4i})^{1-y_i}\right] \end{array} \right\}.
\end{aligned}$$

Because the phase ambiguous state c_{si} and c_{ti} are not observable, we treat them as missing data and use EM algorithm to estimate them iteratively (See below).

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 apply the adaptive LASSO to our model for its oracle property; namely, it performs variable selection and parameter estimation as if the true underlying model is known in advance [23]. Instead of maximizing the above log likelihood, we estimate the parameters by maximizing the log likelihood with the adaptive LASSO penalty.

$$L' = -2L + \lambda\sum_i w_i \mid \beta_i \mid \tag{3}$$

where λ is a tuning parameter for the likelihood and penalty term, and is chosen by the minimum Bayesian Information Criterion (BIC); $\omega = (w_1, w_2, ..., w_8)$ is a weight vector for the genetic effects β. When $w_j = 1$ for every j, this leads to a general LASSO penalty. Although the general LASSO estimator may not be consistent, some data dependent weight vector ω is able to warrant the oracle property for the corresponding adaptive LASSO estimator. Specifically, one choice of ω is $\omega = 1/\beta_{OLS}$, where β_{OLS} is the ordinary least square (OLS) estimator. This makes the adaptive LASSO estimate 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:

1) Initialize β, γ, and calculate $\pi_i = p(y_i = 1 \mid x_{ig}, x_{ie}) = \frac{\exp(x_{ig}\beta + x_{ie}\gamma)}{1+\exp(x_{ig}\beta + x_{ie}\gamma)}$ for subject i;

2) **E-step**: Estimate c_{si}, c_{ti} for subjects with phase ambiguous genotypes with $E(c_{ji})$ by

$$E(c_{ji}) = \frac{\phi_j \pi_{M_k 1i}^{y_i}(1-\pi_{M_k 1i})^{1-y_i}}{\phi_j \pi_{M_k 1i}^{y_i}(1-\pi_{M_k 1i})^{1-y_i} + (1-\phi_j)\pi_{M_k 0i}^{y_i}(1-\pi_{M_k 0i})^{1-y_i}},$$

for $i \in M_k$ $(k, j) \in \{(2, s\}, (3,t)\}$.

For i ∈ M_4, we have

$$E(c_{si}) = \frac{\phi_s\phi_t\pi_{M_4 1i}^{\gamma_i}(1-\pi_{M_4 1i})^{1-\gamma_i}+\phi_s(1-\phi_t)\pi_{M_4 2i}^{\gamma_i}(1-\pi_{M_4 2i})^{1-\gamma_i}}{\Pi},$$

$$E(c_{ti}) = \frac{\phi_s\phi_t\pi_{M_4 1i}^{\gamma_i}(1-\pi_{M_4 1i})^{1-\gamma_i}+(1-\phi_s)\phi_t\pi_{M_4 3i}^{\gamma_i}(1-\pi_{M_4 3i})^{1-\gamma_i}}{\Pi},$$

where $\begin{aligned}\Pi = {} & \phi_s\phi_t\pi_{M_4 1i}^{\gamma_i}(1-\pi_{M_4 1i})^{1-\gamma_i} + \phi_s(1-\phi_t)\pi_{M_4 2i}^{\gamma_i}(1-\pi_{M_4 2i})^{1-\gamma_i} \\ & +(1-\phi_s)\phi_t\pi_{M_4 3i}^{\gamma_i}(1-\pi_{M_4 3i})^{1-\gamma_i} + (1-\phi_s)(1-\phi_t)\pi_{M_4 4i}^{\gamma_i}(1-\pi_{M_4 4i})^{1-\gamma_i}\end{aligned}$

3) **M-step**: Update β,γ by maximizing the penalized log likelihood function (3);

4) Repeat step 1)-3) 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 *et al.* [29]. To simplify the notations, we explain our method without environmental covariates.

