Published in final edited form as:
Genet Epidemiol. 2013 November ; 37(7): 695–703. doi:10.1002/gepi.21749.

A Kernel Regression Approach to Gene-Gene Interaction Detection for Case-Control Studies

Nicholas B. Larson[1,§] and **Daniel J. Schaid**[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Gene-gene interactions are increasingly being addressed as a potentially important contributor to the variability of complex traits. Consequently, attentions have moved beyond single locus analysis of association to more complex genetic models. While several single-marker approaches toward interaction analysis have been developed, such methods suffer from very high testing dimensionality and do not take advantage of existing information, notably the definition of genes as functional units. Here we propose a comprehensive family of gene-level score tests for identifying genetic elements of disease risk, in particular pair-wise gene-gene interactions. Using kernel machine methods, we devise score-based variance component tests under a generalized linear mixed model framework. We conducted simulations based upon coalescent genetic models to evaluate the performance of our approach under a variety of disease models. These simulations indicate that our methods are generally higher powered than alternative gene-level approaches and at worst competitive with exhaustive SNP-level analyses. Furthermore, we observe that simulated epistatic effects resulted in significant marginal testing results for the involved genes regardless of whether or not true main effects were present. We detail the benefits of our methods and discuss potential genome-wide analysis strategies for gene-gene interaction analysis in a case-control study design.

Keywords

kernel methods; gene-gene interaction; epistasis; score tests; variance component

INTRODUCTION

Genome wide association studies (GWAS) are a popular approach toward investigating the genetic component of complex diseases. Through the use high-throughput genotyping chips, GWAS can simultaneously characterize hundreds of thousands of single nucleotide polymorphisms (SNPs) for a given subject. Analysis of GWAS data typically involves the isolated evaluation of individual SNPs for association with a given phenotype. Despite much success in identification of associated loci [Hindorff, et al. 2009], such findings generally are of modest effect and often explain only a small proportion of heritability in complex phenotypes [Manolio, et al. 2009]. This “missing heritability” has prompted investigators to consider alternative sources of genetic variation in association analysis.

[§]Corresponding author Nicholas B. Larson, Ph.D., Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Larson.nicholas@mayo.edu, (507) – 293 – 1700 (phone), (507) – 284 – 1516 (fax).

The authors declare no conflict of interest.

It is well established that coding products of some genes interact with one another molecularly in complex networks, such as enzymatic reactions and signaling cascades [Bonetta 2010]. Such interactions may contribute to the genetic variation of complex traits [Moore 2003], with multiple examples documented [Howard, et al. 2002; Li, et al. 2012; Moore and Williams 2002; Sima, et al. 2012]. Statistically, gene-gene interactions are defined as deviations from additive marginal effects of individual genes [Kempthorne 1954], and our reference of gene-gene interactions hereafter is with respect to such. In regard to genotyping data, pair-wise gene-gene interactions can be considered at the SNP-level as statistical interactions between two SNPs in respective genes of interest. Similar to single marker regression analysis, SNP-SNP interaction analysis can be framed as a traditional regression-based analysis by including interaction terms into a pair-wise generalized linear model. It is important to note that this definition of interaction does necessarily coincide with the biological interpretation of interaction, and that one does not necessarily imply the other [Greenland 2009]. Although the utility of identifying such interactions with respect to explaining missing heritability is contentious [Aschard, et al. 2012; Moore and Williams 2009], such interactions can at the very least contribute to our understanding of complex disease etiology.

Advancements in both genotyping technology and imputation methodology have increased the density of genotyped markers in the coding regions of genes. Moreover, large scale next-generation sequencing technologies, such as whole exome/genome sequencing, interrogate all genetic variation within regions of interest. Unlike traditional GWAS, these tools yield dense genotype data. Under such conditions, exhaustive genome-wide evaluation of SNP-level pair-wise interaction is computationally burdensome [Moore and Ritchie 2004]. Thus, the development of statistically powerful and computationally efficient algorithms for detecting these interactions is of great interest. A comprehensive review of gene-gene interaction analysis can be found by Cordell [Cordell 2009].

Gene-level testing has recently grown in popularity due to its dimensional reduction and biological interpretability [Jorgenson and Witte 2006; Neale and Sham 2004]. In contrast to single-SNP analyses, such tests allow for all of the SNPs within the region of a gene to be modeled jointly as a set and can take into account the linkage disequilibrium (LD) structure within the gene. By grouping SNPs based upon prior biological information, SNP-set testing may improve power and increase the chance of reproducible significant findings [Wu, et al. 2010], particularly when multiple causal SNPs are present in a given gene. While SNP-set approaches are not necessarily restricted to gene-level definition, the gene as a functional unit is a natural choice and provides an intuitive decomposition of the genome.

Kernel machine methods in particular have provided a successful tool in SNP-set association testing [Kwee, et al. 2008; Wu, et al. 2010; Wu, et al. 2011]. Such approaches determine genetic association through representations of pair-wise genomic similarity between pairs of subjects [Schaid 2010a; Schaid 2010b]. Recently, Li and Cui presented a gene-level interaction approach for continuously-valued quantitative traits using a kernel machine smoothing-spline ANOVA model, which they refer to as SPA3G [Li and Cui 2012]. An application of this method for a binary response, such as disease status, presents unique challenges which preclude a direct application of SPA3G, notably that the response can no longer be assumed to be Gaussian distributed. These challenges motivated our work to adapt the methods within SPA3G to be applicable to case-control studies.

