Multivariate detection of gene-gene interactions

Indika Rajapakse[1,2*], Michael D. Perlman[3*], John A. Hansen[4,5], and Charles Kooperberg[2]

*Contributed equally to this work

[1]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA. [2] Public Health Sciences, Fred Hutchinson Cancer Research Center. [3]Department of Statistics, University of Washington, Seattle, WA 98195. [4]Clinical Research Division, Fred Hutchinson Cancer Research Center, [5]Department of Medicine, University of Washington, Seattle, WA 98195.

Abstract

Unraveling the nature of genetic interactions is crucial to obtaining a more complete picture of complex diseases. Accumulating evidence suggests that gene-gene interactions play an important role in the etiology of cancer, cardiovascular and immune-mediated disease. Interactions among genes are defined as phenotypic effects that differ from those observed for independent contributions of each gene, usually detected by univariate logistic regression methods. Using a multivariate extension of linkage disequilibrium, we have developed two novel methods, based on distances between sample covariance matrices for groups of SNPs, to test for gene-gene interactions associated with a disease phenotype. Since a disease-associated interacting locus will often be in linkage disequilibrium with more than one marker in the region, methods that examine a set of markers in a region collectively offer greater power than traditional methods. Our methods effectively identify interaction effects in simulated data, as well as in data on the genetic contributions to the risk for graft-versus-host disease following hematopoietic cell transplantation.

Introduction

Many complex diseases are influenced by both genetic and environmental factors. Determining the underlying genetic etiology can be difficult, as it may involve single genes as well as interactions between two or more genes. While

initial and ongoing efforts have centered on disease associations with single genes (a single nucleotide polymorphism (SNP) or haplotypes/diplotypes of multiple SNPs from single genes or regions), recent interest has expanded to include examination of gene-gene interactions regardless of their location within the genome $[1-3]$, which is the focus of our present research.

A gene-gene interaction is typically detected by testing for phenotypic effects that differ from those observed when each gene contributes independently, e.g. departure from additivity in a logistic regression model. In most genetic association studies the "causal" SNP is not genotyped, but rather inference about a functional variant is made indirectly because a SNP that is in linkage disequlibrium (LD) with the causal SNP will show association with phenotype. When the causal SNP is part of an LD group, multiple nearby SNPs may show an association. Similarly, we may expect that if there is an interaction effect on a disease of two causal SNPs, pairs of SNPs in the LD group adjacent to either of the two causal SNPs may show some association. In a traditional logistic regression analysis, this adjacent LD is not used as each pair of SNPs is tested separately for possible interactions, so we expect to lose power if nearby SNPs are not considered. Here we propose to test for interaction effects between *groups* of SNPs, thereby possibly gaining power.

Chatterjee et al. [2] developed a procedure to identify main effects and interactions of groups of SNPs simultaneously using the Tukey one degree of freedom test. However, the goal of [2] is to increase the power to identify SNPs that have a marginal effect using interactions, rather than to identify the interactions themselves. Zhao et al. [3] introduced a test for the interaction between two unlinked loci and defined interaction as deviation from penetrance "for a haplotype at two loci from the product of the marginal penetrance of the individual alleles that span the haplotype" [4]. The disadvantage of this method is that the haplotype cannot be determined with certainty. If the joint distribution of genotype markers can be shown to depend on disease status, then there is evidence that these markers (or variants in high LD with these markers) combine to affect disease risk.

Our method summarizes and contrasts the difference in LD between cases and controls. To measure the LD we use the composite LD (CLD), which is advantageous because it is not necessary to phase the genotype data. If the CLD patterns are different between cases and controls, we conclude that there is an interaction. A disease-associated interacting locus will often be in LD with more than one genotyped marker in the region. Therefore methods like ours that examine a set of markers in a region collectively can potentially

offer greater power than the traditional method of examining 2-way or 3-way interactions in univariate logistic regression models.

Linkage disequilibrium and composite linkage disequilibrium

Linkage disequilibrium indicates that particular alleles at nearby sites co-occur on the same haplotype more often than is expected by chance. Lewontin [6] defined the *gametic LD coefficient* as $D_{AB} = p_{AB} - p_A p_B$, or the simple difference between the haplotype probability and the product of the allele frequency, when data are collected on haplotypes for diallelic loci. Weir [7] and Weir & Cockerham [8] defined the *non-gametic digenic disequilibrium coefficient* $D_{A/B} = p_{A/B} - p_A p_B$, where the slash indicates that the two alleles occur on different chromosomes. For the phase-unknown situation where random mating cannot be assumed, these papers introduce the *composite linkage disequilibrium* (CLD)

$$\Delta_{AB} = D_{AB} + D_{A/B} = p_{AB} + p_{A/B} - 2p_A p_B.$$

In the context of association mapping, Nielsen et al. [9] presented a direct LD comparison approach involving two bi-allelic loci and noted that a test that directly compares the LD between the case and the control groups can be a powerful alternative to either haplotype-based or single marker approaches. They considered only the case of unambiguous haplotype phase. When the haplotype phase is unknown, computational algorithms can be used to infer frequencies of haplotypes and, ultimately, to assess LD. Typically this requires the assumption of Hardy-Weinberg equilibrium (HWE) for the haplotypes. Schaid [10] showed that LD estimation with use of the composite linkage disequilibrium approach provides results similar to the haplotype reconstruction method under HWE, is computationally simpler, and avoids the assumption of HWE for the haplotypes. Therefore we use CLD rather than using LD to characterize the relation between SNPs.

