The Annals of Applied Statistics
2012, Vol. 6, No. 3, 1134–1161
DOI: 10.1214/12-AOAS545

GENE-CENTRIC GENE–GENE INTERACTION: A MODEL-BASED KERNEL MACHINE METHOD[1]

BY SHAOYU LI AND YUEHUA CUI

Michigan State University and St. Jude Children's Research Hospital

Much of the natural variation for a complex trait can be explained by variation in DNA sequence levels. As part of sequence variation, gene–gene interaction has been ubiquitously observed in nature, where its role in shaping the development of an organism has been broadly recognized. The identification of interactions between genetic factors has been progressively pursued via statistical or machine learning approaches. A large body of currently adopted methods, either parametrically or nonparametrically, predominantly focus on pairwise single marker interaction analysis. As genes are the functional 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 framework for genome-wide gene–gene interaction detection. We treat each gene as a testing unit and derive a model-based kernel machine method for two-dimensional genome-wide scanning of gene–gene interactions. In addition to the biological advantage, our method is statistically appealing because it reduces the number of hypotheses tested 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 method provides a conceptual framework for the identification of gene–gene interactions which could shed novel light on the etiology of complex diseases.

1. Introduction. Accumulative evidence shows 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 effect sizes is one of the primary goals and challenges for modern genetics. The recent breakthroughs in high-throughput genotyping technologies and the completion of the International HapMap project provide unprecedented opportunities to characterize the genetic machinery of living organisms. Genetic association analyses focusing on single nucleotide polymorphisms (SNPs) or haplotypes have led to the identification of many novel genetic determinants of complex disease traits. However, despite enormous success

Received October 2011; revised January 2012.

[1]Supported in part by NSF Grants DMS-0707031, MCB-1121650 and by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, DHHS.

Key words and phrases. Allele matching kernel, association study, gene-clustered SNPs, genomic similarity, reproducing kernel Hilbert space, quantitative traits.

in genome-wide association studies, single SNP or haplotype based studies still suffer from low replication rates because of the infeasibility of dealing with the complex patterns of association, for example, genetic heterogeneity, gene–gene interaction and gene-environment interaction. Many of the genetic components of many diseases remain unaccounted for and only a small proportion of the heritability has been explained.

It is commonly recognized that genes do not function alone, rather they interact constantly with each other. Gene–gene interactions have been broadly considered as important contributors to the unexplained heritability of complex traits [Thornton-Wells, Moore and Haines (2004), Maher (2008), Moore and Williams (2009), Eichler et al. (2010)]. Current methods for gene–gene interactions are mostly focused on single locus interactions, using either parametric methods such as the regression-based tests of interaction [Piegorsch, Weinberg and Taylor (1994)] and Bayesian epistasis mapping [Zhang and Liu (2007)], nonparametric methods such as the entropy-based approaches [Kang et al. (2008)], or data mining methods such as the multifactor dimensionality reduction (MDR) [Ritchie et al. (2001)] and random forests [Breiman (2001)]. Methods based on interaction of haplotypes have also been developed [e.g., Li et al. (2010), Li, Zhang and Yi (2011)]. Due to the issue of haplotype phase-ambiguity, however, haplotype-based interaction analysis is limited to small sized haplotypes. Extension to large sized haplotype interaction is computationally challenging. For a comprehensive review of statistical methods developed for detecting gene–gene interactions, readers are referred to Cordell (2009).

A number of research reports have argued the relative merit of gene-based association analysis [e.g., Neale and Sham (2004), Jorgenson and Witte (2006), Cui et al. (2008), Ma et al. (2010)]. Neale and Sham (2004) 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. As genes are the functional units in a human genome, variants in genes should have high probability of being functionally more important than those that occur outside of a gene [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. Moreover, when multiple variants within a gene function in a complicated manner, the gene-based association test can gain additional power by capturing the joint function of multiple variants simultaneously compared to a single SNP analysis [Cui et al. (2008), Buil et al. (2009)]. In addition, a gene-based analysis is statistically appealing. By considering multiple SNP markers within a gene as testing units, one can reduce the number of tests, hence releasing the multiple testing burden and improving association test power.

The relative advantage of gene-centric analysis motivates us to consider genes as modeling units to identify interactions in a gene level. It is our expectation that the identification of genetic interactions in a gene level should carry the same benefits and gains as it does with gene-based association analysis. We therefore

propose to jointly model the genetic variation of SNPs within a gene, then further test the interaction in a gene level rather than in a single SNP level. This conceptual definition and modeling of gene-centric gene–gene (denoted as 3G) interaction would change the traditional paradigm of gene–gene interaction analysis and help us gain novel insight into the genetic etiology of complex diseases. In addition to its biological merits, 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, the 3G interaction analysis is also statistically appealing.

Following the definition of the gene-centric interaction, we propose a model-based kernel machine method to identify 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 et al. (2005), Wessel and Schork (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 methods in genetic association analysis has been reported in the literature [e.g., Schaid et al. (2005), Kwee et al. (2008), Wu et al. (2010)], but none of them considers gene–gene interactions. In this work, we propose a general 3G interaction framework by applying the smoothing-spline ANOVA model [Wahba (1990)] to model gene–gene interactions. The proposed method, termed Gene-centric Gene–Gene interaction with Smoothing-sPline ANOVA Model (3G-SPA), is implemented through a two-step procedure: (1) an exhaustive two-dimensional genome-wide search for any genetic effects; and (2) assessment of significance of interactions for the identified gene pairs.

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

2. Statistical methods.

2.1. *Smoothing spline-ANOVA model.* We assume n unrelated individuals sampled from a population, each of which possesses a measurement for a quantitative disease trait of interest. The quantitative measurements of n individuals are denoted as $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. Traditional approaches for detecting gene–gene interactions, such as MDR or regression type analysis, identify SNP-SNP interactions. In this work, we focus our attention to pairwise gene–gene interactions by

considering each gene as a unit. Consider two genes, denoted as G_1 and G_2, with L_1 and L_2 SNP markers, respectively. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,L})$ be an $1 \times L$ genotype vector of the gene pair for subject i, where $L = L_1 + L_2$ is the total number of SNP markers in the two genes. We model the relationship between the genotypes of the gene pair ($\mathbf{x}_i$) and the phenotype y_i by the following model:

$$y_i = m(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where m is an unknown function and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ is a random subject-specific error term and independent of $\mathbf{x}_i$. Here σ_i^2 $(=\sigma^2)$ is generally assumed to be homogeneous.

Gu (2002) has discussed the ANOVA decomposition of multivariate functions on generic domains of each single coordinate. Actually, the decomposition can also be defined on nested domains (see Appendix A). Following a similar idea, the genotype vector $\mathbf{x}_i$ is partitioned as $\mathbf{x}_i = [\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}]$, where $\mathbf{x}_i^{(j)}$ represents the L_j SNP predictors for gene j $(j = 1, 2)$. Let a product domain be $\mathcal{X} = \mathcal{X}^{(1)} \otimes \mathcal{X}^{(2)}$ with $\mathbf{x}^{(j)} \in \mathcal{X}^{(j)}$ and A_j be an averaging operator on $\mathcal{X}^{(j)}$, that averages out $\mathbf{x}_i^{(j)}$, $j = 1, 2$. Then a function $m(\cdot)$ defined on the product domain has a functional ANOVA decomposition as in the following:

$$\begin{aligned} m &= \prod_{j=1}^{2} (I - A_j + A_j) m \\ &= \{A_1 A_2 + (I - A_1) A_2 + A_1 (I - A_2) + (I - A_1)(I - A_2)\} m \\ &= \mu + m_1 + m_2 + m_{12}, \end{aligned} \tag{2.2}$$

where μ is the overall mean, m_1, m_2 are the main effects of the two genes and m_{12} describes the interaction effect between them (see Appendix A for more details).

