RM-687 (December, 2010)

A model-based kernel machine method for gene-centric gene-gene interaction

Shaoyu Li and Yuehua Cui*

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824

Abstract

Much of the natural variation for a complex trait can be explained by the structural variation in DNA sequences. As part of the sequence variation, gene-gene interaction or epistasis has been ubiquitously observed in nature where its role in shaping the development of an organism has been broadly recognized. The identification of genetic epistasis has been progressively pursued via statistical or machine learning approaches. A large body of currently adopted methods, either parametrically or non-parametrically, are predominantly focused on pairwise single marker interaction analysis. As genes are the heredity units in living organisms, analysis by focusing on a gene as a system could potentially yield more biologically meaningful results. In this work, we conceptually propose a gene-centric gene-gene interaction framework for genome-wide epistasis detection. We treat each gene as a testing unit and derive a model-based kernel machine method for a two-dimensional genome-wide scanning of gene-gene interactions. In addition to the biological advantage, our method is statistically appealing by reducing the number of hypothesis testing in a genome-wide scan. Extensive simulation studies are conducted to evaluate the performance of the method. The utility of the method is further demonstrated with applications to two real data sets. Our gene-centric gene-gene interaction analysis provides a conceptual framework for epistasis identification.

1 Introduction

Accumulative evidences have shown that much of the genetic variation for a complex trait can be explained by the joint function of multiple genetic factors, as well as environmental contributions. Searching for these contributing genetic factors and further characterizing their effects size, is one of the primary goals and challenges for modern genetics study. The recent rapid breakthrough in high-throughput genotyping technologies and the completion of the International Haplotype Mapping (HapMap) project provide unprecedented opportunities to characterize the genetic machinary of an living organism. Genetic association analyses focusing on single nucleotide polymorphisms (SNPs) or haplotypes have led to the identification of many novel genetic determinants of complex traits. However, despite enormous success in genome-wide association analysis, single SNP or haplotype based studies still suffer from low replication rates because of the infeasibility of dealing with the complex patterns of association, e.g. genetic heterogeneity, epistasis and gene-environment interaction, leaving much of the genetic components of many traits remaining unaccounted for and only a small proportion of the heritability being explained.

It has been broadly recognized that most common human diseases are likely to have complex etiologies (Thornton-Wells et al. 2004). In a recent report, Neale and Sham (2004) discussed the choice of the basic genetic components to be considered for association with a complex trait. They argued that a gene-based approach, in which all variants within a putative gene are considered jointly, have relative advantages over single SNP or haplotype analysis. There are multiple reasons for this. First, it is well known that genes are the functional units of the human genome. Variants in genes have high probability of being functionally important than those that occur outside of genes (Jorgenson and Witte, 2006). Because of this characteristic, gene-based association analysis would provide more biologically interpretable results than the single-SNP or haplotype based analysis. Second, the position, sequence and function of genes are highly consistent across diverse human populations, which makes the gene-based studies more powerful in terms of replication (Neale and Sham, 2004). Third, when there are multiple variants within a gene that function in a complicated manner, the gene-based association test can gain additional power compared to a single SNP analysis by capturing the joint function of multiple variants simultaneously (Cui et al., 2008; Buil et al., 2009). Finally, a gene-based analysis is statistically appealing. By considering multiple SNP markers within a gene as a testing unit, the number of tests would decrease dramatically, hence simplify the multiple testing problem and improve the power of association test.

We all know that genes do not function alone, rather they constantly interact with each other. It has been widely recognized that gene-gene interaction, or epistasis, is an important category that contributes to the unexplained heritability of complex traits (Thornton-Wells et al., 2004; Maher, 2008; Moore and Williams, 2009; Eichler et al., 2010). Methods for detecting gene-gene interaction have been historically pursued on a single locus level, either parametrically such as the regression-based tests of interaction (Piegorsch et al., 1994) and the Bayesian epistasis mapping (Zhang and Liu, 2007), or non-parametrically such as the entropy-based approaches (Kang et al., 2008), and some data mining methods such as the multifactor dimensionality reduction (MDR) (Ritchie et al., 2001) and its extensions (e.g., Lou et al. 2007) and random forests (Breiman, 2001). Methods based on interaction of haplotypes have also been developed (e.g., Li et al. 2010). However, due to the phase-ambiguity problem, the haplotype-based methods are limited to only small size haplotypes. Extensions to interaction of large size haplotypes are challenged by computational cost. For a comprehensive review of statistical methods developed for detecting gene-gene interactions, readers are referred to Cordell (2009).

Given the relative merits of the gene-based association analysis, the identification of genetic interactions by focusing on genes as functional units should carry the same benefits and gains as it does with single gene analysis. Thus we propose to jointly model the genetic variation of SNPs within a gene, then pairwise gene-gene interactions can be carried out in a genome-wide search. The idea of Gene-centric Gene-Gene (denoted as 3G) interaction would conceptually change the way we model epistasis and meantime bring statistical challenges. Through the modeling of the joint variation of a gene pair, we argue that a gene-centric epistasis analysis is biologically attractive. In addition, by focusing on genes as testing units, the number of pairwise interaction tests can be dramatically reduced compared to a single SNP-based pairwise interaction analysis. Thus a 3G interaction analysis is also statistically appealing.

In this work, we propose a model-based kernel machine method for the purpose of identifying significant gene-gene interactions under the proposed 3G analysis framework. Kernel based methods have been proposed to evaluate association of genetic variants with complex traits in the past decades (e.g., Tzeng et al., 2003; Schaid 2005; Wessel et al., 2006; Schaid, 2010a, 2010b). A general kernel machine method can account for complex nonlinear SNP effects within a genetic feature (e.g. a gene or a pathway) by using an appropriately selected kernel function. Generally speaking, a kernel function captures the pairwise genomic similarity between individuals for variants within an appropriately defined feature (Schaid, 2010a). The application of kernel-based method in genetic association analysis has been reported in the literature (e.g., Schaid 2005; Kwee et al., 2008; Wu et al., 2010). However, none of these considers interaction of genes. In this work, we propose a general 3G interaction framework by applying the smoothing-spline ANOVA model (Wahba, 1990) to model gene-gene interaction. The proposed method, termed **G**ene-centric **G**ene-**G**ene interaction with **S**moothing-s**P**line **A**NOVA **M**odel (3G-SPAM), is implemented through a two-step procedure: (1) an exhaustive two-dimensional genome-wide search for pairwise gene-gene interactions; and (2) significance assessment of pairwise interactions.

The rest of the paper is organized as follows. In section 2, we describe the detailed model derivation of our method. We proposed two score statistics for testing the overall genetic effect and the interaction effect based on the 3G-SPAM. To evaluate the performance of the proposed method, Monte Carlo simulations are performed in section 3. The utility of the method are demonstrated by analyzing two real data sets in section 4 followed by a discussion in section 5.

2 Methods

2.1 Smoothing Spline-ANOVA Model

Given n unrelated individuals sampled from a population, each of which possessing a measurement for certain quantitative trait of interest. The quantitative measurements of all the n individuals are denoted as $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$. In searching for gene-gene interactions, traditional approaches such as MDR or regression type analysis start with a two-dimensional pairwise SNP interaction analysis. In this work, we focus our attention to pairwise gene-gene interactions by considering each gene as a unit. Considering two genes denoted as G_1 and G_2, the number of SNP markers within each gene is denoted as L_1 and L_2, respectively. Let $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,L})^T$ denote an $L \times 1$ vector of all SNP genotypes for the two genes for subject i. Here $L = L_1 + L_2$ is the total number of SNP markers in the two genes considered. Considering $\boldsymbol{x}_i$ as an L dimensional vector of random variables, a natural model for studying the relationship between the SNPs ($\boldsymbol{x}_i$) and the phenotype y_i is by a regression model

$$y_i = m(\boldsymbol{x}_i) + \epsilon_i, i = 1, 2, \cdots, n \tag{1}$$

where m is an unknown function of SNPs $\boldsymbol{x}_i = (x_{1i}, \cdots, x_{Li})^T$, and $\epsilon_i \sim (0, \sigma^2)$ is a random subject-specific error term which is generally assumed to be normally distributed with mean 0 and variance σ^2 and be independent of $\boldsymbol{x}_i$.

To explore the relationship of the genetic contributions of each gene and their interaction to the total variation of a trait, we decompose the function m into main effects m_j and interaction m_{jk} between the two genes, following the functional analysis of variance framework (Wahba 1990; Wahba et al. 1995; Gu and Wahba 1993). To do so, we partition $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = [\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}]$, where $\boldsymbol{x}_i^{(j)}$ represents the L_j SNP predictors for gene j (j=1, 2). Then the function $m(\cdot)$ in model (1) can be represented by

$$m(\boldsymbol{x}_i) = \mu + m_1(\boldsymbol{x}_i^{(1)}) + m_2(\boldsymbol{x}_i^{(2)}) + m_{12}(\boldsymbol{x}_i^{(1)}, \mathbf{x}_i^{(2)}) \tag{2}$$

where μ is the intercept, functions m_1 and m_2 represent the main effects of the two genes, and m_{12} describe the interaction between the two genes. We assume that m is a member of some "smooth" class of functions of $\boldsymbol{x}$, and estimate then as the minimizer of some objective function in an appropriate function space, for example, minimizing a penalized sum of squares

$$\mathcal{L}(y, m) = \sum_{i=1}^{n} (y_i - m(\boldsymbol{x}_i))^2 + \lambda J(m) \tag{3}$$

where J is a roughness penalty.