We first derive the first order optimality conditions for the penalized likelihood (3). It is noticed that the penalized likelihood *L'* is piecewise differentiable. Following the notation in Shevade [29], denote $F_j = \partial(2\ L)/\partial\beta_j$. The first order optimality conditions $\partial L'/\partial\beta_j = 0$ could be achieved as follows:

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

For the phase known genotypes, F_j will have an explicit form as:

$$F_j = \sum_{i \in M_1} [\gamma_i x_{ij} - \frac{\exp(\sum_k x_{ik}\beta_k)}{1+\exp(\sum_k x_{ik}\beta_k)} x_{ij}]$$

With the phase ambiguous genotypes, F_j can be calculated accordingly with the mixture proportion $E(c_{si})$and $E(C_{ti})$that are estimated from E-step.

Based on the above conditions, we define

$$\begin{aligned} Viol_j & = \mid F_j \mid && if\ j = 0 \\ & = \mid w_j\lambda - F_j \mid && if\ \beta_j > 0, j > 0 \\ & = \mid w_j\lambda + F_j \mid && if\ \beta_j < 0, j > 0 \\ & = \max(F_j - w_j\lambda, -F_j - w_j\lambda, 0) && if\ \beta_j = 0, j > 0 \end{aligned}$$

Therefore, the optimal conditions could be achieved when $Viol_j = 0$ for $\forall j$. For a given λ and w_j, j = 1.....p, we further define I_z = {j: β_j = 0, j > 0}; and I_{nz} = {0}∪{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 $Viol_j > 0$ in I_z,

```
        Find the maximum violator $V_k$,
            Update $\beta_k$ by optimizing $L'$;
    While any $Viol_j > 0$ in $I_{nz}$,
        Find the maximum violator $V_l$,
            Update $\beta_l$ by optimizing $L'$,
            Until no violator exists in $I_{nz}$;
Until no violator exists in $I_z$
```

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

This method is based on the convexity of the likelihood function. The computation procedure updates one β_j at a time until all the optimality conditions are achieved. The algorithm is relatively efficient because it does not involve matrix inverse. The convexity condition warrants one and only one solution for each update (See additional file 1). Similar algorithm has been used in linear regression setting, commonly referred to as 'the shooting algorithm' [26], and in logistic regression setting for general LASSO [29]. The asymptotic convergence of this method for non-linear optimization problem has been proven in [[28], Ch.3Prop 4.1].

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

Results

Simulation study

We conducted a series of simulation with various scenarios to evaluate the statistical property of the proposed method. Within each block, the minor allele frequencies of the two SNPs were assumed to be 0.3

and 0.4 with a linkage disequilibrium D = 0.02. The simulation was conducted under different sample sizes (i.e., n = 200, 500, 1000)

Data were simulated by assuming one haplotype was 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 assumed no genetic effect at all. Other scenarios assumed different structure of genetic effects. Scenario S1 was an extreme case where all parameters were significant. The purpose of this simulation was to compare the selection power of different genetic parameters. Scenario S2 assumed that only one haplotype block has effects; Scenario S3 assumed both blocks had a genetic contribution to the disease phenotype without interaction between them; and Scenario S4 assumed 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 showed 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 corresponded to Scenario S0, in which the proportion of selection was equivalent to the false positive (or selection) rate. It can be seen that the false selection rates for all parameters were all under the nominal level of 0.05, indicating a good false positive control. For the other scenarios (S1-S4), the selection power increased as the sample size increased. Compared to S0, the selection rates for true negatives increased, but were also under reasonable control. Also as we expected, the selection power for the main effects was generally larger than the interaction effect (S1). Among the four interaction effects, the dominance×dominance effect performed the worst (S1 and S4). The simulation results also indicated that small sample size (n = 200) generally performed badly given the large number of genetic parameters to be estimated. Generally, at least 500 samples were required to achieve reasonable power to detect interactions.