2.2. *Reproducing kernel Hilbert space and the dual representation.* Based on the ANOVA decomposition, a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ can be constructed [Gu and Wahba (1993) and Wahba et al. (1995)]. Let $\mathcal{H}^{(j)}$ be an RKHS of functions on $\mathcal{X}^{(j)}$, $j = 1, 2$, and $\mathbf{1}^{(j)}$ be a space of constant functions on $\mathcal{X}^{(j)}$, then

$$\begin{aligned} \mathcal{H} &= \prod_{j=1}^{2} (\mathbf{1}^{(j)} \oplus \mathcal{H}^{(j)}) \\ &= [\mathbf{1}] \oplus [\mathcal{H}^{(1)} \otimes \mathbf{1}^{(2)}] \oplus [\mathbf{1}^{(1)} \otimes \mathcal{H}^{(2)}] \oplus (\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}) \\ &= [\mathbf{1}] \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^3, \end{aligned} \tag{2.3}$$

where $\oplus$ refers to direct sum and $\otimes$ refers to tensor product. Equation (2.3) provides an orthogonal decomposition of the entire functional space $\mathcal{H}$. So $\mathcal{H}$ is a

RKHS with the associated reproducing kernel as the sum of the reproducing kernels of these component subspaces. Each functional component in (2.2) lies in a subspace in (2.3), and is estimated in the corresponding RKHS. The identifiability of the components is assured by side conditions: $\int_{\mathcal{X}^{(j)}} m_j(\mathbf{x}^{(j)})\, d\mu_j = 0$, $j = 1, 2$.

We assume that function m is a member of the RKHS $\mathcal{H}$ and can be estimated as the minimizer of the following penalized sum of squares:

$$\mathcal{L}(\mathbf{y}, m) = \sum_{i=1}^{n} \big(y_i - m(\mathbf{x}_i)\big)^2 + \lambda J(m), \tag{2.4}$$

where $J(\cdot)$ is a roughness penalty. With the orthogonal decomposition of space $\mathcal{H}$, the penalty function $J(\cdot)$ can be decomposed such that equation (2.4) becomes

$$\mathcal{L}(\mathbf{y}, m) = \sum_{i=1}^{n} \big(y_i - m(\mathbf{x}_i)\big)^2 + \sum_{l=1}^{3} \lambda_l \| P^l m(\cdot) \|_{\mathcal{H}}^2, \tag{2.5}$$

where P^l is the orthogonal projector in $\mathcal{H}$ onto $\mathcal{H}^l$, and the λ_l's are the tuning parameters which balance the goodness of fit and complexity of the model. The minimizer of the objective function (2.5) is known to have a representation [Wahba (1990), Chapter 10] in terms of a constant and the associated reproducing kernels $\{k_l(s,t)\}$ of the $\mathcal{H}^l, l = 1, 2, 3$, that is,

$$\begin{aligned} m(\mathbf{x}) &= \mu + \sum_{i=1}^{n} c_i \sum_{l=1}^{3} \theta_l k_l(\mathbf{x}_i, \mathbf{x}) \\ &= \mu + \sum_{l=1}^{3} K_l^T(\mathbf{x}) C_l, \end{aligned} \tag{2.6}$$

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

Substituting the representation of $m(\cdot)$ into (2.5), we get

$$\begin{aligned} \mathcal{L}(\mathbf{y}, m) &= \sum_{i=1}^{n} \big(y_i - m(\mathbf{x}_i)\big)^2 + \sum_{l=1}^{3} \lambda_l \| P^l m(\cdot) \|_{\mathcal{H}}^2 \\ &= \big(\mathbf{y} - m(\mathbf{X})\big)^T \big(\mathbf{y} - m(\mathbf{X})\big) + \sum_{l=1}^{3} \lambda_l C_l^T \mathbf{K}_l C_l \\ &= \Big(\mathbf{y} - \mu\mathbf{1} - \sum_{l=1}^{3} \mathbf{K}_l C_l\Big)^T \Big(\mathbf{y} - \mu\mathbf{1} - \sum_{l=1}^{3} \mathbf{K}_l C_l\Big) + \sum_{l=1}^{3} \lambda_l C_l^T \mathbf{K}_l C_l, \end{aligned} \tag{2.7}$$

where $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ and

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

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

$$\frac{\partial \mathcal{L}}{\partial \mu} = 2\mathbf{1}^T \left(\mathbf{y} - \mu \mathbf{1} - \sum_{l=1}^{3} \mathbf{K}_l C_l \right)$$

and

$$\frac{\partial \mathcal{L}}{\partial C_l} = 2 \left\{ \mathbf{K}_l^T \left(\mathbf{y} - \mu \mathbf{1} - \sum_{l=1}^{3} \mathbf{K}_l C_l \right) + \lambda_l \mathbf{K}_l C_l \right\}.$$

Therefore, the first order condition is satisfied by the system

$$\begin{aligned} &\begin{bmatrix} n & \mathbf{1}^T\mathbf{K}_1 & \mathbf{1}^T\mathbf{K}_2 & \mathbf{1}^T\mathbf{K}_3 \\ \mathbf{K}_1^T\mathbf{1} & \mathbf{K}_1^T\mathbf{K}_1 + \lambda_1\mathbf{K}_1 & \mathbf{K}_1^T\mathbf{K}_2 & \mathbf{K}_1^T\mathbf{K}_3 \\ \mathbf{K}_2^T\mathbf{1} & \mathbf{K}_2^T\mathbf{K}_1 & \mathbf{K}_2^T\mathbf{K}_2 + \lambda_2\mathbf{K}_2 & \mathbf{K}_2^T\mathbf{K}_3 \\ \mathbf{K}_3^T\mathbf{1} & \mathbf{K}_3^T\mathbf{K}_1 & \mathbf{K}_3^T\mathbf{K}_2 & \mathbf{K}_3^T\mathbf{K}_3 + \lambda_3\mathbf{K}_3 \end{bmatrix} \begin{bmatrix} \mu \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} \\ &\quad = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{K}_1^T \\ \mathbf{K}_2^T \\ \mathbf{K}_3^T \end{bmatrix} \mathbf{y}. \end{aligned} \tag{2.8}$$

The connection between smoothing splines and the linear mixed effects model has been previously established [Wahba (1990), Speed (1991)]. For the two-way ANOVA decomposition model considered in this paper, we show that the first order system above is equivalent to Henderson's normal equation of the following linear mixed effects model (see Appendix B for details):

$$\mathbf{y} = \mu \mathbf{1} + m_1 + m_2 + m_{12} + \varepsilon, \tag{2.9}$$

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 $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$ is independent of m_1, m_2 and m_{12}. This connection indicates that the estimators of functions m_1, m_2, m_{12} are just the BLUPs of the linear mixed effects model [see also Liu, Lin and Ghosh (2007)]. Tuning parameters $\lambda_l, l = 1, 2, 3$, are functions of the variance components, which can be estimated either by the maximum likelihood method or by the restricted maximum likelihood (REML) method. Since the REML method produces unbiased estimators for the variance components, we adopt REML estimation in this work. The dual representation of the linear mixed

effects model obtained for the SS-ANOVA model makes it feasible to do inferences about the main and interaction components under the mixed effects model framework.