2.2 Gene-gene Interaction in the Reproducing Kernel Hilbert Space

A smoothing spline ANOVA (SS-ANOVA) model (Wahba et. al. 1995; Gu and Wahba, 1993) provides a unique ANOVA-like decomposition of m of the form as in model (2). Base on the decomposition, an Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of functions on $\mathcal{F}$ could be constructed, in the sense that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of m in $\mathcal{H}$. Here $\mathcal{F}$ is a measurable space, $[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}] = \boldsymbol{x} \in \mathcal{F} = \mathcal{F}^{(1)} \otimes \mathcal{F}^{(2)}$, where $\otimes$ refers to the direct product. Then RKHS methods can be used to impose smoothness penalties on each component with the form $\lambda_1 J_1(m_1) + \lambda_2 J_2(m_2) + \lambda_3 J_3(m_{12})$ (Wahba et al. 1995).

Let $\mathcal{H}^{(j)}$ be an RKHS of functions on $\mathcal{F}^{(j)}$, $j = 1, 2$ with $\int_{\mathcal{F}^{(j)}} m_j(\boldsymbol{x}^{(j)}) d\mu_j = 0$, for $m_j(\boldsymbol{x}^{(j)}) \in \mathcal{H}^{(j)}$, and let $[\mathbf{1}^{(j)}]$ be the space of constant functions on $\mathcal{F}^{(j)}$. Following Wahba et al. (1995), construct $\mathcal{H}$ as

$$\begin{aligned}\mathcal{H} &= \prod_{j=1}^{2}([\mathbf{1}^{(j)}] \oplus \mathcal{H}^{(j)}) \\ &= [\mathbf{1}] \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)} \oplus (\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}) \\ &= [\mathbf{1}] \oplus \mathcal{H}^{1} \oplus \mathcal{H}^{2} \oplus \mathcal{H}^{3}\end{aligned}$$

where $\oplus$ refers to the direct sum.

As an element of the RKHS, function m can be estimated as the function in $\mathcal{H}$ which minimizes the following objective function

$$\mathcal{L}(y, m) = \sum_{i=1}^{n}(y_i - m(\boldsymbol{x}_i))^2 + \frac{1}{2}\sum_{l=1}^{3}\lambda_l \parallel P^l m(.) \parallel_{\mathcal{H}}^2 \tag{4}$$

where P^l is the orthogonal projector in $\mathcal{H}$ onto $\mathcal{H}^l$, λ_l are the tuning parameters which balance the goodness of fit and complexity of the model. The minimizer of (4) is known to have a representation (Wahba, 1990, Chapter 10) in terms of a constant and the reproducing kernels $\{k_l(s,t)\}$ for the $\mathcal{H}^l$. Letting $Q_\theta(s,t) = \sum_{l=1}^{3}\theta_l k_l(s,t)$, then

$$\begin{aligned}m_\theta(\boldsymbol{x}) &= \mu\mathbf{1} + \sum_{i=1}^{n} c_i Q_\theta(\boldsymbol{x}_i, \boldsymbol{x}) \\ &= \mu\mathbf{1} + \sum_{i=1}^{n} c_i \sum_{l=1}^{3}\theta_l k_l(\boldsymbol{x}_i, \boldsymbol{x}) \qquad (5) \\ &= \mu\mathbf{1} + \sum_{l=1}^{3}\mathbf{K}_l^T(\boldsymbol{x})C_l\end{aligned}$$

where $\mathbf{K}_l^T(\boldsymbol{x}) = (k_l(\boldsymbol{x}_1, \boldsymbol{x}), \cdots, k_l(\boldsymbol{x}_n, \boldsymbol{x}))$, $C_l = (c_1, \cdots, c_n)\theta_l$. Details on the choice of the reproducing kernel functions corresponding to the three subspaces will be discussed in later sections.

Substituting the representation of $m(\cdot)$ into equation (4), we get:

$$\begin{aligned}
\mathcal{L}(y,m) &= \sum_{i=1}^{n}(y_i - m(\boldsymbol{x}_i))^2 + \frac{1}{2}\sum_{l=1}^{3}\lambda_l \parallel P^l m(.) \parallel_{\mathcal{H}}^2 \\
&= (\boldsymbol{y} - m(\boldsymbol{X}))^T(\boldsymbol{y} - m(\boldsymbol{X})) + \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l \qquad (6)\\
&= (\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l)^T(\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l) + \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l
\end{aligned}$$

where

$$\mathbf{K}_l = \begin{bmatrix} K_l^T(\boldsymbol{x}_1) \\ K_l^T(\boldsymbol{x}_2) \\ \vdots \\ K_l^T(\boldsymbol{x}_n) \end{bmatrix}$$

The gradients of $\mathcal{L}$ with respect to the coefficients (μ, $C_l : l = 1,2,3$) are

$$\frac{\partial \mathcal{L}}{\partial \mu} = \mathbf{1}^T(y - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l) = 0 \qquad (7)$$

and

$$\frac{\partial \mathcal{L}}{\partial C_l} = \mathbf{K}_l^T(y - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l) + \lambda_l \mathbf{K}_l C_l = 0 \qquad (8)$$

The first order condition is satisfied by the system

$$\begin{bmatrix} n & \mathbf{1}^T\mathbf{K}_1 & \mathbf{1}^T\mathbf{K}_2 & \mathbf{1}^T\mathbf{K}_3 \\ \mathbf{K}_1^T\mathbf{1} & \mathbf{K}_1^T\mathbf{K}_1 - \lambda_1\mathbf{K}_1 & \mathbf{K}_1^T\mathbf{K}_2 & \mathbf{K}_1^T\mathbf{K}_3 \\ \mathbf{K}_2^T\mathbf{1} & \mathbf{K}_2^T\mathbf{K}_1 & \mathbf{K}_2^T\mathbf{K}_2 - \lambda_2\mathbf{K}_2 & \mathbf{K}_2^T\mathbf{K}_3 \\ \mathbf{K}_3^T\mathbf{1} & \mathbf{K}_3^T\mathbf{K}_1 & \mathbf{K}_3^T\mathbf{K}_2 & \mathbf{K}_3^T\mathbf{K}_3 - \lambda_3\mathbf{K}_3 \end{bmatrix} \begin{bmatrix} \mu \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{K}_1^T \\ \mathbf{K}_2^T \\ \mathbf{K}_3^T \end{bmatrix} y$$

Simple calculation shows that the above system is equivalent to

$$\begin{bmatrix} n & \mathbf{1}^T & \mathbf{1}^T & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} - \lambda_1\mathbf{K}_1^{-1} & \mathbf{I} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} - \lambda_2\mathbf{K}_2^{-1} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} & \mathbf{I} - \lambda_3\mathbf{K}_3^{-1} \end{bmatrix} \begin{bmatrix} \mu \\ m_1 \\ m_2 \\ m_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{I} \\ \mathbf{I} \\ \mathbf{I} \end{bmatrix} y$$

where $m_l = \mathbf{K}_l C_l$, $l = 1,2$ and $m_{12} = \mathbf{K}_3 C_3$.

Let $\tau_l^2 = \sigma^2/\lambda_l, l = 1,2,3$, the system is exactly the Henderson's normal equation of the following linear mixed model

$$y = \mu\mathbf{1} + m_1 + m_2 + m_{12} + \epsilon \qquad (9)$$

where m_1, m_2, m_{12} are independent $n \times 1$ vector of random effects; $m_1 \sim N(\mathbf{0}, \tau_1^2 \mathbf{K}_1)$, $m_2 \sim N(\mathbf{0}, \tau_2^2 \mathbf{K}_2)$, $m_{12} \sim N(\mathbf{0}, \tau_3^2 \mathbf{K}_3)$, and $\epsilon \sim N(0, \sigma^2 I)$ is independent of m_1, m_2 and m_{12}. This dual representation of linear mixed model for the ANOVA model makes it feasible to do inferences about the main and interaction components under the mixed effects model framework. Estimation of the parameters can be done by using maximum likelihood method or the restricted maximum likelihood (REML) method. Since REML method gives unbias estimates for the variance components, we adopt the REML estimation in this work.

2.3 Choice of the Kernel Function for Genetic Similarity

The choice of reproducing kernel is not arbitrary in the sense that the kernel function must be positive-definite. By theorem 1.1.1 (Wahba 1990, p2), given a positive-definite function k on $\mathcal{F} \times \mathcal{F}$, we can construct a unique RKHS of real-valued functions on $\mathcal{F}$ with k as its reproducing kernel. In a genetic association study, a kernel function captures the pairwise genomic similarities across multiple SNPs in a gene. It projects the genotype data from the original space, which can be high dimensional and nonlinear, to a one-dimensional linear space. The Allele Matching (AM) kernel is one of the most popularly used kernels for measuring genotype similarity. This type of kernel measure has been used in linkage analysis (Weeks and Lange, 1988) and in association studies (Tzeng et al, 2003; Schaid, et al., 2005; Wessel and Schork, 2006; Kwee, et al., 2008 and Mukhopadhyay et al., 2010). For a review of genomic similarity and kernel methods, readers are referred to Schaid (2010a, b). With the notable strength that it does not require knowledge of the risk allele for each SNP, AM kernel is chosen as the kernel function in this study. This similarity kernel counts the number of matches among the four comparisons between two genotypes $g_{i,s}$ (with two alleles A and B) and $g_{j,s}$ (with two alleles C and D) of two individuals i and j at locus s, and can be expressed as

$$AM(g_{i,s} = A/B, g_{j,s} = C/D) = I(A \equiv C) + I(A \equiv D) + I(B \equiv C) + I(B \equiv D) \qquad (10)$$

where I is the indicator function and "≡" means the two alleles are in identical-by-state (IBS). The kernel function based on AM similarity measure then takes the form

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} AM(g_{i,s}, g_{j,s})}{4S} \tag{11}$$

where S is the number of SNPs considered for each kernel function.

To incorporate valuable SNP-specific information into analysis to potentially improve performance, a weighted-AM kernel can be applied which has the form

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} w_s AM(g_{i,s}, g_{j,s})}{4\sum_{s=1}^{S} w_s} \tag{12}$$

where w_s is the weighting function which can incorporate information about minor allele frequency or p-values depends on the underlying study purpose to gain extra power. For example, when a study is trying to identify the function of rare variants, the weight function can be taken as the inverse of the minor allele frequency to boost the signal for rare variants (Schaid, 2010b).