Following Weir et al. [11] we show the relationship between LD and CLD as follows. Let m and n be the number of cases and controls, respectively. Let $x_{ijk} = 1$ if the k^{th}, $k = 1, 2$, haplotype in the j^{th}, $j = 1, 2, \ldots, p$, SNP for case $i = 1, 2, \ldots, m$, carries major allele A and 0 if it carries minor allele a. The LD between SNPs j and j' is the covariance of x_{ijk} and $x_{ij'k}$ whereas

the CLD between SNPs j and j' is the covariance of

$$X_{ij} = \frac{x_{ij1} + x_{ij2}}{2} \qquad \text{and} \qquad X_{ij'} = \frac{x_{ij'1} + x_{ij'2}}{2}.$$

The quantities X_{ij} and $X_{ij'}$ are the proportions of the alleles a subject in the case group carries at SNP j and j'. Let $\mathbf{X}$ denote the $m \times p$ matrix $\{X_{ij}\}$. Similarly, define y_{ijk}, Y_{ijk}, and $\mathbf{Y}$ for the control group, where $\mathbf{Y}$ is $n \times p$. Thus, for genotype data we can estimate the CLD by the sample covariance between the genotypes $(X_{ij}, X_{ij'})$ without using phase information. Note that CLD does not require HWE to hold, but when HWE holds, CLD is equal to LD [10] , [11]. The CLD does not distinguish between the two possible phases of the double heterozygotes, so CLD can be defined for SNPs within the same chromosome (in *cis*) or between chromosomes (in *trans*).

Tests for equality of block interactions

In order to compare CLDs between two *groups* of SNPs in cases and controls, rather than only between single pairs of SNPs, we propose two multivariate statistics that measure differences between *blocks* of pairwise CLDs in cases and controls. Let group 1 have p_1 SNPs and group 2 have p_2 SNPs, where $p_1 + p_2 = p$, and let S and T be the $(p_1 + p_2) \times (p_1 + p_2)$ sample covariance matrices for the two groups of SNPs for cases and controls, based on $\mathbf{X}$ and $\mathbf{Y}$ respectively. Partition S as

$$S = \begin{array}{c} \\ p_1 \\ p_2 \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p_2 \end{array} \\ \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \end{array}, \tag{1}$$

and partition T similarly. Here S_{11} and S_{22} are the sample intra-group covariance matrices for group 1 and for group 2 respectively, and $S_{12} (= S'_{21})$ is the inter-group sample covariance matrix. Denote the corresponding quantities for the controls as T_{11}, T_{22}, and $T_{12} (= T'_{21})$. Note that if $p_1 = p_2 = 1$, then S_{12} and T_{12} both reduce to CLD as defined above.

Let Σ (cases) and Ω (controls) be the population covariance matrices that correspond to S and T respectively, partitioned according to (1). We propose to test whether the interaction effects (= covariances) between the two groups of SNPs are different for cases than for controls, that is, to test

equality of the *block interactions*, i.e., test

$$H_{12} : \Sigma_{12} = \Omega_{12}, \tag{2}$$

rather than to test for differences between single pairs of corresponding elements in Σ_{12} and Ω_{12}. (In fact we shall test $H_{12} \mid (H_1 \cap H_2)$ – see (7).)

To motivate our proposed multivariate test statistics, suppose for the moment that the underlying data matrices $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed, so that $U \equiv mS$ and $V \equiv nT$ are independent Wishart random matrices:

$$U \sim W_{p_1+p_2}(m, \Sigma), \qquad V \sim W_{p_1+p_2}(n, \Omega),$$

with m and n degrees of freedom, respectively.

First consider the classical problem of testing the hypothesis

$$H_0 : \Sigma = \Omega \quad \text{vs} \quad K : \Sigma \neq \Omega \tag{3}$$

based on U and V. If

$$\min(m, n) \geq p_1 + p_2 =: p, \tag{4}$$

so that U and V are nonsingular with probability one, the likelihood ratio test (LRT, also known as Bartlett's test; cf. Anderson [12] rejects H_0 if

$$\lambda^2 := \frac{|U+V|^{m+n}}{|U|^m |V|^n}$$

is sufficiently large. It has been noted by several authors (e.g., Chaudhuri and Perlman [13]) that λ^2 can be decomposed as follows. If we partition U and V according to (1), define $U_{11\cdot 2} = U_{11} - U_{12}U_{22}^{-1}U_{21}$, and define $V_{11\cdot 2}$, $\Sigma_{11\cdot 2}$, and $\Omega_{11\cdot 2}$ similarly, then

$$\begin{aligned} \lambda^2 &= \frac{|U_{11\cdot 2} + V_{11\cdot 2}|^{m+n}}{|U_{11\cdot 2}|^m |V_{11\cdot 2}|^n} \cdot \frac{|U_{22} + V_{22}|^{m+n}}{|U_{22}|^m |V_{22}|^n} \cdot \frac{|U_{11\cdot 2} + V_{11\cdot 2} + \Delta|^{m+n}}{|U_{11\cdot 2} + V_{11\cdot 2}|^{m+n}} \\ &\equiv \lambda_{1\cdot 2} \cdot \lambda_2^2 \cdot \lambda_{1|2}^2 \,, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \Delta = &\big(U_{12}U_{22}^{-1} - (U_{12}+V_{12})(U_{22}+V_{22})^{-1}\big)U_{22}\big(\cdots\big)' \\ &+ \big(V_{12}V_{22}^{-1} - (U_{12}+V_{12})(U_{22}+V_{22})^{-1}\big)V_{22}\big(\cdots\big)'. \end{aligned}$$

Here $\lambda_{1\cdot 2}$ is the LRT statistic for testing $H_{1\cdot 2} : \Sigma_{11\cdot 2} = \Omega_{11\cdot 2}$, λ_2 is the LRT statistic for testing $H_2 : \Sigma_{22} = \Omega_{22}$, and $\lambda_{1|2}$ is the LRT statistic for testing $H_{1|2} : \Sigma_{12}\Sigma_{22}^{-1} = \Omega_{12}\Omega_{22}^{-1}$ assuming that $H_{1\cdot 2}$ holds (denoted by $H_{1|2}|H_{1\cdot 2}$). We are particularly interested in $H_{1|2}|H_{1\cdot 2}$, which, like H_{12}, can be interpreted to indicate that the interaction effects of the two genes in case and control are identical. Under the overall null hypothesis that $\Sigma = \Omega$, $\lambda_{1\cdot 2}$, λ_2, and $\lambda_{1|2}$ are mutually independent with known null distributions that do not depend on the common value of $\Sigma = \Omega$, so these three statistics can be applied to test $H_{1\cdot 2}$, H_2, and $H_{1|2}|H_{1\cdot 2}$. In fact, that $H_0 = H_{1\cdot 2} \cap H_2 \cap H_{1|2}$.