2.3. *Choice of kernel function for genotype similarity.* The choice of a reproducing kernel is not arbitrary in the sense that the kernel function must be nonnegative definite. By Theorem 2.3 [Gu (2002)], given a nonnegative definite function k on $\mathcal{X}$, we can construct a unique RKHS of real-valued functions on $\mathcal{X}$ with k as its reproducing kernel. In genetic association studies, 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 genomic similarity. This type of kernel measure has been used in linkage analyses [Weeks and Lange (1988)] and in association studies [Tzeng et al. (2003), Schaid et al. (2005), Wessel and Schork (2006), Kwee et al. (2008), Mukhopadhyay et al. (2010) and Wu et al. (2010)]. For a review of genomic similarity and kernel methods, readers are referred to Schaid (2010a, 2010b). With the notable strength that it does not require knowledge of the risk allele for each SNP, the 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

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

where I is the indicator function and "$\equiv$" means the two alleles are identical-by-state (IBS). The kernel function based on this AM similarity measure then takes the following form:

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} \mathrm{AM}(g_{i,s}, g_{j,s})}{4S}, \tag{2.10}$$

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

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

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} w_s \mathrm{AM}(g_{i,s}, g_{j,s})}{4\sum_{s=1}^{S} w_s}, \tag{2.11}$$

where w_s is the weighting function which can be adopted to incorporate prior knowledge in order to gain extra power. For example, when a study is trying to identify the effect 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)]. An example illustrating the calculation of the kernel matrix is given in Appendix C.

We use the AM kernel as the reproducing kernel for the two subspaces $\mathcal{H}^1$ and $\mathcal{H}^2$ corresponding to the main effects. Utilizing the fact that the reproducing kernel of a tensor product of two reproducing kernel Hilbert spaces is the product of the two reproducing kernels [Aronszajn (1950)], the associated reproducing kernel for $\mathcal{H}^3$ can be taken as the product of the reproducing kernels of the two subspaces: $\mathcal{H}^1$ and $\mathcal{H}^2$.

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 search for gene pairs with overall contribution to the phenotypic variation and then tests those contributing gene pairs for interaction effect. Under the SS-ANOVA framework, testing the overall genetic effect of a gene pair is equivalent to testing $H_0 : m_1 = m_2 = m_{12} = 0$. Similarly, testing for interaction effect can be formulated as $H_0 : m_{12} = 0$. With the linear mixed effects model representation, the aforementioned two tests are equivalent to $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ and $H_0^2 : \tau_3^2 = 0$, respectively. Here, $\tau_1^2, \tau_2^2, \tau_3^2$ are the variance components in model (2.9).

A well-known 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 $\mathbf{K}_\ell$'s are not block-diagonal. Thus, the asymptotic distribution of a likelihood ratio test (LRT) statistic does not follow a central chi-square distribution under the null hypothesis. The mixture chi-square distribution proposed by Self and Liang (1987) under irregular conditions also does not apply in our case. In this paper, we construct score test statistics based on the restricted likelihood. Consider the linear mixed model in (2.9), $\mathbf{y} \sim N(\mu\mathbf{1}, V(\beta))$. The restricted log-likelihood function can be written as

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

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

$$\frac{\partial \ell_R}{\partial \beta_i} = -\frac{1}{2}\operatorname{tr}(RV_i) + \frac{1}{2}(\mathbf{y} - \hat{\mu}\mathbf{1})^T V^{-1}(\beta) V_i V^{-1}(\beta)(\mathbf{y} - \hat{\mu}\mathbf{1}), \tag{2.12}$$

where $V_i = \frac{\partial V(\beta)}{\partial \beta_i}$, $i = 1, \ldots, 4$, so that $V_1 = I$, $V_2 = \mathbf{K}_1$, $V_3 = \mathbf{K}_2$, $V_4 = \mathbf{K}_3$ 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 given by

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

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 using the following score statistic:

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

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

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

Denoting the true value of σ^2 under the null by σ_0^2, $S(\sigma_0^2)$ is a quadratic form in $\mathbf{y}$. Following Liu, Lin and Ghosh (2007), we use the Satterthwaite method to approximate the distribution of $S(\sigma_0^2)$ by a scaled chi-square distribution, that is, $S(\sigma_0^2) \sim a\chi_g^2$, 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

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

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 replace it by its MLE under the null model, denoted by $\hat{\sigma}_0^2$. 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 this substitution, we estimate a and g by replacing ν by $\tilde{\nu}$ based on the efficient information. The elements of the Fisher information matrix of $\boldsymbol{\tau} = (\tau_1^2, \tau_2^2, \tau_3^2)$ are given by

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

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

and $I_{\sigma^2\sigma^2} = \frac{1}{2}\operatorname{tr}(P_0P_0)$. Then the efficient information $\tilde{I}_{\boldsymbol{\tau}\boldsymbol{\tau}} = I_{\boldsymbol{\tau}\boldsymbol{\tau}} - I_{\boldsymbol{\tau}\sigma^2}^T I_{\sigma^2\sigma^2}^{-1} I_{\boldsymbol{\tau}\sigma^2}$ and $\tilde{\nu} = \operatorname{Var}[S(\hat{\sigma}^2)] \approx \operatorname{SUM}[\tilde{I}_{\boldsymbol{\tau}\boldsymbol{\tau}}]$, where operator "SUM" indicates the sum of every element of the matrix.

2.4.2. *Testing for G×G interaction.* For testing the interaction effect, that is, testing $H_0^2 : \tau_3^2 = 0$, we also apply a score test. Let $\Sigma = \sigma^2 I + \tau_1^2 \mathbf{K}_1 + \tau_2^2 \mathbf{K}_2$. The score function (2.12) under this null hypothesis becomes

$$\begin{aligned}\left.\frac{\partial \ell_R}{\partial \tau_3^2}\right|_{\tau_3^2=0} &= -\frac{1}{2}[\operatorname{tr}(P_{01}\mathbf{K}_3) - (\mathbf{y} - \hat{\mu}\mathbf{1})^T \Sigma^{-1}\mathbf{K}_3\Sigma^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1})] \\ &= -\frac{1}{2}(\operatorname{tr}(P_{01}\mathbf{K}_3) - \mathbf{y}^T P_{01}\mathbf{K}_3 P_{01}\mathbf{y}),\end{aligned}$$