This AM kernel can be used as the reproducing kernel for the two subspaces $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ corresponding to the main effect of the two genes. Utilizing the fact that the reproducing kernel for a tensor product of two reproducing kernel spaces is the product of the two reproducing kernels (Aronszajn, 1950), the reproducing kernel for subspace $\mathcal{H}^{(3)}$ corresponding to the interaction effect of the two genes can be taken as the product of the reproducing kernels of the two main subspaces.

2.4 Hypothesis Testing

2.4.1 Testing overall genetic effect

In a gene-based genetic association study, one is interested in whether a gene as a system is associated with a disease trait. In the proposed 3G interaction study, we are interested in the association of each gene with a quantitative trait as well as the interaction between genes if any. The analysis starts with a two-dimensional pairwise G×G interaction search. Testing the overall contribution of a gene pair to a phenotypic trait is equivalent to test $H_0 : m_1(\boldsymbol{x}^{(1)}) = m_2(\boldsymbol{x}^{(2)}) = m_{12}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = 0$. Similarly testing for interaction effect can be formulated as $H_0 : m_{12}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = 0$. With the linear mixed model representation, parameters $\tau_1^2, \tau_2^2, \tau_3^2, \sigma^2$ are treated as the variance components in model (9). Correspondingly, the aforementioned two tests for the overall and interaction effects can be defined as (I) $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ and (II) $H_0^2 : \tau_3^2 = 0$, respectively.

A typical issue in variance component analysis is that the parameters under the null hypotheses are on the boundary of the parameter space. Moreover, the kernel matrices $\mathbf{K}_s$'s are not block-diagonal. Thus, the asymptotic null distribution for testing H_0^1 by using the likelihood ratio test (LRT) does not follow a central chi-square distribution under the null. The mixture chi-square distribution proposed by Self and Liang (1987) under irregular conditions does not apply in our case either. To ease the problem, we consider a score test based on the restricted likelihood. Consider the linear mixed model in (9) in which $y \sim N(\mu\mathbf{1}, V(\beta))$, the restricted log-likelihood function can be written as

$$\ell_R \propto -\frac{1}{2}ln(|V(\beta)|) - \frac{1}{2}ln(|\mathbf{1}^T V^{-1}(\beta)\mathbf{1}|) - \frac{1}{2}(y - \hat{\mu}\mathbf{1})^T V(\beta)^{-1}(y - \hat{\mu}\mathbf{1}) \tag{13}$$

where $\beta = (\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$, $V(\beta) = \sigma^2 I + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2 + \tau_3^2\mathbf{K}_3$. The first derivative of the restricted log-likelihood function with respect to each variance component is given by

$$\frac{\partial \ell_R}{\partial \beta_i} = -\frac{1}{2}tr(RV_i) + \frac{1}{2}(y - \hat{\mu}\mathbf{1})^T V^{-1}(\beta) V_i V^{-1}(\beta)(y - \hat{\mu}\mathbf{1}) \tag{14}$$

where $V_i = \frac{\partial V(\beta)}{\partial \beta_i}, i = 1, \cdots, 4$, so $V_1 = I, V_2 = \mathbf{K}_1, V_3 = \mathbf{K}_2, V_4 = \mathbf{K}_3$ and $R = V^{-1} - V^{-1}\mathbf{1}(\mathbf{1}^T V^{-1}\mathbf{1})^{-1}\mathbf{1}^T V^{-1}$.

The restricted score function under the null hypothesis: $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ is

$$\frac{\partial \ell_R}{\partial \beta_i}|_{\tau_1^2=\tau_2^2=\tau_3^2=0} = -\frac{1}{2\sigma^2}tr(P_0 V_i) + \frac{1}{2\sigma^4}(y - \hat{\mu}\mathbf{1})^T V_i (y - \hat{\mu}\mathbf{1})$$

where $P_0 = \mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ is the projection matrix under the null. Thus, H_0^1 can be tested by the following score statistic

$$S(\sigma^2) = \frac{1}{2\sigma^2}(y - \hat{\mu}_0\mathbf{1})^T \sum_{l=1}^{3} \mathbf{K}_l(y - \hat{\mu}_0\mathbf{1})$$

where $\hat{\mu}_0 = (\mathbf{I} - P_0)y$ is the MLE of μ under the null. This leads to

$$S(\sigma^2) = \frac{1}{2\sigma^2} y^T P_0 \sum_{l=1}^{3} \mathbf{K}_l P_0 y \tag{15}$$

Denoting σ_0^2 as the true value of σ^2 under the null, then $S(\sigma_0^2)$ is a quadratic form in y. Following Liu and Lin (2007), we use the satterthwaite method to approximate the distribution of $S(\sigma_0^2)$ by a scaled chi-square distribution, i.e., $S(\sigma_0^2) \sim a\chi_g^2$, where the scale parameter a and the degrees of freedom g can be estimated by the method of moments (MOM). By equating the mean and variance of the test statistic $S(\sigma_0^2)$ with those of $a\chi_g^2$, we have

$$\begin{cases} \delta = E[S(\sigma_0^2)] = tr(P_0 \sum_{i=1}^{3} \mathbf{K}_i)/2 = E[a\chi_g^2] = ag \\ \nu = Var[S(\sigma_0^2)] = tr(\sum_{i=1}^{3}(P_0\mathbf{K}_i) \sum_{i=1}^{3}(P_0\mathbf{K}_i))/2 = Var[a\chi_g^2] = 2a^2 g \end{cases} \tag{16}$$

Solving for the two equations leads to $\hat{a} = \nu/2\delta$ and $\hat{g} = 2\delta^2/\nu$.

In practice, we do not know the true value σ_0^2 and we usually estimate it by its MLE under the null model, denoted as $\hat{\sigma}_0^2$. The asymptotic distribution of $S(\hat{\sigma_0}^2)$ can still be approximated by the scaled chi-square distribution because the MLE is $\sqrt{n}$ consistent. To account for the fact that σ_0^2 is estimated by MLEs, we estimate a and g by replacing ν with $\tilde{\nu}$ based on the efficient information. The Fisher's information matrix of $\boldsymbol{\tau} = (\tau_1^2, \tau_2^2, \tau_3^2)$ is given by

$$I_{\boldsymbol{\tau\tau}} = \frac{1}{2}\begin{bmatrix} tr(P_0\mathbf{K}_1P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_1P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_1P_0\mathbf{K}_3) \\ tr(P_0\mathbf{K}_2P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_2P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_2P_0\mathbf{K}_3) \\ tr(P_0\mathbf{K}_3P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_3P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_3P_0\mathbf{K}_3) \end{bmatrix}$$

$$I_{\boldsymbol{\tau}\sigma^2} = \frac{1}{2}\begin{pmatrix} tr(P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_3) \end{pmatrix}^T$$

and $I_{\sigma^2\sigma^2} = \frac{1}{2}tr(P_0P_0)$. Then the efficient information $\tilde{I}_{\boldsymbol{\tau\tau}} = I_{\boldsymbol{\tau\tau}} - I^T_{\boldsymbol{\tau}\sigma^2} I^{-1}_{\sigma^2\sigma^2} I_{\boldsymbol{\tau}\sigma^2}$ and

$$\tilde{\nu} = Var[S(\hat{\sigma}^2)] \approx SUM[\tilde{I}_{\boldsymbol{\tau\tau}}] \tag{17}$$

where operator "SUM" indicates the sum of every elements of the matrix.

2.4.2 Testing G×G interaction

For testing the significance of the interaction term, i.e., testing $H_0^2 : \tau_3^2 = 0$, we also apply a score test. Denote $\Sigma = \sigma^2 I + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2$. The score function (14) under the null becomes:

$$\begin{aligned} \frac{\partial \ell_R}{\partial \tau_3^2}|_{\tau_3^2=0} &= -\frac{1}{2}[tr(P\mathbf{K}_3) - (y - \hat{\mu}\mathbf{1})^T\Sigma^{-1}\mathbf{K}_3\Sigma^{-1}(y - \hat{\mu}\mathbf{1})] \\ &= -\frac{1}{2}(tr(P_{01}\mathbf{K}_3) - y^TP_{01}\mathbf{K}_3Py) \end{aligned} \tag{18}$$

where $P_{01} = \Sigma^{-1} - \Sigma^{-1}\mathbf{1}(\mathbf{1}^T\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Sigma^{-1}$ is the projection matrix under the null, then

$$S_I = \frac{1}{2}y^TP_{01}\mathbf{K}_3P_{01}y \tag{19}$$

Similarly, Satterthwaite approximation can be used to approximate the distribution of S_I by $a_I\chi^2_{g_I}$. Parameters a_I and g_I are estimated by MOM. Specifically, $\hat{a}_I = \nu_I/2\delta_I$ and $\hat{g}_I = 2\delta_I^2/\nu_I$, where $\delta_I = \frac{1}{2}tr(P_{01}\mathbf{K}_3)$ and $\nu_I = \frac{1}{2}tr({}_{01}P\mathbf{K}_3P_{01}\mathbf{K}_3) - \frac{1}{2}\Phi^T\Delta^{-1}\Phi$, where $\Phi = [tr(P_{01}^2\mathbf{K}_3), tr(P_{01}\mathbf{K}_3P_{01}\mathbf{K}_1), tr(P_{01}\mathbf{K}_3P_{01}\mathbf{K}_2)]^T$ and

$$\Delta = \begin{bmatrix} tr(P_{01}^2) & tr(P_{01}^2\mathbf{K}_1) & tr(P_{01}^2\mathbf{K}_1) \\ tr(P_{01}^2\mathbf{K}_1) & tr(P_{01}\mathbf{K}_1P_{01}\mathbf{K}_1) & tr(P_{01}\mathbf{K}_1P_{01}\mathbf{K}_2) \\ tr(P_{01}^2\mathbf{K}_2) & tr(P_{01}\mathbf{K}_2P_{01}\mathbf{K}_1) & tr(P_{01}\mathbf{K}_2P_{01}\mathbf{K}_2) \end{bmatrix}$$

3 Simulation Study

3.1 Simulation Design

Monte Carlo simulations were conducted to evaluate the performance of the proposed approach for detecting genetic effects as well as gene-gene interaction in an association study. The genotype data were simulated using two approaches introduced in Cui et al. (2008). In the following, we described the details of the two genotype generating methods: MS program and LD-based simulation.