An advantage of this approach is that if H_0 is rejected, the source of the difference between Σ and Ω is exhibited more precisely. A disadvantage is that it presumes an asymmetric relationship between genes 1 and 2, i.e., it presumes causal (directional) effects of SNP group 2 (from gene 2) on SNP group 1 (from gene 1). This is because $\Sigma_{12}\Sigma_{22}^{-1}$ and $\Omega_{12}\Omega_{22}^{-1}$ are the coefficients of the regression of the group 1 variables on the group 2 variables in cases and controls respectively. Thus this method is also applicable if the reverse causal relationships are presumed and may lead to a different conclusion, clearly an undesirable property. In the application considered here, however, there is no presumption of an asymmetric relation between the two genes. Therefore we seek methods that test the hypothesis H_{12} without presuming an asymmetric relationship between the genes.

Method 1: an alternative decomposition of the LRT statistic.

Our first approach is to modify the decomposition in (5) as follows:

$$\begin{aligned}\lambda^2 &= \frac{|U_{11}+V_{11}|^{m+n}}{|U_{11}|^m|V_{11}|^n} \cdot \frac{|U_{22}+V_{22}|^{m+n}}{|U_{22}|^m|V_{22}|^n} \cdot \frac{\left[\frac{|U+V|}{|U_{11}+V_{11}||U_{22}+V_{22}|}\right]^{m+n}}{\left[\frac{|U|}{|U_{11}||U_{22}|}\right]^m \left[\frac{|V|}{|V_{11}||V_{22}|}\right]^n} \\ &\equiv \lambda_1^2 \cdot \lambda_2^2 \cdot \lambda_{12}^2, \end{aligned} \tag{6}$$

where λ_1 is the LRT statistic for testing $H_1 : \Sigma_{11} = \Omega_{11}$ and λ_2 is again the LRT statistic for testing $H_2 : \Sigma_{22} = \Omega_{22}$. This suggests that λ_{12} is a reasonable statistic for testing

$$H_{12} \mid (H_1 \cap H_2) : \quad \Sigma_{12} = \Omega_{12} \ \text{ given that } \Sigma_{11} = \Omega_{11},\ \Sigma_{22} = \Omega_{22}. \tag{7}$$

We now express λ_{12} in a form that justifies its suitability as a test statistic for (7). Set

$$\begin{aligned} \mu &= \frac{m}{m+n}, \quad \nu = \frac{n}{m+n} \quad (\text{so } \mu+\nu=1), \\ W &= \frac{U+V}{m+n} = \mu S + \nu T, \end{aligned} \tag{8}$$

the pooled estimate of $\Sigma = \Omega$ under H_0, The statistic $\lambda_{12} \equiv \lambda_{12}(S,T)$ can be expressed as follows:

$$\begin{aligned} \lambda_{12}^2(S,T) &= \frac{|I - W_{11}^{-1}W_{12}W_{22}^{-1}W_{21}|^{m+n}}{|I - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}|^m |I - T_{11}^{-1}T_{12}T_{22}^{-1}T_{21}|^n} \\ &= \frac{|I - R_W R_W'|^{m+n}}{|I - R_S R_S'|^m |I - R_T R_T'|^n}, \\ &= \frac{\prod_{i=1}^{p_1 \wedge p_2}(1-w_i^2)^{m+n}}{\prod_{i=1}^{p_1 \wedge p_2}(1-s_i^2)^m \prod_{i=1}^{p_1 \wedge p_2}(1-t_i^2)^n}, \end{aligned} \tag{9}$$

where, using symmetric matrix square roots,

$$\begin{aligned} R_S &:= S_{11}^{-1/2} S_{12} S_{22}^{-1/2}, \\ R_T &:= T_{11}^{-1/2} T_{12} T_{22}^{-1/2}, \\ \text{and} \quad R_W &:= W_{11}^{-1/2} W_{12} W_{22}^{-1/2} \end{aligned}$$

are the matrix-valued correlations and s_i^2, t_i^2, and w_i^2 are the squared canonical correlations, both based on S, T, and W respectively, between the two groups of SNPs.

The form (9) suggests the following justification for using λ_{12}. Under $H_1 \cap H_2$, $\Sigma_{11} = \Omega_{11} =: \Xi_{11}$ and $\Sigma_{22} = \Omega_{22} =: \Xi_{22}$, so the population counterpart $\lambda_{12}^2(\Sigma,\Omega)$ of $\lambda_{12}^2(S,T)$ assuming $H_1 \cap H_2$ is the $(m+n)$-th power of

$$\frac{\left| I - \left[\Xi_{11}^{-\frac{1}{2}} \left(\mu\Sigma_{12} + \nu\Omega_{12} \right) \Xi_{22}^{-\frac{1}{2}} \right] \left[\Xi_{11}^{-\frac{1}{2}} \left(\mu\Sigma_{12} + \nu\Omega_{12} \right) \Xi_{22}^{-\frac{1}{2}} \right]' \right|}{\left| I - \left(\Xi_{11}^{-\frac{1}{2}} \Sigma_{12} \Xi_{22}^{-\frac{1}{2}} \right) \left(\Xi_{11}^{-\frac{1}{2}} \Sigma_{12} \Xi_{22}^{-\frac{1}{2}} \right)' \right|^{\mu} \left| I - \left(\Xi_{11}^{-\frac{1}{2}} \Omega_{12} \Xi_{22}^{-\frac{1}{2}} \right) \left(\Xi_{11}^{-\frac{1}{2}} \Omega_{12} \Xi_{22}^{-\frac{1}{2}} \right)' \right|^{\nu}}.$$