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

$$S_I = \tfrac{1}{2}\mathbf{y}^T P_{01}\mathbf{K}_3 P_{01}\mathbf{y}.$$

Similarly, the Satterthwaite method is 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 $\delta_I = \frac{1}{2}\operatorname{tr}(P_{01}\mathbf{K}_3)$ and $\nu_I = \frac{1}{2}\operatorname{tr}(P_{01}\mathbf{K}_3 P_{01}\mathbf{K}_3) - \frac{1}{2}\Phi^T\Delta^{-1}\Phi$, in which

$$\Phi = [\operatorname{tr}(P_{01}^2\mathbf{K}_3) \quad \operatorname{tr}(P_{01}\mathbf{K}_3 P_{01}\mathbf{K}_1) \quad \operatorname{tr}(P_{01}\mathbf{K}_3 P_{01}\mathbf{K}_2)]^T$$

and

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

3. Simulation study.

3.1. *Simulation design.* Monte Carlo simulations were conducted to evaluate the performance of the proposed method for detecting overall genetic effects as well as interaction between two genes. Genotype data were simulated using the MS program developed by Hudson (2002). The MS program 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. Parameters of the coalescent model were set as follows: (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 were $\rho = 4N_0 r = 4.0 \times 10^{-3}/bp$ and $\rho = 8 \times 10^{-3}/bp$ for the two samples. In each sample, 100 haplotypes were simulated for a locus with 10 kb long and the number of SNP sequences was 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. Finally, we had

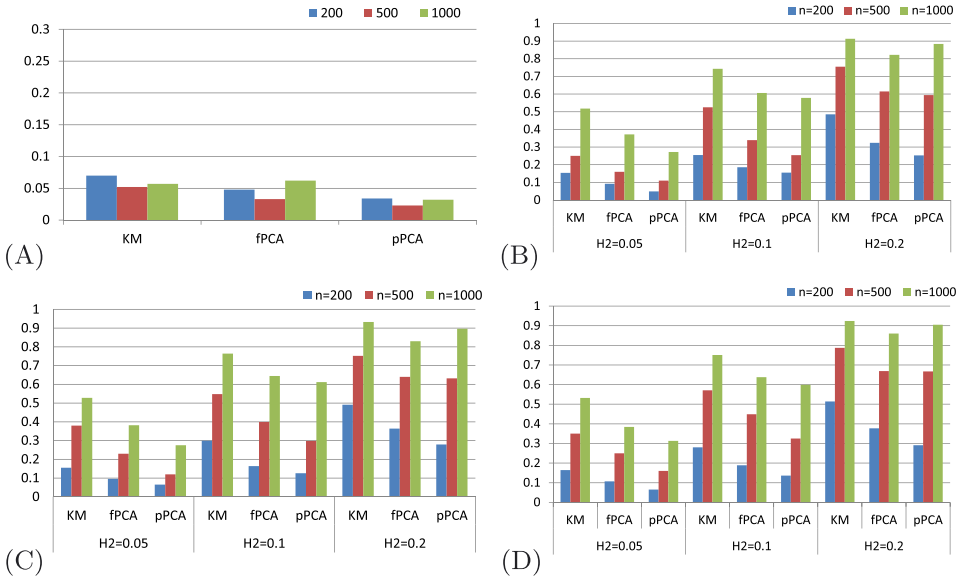

FIG. 1. *The empirical type I error* (A) *and power* (B)–(D) *of the three methods, where KM, fPCA and pPCA refer to the proposed kernel machine method, the full and partial PCA methods, respectively. Different heritability levels* (H2) *were assumed. The variance terms* $(\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$ *corresponding to different heritability levels* (H2 = 0.05, 0.1, 0.2) *were given as in*: (B) (0.8, 0.021, 0.021, 0), (0.8, 0.044, 0.044, 0), (0.8, 0.1, 0.1, 0); (C) (0.8, 0.016, 0.016, 0.088), (0.8, 0.036, 0.036, 0.018), (0.8, 0.08, 0.08, 0.04); *and* (D) (0.8, 0.011, 0.011, 0.022), (0.8, 0.022, 0.022, 0.044), (0.8, 0.05, 0.05, 0.1).

genotypes for n individuals for two separate genes with 10 SNPs each, and the two genes were independent.

When simulating phenotypes, four scenarios were considered (Figure 1). In scenario I [Figure 1(A)], the genetic effects were all set to zero so that we could assess the false positive control. In scenario II [Figure 1(B)], we considered the main effects for the two genes, but set the interaction effect as zero. In scenarios III [Figure 1(C)] and IV [Figure 1(D)], both main effects and interaction effect were considered. The difference between scenarios 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 the main effects. Quantitative traits 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 τ_1^2, τ_2^2, τ_3^2 took different values under different scenarios; $\mathbf{K}_\ell$, $\ell = 1, 2, 3$, are the kernel matrices using the allele matching method described before. Different sample sizes ($n = 200$, 500 and 1000) and different heritability ($H^2 = 0.05, 0.1, 0.2$) were assumed. Let $\sigma_G^2 = \tau_1^2 + \tau_2^2 + \tau_3^2$. The heritability was defined as $H^2 = \sigma_G^2/(\sigma_G^2 + \sigma^2)$. In all simulation scenarios, we fixed the residual variance $\sigma^2 = 0.8$, and τ_1^2 and τ_2^2 were set to be equal.

3.2. *Model comparison*. We compared our simulation results with two other methods described in the following. Wang et al. (2009) proposed an interaction method using a partial least squares approach which was 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 $G_j = (S_1, S_2, \ldots, S_{L_j})$ 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 $U_j^1, U_j^2, \ldots, U_j^{m_j}$ ($m_j \le L_j$ is the rank for G_j). An interaction model can be expressed as

$$\mathbf{y} = \mu\mathbf{1} + \sum_{l_1=1}^{L_1} \beta_{l_1} S_{l_1} + \sum_{l_2=1}^{L_2} \beta_{l_2} S_{l_2} + \gamma U_1^1 U_1^2, \tag{3.1}$$

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 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. (2009). We followed Wang et al. (2009) and compared the performance of our model with this one.

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 (3.1). Thus, if we replace the main effect of each gene in model (3.1) with PCs rather than single SNPs to reduce the model degrees of freedom, model (3.1) then becomes

$$\mathbf{y} = \mu\mathbf{1} + \sum_{p_1=1}^{P_1} \beta_{p_1} U_{p_1}^1 + \sum_{p_2=1}^{P_2} \beta_{p_2} U_{p_2}^2 + \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \gamma_{p_1 p_2} U_{p_1}^1 U_{p_2}^2, \tag{3.2}$$

where $U_{p_j}^j$, $j = 1, 2$, represents the PCs for gene j, and P_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 interactions of the selected PCs. G×G interaction was assessed by testing $H_0 : \gamma_{p_1 p_2} = 0$, for all p_1 and p_2. This model was applied by He et al. (2010) in their gene-based interaction analysis.

In addition to the two models above, we also compared our gene-centric approach to a simple pairwise SNP interaction model. Details of the comparison are given in Section 3.3. 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*.

3.3.1. *Comparison of the three methods for the overall genetic test*. We first evaluated the type I error rate and the power of the three methods for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$). Figure 1 summarizes the comparison results between our kernel machine (KM) method and the partial PCA (pPCA) [model (3.1)] and the full PCA (fPCA) [model (3.2)] methods. In Figure 1(A), we can see that our method has the empirical type I error rate reasonably controlled for the overall genetic effect test. The partial PCA-based interaction model generates very conservative results.