In this paper, we outline a comprehensive approach toward hypothesis testing for marginal and interaction effects of genes in association analysis for dichotomous responses using regression-based score tests. In addition to detailing omnibus and marginal tests, we define a kernel regression approach toward gene-gene interaction detection for a dichotomous

response under a generalized linear mixed model (GLMM) framework. We evaluate the performance of these testing approaches using coalescent simulation data under a variety of experimental conditions, and investigate their relation to one another within the context of multiple epistatic models. We also compare our approach to exhaustive SNP-SNP logistic regression and two leading gene-level gene-gene interaction methods. Finally, we discuss the implications of our findings and suggest future directions for further development.

METHODS

Consider a case-control association study involving N individuals, such that N is composed of N_{case} cases and N_{cont} controls. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^T$ be a binary representation of case-control status, such that $y_j = 1$ if the j^{th} subject is designated a case and 0 otherwise. Let $\boldsymbol{X}$ be an $N \times p$ set of any additional covariate data, and $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ be respective $N \times q_1$ and $N \times q_2$ matrices of genotypes for markers contained within the regions of genes 1 and 2, where q_1 and q_2 correspond to the number of respective markers within each gene. It is assumed that these regions are defined *a priori* based upon some relevant biological criteria. We define genotypes under an additive model, such that $\boldsymbol{G}_i(j, k) \equiv g_{ijk} \in \{0,1,2\}$ is the integer count of minor alleles observed at marker k in gene i for subject j.

Using a positive-definite kernel function, $\kappa(\cdot,\cdot)$, we can map $\boldsymbol{G}_i$ to some Hilbert space through the mapping $\varphi: G \rightarrow F$ such that F is an inner product space. This is accomplished through the "kernel trick" [Schölkopf and Smola 2002] which calculates inner products in F through the given kernel function, such that

$$\kappa\left(\boldsymbol{G}_i(j,\cdot), \boldsymbol{G}_i(j',\cdot)\right) = \left\langle \phi(\boldsymbol{G}_i(j,\cdot)), \phi(\boldsymbol{G}_i(j',\cdot)) \right\rangle_F$$

where $\boldsymbol{G}_i(j,\cdot)$ represents all of the marker genotypes for in gene i for subject j. The kernel function circumvents the necessity to calculate the explicit mappings $\varphi(\boldsymbol{G}_i(j,\cdot))$, yielding the kernel space mapping $\boldsymbol{K}_i$ of the respective original genotype matrix $\boldsymbol{G}_i$. This kernel matrix $\boldsymbol{K}_i$ is an $N \times N$ full Gram matrix, such that the element-wise definition is given as $\boldsymbol{K}_i(j, j') = \kappa(\boldsymbol{G}_i(j,\cdot), \boldsymbol{G}_i(j',\cdot))$, for $j, j' = 1, \ldots, N$. From Aronszajn [Aronszajn 1950], we also define the interaction kernel matrix $\boldsymbol{K}_3$ as $\boldsymbol{K}_1 \circ \boldsymbol{K}_2$, where the operator $\circ$ represents the Hadamard, or element-wise, product. Through $\boldsymbol{K}_1$, $\boldsymbol{K}_2$, and $\boldsymbol{K}_3$, the genetic effect of the two genes of interest on the phenotypic variation is decomposed into main and interaction effects. These matrices in turn can be applied in a mixed-model context as underlying covariance structures for variance components. Let μ_i represent the probability that the i^{th} observation is a case, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^T$. We consider a mixed effects logistic model for $\boldsymbol{\mu}$, such that

$$logit(\mu) = \tilde{\boldsymbol{X}}\beta + m_1 + m_2 + m_3$$

where $m_1 \sim N(\boldsymbol{0}, \tau_1^2 \boldsymbol{K}_1)$, $m_2 \sim N(\boldsymbol{0}, \tau_2^2 \boldsymbol{K}_2)$, and $m_3 \sim N(\boldsymbol{0}, \tau_3^2 \boldsymbol{K}_3)$ are independent $N \times 1$ random effect vectors, and $\tilde{\boldsymbol{X}} = [\boldsymbol{1}\ \boldsymbol{X}]$.

Global Hypothesis Test

Define the omnibus, or global, hypothesis of no genetic effect such that $H_0: \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$. The score statistic is defined as $Q_0 = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}_{aU} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, where $\boldsymbol{K}_{all} = \boldsymbol{K}_1 + \boldsymbol{K}_2 + \boldsymbol{K}_3$ and $\hat{\boldsymbol{\mu}}$ are the fitted values of μ on $\tilde{\boldsymbol{X}}$ under H_0. Under the null hypothesis, Q_0 is asymptotically distributed as a weighted mixture of chi-square distributions [Liu, et al. 2008]. While there are a number of methods to characterize this distribution for purposes of hypothesis testing,

we employ Pearson's three-moment approach [Imhof 1961], since the approximation error can be bounded.

Marginal and Interaction Hypothesis Tests

It is possible to test for the presence of marginal effects of each gene individually by using the respective kernel matrix in the framework of the score statistic, such that $Q_i = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}_i (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ for $i = 1,2$. This is equivalent to the sequence kernel association test (SKAT) [Wu, et al. 2011]. If there are no marginal effects present ($\tau_1^2=0$, $\tau_2^2=0$), we can also test specifically for a statistical interaction between genes 1 and 2 via the score statistic $Q_3 = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}_3 (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, which we refer to as the interaction test. For any of these tests, we again approximate the null distribution of Q_i by the Pearson's approximation.