Because

$$\log |I - ZZ'| \equiv \log \begin{vmatrix} I & Z \\ Z' & I \end{vmatrix}$$

is strictly concave in Z provided that $I - ZZ'$ is positive definite, it follows that $\log \lambda_{12}(\Sigma, \Omega) = 0$ when $\Sigma_{12} = \Omega_{12}$ and is > 0 when $\Sigma_{12} \neq \Omega_{12}$, so $\log \lambda_{12}(\Sigma, \Omega)$ provides a measure of the distance between Σ_{12} and Ω_{12}. Because $\log \lambda_{12}(S, T)$ provides an estimate of this distance, λ_{12} appears to be a reasonable statistic for detecting departures from the null hypothesis $H_{12} \mid (H_1 \cap H_2)$. Note also that by (9), λ_{12} is invariant under all nonsingular matrix scale transformations of the form $A = \mathrm{diag}(A_1, A_2)$, $A_i : p_i \times p_i$, i.e., those linear transformations that act separately on the two groups of SNPs, that is,

$$\lambda_{12}(S, T) = \lambda_{12}(ASA',\, ATA'). \tag{10}$$

Method 2: a quadratic distance-based method.

Our second approach uses the Nagao [14] *normalized quadratic distance (NQD)*

$$\delta^2 \equiv \delta(\tilde{S}, \tilde{T}) := \mathrm{tr}[(\tilde{S} - \tilde{T})W^{-1}(\tilde{S} - \tilde{T})W^{-1}]$$

applied to $\tilde{S}$ and $\tilde{T}$, where

$$\tilde{S} = \begin{pmatrix} W_{11} & S_{12} \\ S_{21} & W_{22} \end{pmatrix}, \qquad \tilde{T} = \begin{pmatrix} W_{11} & T_{12} \\ T_{21} & W_{22} \end{pmatrix}.$$

Here W_{11} (resp., W_{22}) is the pooled estimate of Ξ_{11} (Ξ_{22}) under H_1 (H_2) based on S_{11} and T_{11} (S_{22} and T_{22}). From (8), to insure that W is nonsingular with probability 1, it is only required that

$$m + n \geq p_1 + p_2 =: p,$$

which is a weaker requirement than (4). Note too that (compare to (8))

$$W = \mu\tilde{S} + \nu\tilde{T}.$$

In general, neither $\tilde{S}$ nor $\tilde{T}$ need be positive definite. Nonetheless, δ^2 is a valid measure of distance between S_{12} and T_{12} under $H_1 \cap H_2$ because

$$\begin{aligned}\delta^2 &= \mathrm{tr}\begin{pmatrix} 0 & S_{12} - T_{12} \\ (S_{12} - T_{12})' & 0 \end{pmatrix} W^{-1} \begin{pmatrix} 0 & S_{12} - T_{12} \\ (S_{12} - T_{12})' & 0 \end{pmatrix} W^{-1} \\ &= \mathrm{tr}(S_{12} - T_{12})' W_{11\cdot 2}^{-1}(S_{12} - T_{12}) + \mathrm{tr}(S_{12} - T_{12}) W_{22\cdot 1}^{-1}(S_{12} - T_{12})'.\end{aligned}$$

Thus $\delta^2 = 0$ iff $S_{12} = T_{12}$, a property not shared by $\log \lambda_{12}$. Furthermore we have the equivalent expressions

$$\begin{aligned}\delta^2 &= \operatorname{tr} \begin{pmatrix} 0 & Q \\ Q' & 0 \end{pmatrix} \begin{pmatrix} I & R \\ R' & I \end{pmatrix}^{-1} \begin{pmatrix} 0 & Q \\ Q' & 0 \end{pmatrix} \begin{pmatrix} I & R \\ R' & I \end{pmatrix}^{-1} \\ &= \operatorname{tr}(L^2),\end{aligned}$$

where, using symmetric matrix square roots,

$$\begin{aligned} Q &= W_{11}^{-1/2} \, (S_{12} - T_{12}) \, W_{22}^{-1/2}, \\ R &= W_{11}^{-1/2} (\mu S_{12} + \nu T_{12}) W_{22}^{-1/2} \quad (= R_W), \\ L &= \begin{pmatrix} I & R \\ R' & I \end{pmatrix}^{-1/2} \begin{pmatrix} 0 & Q \\ Q' & 0 \end{pmatrix} \begin{pmatrix} I & R \\ R' & I \end{pmatrix}^{-1/2}. \end{aligned}$$

Note that L is a symmetric matrix and that

$$\delta^2 = \sum_{i=1}^{p} l_i^2,$$

where $l_1 \geq \cdots \geq l_p$ are the ordered eigenvalues of L, equivalently, the ordered eigenvalues of

$$(\tilde{S} - \tilde{T}) W^{-1} \equiv \begin{pmatrix} 0 & S_{12} - T_{12} \\ (S_{12} - T_{12})' & 0 \end{pmatrix} W^{-1}.$$

Furthermore, like λ_{12}, δ^2 is invariant under all nonsingular matrix scale transformations of the form $A = \operatorname{diag}(A_1, A_2)$, $A_i : p_i \times p_i$ (recall (10)). Thus δ^2 is another reasonable statistic for detecting departures from the null hypothesis $H_{12} \mid (H_1 \cap H_2)$.

Because CLD data is not normally distributed, the significance levels for the test statistics λ_{12} and δ^2 must be determined using a permutation method.