In all simulations we fixed the residual variance σ^2 to 0.8, and changed the three genetic effects to get different heritability levels. For scenario II [Figure 1(B)], data were simulated assuming no interaction (i.e., $\tau_3^2 = 0$). Then τ_1^2 and τ_2^2 were calculated for a given heritability level. For example, when $H^2 = 0.05$, $\tau_1^2 = \tau_2^2 = 0.021$. In scenarios III and IV, the interaction effect was set as either half of the main effects (scenario III) or twice the main effects (scenario IV). As we expected, the testing power increases as the heritability level and sample size increased [Figure 1(B)–(D)]. For a fixed heritability level, the KM method always outperformed the other two under different sample sizes. The power difference of the KM method over the other two is more striking under a small heritability level (say, 0.05) or when the sample size is small. For most complex diseases in humans, the heritability of a genetic variant is generally small. Thus, the KM method is preferable over the other two in the first stage of the interaction analysis. In addition, we also observed that the KM method is insensitive to whether the genetic effect is due to the main effect or the interaction, whereas the PCA-based method gains power as more of the genetic effect is due to the interaction.

3.3.2. *Comparison of the three methods for the interaction test*. Interaction may be due to a variety of underlying mechanisms. Some genes might have both significant main effects and interaction effect, while others might only incur interaction effect without main effects. Simulation studies were designed to compare the performance of the proposed KM method in discovering gene × gene interactions over the other two methods, considering different interaction effect sizes. Since the power of an interaction test is largely determined by the size of the interaction effect, we simulated data assuming different proportions of interaction effects among the total genetic variance. This proportion is defined by $\eta = \tau_3^2/\tau^2$, where τ^2 $(=\tau_1^2 + \tau_2^2 + \tau_3^2)$ refers to the total genetic variance. For a fixed total genetic variance, the value of η indicates the strength of the interaction effect between two genes. For a fixed residual variance ($\sigma^2 = 0.8$), the total genetic variance is set to 0.2 when $H^2 = 0.2$ and to 0.53 when $H^2 = 0.4$. The variance size for the two main effects were set to be equal, so we could calculate the interaction variance. For example, $(\tau_1^2, \tau_2^2, \tau_3^2) = (0.08, 0.08, 0.04)$ when $\eta = 0.2$, and

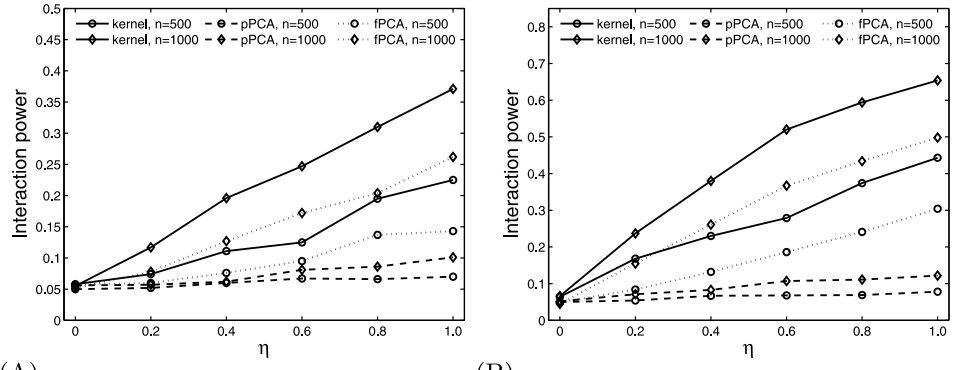

FIG. 2. *Power comparison of the proposed KM model* (*solid line*), *pPCA model* (3.1) (*dashed line*) *and fPCA model* (3.2) (*dotted line*) *under different sample sizes* (n) *and different interaction sizes* (η), *where* η *refers to the proportion of interaction effect among the total genetic effect*. (A) $H^2 = 0.2$ *and* (B) $H^2 = 0.4$.

$(\tau_1^2, \tau_2^2, \tau_3^2) = (0.02, 0.02, 0.16)$ when $\eta = 0.8$ under $H^2 = 0.2$. Six values of the proportion $\eta = (0, 0.2, 0.4, 0.6, 0.8, 1.0)$ were considered, including the two extreme cases: no interaction at all ($\eta = 0$) and pure interaction ($\eta = 1$). The method was compared with the other two PCA-based interaction analyses under two different sample sizes, 500 and 1000.

Figure 2 shows power comparison (based on 1000 replicates) under two different heritability levels [Figure 2(A) for $H^2 = 0.2$ and Figure 2(B) for $H^2 = 0.4$]. The type I error (when $\eta = 0$) for the interaction test is reasonably controlled for the three methods. As we expected, the interaction power increases as the interaction effect size (η) increases. Among the three methods, our KM method has the highest power. The partial PCA model (3.1) has the lowest power. This demonstrates that only considering interaction of the first principle component in each gene is not enough to capture the interaction between two genes. The effect of sample size on the interaction power is also significant. Large sample size always leads to large power.

A common concern in detecting gene–gene interactions is the computational burden. As we focus on genes as testing units, the total number of tests reduces dramatically in a gene-centric analysis compared to a single-marker based analysis. In addition, computation time can be saved by implementing the two-stage analysis: (1) doing score tests in the first stage to assess the significance of all genetic components; and (2) testing interaction only for those pairs showing statistical significance after multiple testing adjustment in the first stage. The computation for the first step is very fast since only parameters under the null model, which is a regular linear regression model, need to be estimated. In general, only a small fraction of signals can pass the significant threshold in the first stage, leaving the second stage interaction test with less computational burden. For a quick

comparison of the KM method with the PC-based approaches, when $n = 500$, the computation time with 1000 simulation replicates for KM, pPCA and fPCA are 28 min, 24.5 min and 24 min, respectively, for the overall test. When testing the interaction term, the KM method takes about 9sec for a single run on average, and is slower than the other two methods. However, this should not be a concern in real applications since the number of interactions is generally not very large after the first stage screening.

In summary, our KM method outperforms the other two methods in the overall genetic test as well as in the interaction test under different simulation scenarios. The results also indicate that large sample sizes are needed for the detection of the interaction term compared to the detection of main effects. Even though the PCA-based analysis has been applied for candidate gene-based association analyses [Wang and Abbott (2008), Wang et al. (2009)], our simulation studies show that it may not be suitable for interaction analysis.

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

$$y_i = \beta_0 + \beta_1 S_{1i} + \beta_2 S_{2i} + \beta_{12} S_{1i} S_{2i} + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{3.3}$$

where β_0 is the intercept; β_1, β_2 and β_{12} represent the effects of SNP S_1 in gene 1, SNP S_2 in gene 2 and the interaction effect between the two, respectively; $\varepsilon_i \sim N(0, \sigma^2)$. We simulated data according to model (3.3) assuming a MAF $p_A = 0.3$. Different heritabilities and different sample sizes were assumed. 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 (3.3) are subject to both the single SNP interaction and the proposed kernel interaction analysis. With this simulation we tried to assess how robust the KM method is when there is only one pair of functional SNPs in two genes.

Table 1 shows that both models show comparable type I error control for the overall genetic test (see P_o in the table). For the interaction test, it looks like the kernel approach generates more conservative results. Here 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 actually produces more inflated false positives compared to our kernel approach when no genetic effect is involved. When data were simulated assuming only main effects but no interaction (case $\beta_{12} = 0$), the two approaches yield very similar false positive rates, indicating reasonable performance of the kernel approach for false positive control.