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

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 was a primary genetic resource for SGA neonates. So we focused 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

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

Results of the real data analysis were summarized in Table 3. Among the identified pairs, genes *HPGD* and *MMP9* only showed main block effects. All the other five showed significant interaction effect. Permutation p-values confirmed 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 was more interesting, thus used this as an example.

Our approach conducts the variable selection and effect estimation simultaneously, which allows us to

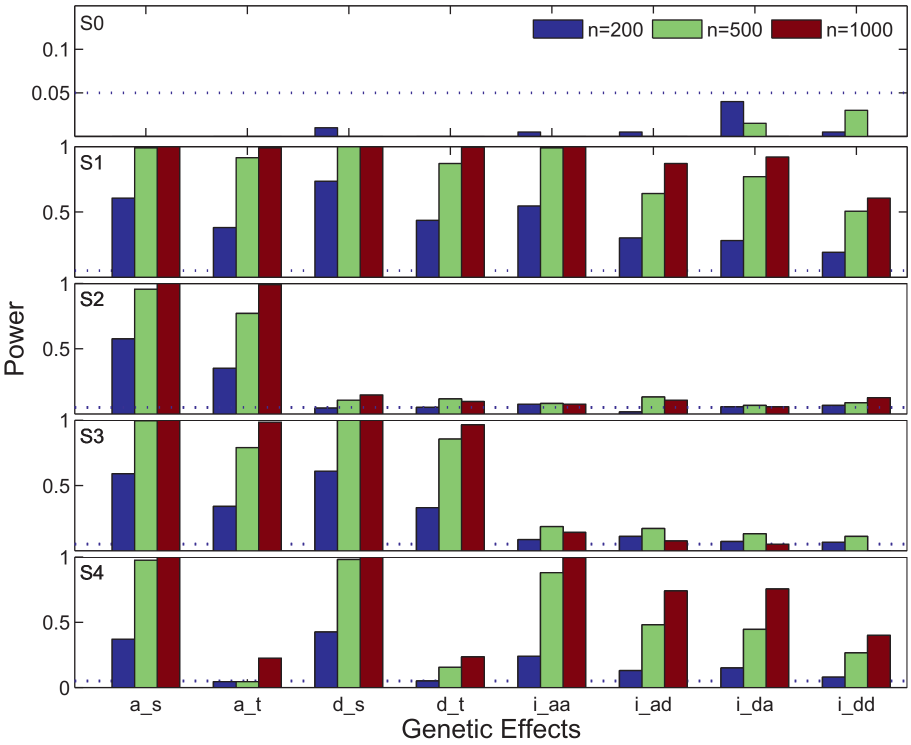

Figure 1 The bar plot of variable selection results under different simulation scenarios. Parameter values are listed in Table 2. The three sets of colored bars correspond to different sample sizes (Blue:200; Green:500; Red:1000). The horizontal dashed line indicates the nominal level of 0.05.

have a direct biological interpretation for the mode of gene action. Here, we use gene *PON1* as an example to illustrate the implementation of our model. In gene *PON1*, the selected risk haplotypes are $[TC]$ for the mother and $[CC]$ for the offspring. We find significant additive × dominant haplotype interaction effect. The two haplotypes separate all the mother-offspring pairs into three 'risk' groups with respect to the development of SGA:

$$\begin{aligned}
R_1 = \{i \mid (G_i^M, G_i^O) = {} & ([TC]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [CC]^{\mathrm{O}}[CC]^{\mathrm{O}}) \\
& \mid ([TC]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [CC]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}}) \\
& \mid ([TC]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [\,\overline{CC}\,]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}})\} \\
R_2 = \{i \mid (G_i^M, G_i^O) = {} & ([TC]^{\mathrm{M}}[TC]^{\mathrm{M}}, [CC]^{\mathrm{O}}[CC]^{\mathrm{O}}) \\
& \mid ([TC]^{\mathrm{M}}[TC]^{\mathrm{M}}, [\,\overline{CC}\,]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}}) \\
& \mid ([\,\overline{TC}\,]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [CC]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}})\} \\
R_3 = \{i \mid (G_i^M, G_i^O) = {} & ([TC]^{\mathrm{M}}[TC]^{\mathrm{M}}, [CC]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}}) \\
& \mid ([\,\overline{TC}\,]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [CC]^{\mathrm{O}}[CC]^{\mathrm{O}}) \\
& \mid ([\,\overline{TC}\,]^{\mathrm{M}}[\,\overline{TC}\,]^{\mathrm{M}}, [\,\overline{CC}\,]^{\mathrm{O}}[\,\overline{CC}\,]^{\mathrm{O}})\}
\end{aligned}$$