Simulation study

We compared our proposed tests based on λ_{12} and δ^2 to tests based on logistic regression (described below) in a simulation study. Because we wish

to test whether multiple SNPs in two genetic regions have a non-null interaction effect on a phenotype, the univariate logistic regression approaches discussed in the Introduction are not applicable. To generate our simulated data we created an artifical population using genotype data obtained from the hapmap project caucasian population [15]. We used PHASE [16] to estimate haplotypes for rs7130285, rs2074040, rs3740878, rs7935586, and rs6485533 (denoted as $A_1, \ldots, A_5$) from the *EXT2* gene and rs2713813, rs7951391, rs7480010, rs906625, and rs6485316 (denoted as $B_1, \ldots, B_5$) from the intergenic region of the *LRRC4CX2* gene (the haplotypes and their frequencies are listed in the Appendix 1). We then randomly paired haplotypes to create our population. We used interaction models developed by Marchini et al. [17] to assign case and control status, which we have denoted IM1 (for Interaction Model 1), IM2, and IM3. IM1 has main effects, but no interaction, IM2 has a multiplicative interaction, and IM3 has a threshold interaction where the risk is increased if both SNPs have at least one copy of the minor allele. Note that we can write the probability of being a case ($D = 1$) for each of these three models in a logistic regression form:

$$\begin{aligned}\operatorname{logit}\left(P\left(D=1 \mid G\right)\right) = \beta_{0,0} &+ \beta_{0,1}\left(g_2=1\right) + \beta_{1,0}\left(g_1=1\right) \\ &+ \beta_{0,2}\left(g_2=2\right) + \beta_{2,0}\left(g_1=2\right) \\ &+ \beta_{1,1}\left(g_1=1\right)\left(g_2=1\right) + \beta_{1,2}\left(g_1=1\right)\left(g_2=2\right) \\ &+ \beta_{2,1}\left(g_1=2\right)\left(g_2=1\right) + \beta_{2,2}\left(g_1=2\right)\left(g_2=2\right).\end{aligned}$$

Here $\beta_{*,0}, \beta_{0,*}$ quantify the additive effects, $\beta_{*,*}$ measures the interactions between two loci, and $\beta_{0,0}$ defines the intercept, and g_1 and g_2 are the number of copies of the rare allele for the two genes. The three interaction models are obtained by

IM1: $\beta_{0,1} = 2\beta_{0,2}, \beta_{1,0} = 2\beta_{2,0}, \beta_{1,1} = \beta_{1,2} = \beta_{2,1} = \beta_{2,2} = 0$
IM2: $\beta_{0,1} = \beta_{1,0}, = \beta_{0,2} = \beta_{2,0} = 0, \beta_{1,1} = 4\beta_{2,2}, \beta_{1,2} = 2\beta_{2,2}, \beta_{2,1} = \beta_{2,2},$
IM3: $\beta_{0,1} = \beta_{1,0}, = \beta_{0,2} = \beta_{2,0} = 0, \beta_{1,1} = \beta_{2,2}, \beta_{1,2} = \beta_{2,2}, \beta_{2,1} = \beta_{2,2}.$

In our simulations for IM1 we take $\beta_{0,1} = \beta_{1,0}$, and we take $\beta_{0,0} = 0.01$ in all models, so that each model only has one parameter β. Note that $\beta_{0,0} = 0.01$ corresponds to a moderately rare disease. We show results for a sample size of 1000 cases and 1000 controls. We examined smaller sample

sizes, and found the results qualitatively similar. In our simulations, we used SNPs A_3 and B_3 as the casual SNPs. The minor allele frequencies of A_3 and B_3 are 0.2303 and 0.3090, respectively. In our simulations we consider three scenarios.

Case 1: Only A_3 and B_3 are observed. This is a standard scenario investigated in the literature, where the SNPs that are interacting are assumed to be observed.

Case 2: We observe $A_1, \ldots, A_5$ and $B_1, \ldots, B_5$. This is the scenario in which we observe blocks of SNPs, including the SNPs that we generated to be causal. In this scenario we expect some power increase because the additional SNPs are in LD with A_3 and B_3, but some decrease of power because of multiple comparisons.

Case 3: We observe A_1, A_2, A_4, A_5 and B_1, B_2, B_4, B_5. We believe that this is the most interesting scenario, as we do not observe the causal SNP, but observe the interaction through multiple SNPs that are in LD with the casual SNP. Our methods are specifically designed with this situation in mind.

We compare four testing methods: the likelihood ratio statistic (λ_{12}), the quadratic distance-based statistic (δ^2), and statistics arising from two logistic models (LM_1, LM_2) in which all SNPs that are considered are present in the model, coded additively. For LM_1 we consider all pairwise interactions simultaneously, testing them using an F-test, and for LM_2 we consider each of the pairwise interactions separately, selecting the most significant one. For all four methods, significance levels are determined using $10,000$ permutations of case-control status. We ran each simulation scenario $1,000$ times.

The power results for Case 1, when the matrix size is 2×2 and equality of a single off-diagonal covariance pair is tested, are shown in Table 1. Note that for this situation the two logistic regression statistics, LM_1 and LM_2, are identical. For IM1, where there are additive effects, but there is no interaction, we note that all approaches maintain the correct Type 1 error of 5%. For IM2, where there is a multiplicative interaction, and M3, where there is an interaction with threshold (dominant $\times$ dominant) effects, all approaches have approximately the same power.

The power results for Case 2, when the matrix size is 10×10 and we test equality of the two off-diagonal 5×5 sub-matrices, are shown in Table

2. In this table and in Table 3 we omit the results for IM1, where there is no interaction; as in Case 1, all approaches maintain the correct Type 1 error. For this case we note that for both IM2 and IM3, our two proposed test statistics, λ_{12} and δ^2, have considerably more power than both logistic regression statistics, which have approximately the same power. It appears that δ^2 has slightly more power than λ_{12} but the difference is small. Compared to Case 1 we notice that both logistic regression statistics have less power because of the larger multiple comparisons penalty (note that we correct using a permutation approach, and not using a Bonferroni correction, which would have led to even lower power). On the other hand, the power of λ_{12} and δ^2 increases from Case 1 to Case 2, because these statistics exploit the entire block of CLDs between the SNPs.