Composite Hypothesis Test

We also define a test specifically for an interaction effect adjusting for the presence of marginal gene effects ($\tau_1 \neq 0$, $\tau_2 \neq 0$), such that $H_0{:}\tau_3^2=0$. This requires fitting the null GLMM that includes the main effects of the two genes, which may be conducted using penalized quasi-likelihood [Breslow and Clayton 1993] (PQL). Maximum likelihood approaches toward fitting GLMMs involve intractable integration of high dimension, and PQL utilizes Laplace approximation in order to accommodate this integration through iterative estimation of the fixed and random model components. For our purposes, we fit this model using the glmmPQL function from the MASS library in R [Venables and Ripley 2002].

Definition of the corresponding score statistic is complicated by the fact that the covariance matrix is no longer diagonal, but includes off-diagonal binomial covariances, which are difficult to obtain. One remedy is to adapt work by Lin [Lin 1997], which outlines score statistics for variance component testing in GLMMs as follows. Define $\boldsymbol{\Delta}$ and $\boldsymbol{W}$ to be diagonal $N \times N$ matrices with corresponding diagonal elements

$$\delta_j = 1/g'(\mu_j) \quad \text{and} \quad w_j = \left[V(\mu_j) \left(g'(\mu_j) \right)^2 \right]^{-1} = (V(\mu_j))^{-1} \delta_j^2$$

where $g(\cdot)$ is the link function in the GLMM, $g'(\mu_j)$ denotes the first derivative of $g(\mu_j)$ with respect to μ_j, $V(\cdot)$ is the corresponding variance function, and μ_j is the mean for the j^{th} subject under the null model. Since we apply the canonical logit link function, it follows that $\delta_j = w_j = \mu_j(1 - \mu_j)$. From Lin [Lin 1997], we define $\boldsymbol{y}^*$ to be the PQL working vector under the null GLMM, such that

$$\boldsymbol{y}^* = \tilde{\boldsymbol{X}} \hat{\beta} + \hat{\eta}_1 + \hat{\eta}_2 + \boldsymbol{\Delta}^{-1} (\boldsymbol{y} - \hat{\mu})$$

Then we define restricted maximum likelihood (REML) version of our composite score statistic to be

$$Q_C (\hat{\beta}, \hat{\eta}_1, \hat{\eta}_2) = (\boldsymbol{y}^* - \tilde{\boldsymbol{X}} \hat{\beta})^T \boldsymbol{P}_C \boldsymbol{K}_3 \boldsymbol{P}_C (\boldsymbol{y}^* - \tilde{\boldsymbol{X}} \hat{\beta})$$

where $\boldsymbol{P}_C = \boldsymbol{V}_C^{-1} - \boldsymbol{V}_C^{-1} \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{V}_C^{-1} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{V}_C^{-1}$ is the null projection matrix and $\boldsymbol{V}_C = \boldsymbol{W}^{-1} + \hat{\tau}_1^2 \boldsymbol{K}_1 + \hat{\tau}_2^2 \boldsymbol{K}_2$ is the estimated null covariance matrix with variance component

parameter estimates $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$. While Lin goes on to define a normalized version of the score statistic, our early findings indicated strong biases for a dichotomous response under the null. Similar to the global and marginal score tests, we derive the null distribution for Q_C using the Pearson's approximation.

Computational Considerations

Fitting the composite null model using PQL requires that $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ be decomposed into corresponding square-root matrices $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, such that $\boldsymbol{Z}_1\boldsymbol{Z}_1^T=\boldsymbol{K}_1$ and $\boldsymbol{Z}_2\boldsymbol{Z}_2^T=\boldsymbol{K}_2$. When a linear (or weighted linear) kernel is used, this is easily accommodated since $\boldsymbol{K}_1=\boldsymbol{G}_1\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{G}_1^T$, where $\boldsymbol{A}$ is a diagonal weight matrix, such that $\boldsymbol{Z}_1=\boldsymbol{G}_1\boldsymbol{A}^{\frac{1}{2}}$. If a nonlinear kernel function, such as the Gaussian kernel, is used, then this may be completed using the incomplete Cholesky decomposition [Kershaw 1978] of $\boldsymbol{K}_i$, whereby $\boldsymbol{Z}_i$ is the lower triangle matrix. Then, the random effects $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are modeled as $\boldsymbol{Z}_1\,\boldsymbol{b}_1$ and $\boldsymbol{Z}_2\,\boldsymbol{b}_2$, such that $\boldsymbol{b}_1 \sim N(0,\tau_1^2\,\boldsymbol{I})$ and $\boldsymbol{b}_2 \sim N(0,\tau_2^2\,\boldsymbol{I})$. Since such decompositions can be computationally intensive, there is initial appeal to the use of some form of linear kernel for this application, particularly when the number of markers per gene is relatively small.

Algorithms for approximating the null distribution of the score statistics (Q_1, Q_2, Q_3, Q_{all}, Q_C) are dependent upon deriving the eigenvalues of $\boldsymbol{U}=\boldsymbol{KP}$ for the respective kernel matrix $\boldsymbol{K}$ and projection matrix $\boldsymbol{P}$ of each test, which always will be $N\times N$. This can be computationally demanding, as such decompositions are in practice $O(N^3)$. However, equivalent eigenvalues can be derived from $\boldsymbol{U^*}=\boldsymbol{Z^T\,PZ}$. This form is more appealing for two reasons: (1) it is it is guaranteed to be positive definite, which can be exploited by decomposition algorithms; and (2) if $rank(\mathbf{Z}) \ll N$, the computational burden of this eigendecomposition is greatly reduced. This can motivate the use of low-rank approximations of $\boldsymbol{Z}$, although we leave this topic to future research.