Following Eq. (1), we can see that R_1 corresponds to the baseline reference group, R_2 corresponds to the risk group with -1/2 interaction coefficient, and R_3 corresponds to the risk group with 1/2 interaction coefficient. Correspondingly, the log odds of the disease development in each 'risk' group and the odds ratio (OR) between groups can be estimated by:

$$\log(odds) = \log\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = \begin{cases} \mu & i \in R_1 \\ \mu - i_{ad}/2 & i \in R_2\,, \\ \mu + i_{ad}/2 & i \in R_3 \end{cases}$$

$$OR = \begin{cases} \text{Re ference} & i \in R_1 \\ \exp(-i_{ad}/2) = 1.25 & i \in R_2 \\ \exp(i_{ad}/2) = 0.80 & i \in R_3 \end{cases}$$

Other non-parametric methods, such as multifactor dimensionality reduction (MDR), have been shown to be successful for the identification of interaction effects in many studies. Because MDR can only be applied to studies with balanced case/control design, generalized MDR (GMDR) has been proposed as an extension to MDR [31]. GMDR maps phenotypic traits into residual scores through certain link functions under the generalized liner model setting, and further conducts SNP selection and testing based on the residual scores. To compare with our method,

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)	PON1 (intron 1)	$[TC]^M$	0	0	0	0	0	-0.45	0	0
20209376 (C/T)	PON1 (intron 5)	$[CC]^O$						p* = 0.001		
659435566 (C/T)	NFKB1 (exon 12)	$[CC]^M$	0	0	0	0	-0.33	0	0	0
659435702 (C/G)	NFKB1 (intron 22)	$[TC]^O$					p* = 0.001			
22767327 (A/T)	FLT4 (intron 7)	$[AT]^M$	0	0	0	0	0	-0.30	0	0
22175087 (C/T)	FLT4 (intron 8)	$[TC]^O$						p* < 0.001		
1125300 (G/T)	SPARC (intron 3)	$[TT]^M$	0	-0.38	0	0	0	0	0	0.245
1125290 (G/T)	SPARC (intron 5)	$[TT]^O$		p* = 0.001						p* < 0.001
634841108 (A/C)	TIMP2 (intron 2)	$[AG]^M$	0	0	0	0	0	0	0	0.68
634841123 (A/G)	TIMP2 (exon 3)	$[CG]^O$								p* < 0.001
634018768 (A/G)	HPGD (promoter)	$[AG]^M$	0	0	0.44	0	0	0	0	0
636105057 (A/G)	HPGD (promoter)	$[GA]^O$			p* < 0.001					
17252653 (G/T)	MMP9 (intron)	$[GC]^M$	0	0	0.53	0	0	0	0	0
17254821 (C/G)	MMP9 (exon 10)	$[TC]^O$			p* < 0.001					

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

we applied GMDR to the data. The mother-offspring paired genotype data were used as input for GMDR, and a logistic link was used to calculate the residual scores.