For the power analysis, we found little difference between the two methods for the overall genetic test (P_o), especially under large sample size and high heritability level. For the interaction test (P_i), the power increases as sample size and heritability level increase. We observed relatively large power differences between the models when the sample size is small (<500). As sample size increases, the

TABLE 1
List of empirical type I error and power based on 1000 *simulation runs when data were simulated with model* (3.3)

H^2	Coefficients $(\beta_0, \beta_1, \beta_2, \beta_{12})$	Sample size (n)	Single SNP interaction		Kernel interaction	
			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.

difference diminishes. For example, the power difference is 0.2 under $n = 200$ and $H^2 = 0.4$, which is reduced to 0.032 when n increases to 500 and to zero when $n = 1000$. When more than one functional SNP within each gene is involved in interacting with one another to affect a trait variation, 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 sizes and high heritability (Table 1).

4. Applications to real data.

4.1. *Application* I: *Analysis of birth weight data*. A candidate gene study was initially conducted in order to study genetic effects associated with being 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. Individuals were genotyped at 797 SNP markers in 186 unique candidate genes. Missing genotypes were imputed using a conditional probability approach as described in Cui et al. (2008). 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. The case/control classification in the two data sets was based on baby's birth weight together with mother's gestational age. A total number of 1511 individuals in the case and control combined data set were used for analysis after removing 14 individuals with birth weight $3 \times$ IQR (inter-quartile range) above Q_3 or below Q_1. After removing these extreme observations and a Box–Cox transformation, the distribution of the birth weight data in the combined sample was approximately normal.

A two-dimensional pairwise G $\times$ G interaction search was conducted (total 17,205 gene pairs). The score test for testing $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ was determined and p-values were obtained for all gene pairs. For a two-dimensional search, it is not clear how to set up a genome-wide threshold to correct for multiple testings. Obviously, the 17,205 tests are not independent and it is too stringent to use the Bonferroni correction method. Thus, we used an arbitrary threshold of 0.001 as a cutoff. Totally, 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 2. 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 (indicated by bold fonts in Table 2). If we increase the overall testing p-value threshold (to 0.0005), only 9 gene pairs will remain in the list and only 2 gene pairs will show significance.

The results indicate a strong 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, Varhaug and Lillehaug (2005)]. Even though its main effect is very strong, the only gene found to have significant interaction effect with this gene was gene PTGER3 (p-value$_i = 0.0434$).

Among the five interacting gene pairs, the interaction between genes ANG (Angiogenin) and EDN1 (Endothelin 1) shows the 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 preeclampsia and nonpreeclampsia women, and found that EDN1

TABLE 2

List of gene pairs with p-value less than 0.001 *in the overall genetic effect test. Gene pairs with significant interaction effect* (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**	0.0340	9.1E–08	0.0024	0.3199	0.000656	8.07E–06
PDGFC	COL5A2	0.0095	2.6E–06	0.0051	0.3232	0.000494	0.2410
	F3	0.0056	1.5E–06	0.0061	0.3242	0.000781	0.0506
	GP1BA	0.0118	0.0362	$<10^{-6}$	0.3229	0.000283	0.7462
	IGF1	0.0124	0.0090	$<10^{-6}$	0.3234	0.000259	0.5133
	IL1B	0.0118	0.0049	$<10^{-6}$	0.3227	0.000554	0.6228
	IL9	1.0E–07	0.0151	0.0129	0.3210	0.000066	0.3849
	LPA	0.0130	0.0076	$<10^{-6}$	0.3226	0.000294	0.5231
	MMP7	0.0122	0.0051	$<10^{-6}$	0.3236	0.000518	0.6548
	OXTR	0.0007	1.1E–06	0.0124	0.3218	0.000447	0.2107
	PLAUR	0.0128	0.0396	$<10^{-6}$	0.3194	0.000536	0.5460
	PTGER3	0.0057	1.7E–07	0.0051	0.3240	0.000279	0.0434
	PTGS2	0.0127	0.0058	$<10^{-6}$	0.3226	0.000514	0.7092
	TIMP2	0.0075	1.1E–07	0.0043	0.3238	0.000916	0.2351
	TLR4	0.0122	0.0115	$<10^{-6}$	0.3239	0.000581	0.5606
PTGS2	ANG	0.0062	0.0182	$<10^{-6}$	0.3218	0.000416	0.7016
	EDN1	0.0054	0.0030	$<10^{-6}$	0.3243	0.000730	0.7966
	LPA	0.0055	0.0062	$<10^{-6}$	0.3239	0.000988	0.5328
	PDGFB	0.0010	1.5E–06	0.0042	0.3246	0.000782	0.2739
	PGF	0.0261	7.4E–08	0.0044	0.3240	0.000850	0.0062
	PLAU	0.0004	3.0E–06	0.0057	0.3231	0.000260	0.0207
IL9	GP1BA	1.1E–06	0.0082	0.0274	0.3220	0.000936	0.4592
	IGF1	0.0174	2.8E–08	0.0075	0.3227	0.000540	0.0009

σ^2 is the residual variance; p-value is for testing $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ and p-value$_i$ is for testing $H_0 : \tau_3^2 = 0$. Note that the estimated interaction effect size does not necessarily indicate how strong an interaction is. The strength of an interaction is rather reflected by the interaction testing p-value (denoted by p-value$_i$).

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 of the two genes with other genes provides important biological hypotheses for further lab verification.

4.2. *Application* II: *Analysis of yeast eQTL 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 progenies 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 gene–gene interactions that regulate the expression of this gene. Note that one of the parental strains, RM11-1a, is a LEU2 gene knockout strain. Thus, we expect strong segregation of this gene in the mapping population. Since BAT2 is in the downstream of the Leucine Biosynthesis Pathway (LEU2 is in the upstream of the same pathway), this motivates us to pick BAT2 as the response [see Figure 5(a) in Sun, Yuan and Li (2008)]. The Bonferroni correction was applied to adjust multiple testings for the 1,072,380 gene pairs. An overall test for pairs of gene effects was conducted followed by the score test for interaction if the overall test is significant.

There were 1465 genes with some containing a single SNP marker. All the genes were subject to the proposed kernel interaction analysis. Figure 3(A) shows the pairwise interaction plot for $-\log 10$ transformed p-values associated with the overall genetic test. 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 chromosomes 3 and 13, which implies that the two locations are potential regulation hotspots. In checking the recent literature,

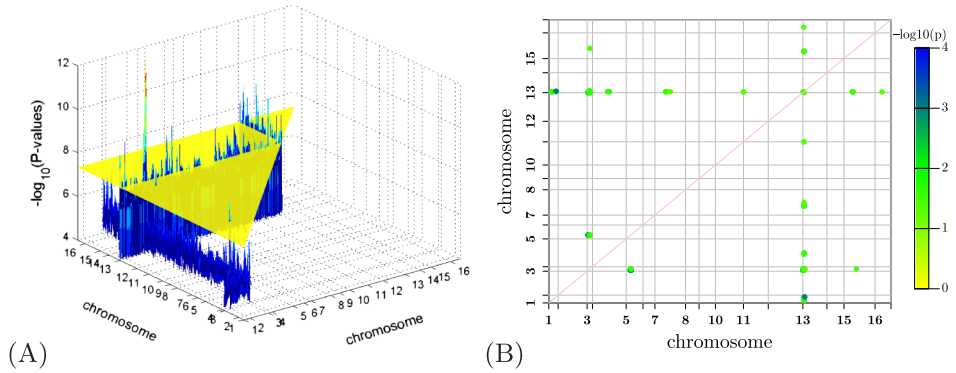

FIG. 3. *Analysis of the expression profile for BAT*2. *The* $-\log 10$ *transformed p-value profile plot of all gene pairs for the overall test* (A) *and the interaction test* (B). *The yellow hyperplane in* (A) *represents the Bonferroni cutoff.*