Kernel Selection

There are multiple options for which kernel function to apply to the marker data [Schaid, et al. 2005]. We used a polygenic kernel, which is a linear kernel applied to standardized genotype data. We define the polygenic kernel representation for gene i to be $\boldsymbol{K}_i=\frac{1}{q_i}\tilde{\boldsymbol{G}}_i\tilde{\boldsymbol{G}}_i^T$ where

$$\tilde{g}_{ijk}=\frac{g_{ijk}-2\pi_{ik}}{\sqrt{2\pi_{ik}\,(1-\pi_{ik})}}$$

Since this is a type of linear kernel, it affords some computational benefits mentioned previously. However, there may be gains in statistical power in utilizing nonlinear kernel functions, such as the Gaussian kernel, which may be capable of detecting nonlinear interactions.

Simulation Study

In order to assess the properties of type I error rate control and statistical power for our hypothesis tests, we devised a comprehensive simulation study. Our basic simulation strategy was to simulate haplotypes and randomly combine haplotypes to create a large population of genotypes. Then, under a given genetic disease model and prevalence, we simulated disease status and perform case-control sampling to obtain our test data. The details of our simulation are given below.

To simulate genotypic data, we used the calibrated coalescent model simulation software COSI [Schifano, et al. 2012] to generate two independent sets of 10,000 50kb regions, each representative of a distinct gene. Recombination maps were based upon observed LD structure in samples of European ancestry. A derived minor allele frequency (dMAF) was calculated for each marker based upon its frequency in the haplotype population to represent a population-based value. From these pools of haplotypes we generated a large population of N_{pop} genotype profiles for simulated individuals by combining two randomly selected haplotypes. The two gene-wise data sets had 1017 and 1040 polymorphic sites, respectively, with 116 and 164 being common SNPs (dMAF ≥ 0.05). We then selected a subset of common SNPs for each gene to represent our simulation genotyped marker data, such that the maximum pairwise Pearson correlation between any two SNPs in a given gene was ≤ 0.50. This resulted in 12 and 25 genotyped SNPs for genes 1 and 2, respectively, ranging in dMAF from 0.05 to 0.49. LD plots of both SNP-sets are found in Figure 1.

To simulate disease status for given genotypes, we adopted a model parameterization applied by Aschard et al. [Aschard, et al. 2012], which used a log-additive approach such that the marginal and interaction effects are independent, in order to directly control the marginal and interaction effect sizes. This approach uses a recoding of the genotype values g_{ijk} to corresponding genotype weights, g_{ijk}, which are based upon the dMAF of the respective SNPs. Let Ω_1 and Ω_2 respectively define the subsets of gene 1 and gene 2 SNPs selected to be causal. Dichotomous phenotypes are then simulated via a log-linear model with probability of occurrence μ_j, such that for subject j

$$\log(\mu_j) = a_0 + \sum_{l \in \Omega_1} \beta_{1l}\, g_{1jl} + \sum_{m \in \Omega_2} \beta_{2m}\, g_{2jm} + \sum_{l \in \Omega_1, m \in \Omega_2} I_{lm}\, g_{1jl}\, g_{2jm}\, \gamma_{lm}$$

where log indicates the natural logarithm, a_0 is the population average prevalence, β_{1l} and β_{2m} are marginal effects for the respective SNPs, and γ_{lm} is the interaction effect between SNP l in gene 1 and SNP m in gene 2, with I_{lm} (0 or 1) an indicator for the presence of that specific interaction. The genotype weights g_{ijk} are functions of the population-level MAF (dMAF) of the respective SNPs, and are defined such that the expected effect of each interaction term conditional on a specific genotype at one locus is always equal to 0 (see Aschard et al. for details). We let all marginal effects be randomly selected uniformly between log(1.1) and log(1.3) to reflect realistic relative risk (RR) values observed in GWAS. By setting various effect components to be null, we also control which genetic effects are present in our disease model. For each simulation, we generated a population of N_{pop} = 100,000 genotypes and performed case-controls sampling, with disease prevalence fixed at 0.10 for each simulation. All causal SNPs were randomly selected for each simulation replication.

Finally, given that gene-gene interaction analysis is an active area of research, we compared the power of our testing procedures to gene-based Bonferroni-adjusted single SNP-SNP logistic regression, along with two leading gene-level approaches: KCCA [Larson, et al. 2013; Yuan, et al. 2012] and principal component (PC) analysis-based logistic regression modeling (PC-LR). KCCA is an LD-based procedure which uses kernelized canonical correlation analysis to test for differences in association between genes across case-control status using a Gaussian kernel function. Variations of PC-LR [Bhattacharjee, et al. 2010; He, et al. 2011; Wang and Abbott 2008] have been shown to be powerful approaches for gene-level interaction analysis by reducing the marker data for a given gene to a few leading PCs. For our PC-LR analysis, we derive the lead PC term from each gene and test the statistical significance of their interaction in the presence of their marginal effects within a basic logistic regression model.

RESULTS

Type I Error

We examined Type I error rate control for sample sizes of 1000, 1500, and 2000, with balanced numbers of cases and controls. For the global, marginal, and interaction tests, a total of 100,000 simulation runs were run for each sample size, with Type I error rates evaluated at α levels of 0.001, and 0.0001. Table I presents the Type I error simulation results for these tests, along with Figure 2 presenting QQ-plots of the respective −log10 transformed p-values. These tests exhibit near nominal type I error rates across all α levels, with the interaction test tending toward being more conservative for smaller sample sizes.

We also examined Type I error rate control for composite test when marginal effects are present in both genes but there is no interaction ($I_{lm} = 0$), and contrast it with that of the interaction test where such marginal effects are not taken into account. We considered disease models where the number of causal markers per gene was 1 or 2, and ran 4000 replications. Results for the error rates of the two tests can be found in Table II at α levels of 0.05 and 0.01. Interestingly, the findings indicate that both the interaction test and composite test control the Type I error rate under both models despite the lack of marginal effect adjustment for the interaction test.