In the example of *PON1*, SNP 20209376 (C/T) in the fetal genome was first selected by GMDR (p-value = 0.0107). SNPs were then paired with each other to identify potential significant pairwise interactions. Only SNP 9508994 (C/T) in the mother genome was found to interact with SNP 20209376 with marginal significance (p-value = 0.0547). More complex model were found to be non-significant (p-value = 0.1719 and p-value = 0.3770 for 3 SNP and 4 SNP model, respectively). Even though GMDR indicated a maternal-fetal interaction between these two SNPs, it did not provide an estimation of the genetic effect and the underlying interaction mechanism between the SNPs.

Model extension

Our method has been 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 will, however, definitely increase the programming challenge and the computing burden. Fortunately, the increaes 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 and Conclusions

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 because single SNPs cannot 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 [12]. 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 that our method has 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 the modeling of the "risk" haplotype, which leads to the partition of composite diplotypes. No matter how many SNPs are involved, 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. The only cost is the 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 of 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 [32]. Gene *FLT4* had been found to be association with the growth of human fetal endothelia cells and early human development [33,34]. Gene *HPGD* was also reported being involved in human intrauterine growth restriction [35]. Gene *MMP9* had been suggested to be related with placenta function [36]. These evidences strongly indicated the biological relevance of our method.

We also identified potential interaction effects for several additional genes, including *NFKB1*, *SPARC* and *TIMP2*. To our knowledge, no experimental evidence has been reported for these genes regarding the biological function related to fetal development or SGA. However, we found that each of these genes had been suggested to be involved in many biological pathways. Studies indicated that gene *NFKB1* was functionally related to stress-impaired neurogenesis and depressive behavior [37], myelin formation [38], and adipose tissue growth [39]. Gene *SPARC* had been suggested to be associated with angiogenesis and tumor growth [40] and the progression of crescentic glomerulonephritis [41]. Gene *TIMP2* was reported to be related to myogenesis [42] and the progression of cerebral aneurysms [43]. Further replicate studies are needed to confirm the biological relevance of these genes to SGA.

Additional material

Additional file 1: Strict convexity of the log likelihood function. The file contains the proof of strict convexity of the log likelihood function.

Acknowledgements
The authors wish to thank the two anonymous referees for their helpful comments that improved the manuscript, and thank Dr. Kelian Sun for helping data processing. 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.

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 the model and wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Epidemiology, Michigan State University, East Lansing, Michigan 48824, USA. [2]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, USA. [3]The Perinatology Research Branch, NICHD, NIH, DHHS, Bethesda, MD, and Detroit, MI 48201, USA.