The power results for Case 3, when the matrix size is 8×8 and equality of the two off-diagonal 4×4 sub-matrices is tested, are shown in Table 3. For this case the causal SNPs are not part of the data that are analyzed. As a result, the logistic regression methods lose almost all the power they had in Case 2. Our proposed statistics λ_{12} and δ^2 also lose power but the loss is smaller, and these statistics still maintain reasonable power, especially for IM2, where the power is not much lower than in Case 1. It apperas that for all cases and all models δ^2 is slightly more powerful than λ_{12}.

An application to genetic data

The *IL10* and *IL10B* receptor genes are involved in immune regulation and suppression. A genetic polymorphism in the promoter region of the *IL10* gene has a significant impact on graft versus host disease (GVHD) after allogeneic hematopoietic stem cell transplantation (HCT) with human leukocyte antigen (HLA) identical sibling donors. In a previous study of SNPs among 953 HLA-identical sibling transplants (18), the presence of the *IL10*/-592*A allele in the patient or the *IL10RB**G allele in the donor was significantly associated with lower risk of severe acute GVHD and non-relapse mortality. It is thought that *IL10* may facilitate immune tolerance after allogeneic transplantation. Higher *IL10* production by ex-vivo stimulation of recipient's cells before transplantation is associated with reduced risk of acute GVHD and non-relapse mortality (19). In this example our goal was to see whether an interaction between *IL10* and *IL10RB* has a synergistic effect on the risk for GVHD. We tested this hypothesis using a dataset with two groups, one of which developed GVHD (case) while the other did not (control), with a

		Parameter β in the model					
		0	0.1	0.5	1	2	4
IM1	λ_{12}	0.048	0.051	0.050	0.051	0.052	0.053
	δ^2	0.047	0.048	0.049	0.050	0.053	0.052
	$LM_1 = LM_2$	0.048	0.049	0.051	0.051	0.050	0.053
IM2	λ_{12}	0.048	0.068	0.104	0.178	0.644	1.000
	δ^2	0.053	0.068	0.106	0.184	0.653	1.000
	$LM_1 = LM_2$	0.049	0.061	0.102	0.176	0.615	1.000
IM3	λ_{12}	0.053	0.056	0.081	0.116	0.551	0.942
	δ^2	0.051	0.057	0.084	0.120	0.559	0.953
	$LM_1 = LM_2$	0.050	0.055	0.080	0.112	0.547	0.935

Table 1: Power of the proposed test statistics for Case 1. Here $p_1 = p_2 = 1$ so we test for equality of a single pair of covariances. IM1, multiplicative within and between loci – no interaction; IM2, multiplicative model; IM3, the threshold model. For this set of simulations, 1000 cases and 1000 controls were sampled for each of 1000 simulation runs. We completed 10000 permutations for each data set, and controlled the significance level at $\alpha = 0.05$.

		Parameter β in the model					
		0	0.1	0.5	1	2	4
IM2	λ_{12}	0.048	0.069	0.125	0.201	0.745	1.000
	δ^2	0.049	0.072	0.134	0.225	0.821	1.000
	LM_1	0.051	0.059	0.089	0.154	0.521	1.000
	LM_2	0.050	0.062	0.095	0.169	0.558	1.000
IM3	λ_{12}	0.051	0.060	0.098	0.180	0.685	1.000
	δ^2	0.051	0.065	0.104	0.195	0.701	1.000
	LM_1	0.050	0.056	0.072	0.104	0.468	0.990
	LM_2	0.050	0.057	0.080	0.119	0.488	1.000

Table 2: Power of the proposed test statistics for Case 2. Here $p_1 = p_2 = 5$ and we test for equality of the two 5×5 blocks Σ_{12} and Ω_{12}. For this set of simulations, 1000 cases and 1000 controls were sampled for each of 1000 simulation runs. We completed 10000 permutations for each data set, and controlled the significance level at $\alpha = 0.05$.

		Parameter β in the model					
		0	0.1	0.5	1	2	4
IM2	λ_{12}	0.049	0.058	0.078	0.133	0.435	1
	δ^2	0.047	0.061	0.089	0.155	0.465	1
	LM_1	0.049	0.054	0.060	0.084	0.226	0.685
	LM_2	0.049	0.055	0.065	0.099	0.242	0.721
IM3	λ_{12}	0.047	0.054	0.063	0.093	0.216	0.561
	δ^2	0.049	0.059	0.068	0.105	0.235	0.611
	LM_1	0.049	0.050	0.054	0.061	0.067	0.128
	LM_2	0.049	0.050	0.055	0.069	0.076	0.141

Table 3: Power of the proposed test statistics for Case 3. Here $p_1 = p_2 = 4$ and the interaction SNPs have been eliminated for the analysis. We test for equality of the two 4×4 blocks Σ_{12} and Ω_{12}. For this set of simulations, 1000 cases and 1000 controls were sampled for each of 1000 simulation runs. We completed 10000 permutations for each data set, and controlled the significance level at $\alpha = 0.05$.

sample size of 159 for each group. This data is part of a study investigating how genetic diversity among patients and donors contributes to differences in individual responses to tissue injury, inflammation, and severity of acute GVHD. The *IL10* gene has three SNPs ($p_1 = 3$), and the *IL10RB* gene has one SNP ($p_2 = 1$) (see Table 4), thus the corresponding covariance matrix is 4×4.