Power

We first considered a set of simulations in which there were single causal interacting SNPs in each gene for sample sizes of $N = 1000$, 1500, and 2000. Since there is specific interest in being able to detect interacting loci in the absence of marginal effects, we considered simulations conditions with and without marginal effects present. We examined four specific values of γ_{12} (log(1.5), log(2.0), log(2.5), log(3.0)) in our simulations, and ran 500 replications for unique set of conditions, reporting empirical power at an α level of 0.05. Figure 3 presents our findings for all of our score-based tests along with the SNP-SNP, PCA, and KCCA approaches under these simulation conditions. The results show that when marginal effects are present, the various score tests generally perform best, especially at lower values of γ_{12}. When marginal effects were absent, KCCA and the global test had the highest power at lower effect sizes as well. Interestingly, the marginal tests indicate power levels above the Type I error rate despite no marginal effects being explicitly modeled.

In all simulations, the SNP-SNP approach tended to be best (or at least competitive) when the interaction effect size was most extreme, regardless of whether or not marginal effects were present. This corroborates previous findings which have found SNP-SNP methods to be competitively powerful when the gene-level interaction is isolated to a single pair of SNPs [He, et al. 2011; Li and Cui 2012].

We also considered an additional set of simulations where two pairs of interacting SNPs were present across genes, and values of γ_{lm} were randomly sampled uniformly from the interval [log(1.5),log(2.0)]. All other simulation conditions were the same as previously defined and 1000 replications were run per unique set of conditions. A barplot of these results can be found in Figure 4. These findings indicate that even in the absence of marginal effects, the global test is the most powerful approach for identifying the presence of interaction. The interaction and composite tests were relatively close in their empirical power, and performed similarly to the SNP-SNP testing. The KCCA approach performed comparably to the previously mentioned test when no marginal effects were present, but was less powerful when marginal effects were included.

It is important to note that under all simulations, the interaction test was more powerful than the composite test regardless of the inclusion of marginal effects.

DISCUSSION

Gene-gene interactions are becoming an increasingly common component to genomic association analysis. Increasing GWAS chip sizes, imputation, and next-generation sequencing platforms will continue to increase the number of genotyped intragenic SNPs, and the need for computationally efficient strategies for exploratory interaction analysis among loci has grown in response. In this paper we have detailed a comprehensive approach toward detecting the presence of genetic effects, specifically gene-gene interactions, for case-control genetic association studies. We have devised a global test for detecting the presence of gene-level associations via kernel matrix representations of marker data. Using a simulation study based upon realistic genotype data, we have demonstrated that it is a powerful approach toward detecting the presence of both main and interaction effects of gene-level risk association. By adapting the work of Li and Cui for quantitative traits to binary traits using GLMMs, we have also defined a score test, the composite test, for detecting gene-gene interactions after adjusting for main effects.

As Figures 3 and 4 indicate, the global test is a powerful approach toward detecting gene-gene interactions even in the absence of marginal effects. Given that the global test only requires fitting a single null regression model, it is a computationally attractive screening procedure for possible interactions and can rapidly be implemented in a genome-wide analysis. Subsequent testing performed on significant findings can then be applied to identify the particular architecture of the genetic association. We also found that marginal tests result in significant findings despite the exclusion of marginal effects from our simulations. Although lower-powered than the global test, conducting solely marginal tests (SKAT) could be an effective alternative strategy in contrast to the testing burden of exhaustive pairwise exploratory analysis.

As per Table II, the interaction test (Q_3) does not incur any quantifiable bias when multiple SNPs with true marginal effects are present in the simulation model. While the included simulations are restricted to a relatively small number of total SNPs per gene as well as marginal effects of modest size, this is a surprising result that raises the question of whether or not the interaction test can be used as a proxy for the composite test. More surprising is that the interaction test is more powerful than the composite test in all of our simulations. While we refrain from recommending the composite test be abandoned for the interaction test, it is computationally appealing prospect which warrants further investigation.

With increasing numbers of polymorphic sites being either genotyped or imputed in association studies, computational burden is of particular importance, especially relative to SNP-level testing. For example, on a modern workstation with an Intel® Core™ i5 3.10 Ghz processor and 4 GB of RAM, running all possible pairwise SNP-SNP tests for our simulation required 7.914 seconds per simulation replication when $N = 1000$. Running the global score test, meanwhile, requires only 2.595 seconds. This discrepancy in computational burden is further evidenced if we increase SNP-level testing burden, as such analyses scale poorly as the number of included SNPs increases. If we consider a simple data simulation where genotypes are independently sampled from a binomial distribution, and set the number of genotyped SNPs per gene to 100, the respective compute times for exhaustive SNP-SNP testing and the global test are 236.54 and 22.00 seconds. It is important to note, however, that the computational burden of the kernel-based tests scales largely with respect to sample size N, as this requires decomposition of larger and larger kernel Gram matrices. Respective compute times for the SNP-SNP tests and the global test when $N = 2000$ on our COSI simulation data are 12.123 and 34.044 seconds. This burden can be mitigated with varying strategies, however, including low-rank decompositions

[Bach and Jordan 2005], which could significantly reduce computational times. More work is necessary to explore the utility of these approaches.