MS program: The MS program developed by Hudson (2002) generates haplotype samples by using the standard coalescent approach in which the random genealogy of a sample is first generated and the mutations are randomly placed on the Genealogy. We first simulated two independent samples of haplotypes by using MS program. Parameters of the coalescent model were set as following: (1) The diploid population size $N_0 = 10,000$; (2) The mutation parameter $\theta = 4N_0\mu = 5.610 \times 10^{-4}/bp$; and (3) The cross-over rate parameters are $\rho = 4N_0 r = 4.0\times10^{-3}/bp$ and $\rho = 8\times10^{-3}/bp$ for the two samples. In each sample, 100 haplotypes were simulated for a locus with 10kb long and the number of SNP sequences were set to be 100. Two haplotypes were then randomly drawn within each simulated haplotype pool and paired to form the genotype on the locus for an individual. For each individual, we randomly selected 10 adjacent SNPs with minor allele frequency (MAF) greater than 5% to form a gene. This was done separately for each simulated haplotype pool and finally we had genotypes for n individuals for two separate genes with 10 SNPs each, and the two genes were independent.

LD-based simulation: Under this scenario, SNP genotypes were simulated by controlling pairwise LD values. Let p_A be the MAF for SNP1. Assuming Hardy-Weinberg equilibrium (HWE), the first SNP marker can be simulated according to a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$ for genotype AA, Aa and aa, respectively. Let the MAF of the next simulated marker (SNP2) as p_B and the LD between SNP1 and SNP2 be D. Assuming HWE, the four haplotype frequencies can be calculated as $p_{AB} = p_A p_B + D$, $p_{Ab} = p_A(1-p_B) - D$, $p_{aB} = (1-p_A)p_B - D$ and $p_{ab} = (1-p_A)(1-p_B) + D$ for haplotype AB, Ab, aB and ab, respectively. The conditional genotype distribution of SNP2 given on

SNP1 can be derived as

$$P(BB|AA) = \frac{P(AABB)}{P(AA)} = \frac{p_{AB}^2}{p_A^2} = \frac{(p_A p_B + D)^2}{p_A^2} \tag{20}$$

Similarly we can get the other 8 conditional genotype distributions (see Table 1 in Cui. et al (2008) for more details). Two genes with 10 SNPs each were simulated by applying the LD-based simulation method. For gene 1, we assume MAF=0.3 and pairwise SNP correlation $r^2 = 0.5$ ($r^2 = \frac{D^2}{p_A p_B (1-p_A)(1-p_B)}$). For gene 2, we assume MAF=0.2, and r^2=0.8.

Phenotype simulation: Four simulation scenarios were considered in simulating the phenotype (Table 1). In Scenario I, the three genetic effects were set as zero, with which we can assess the false positive control of different methods. In Scenario II, we considered the main effects for the two genes, but set the interaction effect as zero. In Scenarios III and IV, both main effects and interaction effect were considered. The difference between the scenario III and IV is that the interaction effect in Scenario III is smaller than the main effect, while in Scenario IV it is larger than both main effects. Quantitative trait of interest were simulated from a multivariate normal distribution with mean $\mu\mathbf{1}_{n\times 1}$ and variance-covariance matrix $V = \sigma^2\mathbf{I} + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2 + \tau_3^2\mathbf{K}_3$, where $\tau_1^2, \tau_2^2, \tau_3^2$ took different values under different scenarios; $\mathbf{K}_i, i = 1, 2, 3$ are the kernel matrices using the allele matching method described before. Different sample sizes (n = 200 and 500) and different heritability (H^2=0.1, 0.2, 0.4) were assumed. Let $\sigma_G^2 = \tau_1^2 + \tau_2^2 + \tau_3^2$, then the heritability is defined as $H^2 = \sigma_G^2/(\sigma_G^2 + \sigma^2)$. For a given value of residual variance σ^2, the main effects of the two genes were set equal. When the interaction effect was considered, it was set as either half of the main effect (Scenario III) or double the main effect (Scenario IV). Thus for a given heritability level, the parameter values were different under different scenarios. Specific values for $\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2$ were given in the first column of Table 1.

3.2 Model Comparison

We mainly compared our simulation results with two other methods described in the follows. Wang et al. (2010) proposed an interaction method using a partial least square approach which is developed specifically for binary disease traits. The method cannot be applied for quantitative traits. However, in Wang et al.'s paper they compared their method with a regression-based principle component analysis method. Specifically, assuming an additive model for each marker in which genotypes AA, Aa and aa are coded as 2,1,0, respectively, the singular value decomposition (SVD) can be applied to both gene matrices. Let G_j be an $n \times L_j$ SNP matrix for gene j $(= 1, 2)$. The SVD for G_j can be expressed as $G_j = U_j D_j V_j^T$, where D_j is a diagonal matrix of singular values, and the elements of the column vector U_j are the principal components $U_j^1, U_j^2, \cdots, U_j^{m_j}$ $(m_j \leq L_j$ is the rank for $G_j)$. An interaction model can be expressed as

$$y = \mu + \sum_{l_1=1}^{L_1} \beta_{l_1} x_{l_1} + \sum_{l_2=1}^{L_2} \beta_{l_2} x_{l_2} + \gamma U_1^1 U_1^2 \tag{21}$$

where γ represents the interaction effect between the first pair of PCs corresponding to the largest eigenvalues in the two genes. The main effect of the each gene is modeled through the sum of all single marker effects. For simplicity, only one interaction effect between the first PC corresponding to the largest eigenvalues in each gene was considered in Wang et al. (2010). We followed Wang et al. (2010) and compared the performance of our model with this model.

In principle, one can select PCs for each gene based on the proportion of variation explained (say $> 85\%$). Then, pairwise interactions can be considered for all selected PCs in model (21). Thus, we replaced the main effect of each gene in model (21) with PCs rather than single SNPs to reduce the model degrees of freedom, model (21) then becomes

$$y = \mu + \sum_{k_1=1}^{K_1} \beta_{k_1} U_{k_1} + \sum_{k_2=1}^{K_2} \beta_{k_2} U_{k_2} + \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{k_1 k_2} U_{k_1}^1 U_{k_2}^2 \tag{22}$$

where $U_{k_j}, j = 1, 2$ represents the PCs for gene j, and $K_j, j = 1, 2$ is chosen based on the proportion of variation explained by the number of PCs in gene j. With this regression model, we considered all possible pairwise PC interactions between the two genes and G×G interaction was done by testing $H_0 : \gamma_{k_1 k_2} = 0$, for all k_1 and k_2. This model was applied by He et al. (2010) in their gene-based interaction analysis.

In addition to the above two models, we also compared our gene-centric approach to a pairwise SNP interaction model. Details of the comparison is given in the following section. For a given simulation scenario, 1000 simulation runs were conducted. Type I error rates and power were examined at the nominal level $\alpha = 0.05$.

3.3 Simulation Results

Table 1 summarizes the comparison results between our kernel method and model (21) and (22). The power of an association test was denoted by $P_1^{\cdot}$, $P_2^{\cdot}$ and $P_3^{\cdot}$ which correspond to the power by using the proposed gene-centric interaction method, model (21) and (22), respectively. The superscript letters o and i denote the power for testing the significance of the overall genetic effects and the interaction effect, respectively. Noted that the power for the interaction test was calculated only when the overall test showed significance. Thus, the power and the false positive rate for the interaction test are smaller than the ones obtained without this constraint.

3.3.1 Comparisons of the proposed method with the two PCA-based methods

The results for Scenario I indicate that our method has reasonable type I error rate control for the overall genetic effect tests under the two genotype simulation scenarios (see Scenario I in Table 1). The two PCA-based interaction models produced a little conservative results when the genotypes were simulated with the MS program. For example, the type I error rates were 0.033 and 0.023 for the two methods when sample size is 500.

Table 1: List of empirical type I error and power based on 1000 simulation runs.