We apply our proposed statistics λ_{12} and δ^2 and the two logistic regression methods, LM_1 and LM_2 for testing whether there is an interaction effect of the *IL10* and *IL10RB* genes on GVHD. Both λ_{12} and δ^2 result in off-diagonal

Gene	Genotype	0	1	2
IL10	rs1800872	CC	AC	AA
	rs1800890	TT	AT	AA
	rs1800896	AA	AG	GG
IL10RB	rs2834167	GG	AG	AA

Table 4: Possible SNP combinations for *IL10* and *IL10RB* genes, by RefSNP (rs) number, shown for the homozygous (0), heterozygous (1) and homozygous variant (2) case.

blocks that are statistically significantly different between cases and controls with $p = 0.0382$ and $p = 0.0374$ respectively. The LM_1 and LM_2 also confirms the interaction with $p = 0.0451$ and $p = 0.0438$ respectively. We can see that the p-values of the proposed test statistics were slightly smaller than those of the logistic regression methods. The small difference between the two statistics is likely due to the fact that the covariance matrix is only 4×4 with just a 1×3 off diagonal matrix.Thus, all methods can successfully detect an interaction effect of *IL10* and *IL10RB* genes on GVHD.

Discussion

Classical methods for identifying disease-susceptibility genes focus on one genomic area or locus at a time. They have worked well for Mendelian disorders but appear insufficient for complex traits because of the presumed multiplicity of genes involved. To facilitate the search for sets of SNPs jointly associated with a disease phenotype, we have developed two new statistics for testing for interaction effects between two blocks of SNPs—two genes—based on defining a distance between sample covariance matrices. We use a multivariate extension of CLD and compute covariance matrices for SNPs separately for cases and controls. A test for equality of the off-diagonal block corresponding to the covariance between the two genes of the two matrices becomes a test of an interaction effect between the two genes on case-control status. Our proposed methods abrogate the need for a multiple comparisons correction as we have a single test for interaction. This offers greater power than the traditional method of individual pairwise testing of SNPs and using a multiple comparisons correction.

Simulation results reveal that our methods perform better than traditional logistic regression-based methods. For the matrix size two 2×2, where the SNPs that are interacting are observed, the power results for the proposed statistics λ_{12} and δ^2 and logistic regression behave approximately equally. When we consider multiple SNPs in a gene, and assume that the true causal interacting SNPs are among them, the power is higher for our statistics λ_{12} and δ^2 than for logistic regression (Table 2). The scenario in Table 3 is the most interesting one, as we eliminate the interaction SNPs for the analysis. Again, here we see that power is much larger for λ_{12} and δ^2 than logistic regression. As in this case we do not observe the causal SNP, but rather we observe the interaction through multiple SNPs that are in LD.

We can easily apply our proposed methods to case only design to explore interactions between two loci, where there is gene-gene independence in the controls (in a population with a rare disease), as we would simply set the off-diagonal sub-matrix for the controls equal to zero. Initial simulations suggest this significantly improves power. We are currently working on an extension of our methods that will allow us to test whether many genes—a network of SNPs—associate with a phenotype by comparing two complete covariance matrices, as in H_0, see (3).

To evaluate performance for detection of interactions between two loci, the proposed λ_{12} and δ^2 statistics were applied to data from hematopoietic cell transplantation (HCT) patients and donors. In this example we wished to distinguish between groups of patients, for example those who developed GVHD and those who did not. Genetic polymorphisms in the promoter region of the *IL10* gene and in a coding region of the *IL10RB* gene at position c238 have been shown to significantly affect risk of GVHD after HCT with an HLA-identical sibling donor [18] , [20]. The *IL10* promoter region regulates production of IL-10, and the *IL102RB c238* SNP has been shown to regulate transcription and cell surface expression of the IL-10 receptor β chain [21]. There is a direct functional relationship between the *IL10RB* gene located on chromosome 1 and the gene encoding its ligand *IL10* located on chromosome 21, and it is therefore highly plausible that there is also a genetic interactions between these 2 genes even though they are on different chromosomes. Our study population, consisting of paired patients and donors, provided a unique opportunity to assess genome-genome interaction between recipient and donor genomes [22] the HCT setting. Using our methods, we confirmed a statistical interaction between these two unlinked loci, a beautiful example of two different chromosomes showing a statistical interaction that aligns with a known biological interaction. This is encouraging evidence that detection of statistical interactions may lead to discovery of novel biological interactions.

While computing test statistics for many blocks of SNPs is computationally intensive, it is reasonably achievable by spreading computations over clusters of computers. In practice we would test for interactions between a limited number of blocks of interest, either because there is biological interest (as was the case for our *IL10* example), or because these blocks suggest the strongest marginal effects (using a similar approach as Kooperberg and LeBlanc [23]). Each of these limited numbers of blocks could then be compared with the complete genome in a sliding window fashion. A computationally intense approach would be to carry out permutation tests separately for

each possible interaction. Rather than separate permutation tests, we would first "rank" all tests, and only carry out the tests for interactions with the largest statistics, for example using the Holm step down procedure [24]. The Box approximation for normally distributed data can be applied to obtain the asymptotic null distribution [12]. This will be applicable when we extend our methods for GWAS. We can obtain the first 50 significant interactions from the parametric form and then apply the permutation-based technique to confirm the interactions.

Novel genomic tools and computational methods have led to a dramatic increase in the rate of discovery of disease genes. While traditional association studies have sought single marker or single gene associations, phenotypes are the result of complex interactions among large numbers of genes. Extensions of the statistical methods we have proposed will allow the investigation of relationships among groups of SNPs in many genes and can discriminate between the genetic signatures of distinct groups of subjects. By identifying interactions among networks of genes, we may further our understanding of how the collective behavior of genes gives rise to phenotypes as well as our ability to predict disease outcome. Detecting interactions among disease associated SNPs may reveal basic biological mechanisms that are critical to understanding development and progression of a disease state [25], and in this way provide a powerful and promising foundation for the development of novel diagnostics and therapeutic strategies.

Acknowledgements

We thank Lindsey Muir for discussion and critical reading of the manuscript. IR is supported by Interdisciplinary Training Grant in Cancer Research grant T32 CA80416 from National Institutes of Health (NIH) and Mentored Quantitative Research Career Development Award (K25) from NIH grant K25DK08279, JH by NIH grants RO1HL087690 and RO1HL105914, and CK by NIH grants R01 CA 90998 and P01 CA53996.