Even with computationally efficient implementations of our gene-level interaction tests, exhaustive pairwise analysis of a genome with 25,000 genes would require $\binom{25,000}{2} = 312,487,500$ hypothesis tests, which is generally infeasible with respect to both computational and multiple testing burdens. Efficient strategies for implementing agnostic genome-wide analysis thus should be dependent in part on prior functional information. One strategy would be to utilize protein-protein interaction databases to define a body of potential gene-gene interaction pairs, greatly reducing the testing space. For example, we downloaded the PINA [Wu, et al. 2009] protein-protein interaction (PPI) dataset for binary interactions in Homo sapiens (accessed February 2013). This information was reduced to the gene level (HUGO designation) and redundant pairs were removed. This resulted in 106,004 unique gene pairs between 14,784 individual genes, a substantially reduced testing multiplicity. Stricter inclusion criteria, such as experimental validation, can further reduce this testing set.

Although there are a number of benefits to gene-level testing, questions remain as to how to interpret replicability of specific findings, since it is possible different sets of interacting SNPs may yield the same significant gene-pair. This requires a paradigm shift in how gene-level association is considered relative to individual SNPs, being more akin to gene-set types of analyses. Moreover, special considerations will be necessary for multiple testing, since there is a clear issue of dependence among test statistics where a given gene is a member of multiple gene-pairs being evaluated. Additional work is necessary to evaluate the effects of such dependence on multiple testing correction.

Power analysis for multi-locus approaches such as gene-level testing is complicated by a number of factors, including the quantity of total and interacting SNPs, their respective MAFs, overall LD structure of the genotyped SNPs themselves, and underlying models of epistasis [Marchini, et al. 2005]. While our random selection of causal SNPs in our simulations averages over a number of these factors, our simulations are by no means exhaustive and systematic influences on power will remain. The kernel function itself may also impact statistical power, as the polygenic kernel is just one of many possible options and alternative selections may behave differently from our findings. While it is not within the scope of this paper to investigate the impact of the kernel function itself, we acknowledge that strategic kernel selection may impact hypothesis testing performance. Influence of kernel selection under differing epistatic models is a focus of future work, particularly with respect to its comparative performance with KCCA, which is specifically capable of nonlinear interaction detection.

While we have presented this work strictly within the context of a dichotomous trait, we note that the theoretical adaptation of our approach from SPA3G could be modified to account for any non-Gaussian response with a presumed exponential family distribution with little difficulty. We also foresee this testing framework being expanded to address pathway analysis applications and higher order interactions through linear combinations of gene-level kernel matrices and their Hadamard products.

Acknowledgments

This research was supported by the U.S. Public Health Service, National Institutes of Health, contract number GM065450. We also thank the anonymous reviewers for their constructive comments.

References

Aronszajn N. Theory of reproducing kernels. Transactions of the American Mathematical Society. 1950; 68

Aschard H, Chen J, Cornelis MC, Chibnik LB, Karlson EW, Kraft P. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. American journal of human genetics. 2012; 90(6):962–72. [PubMed: 22633398]

Bach, FR.; Jordan, MI. Predictive low-rank decomposition for kernel methods. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: ACM; 2005. p. 33-40.

Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N. Using principal components of genetic variation for robust and powerful detection of gene-gene interactions in case-control and case-only studies. American journal of human genetics. 2010; 86(3):331–42. [PubMed: 20206333]

Bonetta L. Protein-protein interactions: Interactome under construction. Nature. 2010; 468(7325):851–4. [PubMed: 21150998]

Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association. 1993; 88(421):9–25.

Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics. 2009; 10(6):392–404.

Greenland S. Interactions in epidemiology: relevance, identification, and estimation. Epidemiology. 2009; 20(1):14–7. [PubMed: 19234397]

He J, Wang K, Edmondson AC, Rader DJ, Li C, Li MY. Gene-based interaction analysis by incorporating external linkage disequilibrium information. European Journal of Human Genetics. 2011; 19(2):164–172. [PubMed: 20924406]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(23):9362–7. [PubMed: 19474294]

Howard TD, Koppelman GH, Xu JF, Zheng SQL, Postma DS, Meyers DA, Bleecker ER. Gene-gene interaction in asthma: IL4RA and IL13 in a Dutch population with asthma. American journal of human genetics. 2002; 70(1):230–236. [PubMed: 11709756]

Imhof JP. Computing the Distribution of Quadratic Forms in Normal Variables. Biometrika. 1961; 48(3/4):419–426.

Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. Nature Reviews Genetics. 2006; 7(11):885–891.

Kempthorne O. The Correlation between Relatives in a Random Mating Population. Proceedings of the Royal Society of London. Series B - Biological Sciences. 1954; 143(910):103–113.

Kershaw DS. Incomplete Cholesky-Conjugate Gradient Method for Iterative Solution of Systems of Linear Equations. Journal of Computational Physics. 1978; 26(1):43–65.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. American journal of human genetics. 2008; 82(2):386–97. [PubMed: 18252219]

Larson NB, Jenkins GD, Larson MC, Vierkant RA, Sellers TA, Phelan CM, Schildkraut JM, Sutphen R, Pharoah PP, Gayther SA, et al. Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. European journal of human genetics : EJHG. 2013

Li S, Cui Y. Gene-centric gene-gene interaction: A model-based kernel machine method. Annals of Applied Statistics. 2012; 6(3):1134–1161.

Li Z, Zhang Y, Wang Z, Chen J, Fan J, Guan Y, Zhang C, Yuan C, Hong W, Wang Y, et al. The role of BDNF, NTRK2 gene and their interaction in development of treatment-resistant depression: Data from multicenter, prospective, longitudinal clinic practice. Journal of psychiatric research. 2012

Lin XH. Variance component testing in generalised linear models with random effects. Biometrika. 1997; 84(2):309–326.