Parameter values			LD-based						MS program					
$(\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$	H^2	n	P_1^{o*}	P_1^{i*}	P_2^{o*}	P_2^{i*}	P_3^{o*}	P_3^{i*}	P_1^{o*}	P_1^{i*}	P_2^{o*}	P_2^{i*}	P_3^{o*}	P_3^{i*}
Scenario I														
(1,0,0,0)	0	200	0.049	0.004	0.045	0.016	0.095	0.025	0.070	0.002	0.048	0.025	0.034	0.011
		500	0.061	0.002	0.044	0.021	0.055	0.012	0.052	0.001	0.033	0.019	0.023	0.008
Scenario II														
(0.8, 0.044, 0.044,0)	0.1	200	0.285	0.019	0.212	0.032	0.209	0.042	0.255	0.016	0.186	0.057	0.115	0.019
		500	0.531	0.026	0.420	0.052	0.374	0.043	0.525	0.036	0.339	0.045	0.254	0.030
(0.8, 0.1, 0.1,0)	0.2	200	0.459	0.029	0.386	0.058	0.387	0.055	0.485	0.044	0.324	0.058	0.253	0.041
		500	0.776	0.048	0.636	0.045	0.686	0.042	0.755	0.041	0.615	0.071	0.594	0.050
(0.8, 0.267, 0.267,0)	0.4	200	0.734	0.072	0.661	0.058	0.684	0.072	0.758	0.080	0.611	0.066	0.604	0.052
		500	0.927	0.065	0.862	0.069	0.939	0.071	0.946	0.066	0.842	0.066	0.917	0.048
Scenario III														
(0.8, 0.036, 0.036, 0.018)	0.1	200	0.289	0.025	0.234	0.051	0.238	0.037	0.299	0.019	0.164	0.041	0.126	0.027
		500	0.565	0.054	0.415	0.059	0.414	0.062	0.548	0.030	0.399	0.069	0.298	0.034
(0.8, 0.08, 0.08, 0.04)	0.2	200	0.486	0.053	0.389	0.065	0.389	0.046	0.491	0.069	0.346	0.056	0.279	0.046
		500	0.806	0.086	0.686	0.085	0.746	0.074	0.752	0.061	0.640	0.089	0.632	0.045
(0.8, 0.21, 0.21, 0.11)	0.4	200	0.765	0.109	0.654	0.087	0.740	0.107	0.766	0.100	0.629	0.091	0.616	0.069
		500	0.946	0.163	0.881	0.131	0.956	0.140	0.941	0.131	0.872	0.128	0.914	0.097
Scenario IV														
(0.8, 0.022, 0.022, 0.044)	0.1	200	0.318	0.047	0.245	0.048	0.253	0.051	0.280	0.027	0.189	0.051	0.136	0.032
		500	0.571	0.064	0.466	0.090	0.432	0.062	0.571	0.038	0.449	0.089	0.325	0.045
(0.8, 0.05, 0.05, 0.1)	0.2	200	0.500	0.074	0.409	0.076	0.443	0.087	0.514	0.053	0.377	0.062	0.291	0.043
		500	0.805	0.141	0.720	0.117	0.755	0.119	0.787	0.111	0.669	0.119	0.667	0.105
(0.8, 0.133, 0.133, 0.266)	0.4	200	0.771	0.172	0.694	0.115	0.750	0.136	0.779	0.153	0.619	0.103	0.680	0.092
		500	0.938	0.304	0.881	0.230	0.961	0.244	0.963	0.256	0.874	0.211	0.955	0.194

* P.^{o} and P.^{i} refer to the power for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$) and for testing interaction effect (i.e., $H_0 : \tau_3^2 = 0$), respectively. P$_1$, P$_2$ and P$_3$ refer to powers by using the proposed gene-centric method, the full PCA-based interaction with model (22) and the partial PCA-based interaction analysis with model (21), respectively.

In Scenarios II-IV, we fixed the residual variance σ^2 to 0.8, and varied the three genetic effects to get different heritability levels. As we expected, the testing power increases as the heritability level and sample size increase. For example, under the LD-based simulation, the overall power increases from 0.565 to 0.946 when H^2 increases from 0.1 to 0.4 with fixed sample size 500 in Scenario III. Under the same Scenario, the overall power increases from 0.486 to 0.806 when sample size increases from 200 to 500 under fixed H^2. We observed a similar trend for genotypes simulated with the MS program (Table 1).

Relatively small interaction power were observed for the three methods (partly due to the way we calculated the interaction power). As sample size or heritability increase, the interaction power also increases. Larger interaction effect (Scenario IV) results in larger interaction power compared to the one obtained with smaller interaction effect (Scenario III). For example, for fixed sample size ($n = 500$) and fixed heritability ($H^2 = 0.4$), the interaction power increases from 16% to 30% under the LD-based simulation when the interaction effect was doubled. We did additional simulation by increasing the sample size to 1000 and achieved reasonable interaction power (data not shown). The simulation results indicate that large sample size is needed in order to obtain reasonable power to detect the interaction effect.

3.3.2 Model performance under different interaction effect sizes

Epistasis may be caused by a variety of underlying mechanisms. Some genes might have both significant marginal and epistatic effects, while others might only incur epistatic effects without main effects. Simulation studies were designed to evaluate the performance of the proposed kernel machine approach in discovering gene × gene interaction under different epistasis effect sizes. We defined the proportion of the epistatic variance among the total genetic variance as $\rho = \tau_3^2/(\tau_1^2 + \tau_2^2 + \tau_3^2)$, which gave us an indication of the strength of the epistatic effect between two genes for a fixed total genetic variance.

Two genes each with 10 SNPs were considered as in previous simulation studies. Genotype data and phenotype data were generated as described in Section 3.1, but with different values for the variance components. For a given heritability level ($H^2 = 0.4$) and a fixed residual error variance ($\sigma^2 = 0.6$), the total genetic variance is calculated as 0.4. We then assumed the same effect size for the two main components, and varied the proportion ρ. For example, we had $(\tau_1^2, \tau_2^2, \tau_3^2) = (0.16, 0.16, 0.08)$ when $\rho = 0.2$, and $(\tau_1^2, \tau_2^2, \tau_3^2) = (0.04, 0.04, 0.32)$ when $\rho = 0.8$. Six values of proportion $\rho = (0, 0.2, 0.4, 0.6, 0.8, 1.0)$ were considered, including the two extreme cases: no epistatic effect at all ($\rho = 0$) and pure epistasis ($\rho = 1$). Comparisons with the other two PCA-based interaction analyses were considered under two different sample sizes, 500 and 1000. Empirical powers was calculated based on testing the interaction effect only.

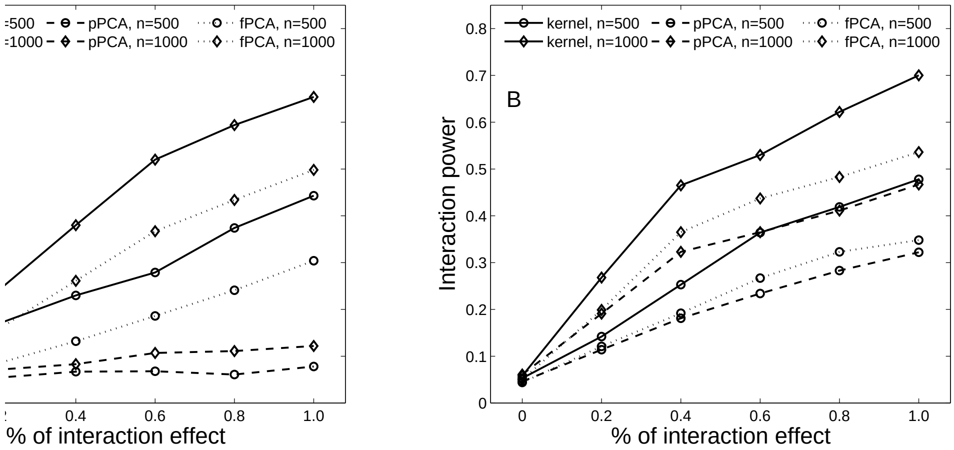

Figure 1: Power comparison of the proposed kernel approach (solid line), the partial PCA-based interaction model (21) (dashed line, denoted as pPCA) and the full PCA-based interaction model (22) (dotted line, denoted as fPCA) under different sample sizes and different proportions (ρ) of epistasis variance. Genotypes were simulated with the MS program (A) and the LD-based algorithm (B).

Results based on 1000 replicates were summarized in Figure 1. All the three methods can reasonably control the type I error ($\rho = 0$). As we expected that the empirical interaction power increases as the interaction effect size increases. When SNPs are correlated (Figure 1B), small number of PCs might be enough to capture the variation of each gene. So the power is larger than MS-based simulation (Figure 1A). Among the three methods, our kernel-based method has the highest power. Model (21) has the lowest power, which implies that only considering one pair of PC interaction is not enough to capture the interaction effect between two genes. The effect of sample size on the interaction power is also significant. Larger sample size always leads to larger power. The results also confirm that detecting gene × gene interactions generally requires relatively larger sample size than it does for detecting main genetic effects.

3.3.3 Comparison with the single SNP interaction model

In a regression-based analysis for interaction, the commonly used approach is the single SNP interaction model with the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{23}$$

where β_0 is the intercept; β_1, β_2 and β_{12} represent the effects of SNP x_1 in gene 1, SNP x_2 in gene 2 and the interaction effect between the two; and $\varepsilon \sim N(0, \sigma^2)$. We simulated data according to model (23) assuming a MAF $p_A = 0.3$. Different heritabilities and different sample sizes were assumed. Obviously it is unfair to compare the two since the single SNP interaction model is the true analytical model and it should have the best performance. However, it is worth to evaluate the performance of our kernel method when there is only one functional pair of SNPs in two genes. For simplicity, we assumed the same effect size for the three coefficients which are calculated under specific heritability ($H^2 = 0.2$ and 0.4) when generating the data. We considered an extreme case in which each gene only contains one single SNP. Data generated with model (23) are subject to both the single SNP interaction and the proposed kernel interaction analysis. The results are summarized in Table 2.

Both models show comparable type I error control for the overall genetic test (see P_o in Table 2). The interaction test is nested within the overall genetic test. If we aggregate the results by dividing P_i by P_o, the single SNP analysis produces more inflated false positives compared to the kernel approach when no genetic effect is involved at all. When data were simulated assuming only main effects but no interaction (case $\beta_{12} = 0$), the two approaches yield very similar false positive rate, indicating reasonable performance of the kernel approach for false positive control.

For the power analysis, we found very minor difference between the two methods for the overall genetic test (P_o), especially under large sample size and high heritability level. In fact, when sample size is 200 and heritability level is 0.2, the kernel method has higher

Table 2: List of empirical type I error and power based on 1000 simulation runs (single SNP interaction model).