Received: 22 March 2010 Accepted: 27 August 2010
Published: 27 August 2010

References

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 CM, McGraw DW, Stack CB, Stephens JC, Judson RS, 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 HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10**:392-404 [http://www.nature.com/nrg/journal/v10/n6/abs/nrg2579.html-a1].
4. Phillips PC, Otto SP, Whitelock MC: **Beyond the average: The evolutionary importance of epistasis and the variability of epistatic effects.** In *Epistasis and the Evolutionary Process.* Edited by: 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 ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121(1)**:185-99,.
8. Kao CH, Zeng ZB, Teasdale RD: **Multiple interval mapping for quantitative trait loci.** *Genetics* 1999, **152(3)**:1203-16.
9. Cui Y, 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, Cambridge 1909.
14. Mani R, St Onge RP, Hartman JL, Giaever G, Roth FP: **Defining genetic interaction.** *Proc Natl Acad Sci* 2008, **105(9)**:3461-6.
15. Wolf JB, Frankino WA, Agrawal AF, Brodie ED, Moore AJ: **Developmental interactions and the constituents of quantitative variation.** *Evolution* 2001, **55(2)**:232-45.
16. Segrè D, DeLuna A, Church GM, Kishony R: **Modular epistasis in yeast metabolism.** *Nat Genet* 2005, **37**:77-83.
17. Moore JH: **The ubiquitous nature of epistasis in determining susceptibility to common human diseases.** *Hum Hered* 2003, **56**:73-82.
18. Nagel RL: **Epistasis and the genetics of human diseases.** *C R Biol* 2005, **328(7)**:606-615.
19. Lin M, Wu RL: **Detecting sequence-sequence interactions for complex diseases.** *Current Genomics* 2006, **7**:59-72.
20. Zhang J, Liang F, Dassen WR, Veldman BA, Doevendans PA, 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 SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: **Detection of gene × gene interactions in genome-wide association studies of human population data.** *Hum Hered* 2007, **63(2)**:67-84.
22. 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.** *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, Tibshirani R: **Least Angle Regression.** *Annals of Statistics* 2004, **32(2)**:407-499.
28. Bertsekas DT, Tsitsiklis JN: **Parallel and Distributed Computation: Numerical Methods.** *Prentice Hall, Englewood Cliffs, NJ, USA* 1989.
29. Shevade SK, Keerthi SS: **A simple and efficient algorithm for gene selection using sparse logistic regression.** *Bioinformatics* 2003, **19(17)**:2246-53.
30. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2)**:263-5.
31. Lou XY, Chen GB, Yan L, Ma J, Zhu J, Elston R, Li MD: **A generalized combinatorial approach for detecting gene-by gene and gene-by-environment interactions with application to Nicotine Dependence.** *Am J Hum Genet* 2007, **80**:1125-1137.
32. Lawlor DA, Gaunt TR, Hinks LJ, Davey SG, Timpson N, Day IN, 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, **20(3)**:244-50.
33. 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, **178(6)**:2077-88.
34. 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, **82(4)**:235-9.
35. 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, **93(1)**:285-92.
36. Kiess W, Chernausek SD, Hokken-Koelega ACS, eds: **Small for Gestational Age. Causes and Consequences.** *Pediatr Adolesc Med Basel, Karger* 2009, **13**:11-25.
37. Koo JW, Russo SJ, Ferguson D, Nestler EJ, Duman RS: **Nuclear factor-kappaB is a critical mediator of stress-impaired neurogenesis and depressive behavior.** *PNAS* 2010, **107(6)**:2669-74.
38. Limpert AS, Carter BD: **Axonal neuregulin 1 type III activates NF-kappaB in Schwann cells during myelin formation.** *J Biol Chem* 2010, **285(22)**:16614-22.
39. Tang T, Zhang J, Yin J, Staszkiewicz J, Gawronska-Kozak B, Jung DY, Ko HJ, Ong H, Kim JK, Mynatt R, Martin RJ, Keenan M, Gao Z, Ye J: **Uncoupling of inflammation and insulin resistance by NF-kappaB in transgenic mice through elevated energy expenditure.** *J Biol Chem* 2010, **285(7)**:4637-44.
40. Bhoopathi P, Chetty C, Gujrati M, Dinh DH, Rao JS, Lakka SS: **The role of MMP-9 in the anti-angiogenic effect of secreted protein acidic and rich in cysteine.** *Br J Cancer* 2010, **102(3)**:530-40.
41. Sussman AN, Sun T, Krofft RM, Durvasula RV: **SPARC accelerates disease progression in experimental crescentic glomerulonephritis.** *Am J Pathol* 2009, **174(5)**:1827-36.
42. Lluri G, Langlois GD, Soloway PD, Jaworski DM: **Tissue inhibitor of metalloproteinase-2 (TIMP-2) regulates myogenesis and beta1 integrin expression in vitro.** *Exp Cell Res* 2008, **314(1)**:11-24.
43. Aoki T, Kataoka H, Moriwaki T, Nozaki K, Hashimoto N: **Role of TIMP-1 and TIMP-2 in the progression of cerebral aneurysms.** *Stroke* 2007, **38(8)**:2337-45.
44. Jon Dattorro : **Convex Optimization & Euclidean Distance Geometry.** Meboo publish 2005.

doi:10.1186/1471-2156-11-79
Cite this article as: Li *et al.*: Mapping Haplotype-haplotype Interactions with Adaptive LASSO. *BMC Genetics* 2010 **11**:79.