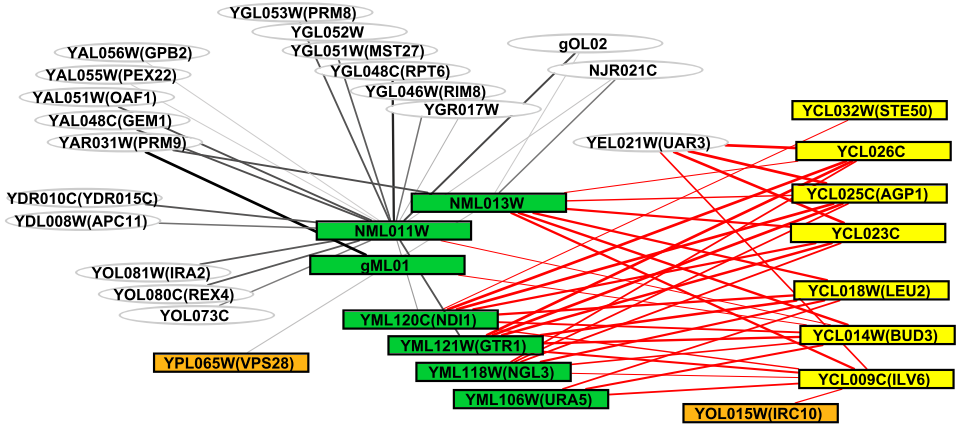

FIG. 4. *The network graph of interacting genes for the eQTL analysis. Each node represents a gene*; *nodes with light oval shape indicate no main effect. Interacting genes are connected with lines* (*the thicker the lines*, *the stronger the interaction*). *Genes with different colors are located in different chromosomes with significant main effects.*

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, Lu and Cui (2010)].

Out of the 1072380 gene pairs, 87 pairs were found to have significant interactions at an experiment-wise significant level of 0.05. Figure 3(B) plots the pairwise significant interactions. Circles correspond 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 chromosomes 5, 13 and 15 to regulate BAT2 expression. We used Cytospace [Shannon et al. (2003)] to generate an interaction network map (see Figure 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 indicate weak or no main effect. We found strong main effects for genes on chromosomes 3 and 13. The 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 main effects (light oval nodes), URA3 is known as a transcription factor [Roy, Exinger and Losson (1990)]. Even though it does not show any main effect, it shows interactions with several other genes on chromosome 3 to regulate the expression of BAT2. The results also imply the important role of several loci on chromosome 13. As their functions are unknown, they could be potential candidate genes for further lab validation.

5. Discussion. The importance of gene–gene interaction in complex traits has stimulated enormous discussion. Fundamental works in statistical methodology development have been broadly pursued in this area [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 [e.g., Cui et al. (2008), Buil et al. (2009), Li et al. (2009), Ma et al. (2010)]. This motivates us to develop a gene-centric approach to understand gene–gene interactions associated with complex diseases. In this paper, we propose a kernel machine method for a gene-centric gene–gene interaction analysis. We adopt a smoothing spline-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. The kernel machine method is shown mathematically 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 and analysis of two real data sets demonstrate the utility of the gene-centric interaction analysis.

He et al. (2010) 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 [e.g., Chatterjee et al. (2006), Chapman and Clayton (2007)]. 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 [Wu et al. (2011)], and will be considered in our follow-up investigation. It is worth mentioning some other methods for a gene-based analysis such as the CTGDR method proposed by Ma et al. (2010) which accounts for the gene and SNP-within-gene hierarchical structure. The method can identify predictive genes as well as predictive SNPs within identified genes. Other methods such as the grouped lasso method can also be applied for a gene-based analysis [Ma, Song and Huang (2007)]. However, their extension to a gene interaction analysis is not straightforward and warrants further investigation.

The advantage of the gene-centric gene–gene interaction analysis was previously discussed in He et al. (2010), including reducing the number of hypothesis tests in a genome-wide scan compared to a single SNP interaction analysis. However, we should not over-emphasize the role of gene-centric analysis. Our simulation study shows that when the underlying truth is a single SNP pair interaction

in two genes, single-SNP interaction analysis performs better. This result agrees with the conclusion made by He et al. (2010). Therefore, we recommend that investigators conduct both types of analyses (single SNP and gene-centric) in real applications, 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, 2010b)]. In this paper, we consider the AM kernel. Other kernel functions can also be applied as discussed in detail in Schaid (2010b). 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 investigations.

The proposed method considers genes as testing 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 is available as an R-package from http://cran.r-project.org/package=SPA3G.

APPENDIX A: ANOVA DECOMPOSITION

Suppose a function f is on domain $\Gamma = \Gamma_1 \otimes \Gamma_2 \otimes \Gamma_3$. Define corresponding averaging operator A_γ on each generic domain $\Gamma_\gamma, \gamma = 1, 2, 3$. An ANOVA decomposition of a function f can be obtained:

$$\begin{aligned} f &= \left\{ \prod_{\gamma=1}^{3} (I - A_\gamma + A_\gamma) \right\} f \\ &= \{(I - A_1)(I - A_2)(I - A_3) + (I - A_1)(I - A_2)A_3 \\ &\quad + (I - A_1)A_2(I - A_3) + A_1(I - A_2)(I - A_3) \\ &\quad + (I - A_1)A_2A_3 + A_1A_2(I - A_3) + A_1(I - A_2)A_3 + A_1A_2A_3\} f. \end{aligned}$$

For a nested domain $(\Gamma_1 \otimes \Gamma_2) \otimes \Gamma_3$, let A_{12} be the averaging operator on domain $(\Gamma_1 \otimes \Gamma_2)$. Then the ANOVA decomposition becomes

$$f = \{(I - A_{12})(I - A_3) + A_{12}(I - A_3) + (I - A_{12})A_3 + A_{12}A_3\} f.$$

Since

$$(I - A_1)(I - A_2) + (I - A_1)A_2 + A_1(I - A_2) = I - A_1A_2$$

by letting $A_{12} = A_1 A_2$,

$$\{(I - A_{12})(I - A_3) + A_{12}(I - A_3) + (I - A_{12})A_3 + A_{12}A_3\} f$$
$$= \left\{ \prod_{\gamma=1}^{3} (I - A_\gamma + A_\gamma) \right\} f.$$

Recursively, it shows that the ANOVA decomposition can also be conducted on products of nested domains.