Heritability	Coefficients	Sample size	Single SNP interaction		Kernel interaction	
(H^2)	($\beta_0, \beta_1, \beta_2, \beta_{12}$)	(n)	P$_o$	P$_i$	P$_o$	P$_i$
0.2	(0.19, 0, 0, 0)	200	0.055	0.019	0.059	0.003
		500	0.058	0.019	0.057	0.003
		1000	0.052	0.017	0.059	0.003
	(0.19, 0.19, 0.19, 0)	200	0.497	0.03	0.534	0.032
		500	0.923	0.045	0.911	0.046
		1000	0.999	0.048	0.997	0.053
	(0.19, 0.19, 0.19, 0.19)	200	1	0.221	1	0.183
		500	1	0.419	1	0.349
		1000	1	0.714	1	0.635
0.4	(0.51, 0, 0, 0)	200	0.053	0.022	0.053	0.003
		500	0.049	0.016	0.062	0.001
		1000	0.054	0.024	0.057	0.008
	(0.51, 0.51, 0.51, 0)	200	1	0.051	1	0.058
		500	1	0.062	1	0.067
		1000	1	0.054	1	0.058
	(0.51, 0.51, 0.51, 0.51)	200	1	0.850	1	0.648
		500	1	0.996	1	0.964
		1000	1	1	1	1

* P$_o$ and P$_i$ refer to the power for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ for the kernel approach and $H_0 : \beta_1 = \beta_2 = \beta_{12} = 0$ for the pairwise SNP interaction analysis) and for testing interaction effect (i.e., $H_0 : \tau_3^2 = 0$ for the kernel approach and $H_0 : \beta_{12} = 0$ for the pairwise SNP interaction analysis), respectively.

power (53.4%) than the single SNP analysis (49.7%). For the interaction test (P$_i$), as we expected that the power increases as sample size and heritability level increase. For example, P$_i$ increases from 0.183 to 0.635 for the kernel approach when sample size increases from 200 to 1000, a 2.5 fold increase in power under a fixed heritability level (H^2=0.2). When heritability level increases from 0.2 to 0.4 under a fixed sample size (say 500), we saw a dramatic power increase from 0.349 to 0.964 for the kernel approach. Overall, the single SNP interaction model (23) yields slightly higher power than the kernel approach. This is not surprising since one would expect to see large power when simulated data are analyzed with the underlying generating model. However, the difference is diminished under large sample and high heritability level ($n > 500$ and $H^2 = 0.4$). We did additional simulations in which more than one functional SNPs within each gene were involved to interact with each other to affect a trait variation. Results showed that the kernel method consistently outperformed the single SNP interaction model (data not shown).

In summary, our model performs reasonably well in different scenarios compared to the other methods. Even when there is only one single SNP pair interacting with each other in two genes, our analysis produces results as good as the ones analyzed with the true model, especially under large sample size and high heritability (Table 2). For the powers obtained under the two genotype simulation methods, the difference is not remarkable. To achieve high power, large sample size (say $n > 500$) is always encouraged.

4 Applications to Real Data

4.1 Analysis of Baby Birth Weight Data

A candidate gene study was initially conducted for the purpose to study genetic effects associated with large for gestational age (LGA) and small for gestational age (SGA). Subjects were recruited through the Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile, and SNPs were selected for genotyping in order to capture at least 90% of the haplotypic diversity of each gene. Each individuals were genotyped at 797 SNP markers on 186 unique candidate genes. Missing genotypes were imputed using a conditional probability approach as we described in the simulation section. We combined the two data sets (LGA and SGA) and used baby's birth weight (in kg) as the response variable to assess if there are any genes or interaction of genes that could explain the normal variation of new born baby's birth weight. Individuals with birth weight 3×IQR (inter-quartile range) above the Q_3 or below Q_1 were treated as outlier and were discarded. There are total 1511 individuals left after removing outliers.

A two-dimensional pairwise G×G interaction search was conducted (total 17205 gene pairs). The score test for testing $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ was done and p-values were obtained for all gene pairs. Figure 2A shows a two-dimensional plot of -log10 transformed p-values. For a two-dimensional search, it is not clear how to set up a genome-wide threshold to correct for multiple testings. Obviously the 17205 tests are not all independent and using Bonferroni correction may be too stringent. Thus, we used a arbitrary threshold of 0.001 as a cutoff. The yellow hyperplane in Fig. 2A shows the 0.001 cutoff. Totally there are 23 gene pairs were found to be significant with this cutoff. A detailed list of these gene pairs, their effect estimates and the p-values for the overall genetic and interaction test are shown in Table 3. Among the 23 gene pairs, five significant $G \times G$ interactions were detected at the 0.05 level. These are gene pairs ANG-EDN1, PDGFC-PTGER3, PTGS2-PGF, PTGS2-PLAU and IL9-IGF1. A two-dimensional interaction p-value plot is shown in Fig. 2B.

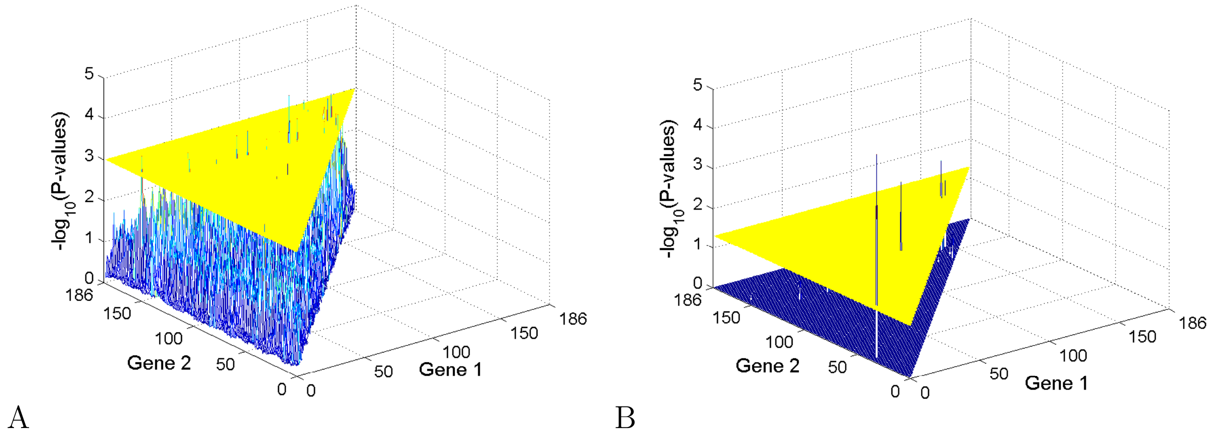

Figure 2: The profile plot of -log10 transformed p-values of all possible gene pairs with the overall genetic effect test (A) and the interaction test (B). The yellow hyperplane represents the 0.001 cutoff for figure A and the 0.05 cutoff for the interaction test for figure B.

The results indicate a strong main genetic effect for gene PDGFC (Platelet-derived growth factor C). This gene is a key component of the PDGFR-α signaling pathway. Studies have shown that PDGFC contributes to normal development of the heart, ear, central nervous system (CNS), and kidney (Reigstad et al. 2005). Even though its main effect is very strong, no strong interaction effects were found between this gene and the rest of the genes. The only gene was found to have significant interaction effect with this gene is gene PTGER3 (p-value=0.0434). No main effect was found for gene PTGER3.

Table 3: List of gene pairs with p-value less than 0.001 in the overall genetic effect test. Genes with significant interactions (p-value$_i$ <0.05) are indicated with bold font.

Gene 1	Gene 2	τ_1^2	τ_2^2	τ_3^2	σ^2	p-value	p-value$_i$
ANG	**EDN1**	0.0341	2.30E-07	0.0025	0.3199	0.000656	8.07E-06
PDGFC	COL5A2	0.0095	2.28E-06	0.0052	0.3232	0.000494	0.2410
	F3	0.0056	7.72E-09	0.0061	0.3243	0.000781	0.0506
	GP1BA	0.0119	0.0364	1.25E-06	0.3229	0.000283	0.7462
	IGF1	0.0125	0.0091	9.61E-08	0.3234	0.000259	0.5133
	IL1B	0.0118	0.0050	8.63E-07	0.3227	0.000554	0.6228
	IL9	0.0113	1.37E-07	0.1049	0.3294	0.000066	0.3849
	LPA	0.0131	0.0077	2.20E-06	0.3226	0.000294	0.5231
	MMP7	0.0123	0.0052	1.08E-06	0.3236	0.000518	0.6548
	OXTR	0.0006	1.43E-06	0.0124	0.3218	0.000447	0.2106
	PLAUR	0.0129	0.0397	5.84E-07	0.3194	0.000536	0.5460
	PTGER3	0.0057	5.47E-07	0.0051	0.3241	0.000279	0.0434
	PTGS2	0.0128	0.0059	2.20E-06	0.3226	0.000514	0.7092
	TIMP2	0.0075	1.68E-08	0.0044	0.3238	0.000916	0.2351
	TLR4	0.0123	0.0116	1.79E-06	0.3239	0.000581	0.5606
PTGS2	ANG	0.0063	0.0183	1.11E-06	0.3218	0.000416	0.7016
	EDN1	0.0055	0.0031	1.70E-07	0.3243	0.000730	0.7966
	LPA	0.0056	0.0063	2.27E-06	0.3239	0.000988	0.5328
	PDGFB	0.0010	1.07E-08	0.0041	0.3246	0.000782	0.2736
	PGF	0.0028	4.45E-08	0.0035	0.3273	0.000850	0.0062
	PLAU	0.0004	7.97E-07	0.0057	0.3231	0.000260	0.0207
IL9	GP1BA	5.89E-07	0.0082	0.0274	0.3220	0.000936	0.4592
	IGF1	4.86E-08	0.0105	0.1597	0.3282	0.000540	0.0009

Among the five interacting gene pairs, the interaction between genes ANG (Angiogenin) and EDN1 (Endothelin 1) shows the most strongest interaction signal (p-value$_i$ $< 10^{-5}$). Study has shown that dysregulation of angiopoietins is associated with low birth weight (Silver et al. 2010). Nezar et al. (2009) studied the role of endothelin 1 in pre-eclampsia and non-pre-eclampsia women, and found that EDN1 correlates with the degree of fetal growth restriction. Although no study has reported the interaction between the two genes, our finding suggests a potential role of interaction between the two genes in affecting fetal growth. Further functional analysis is needed to validate this result.