Liu DW, Ghosh D, Lin XH. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC bioinformatics. 2008; 9

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. [PubMed: 19812666]

Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics. 2005; 37(4):413–7. [PubMed: 15793588]

Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human heredity. 2003; 56(1-3):73–82. [PubMed: 14614241]

Moore JH, Ritchie MD. STUDENTJAMA. The challenges of whole-genome approaches to common diseases. JAMA : the journal of the American Medical Association. 2004; 291(13):1642–3. [PubMed: 15069055]

Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Annals of Medicine. 2002; 34(2):88–95. [PubMed: 12108579]

Moore JH, Williams SM. Epistasis and Its Implications for Personal Genetics. American journal of human genetics. 2009; 85(3):309–320. [PubMed: 19733727]

Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. American journal of human genetics. 2004; 75(3):353–62. [PubMed: 15272419]

Schaid DJ. Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Human heredity. 2010a; 70(2):109–131. [PubMed: 20610906]

Schaid DJ. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Human heredity. 2010b; 70(2):132–140. [PubMed: 20606458]

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. American journal of human genetics. 2005; 76(5):780–793. [PubMed: 15786018]

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. SNP Set Association Analysis for Familial Data. Genetic epidemiology. 2012

Schölkopf, B.; Smola, A. Learning with kernels : support vector machines, regularization, optimization, and beyond. The MIT Press; 2002.

Sima X, Xu J, Li Q, Luo L, Liu J, You C. Gene-gene interactions between interleukin-12A and interleukin-12B with the risk of brain tumor. DNA and cell biology. 2012; 31(2):219–23. [PubMed: 22011063]

Venables, WN.; Ripley, BD. Modern applied statistics with S. New York: Springer; 2002.

Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies. Genetic epidemiology. 2008; 32(2):108–18. [PubMed: 17849491]

Wu J, Vallenius T, Ovaska K, Westermarck J, Makela TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. Nature methods. 2009; 6(1):75–7. [PubMed: 19079255]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. American journal of human genetics. 2010; 86(6):929–42. [PubMed: 20560208]

Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American journal of human genetics. 2011; 89(1):82–93. [PubMed: 21737059]

Yuan ZS, Gao QS, He YG, Zhang XS, Li FY, Zhao JH, Xue FZ. Detection for gene-gene co-association via kernel canonical correlation analysis. BMC genetics. 2012; 13

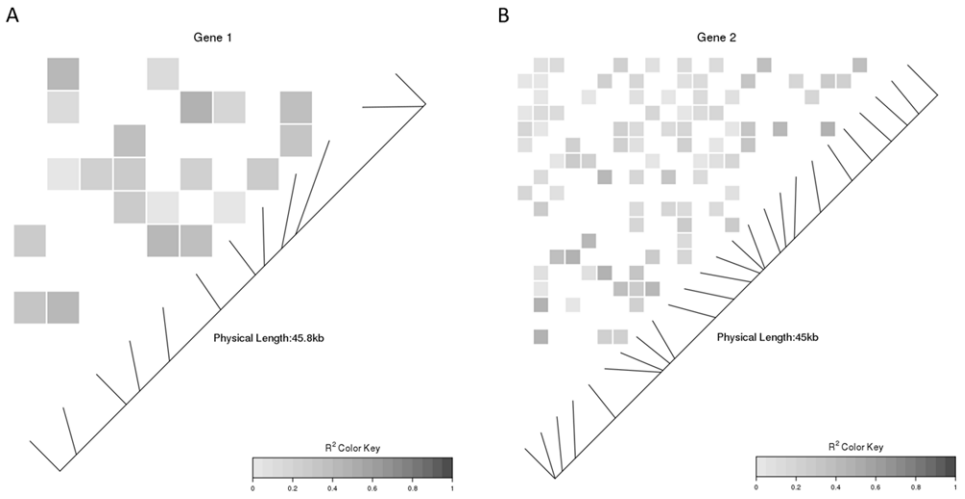

Figure 1.
Pairwise linkage disequilibrium plots of the simulation SNPs for (A) gene 1 and (B) gene 2.

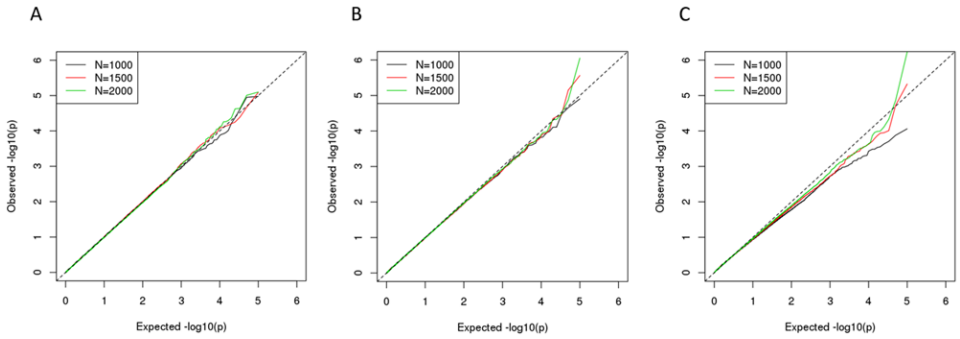

Figure 2.
QQ-plots of the −log10 transformed p-values for the (A) global test and (B) marginal test under the complete null model, for sample sizes of 200, 500, and 1000.