APPENDIX B: THE DUAL REPRESENTATION

Consider the linear mixed effect model:

$$y = \mu \mathbf{1} + m_1 + m_2 + m_{12} + \varepsilon,$$

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 $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$ is independent of m_1, m_2 and m_{12}. Henderson's normal equation for obtaining the BLUPs of the random effects is

$$\text{(B.1)} \quad \begin{bmatrix} n & \mathbf{1}^T & \mathbf{1}^T & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \dfrac{\sigma^2}{\tau_1^2}\mathbf{K}_1^{-1} & \mathbf{I} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} + \dfrac{\sigma^2}{\tau_2^2}\mathbf{K}_2^{-1} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} & \mathbf{I} + \dfrac{\sigma^2}{\tau_3^2}\mathbf{K}_3^{-1} \end{bmatrix} \begin{bmatrix} \mu \\ m_1 \\ m_2 \\ m_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{I} \\ \mathbf{I} \\ \mathbf{I} \end{bmatrix} \mathbf{y}.$$

It can be shown that this normal equation is equivalent to the first order condition for estimating function m, equation (2.8). Multiplying both sides of (2.8) by the matrix,

$$\begin{bmatrix} 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_1^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{K}_3^{-1} \end{bmatrix},$$

one can get

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

TABLE 3
Calculation of the allele matching score

	$g_{j,s}$		
$g_{i,s}$	**AA(2)**	**Aa(1)**	**aa(0)**
AA(2)	4	2	0
Aa(1)	2	2	2
aa(0)	0	2	4

Letting $m_\ell = \mathbf{K}_\ell C_\ell$ $(\ell = 1, 2)$, $m_{12} = \mathbf{K}_3 C_3$ and $\tau_\ell^2 = \sigma^2/\lambda_\ell$ $(\ell = 1, 2, 3)$, the system is exactly equation (B.1), which is Henderson's normal equation of linear mixed effects model (2.9).

APPENDIX C: CALCULATION OF THE KERNEL MATRIX: AN EXAMPLE

As an example, we consider three unrelated individuals. Suppose there are 10 SNP markers in one gene. The genotypes for these markers are numerically coded as 0, 1 or 2 depending on the copy number of a certain allele (e.g., A). So genotypes aa, Aa and AA are coded as 0, 1 and 2, respectively. See Table 4 for the numerical genotype coding of the three individuals at the ten marker position.

At a certain SNP position, the allele matching score of the genotypes for two individuals i and j (denoted as AM_{ij}), is calculated as the total number of alleles which are identical by state (see Table 3). Then the genomic similarity score across the whole gene between two individuals i and j (denoted as GSS_{ij}) is then calculated as the weighted sum of allele matching scores for all SNP markers in that gene (see Table 4).

TABLE 4
Illustration on the calculation of genomic similarity score

	$\mathbf{SNP_1}$	$\mathbf{SNP_2}$	$\mathbf{SNP_3}$	$\mathbf{SNP_4}$	$\mathbf{SNP_5}$	$\mathbf{SNP_6}$	$\mathbf{SNP_7}$	$\mathbf{SNP_8}$	$\mathbf{SNP_9}$	$\mathbf{SNP_{10}}$
Individual 1	2	0	2	1	1	0	1	1	1	1
Individual 2	0	0	0	0	0	0	0	1	0	0
Individual 3	0	0	0	1	1	0	1	0	1	1
AM_{12}	0	4	0	2	2	4	2	2	2	2
GSS_{12}	$(0+4+0+2+2+4+2+2+2+2)/(4\times 10) = 0.5$									
AM_{13}	0	4	0	2	2	4	2	2	2	2
GSS_{13}	$(0+4+0+2+2+4+2+2+2+2)/(4\times 10) = 0.5$									
AM_{23}	4	4	4	2	2	4	2	2	2	2
GSS_{23}	$(4+4+4+2+2+4+2+2+2+2)/(4\times 10) = 0.7$									

With the genomic similarity measurements between all possible individual pairs [in the example, $(1,2)$, $(1,3)$ and $(2,3)$], the 3×3 kernel matrix based on the example data is given as

$$\mathbf{K} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.7 \\ 0.5 & 0.7 & 1 \end{bmatrix}.$$

Acknowledgments. We thank Editor K. Lange, the Associate Editor and two anonymous referees for their insightful comments that greatly improved the manuscript. We also thank Dr. R. Romero for providing the birth weight data, Professor J. Stapleton for his careful reading of the manuscript, and Mr. B. Rosa and Dr. J. Chen for helpful discussion on using the Cytoscape program. 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. and KRUGLYAK, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296** 752–755.

BUIL, A., MARTINEZ-PEREZ, A., PERERA-LLUNA, A. et al. (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** 894–909.

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** 1002–1016.

CORDELL, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10** 392–404.

CUI, Y., KANG, G., SUN, K., QIAN, M., ROMERO, R. and FU, W. (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. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. MR1876599

GU, C. and WAHBA, G. (1993). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Comput. Graph. Statist.* **2** 97–117. MR1272389

HE, J., WANG, K., EDMONDSON, A. C. et al. (2010). Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur. J. Hum. Genet.* **19** 164.

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

JORGENSON, E. and WITTE, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.* **7** 885–891.

KANG, G., YUE, W., ZHANG, J., CUI, Y., ZUO, Y. and ZHANG, D. (2008). An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theoret. Biol.* **250** 362–374.

KWEE, L. C., LIU, D., LIN, X., GHOSH, D. and EPSTEIN, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.* **82** 386–397.

LI, S., LU, Q. and CUI, Y. (2010). A systems biology approach for identifying novel pathway regulators in eQTL mapping. *J. Biopharm. Statist.* **20** 373–400. MR2752212

LI, J., ZHANG, K. and YI, N. (2011). A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. *Hum. Hered.* **71** 148–160.

LI, S., LU, Q., FU, W., ROMERO, R. and CUI, Y. (2009). A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. *Stat. Appl. Genet. Mol. Biol.* **8** Art. 45, 28. MR2578777

LI, M., ROMERO, R., FU, W. J. and CUI, Y. (2010). Mapping haplotype-haplotype interactions with adaptive LASSO. *BMC Genet.* **11** 79.

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088, 1311. MR2414585

MA, S., SONG, X. and HUANG, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 60.

MA, S., ZHANG, Y., HUANG, J., HAN, X., HOLFORD, T., LAN, Q., ROTHMAN, N., BOYLE, P. and ZHENG, T. (2010). Identification of non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. *Bioinformatics* **26** 15–21.

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 genotype 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.-S., EL BAKY, A. M. A., SOLIMAN, O. A.-S., ABDEL-HADY, H. A.-S., HAMMAD, A. M. and 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** 15–26.

PERLSTEIN, E. O., RUDERFER, D. M., ROBERTS, D. C., SCHREIBER, S. L. and 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. and LILLEHAUG, J. R. (2005). Structural and functional specificities of PDGF-C and PDGF-D, the novel members of the platelet-derived growth factors 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. and 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. Amer. Statist. Assoc.* **82** 605–611.

SHANNON, P., MARKIEL, A., OZIER, O. et al. (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. F., TAYLOR, D. W. and KAIN, K. C. (2010). Dysregulation of angiopoietins is associated with placental malaria and low birth weight. *PLoS ONE* **5** e9481.

SPEED, T. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 42–44.

SUN, W., YUAN, S. and LI, K.-C. (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** 640–647.

TORRY, D. S., MUKHERJEA, D., ARROYO, J. and 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* **59**. SIAM, Philadelphia, PA. MR1045442

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.* **23** 1865–1895. MR1389856

WANG, K. and ABBOTT, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32** 108–118.

WANG, T., HO, G., YE, K., STRICKLER, H. and ELSTON, R. C. (2009). 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.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

ZHANG, Y. and LIU, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **39** 1167–1173.

DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MICHIGAN 48824
USA
AND
DEPARTMENT OF BIOSTATISTICS
ST. JUDE CHILDREN'S RESEARCH HOSPITAL
MEMPHIS, TENNESSEE 38105
USA
E-MAIL: Shaoyu.Li@stjude.org

DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
A432 WELLS HALL
EAST LANSING, MICHIGAN 48824
USA
E-MAIL: cui@stt.msu.edu