Interactions were also found between gene PTGS2 (Prostaglandin-endoperoxide synthase 2) and genes PLAU (Urokinase-type plasminogen activator) and PGF (Placental growth factor), and between gene IL9 (Interleukin 9) and IGF1 (Insulin-like growth factor 1). It has been recognized that genes PGF and IGF1 are associated with fetal growth (Torry et al. 2003; Osorio et al. 1996). The identification of interactions between the two genes with other genes provides important biological hypothesis for further lab verification.

4.2 Analysis of Yeast eQTL Mapping Data

The second data set we analyzed with our model is a well studied yeast eQTL mapping data set generated to understand the genetic architecture of gene expression (Brem and Kruglyak 2005). The data were generated from 112 meiotic recombinant progeny of two yeast strains: BY4716 (BY: a laboratory strain) and RM11-1a (RM: a natural isolate). The data set contains 6229 gene expression traits and 2956 SNP marker genotype profiles. As an example to show the utility of our approach to an eQTL mapping study, we picked the expression profile of one gene (BAT2) as the quantitative response to identify potential genes or epistasis that regulate the expression of this gene. Noted that the parental strain RM11-1a is a LEU2 knockout strain. We expect strong segregation of this gene in the mapping population. Thus we picked this gene which is in the downstream of Leucine Biosynthesis Pathway (see Fig. 5(a) in Sun et al. 2008) as the response. A two-dimensional pairwise interaction search was done. Due to strong signals, Bonferroni correction was applied to adjust multiple testings for the 1072380 gene pairs. Overall test for pairs of gene effects was conducted followed by the score test for interaction if the overall test is significant.

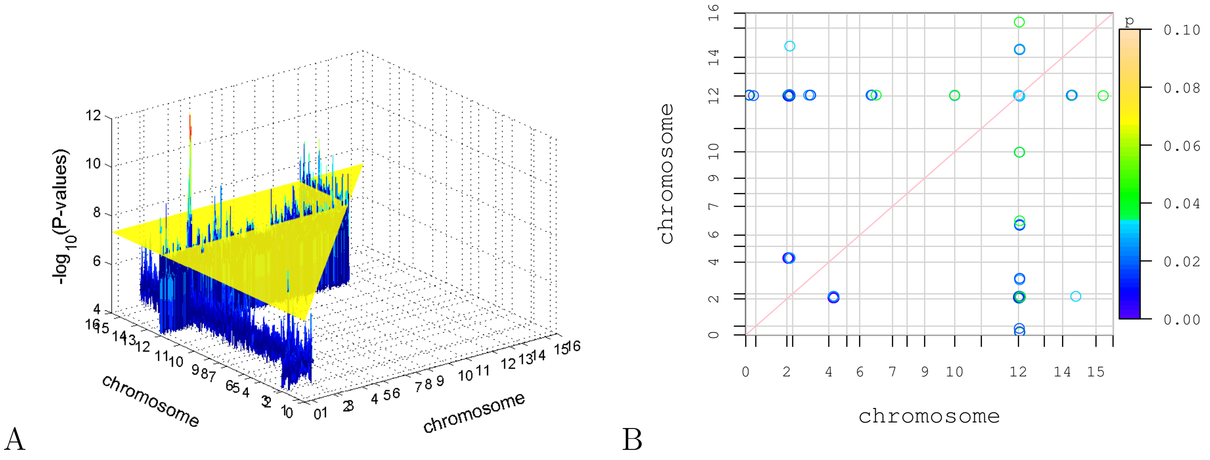

Figure 3: The -log10 transformed p-value profile plot of all gene pairs for the overall test (A) and the interaction test (B). The yellow hyperplane in A represents the Bonferroni cutoff.

There are total 1465 genes with some containing a single SNP marker. All the genes were subject to the proposed kernel interaction analysis. Figure 3A shows the pairwise interaction plot for -log10 transformed p-values associated with the overall genetic test (I). The yellow hyperplane indicates the Bonferroni correction threshold. Data points with p-values larger than 10^{-4} were masked. The plot indicates a strong genetic effect at chromosome 3 and 13, which implies that the two locations are potential regulation hotspots. In checking the recent literature, we found that the two positions were reported as eQTL hotspots in a number of studies (e.g., Brem et al., 2002; Perlstein et al. 2007; Li et al. 2010).

Out of the 1072380 gene pairs, 87 pairs were found to have significant interaction with each other at the 0.05 level. Figure 3B plots the pairwise significant interactions. Circles corresponds to significant interaction pairs with the darkness of the color indicating the strength of the interaction. We saw a strong interaction pattern on chromosome 13. One or several genes at this location interact with many other genes to affect the transcription of gene

BAT2. Another interaction "hotspot" is at chromosome 3 where genes (containing LUE2 and its neighborhood genes) interact with genes at chromosome 5, 13 and 15 to regulate BAT2 expression. We used Cytospace (Shannon et al. 2003) to generate an interaction network (see Fig. 4). Each node represents a gene and the thickness of the connection line indicates the strength of the interaction effect. Genes at the same chromosome location are clustered together in the plot. Light nodes with oval shapes indicates weak or no marginal effects. We found strong marginal effects for genes on chromosome 3 and 13. The most strongest interaction effect is between genes on chromosome 3 and chromosome 13. We also highlighted (red lines) the interaction between genes on chromosome 3 and others. Among the genes with no marginal effects (light oval nodes), URA3 is one of them and is a known transcription factor (Roy et al. 1990). Even though it does not show any main effect, it interacts with several genes on chromosome 3 to regulate the expression of BAT2. The results also imply the important role of several loci on chromosome 13. Since their functions are unknown, they can be potential candidate genes for further lab validation.

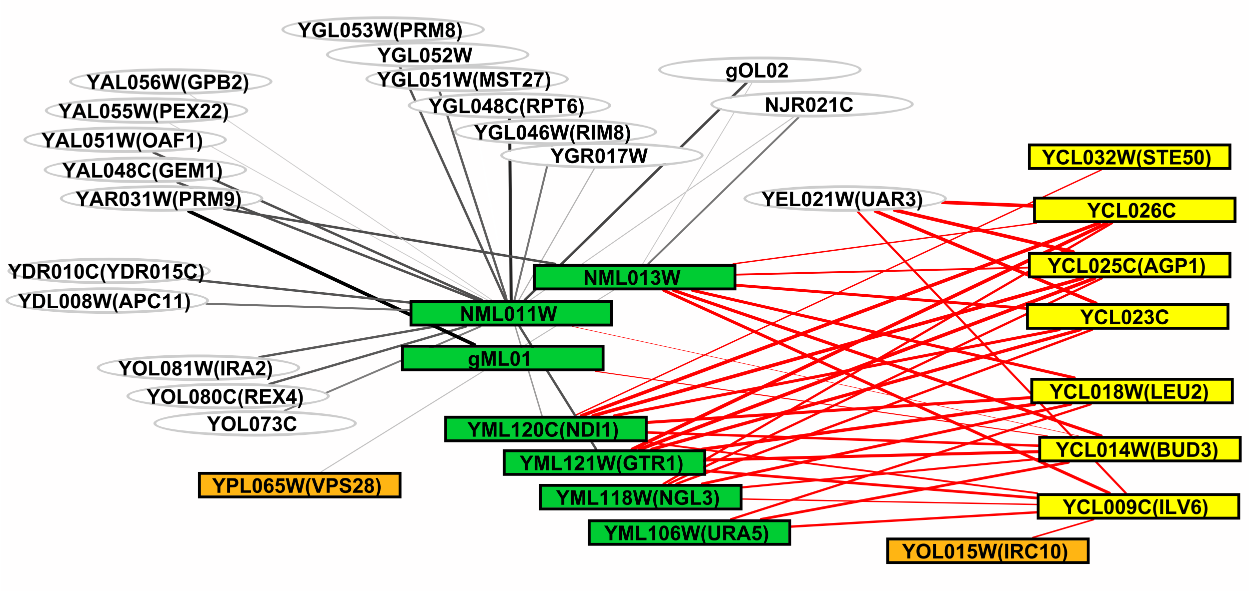

Figure 4: The network graph of interacting genes generated with Cytoscape (Shannon et al. 2003). The thickness of the connection line indicates the strength of the interaction. Nodes with light oval shapes indicate no marginal effect.

5 Discussion

The importance of gene-gene interaction in complex traits has stimulated enormous discussion and fundamental works in statistical methodology development have been broadly pursued (reviewed in Cordell 2009). Previous investigations have demonstrated the importance of a gene-centric approach in genetic association studies by simultaneously considering all markers in a gene to boost association power and reduce the number of tests (e.g., Cui et al. 2008; Buil et al., 2009). This motivates us to develop a gene-centric approach to understand gene-gene interaction associated with complex traits.

In this work, we have proposed a gene-centric kernel machine framework for gene-gene interaction analysis. Our model considers all variants in a gene as a system and adopts a kernel function to model the genomic similarity between SNP variants. The kernel machine method was previous developed for an association test and has been shown to be powerful in association studies (Kwee et al. 2008; Wu et al. 2010). Motivated by these work, we propose a spline-smoothing ANOVA decomposition method to decompose the genetic effects of two genes into separate main and interaction effects, and further model and test the genetic effects in the reproducing kernel Hilbert space. The joint variation of SNP variants within a gene is captured by a properly defined kernel function, which enables one to model the interaction of two genes in a linear reproducing Hilbert space by a cross-product of two kernel functions. Following rigorous derivations, the kernel machine method is shown to be equivalent to a linear mixed effects model. Thus, testing main and interaction effects can be done by testing the significance of different variance components. Extensive simulations under various settings and the analysis of two real data sets demonstrate the advantage of the gene-centric analysis.