References

[1] Cordell, H.J. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 10:392-404.

[2] Chatterjee, N., Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder. 2006. Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. The American Journal of Human Genetics. 79:1002-1016.

[3] Zhao, J., L. Jin, and M. Xiong. 2006. Test for Interaction between Two Unlinked Loci. The American Journal of Human Genetics. 79:831-845.

[4] Chen, X., C.-T. Liu, M. Zhang, and H. Zhang. 2007. A forest-based approach to identifying gene and gene–gene interactions Proceedings of the National Academy of Sciences. 104:19199-19203.

[5] Millstein, J., D.V. Conti, F.D. Gilliland, and W.J. Gauderman. 2006. A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. The American Journal of Human Genetics. 78:15-27.

[6] Lewontin, R.C. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 49:49-67.

[7] Weir, B.S. 1996. Genetic data analysis II. Sunderland, MA: Sinauer Associates.

[8] Weir, B.S., and C.C. Cockerham. 1989. Complete characterization of disequilibrium at two loci. In Mathematical Evolutionary Theory. M. W.Feldman, editor. Princeton University Press. 86-110.

[9] Nielsen, D.M., M.G. Ehm, D.V. Zaykin, and B.S. Weir. 2004. Effect of Two- and Three-Locus Linkage Disequilibrium on the Power to Detect Marker/Phenotype Associations. Genetics. 168:1029-1040.

[10] Schaid, D.J. 2004. Linkage Disequilibrium Testing When Linkage Phase Is Unknown. Genetics. 166:505-512.

[11] Weir, B.S., W.G. Hill, and L.R. Cardon. 2004. Allelic association patterns for a dense SNP map. Genetic Epidemiology. 27:442-450.

[12] Anderson, T.W. 2003. An Introduction to Multivariate Statistical Analysis. Wiley, New York

[13] Chaudhuri, S., and M.D. Perlman. 2006. Two step-down tests for equality of covariance matrices. Linear Algebra and its Applications. 417:42-63.

[14] Nagao, H. 1973. On Some Test Criteria for Covariance Matrix. The Annals of Statistics. 1:700-709.

[15] The International HapMap, C. 2005. A haplotype map of the human genome. Nature. 437:1299-1320.

[16] Stephens, M., N.J. Smith, and P. Donnelly. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. The American Journal of Human Genetics. 68:978-989.

[17] Marchini, J., P. Donnelly, and L.R. Cardon. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 37:413-417.

[18] Lin, M.-T., B. Storer, P.J. Martin, L.-H. Tseng, B. Grogan, P.-J. Chen, L.P. Zhao, and J.A. Hansen. 2005. Genetic variation in the IL-10 pathway modulates severity of acute graft-versus-host disease following hematopoietic cell transplantation: synergism between IL-10 genotype of patient and IL-10 receptor {beta} genotype of donor. Blood. 106:3995-4001.

[19] Holler, E., M.G. Roncarolo, R. Hintermeier-Knabe, G. Eissner, B. Ertl, U. Schulz, H. Knabe, H.J. Kolb, R. Andreesen, and W. Wilmanns. 2000. Prognostic significance of increased IL-10 production in patients prior to allogeneic bone marrow transplantation. Bone marrow transplantation. 25:237-241.

[20] Lin, M.-T., B. Storer, P.J. Martin, L.-H. Tseng, T. Gooley, P.-J. Chen, and J.A. Hansen. 2003. Relation of an Interleukin-10 Promoter Polymorphism to Graft-versus-Host Disease and Survival after Hematopoietic-Cell Transplantation. N Engl J Med. 349:2201-2210.

[21] Frodsham, A., L. Zhang, U. Dumpis, N. Taib, S. Best, A. Durham, B. Hennig, S. Hellier, S. Knapp, M. Wright, M. Chiaramonte, J. Bell, M. Graves, H. Whittle, H. Thomas, M. Thursz, and A. Hill. 2006. Class II cytokine receptor gene cluster is a major locus for hepatitis B persistence. Proceedings of the National Academy of Sciences. 103:9148-9153.

[22] Spilianakis, C.G., M.D. Lalioti, T. Town, G.R. Lee, and R.A. Flavell. 2005. Interchromosomal associations between alternatively expressed loci. Nature. 435:637-645.

[23] Kooperberg, C., and M. LeBlanc. 2008. Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genetic Epidemiology. 32:255-263.

[24] Drton, M., and M. Perlman. 2007. T Multiple testing and error control in Gaussian graphical model selection. J Statist. Sci. 22:430-449.

[25] Hartwell, L., L. Hood, M. Goldberg, N. Reynolds, L. Silver, and R. Veres. 2006. Genetics: from genes to genomes. Columbus, McGraw-Hill.

Appendix 1

Haplotype					Frequency	
					Block 1	Block 2
0	0	0	0	1	0.0544	0
0	0	0	1	1	0.0163	0
0	0	1	0	1	0.0239	0
0	0	1	1	0	0.0258	0.0151
0	0	1	1	1	0.1645	0.0288
0	1	0	0	0	0	0.0123
0	1	0	0	1	0.0066	0
0	1	0	1	0	0	0.0082
0	1	0	1	1	0	0.036
0	1	1	0	0	0	0.0191
0	1	1	1	0	0.0118	0
0	1	1	1	1	0	0.0379
1	0	0	1	1	0.0413	0
1	0	1	0	0	0	0.0679
1	0	1	0	1	0	0.0315
1	0	1	1	0	0.0061	0
1	0	1	1	1	0.0328	0.0645
1	1	0	0	1	0.0589	0.0713
1	1	0	1	0	0.0252	0.1074
1	1	0	1	1	0.0276	0.0737
1	1	1	0	0	0	0.0302
1	1	1	0	1	0.2832	0.1104
1	1	1	1	0	0.0379	0.0994
1	1	1	1	1	0.1837	0.1862