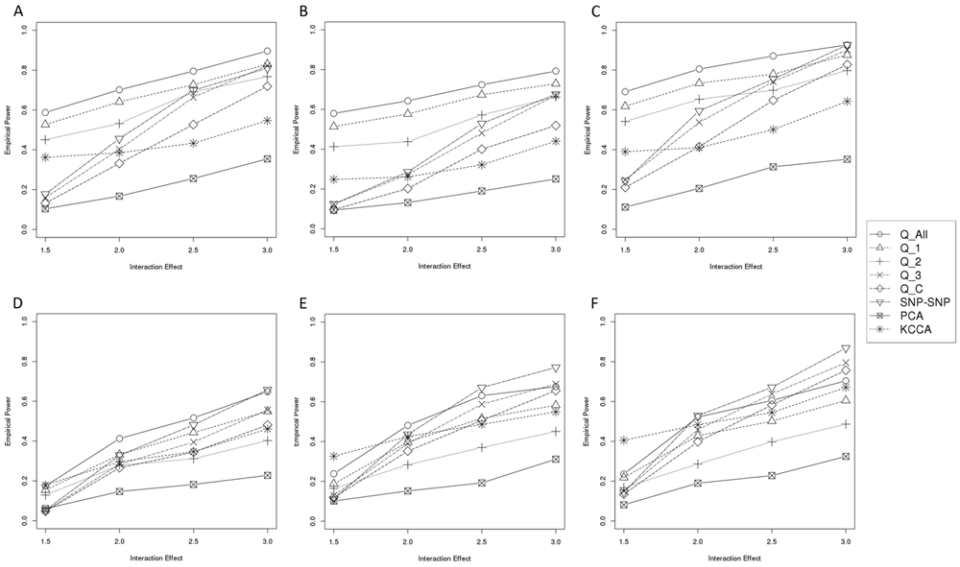

Figure 3.
Empirical power curves ($\alpha = 0.05$) as a function of interaction effect size $\exp(\gamma_{12})$, for the global, marginal, interaction, and composite tests, along with SNP-SNP logistic regression, PCA, and KCCA methods. Results are shown with marginal effects present for sample sizes (A) $N = 1000$, (B) $N = 1500$, and (C) $N = 2000$, and with marginal effects absent for sample sizes (D) $N = 1000$, (E) $N = 1500$, and (F) $N = 2000$.

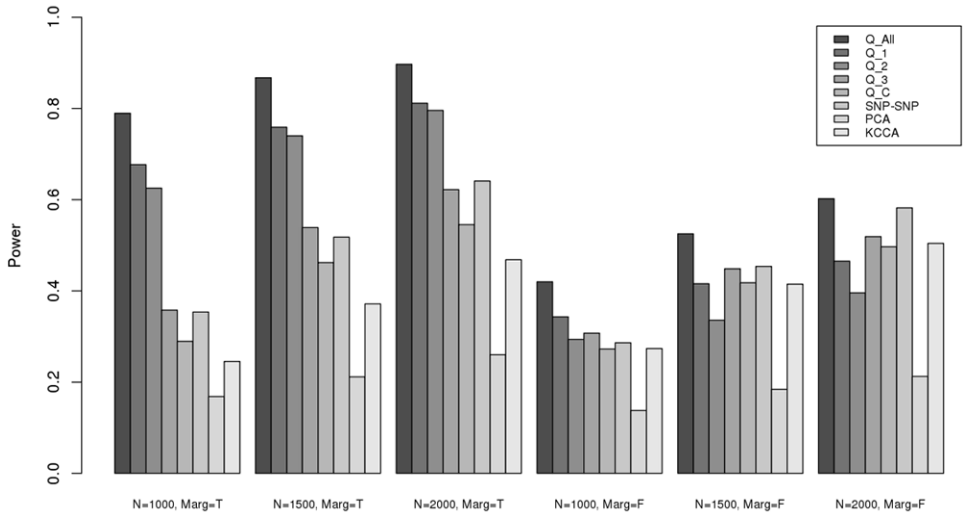

Figure 4.

Barplot of empirical power results ($\alpha = 0.05$) for hypothesis testing when the number of causal SNPs per gene is two, where interaction effects γ_{lm} are uniformly drawn from [log(1.5), log(2.0)] Results are presented for sample sizes of $N = 1000$, 1500, and 2000, with marginal effects either present (Marg = T) or absent (Marg = F).

Table I

Complete Null Type I Error Rates for Global, Marginal, and Interaction Tests

	Global Test		*Marginal Test*		*Interaction Test*	
N	**α = 1e-03**	**α = 1e-04**	**α = 1e-03**	**α = 1e-04**	**α = 1e-03**	**α = 1e-04**
1000	8.3e-04	5.0e-05	9.3e-04	6.0e-05	3.7e-04	1.0e-05
1500	8.0e-04	6.0e-05	1.1e-03	1.1e-04	5.4e-04	3.0e-05
2000	8.7e-04	6.0e-05	1.1e-03	1.2e-04	7.0e-04	4.0e-05

Genet Epidemiol. Author manuscript; available in PMC 2013 December 01.

Table II

Type I Error Rates for Interaction and Composite Tests with Marginal Effects Present

	1 Causal SNP per Gene				*2 Causal SNPs per Gene*			
	Interaction **(Q_3)**		***Composite*** **(Q_C)**		***Interaction*** **(Q_3)**		***Composite*** **(Q_C)**	
N	**$\alpha = 0.05$**	**$\alpha = 0.01$**	**$\alpha = 0.05$**	**$\alpha = 0.01$**	**$\alpha = 0.05$**	**$\alpha = 0.01$**	**$\alpha = 0.05$**	**$\alpha = 0.01$**
1000	0.0390	0.0090	0.0398	0.0088	0.0355	0.0058	0.0378	0.0050
1500	0.0385	0.0065	0.0375	0.0063	0.0408	0.0070	0.0398	0.0070
2000	0.0420	0.0063	0.0438	0.0068	0.0440	0.0108	0.0445	0.0108