He et al. (2009) previously proposed a gene-based interaction method in which each gene is summarized by several principle components and interaction was tested through the modeling of the PC terms rather than single SNPs. The authors proposed a weighted genotype scoring method using pairwise LD information to test gene-gene interaction. Their method is similar to several other methods which jointly consider information contributed by multiple markers (Chatterjee et al. 2007; Chapman and Clayton 2003). Our method is fundamentally different from their approach in which we capture the joint variation of SNP variants within and between genes by kernel functions (see Schaid 2010a for more discussion of the advantage of the kernel methods). Our method can also be extended to test interaction of variants by incorporating various weighting functions to define a kernel measure. Simulation studies demonstrate the advantage of the method over the PC-based regression analysis.

The advantage of the gene-centric gene-gene interaction analysis was previously discussed in He et al. (2009) such as reducing the number of hypothesis tests in a genome-wide scan. However, we should not over-emphasize the role of gene-centric analysis. Our simulation study indicates that when the underlying truth is that interaction only occurs between two single SNPs in two genes, single-SNP interaction analysis performs better. This result agrees with the conclusion made by He et al. (2009). Therefore, we recommend investigators conduct both types of analysis (single SNP and gene-centric) in real application, especially when no prior knowledge is available on how SNPs function within a gene as well as between genes. For a large-scale genome-wide or candidate gene study, one can also use the gene-centric approach as a screening tool, then further target which SNPs in different genes interact with each other.

The choice of kernel function may have potential effects on the testing power (Schaid 2010a, b). In this paper, we consider the allele matching (AM) kernel. Other kernel functions can also be applied such as the additive kernel, linear dosage kernel and product kernel and many others (Mukhopadhyay et al. 2010). Schaid (2010b) gave a very nice summary of various choices of kernel functions and their applications in genetic association studies. It is not the purpose of this paper to compare the performance of difference kernel choices on the power of an association test. A comparison study of different kernel functions on the power of the interaction test will be considered in future investigation.

The proposed method considers two genes as two units to test their interaction. It is easy to extend the idea to incorporate other genomic features such as pathways as testing units to assess pathway-pathway interaction under the proposed framework. The mapping results can then be visualized by some network graphical tools such as the Cytospace software (Shannon et al. 2003) which can help investigators generate important biological hypotheses for further lab validation. The computational code written in R (3G-SPAM) for implementing the work is available at http://www.stt.msu.edu/∼cui.

Acknowledgement

This work was supported in part by NSF grant DMS-0707031 and by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, DHHS. We thank Dr. R Romero for providing the baby birth weight data. The computation of the work is supported by Revolution R (http://www.revolutionanalytics.com/)

References

Aronszajn, N. (1950). Theory of reproducing kernels. Trans. Amer. Math. Soc. 68: 337-404.

Breiman, L. (2001). Random forests. Mach. Learn. 45: 5-32.

Brem, B.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. USA 102: 1572-1577.

Brem, R.B., Yvert, G., Clinton, R., Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752-755.

Buil1, A., Martinez-Perez, A., Perera-Lluna, A., Rib, L., Caminal, P. and Soria, J.M. (2009). A new gene-based association test for genome-wide association studies. BMC Proc. 3: S130.

Chapman, J. and Clayton, D. (2007). Detecting association using epistatic information. Genet. Epidemiol. 31: 894909.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am. J. Hum. Genet. 79: 10021016.

Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human disease. Nat. Rev. Genet. 10: 392-404.

Cui, Y.H., Kang, G.L, Sun, K.L, Qian, M.P., Romero, R. and Fu, W.J. (2008). Gene-centric genomewide association study via entropy. Genetics 179: 637-650.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11: 446-450.

Gu, C. and Wahba, G. (1993). Smoothing Spline ANOVA with Component-Wise Bayesian “Confidence Intervals”. J. Comput. Graph. Statist. 2: 97-117.

Jorgenson, E. and Witte, J.S. (2006). A gene-centric approach to genome-wide association studies. Nat. Rev. Genet. 7: 885-891.

He, J., Wang, K., Edmondson, A.C., Rader, D.J., Li, C., Li, M. (2010). Gene-based interaction analysis by incorporating external linkage disequilibrium information. Eur. J. Hum. Genet. Oct 6 Epub ahead of print, doi:10.1038/ejhg.2010.164.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Kang, G., Yue,, W., Zhang, J., Cui, Y.H., Zuo, Y. and Zhang, D. (2008). An entropy-based approach for testing genetic epistasis underlying complex diseases. J. Theor. Biol. 250: 362-374.

Kwee, L, Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. 82: 386-397.

Li, M., Romero, R. Fu, W.J. and Cui, Y.H. (2010). Mapping haplotype-haplotype interactions with adaptive LASSO. BMC Genet. 11: 79.

Li, S.Y., Lu, Q. and Cui, Y.H. (2010). A systems biology approach for identifying novel pathway regulators in eQTL mapping. J. Biopharm. Stat. 20: 373-400.

Li, S.Y., Lu, Q., Fu, W., Romero, R. and Cui, Y.H. (2009). A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. Stat. Appl. Genet. Mol. Biol. 8: Iss.1, Article 45.

Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. Biometrics 63: 1079-1088.

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, and Li MD (2007) A generalized combinatorial approach for detecting gene by gene and gene by environment interactions with application to nicotine dependence. Amer. J. Hum. Genet. 80: 1125-1137.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature 456: 18-21.

Moore, J.H. and Williams, S.M. (2009). Epistasis and its implications for personal genetics. Am. J. Hum. Genet. 85: 309-320.

Mukhopadhyay, I., Feingold, E., Weeks, D.E. and Thalamuthu, A. (2010). Association Tests

Using Kernel-Based Measures of Multi-Locus Genotypes Similarity Between Individuals. Genet. Epidemiol. 34: 213-221.

Neale, B.M. and Sham, P.C. (2004). The Future of Association Studies: Gene-Based Analysis and Replication. Am. J. Hum. Genet. 75: 353-362.

Nezar, M.A., el-Baky, A.M., Soliman, O.A., Abdel-Hady, H.A., Hammad, A.M., Al-Haggar, M.S. (2009). Endothelin-1 and leptin as markers of intrauterine growth restriction. Indian J. Pediatr. 76: 485-488.

Osorio, M., Torres, J., Moya, F., Pezzullo, J., Salafia, C., Baxter, R., Schwander, J., and Fant, M. (1996). Insulin-like growth factors (IGFs) and IGF binding proteins-1, -2, and -3 in newborn serum: relationships to fetoplacental growth at term. Early Hum. Dev. 46: 1526.

Perlstein, E.O., Ruderfer, D.M., Roberts, D.C., Schreiber, S.L., Kruglyak, L. (2007). Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nat. Genet. 39: 496-502.

Piegorsch, W.W., Weinberg, C.R. and Taylor, J.A. (1994). Non-hierarchical logistic models and case-only designs for accessing susceptibility in population-based case-control studies. Stat. Med. 13: 153-162.

Reigstad, L.J., Varhaug, J.E., Lillehaug, J.R. (2005). Structural and functional specificities of PDGF-C and PDGF-D, the novel members of the platelet-derived growth factor family. FEBS J. 272: 5723-5741.

Ritchie M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69: 138-147.

Roy, A., Exinger, F., Losson, R. (1990). cis- and trans-acting regulatory elements of the yeast URA3 promoter. Mol. Cell Biol. 10: 5257-5270.

Schaid, D.J. (2010a). Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum. Hered. 70: 109-131.

Schaid, D.J. (2010b). Genomic Similarity and Kernel Methods II: Methods for genomic information. Hum. Hered. 70: 132-140.

Schaid, D.J., McDonnell, S.K., Hebbring, S.J., Cunningham, J.M. and Thibodeau, S.N. (2005). Nonparametric tests of association of multiple genes with human disease. Am. J. Hum. Genet. 76: 780-793.

Self, S.G. and Liang, K.Y. (1987). Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. J. Am. Stat. Assoc. 82:605-611.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 3: 2498-2504.

Silver, K.L., Zhong, K., Leke, R.G., Taylor, D.W., Kain, K.C. (2010). Dysregulation of angiopoietins is associated with placental malaria and low birth weight. PLoS One. 5: e9481.

Sun, W., Yuan, S. and Li, K. (2008). Trait-trait dynamic interaction: 2D-trait eQTL mapping for genetic variation study. BMC Genomics 9: 242.

Thornton-Wells, T.A., Moore, J.H. and Haines, J.L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. 20: 640-7.

Torry, D.S., Mukherjea, D., Arroyo, J., Torry, R.J. (2003). Expression and function of placenta growth factor: implications for abnormal placentation. J. Soc. Gynecol. Investig. 10: 178-188.

Tzeng, J.Y., Devlin, B., Wasserman, L. and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am. J. Hum. Genet. 72: 891-902.

Wahba, G. (1990). Spline Models for Observational Data: CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, (Philadelphia: Society of Industrial and Applied Mathematics).

Wahba, G., Wang, Y.D. Gu, C., Klein, R. and Klein, B. (1995). Smoothing Spine Anova for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Ann. Statist. 23:1865-1895.

Wang, T., Ho, G., Ye, K. and Elston, R. (2010). A Partial Least Square Approach for Modeling Gene-gene and Gene-environment Interactions When Multiple Markers Are Genotyped. Genet. Epidemiol. 33: 6-15.

Weeks, D.E. and Lange, K. (1988). The affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. 42: 315-326.

Wessel, J. and Schork, N.J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. Am. J. Hum. Genet. 79: 792-806.

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. and Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. Am. J. Hum. Genet. 86: 929-942.

Zhang, Y. and Liu, J.S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nat. Genet. 39: 1167-1173.