ARTICLE

Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions

Nilanjan Chatterjee, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder

In modern genetic epidemiology studies, the association between the disease and a genomic region, such as a candidate gene, is often investigated using multiple SNPs. We propose a multilocus test of genetic association that can account for genetic effects that might be modified by variants in other genes or by environmental factors. We consider use of the venerable and parsimonious Tukey's 1–degree-of-freedom model of interaction, which is natural when individual SNPs within a gene are associated with disease through a common biological mechanism; in contrast, many standard regression models are designed as if each SNP has unique functional significance. On the basis of Tukey's model, we propose a novel but computationally simple generalized test of association that can simultaneously capture both the main effects of the variants within a genomic region and their interactions with the variants in another region or with an environmental exposure. We compared performance of our method with that of two standard tests of association, one ignoring gene-gene/gene-environment interactions and the other based on a saturated model of interactions. We demonstrate major power advantages of our method both in analysis of data from a case-control study of the association between colorectal adenoma and DNA variants in the *NAT2* genomic region, which are well known to be related to a common biological phenotype, and under different models of gene-gene interactions with use of simulated data.

The identification of a large number of SNPs across the human genome has created great opportunity for fine mapping disease susceptibility loci (DSL) through population-based association studies.[1–5] An increasingly popular design of association studies has been the indirect approach, in which the association between the disease and a genomic region, such as a candidate gene, is studied using a set of marker SNPs that themselves may or may not have causal effects but would be likely to be in linkage disequilibrium (LD) with the underlying causal variants, if any exist. The availability of LD information across the human genome from the International HapMap project[6,7] and a number of other emerging databases[8,9] is now enabling researchers to select informative sets of tagging SNPs that could be used as markers in indirect association studies.[10–13]

A central statistical issue for indirect association studies is how to optimally analyze the association of a disease phenotype with multiple tightly linked SNPs within a genomic region. A locus-by-locus approach could be optimal if one of the genotyped SNPs itself is causal. In contrast, multilocus tests that assess the association of a disease with multiple marker SNPs simultaneously could be superior when several SNPs may be associated with the disease because of either their direct causal effects or their LD with the underlying causal variant(s) in the region. Two classes of multivariate tests, one based on multilocus genotype data[12,14] and the other based on reconstructed haplotype information,[15,16] are now popularly used in practice.

Another important issue for identification of DSL in complex diseases is that the etiologic effects of the underlying causal variants are likely to be complex because of a number of factors, including but not limited to gene-gene and gene-environment interactions. It has been long recognized that failing to account for these sources of heterogeneity could dramatically reduce the power of detecting DSLs in both linkage and association studies. Since the late 1980s, a variety of multipoint methods have been developed to account for gene-gene interaction in linkage analysis.[17–21] Methods for linkage scans accounting for gene-environment interactions have also received some attention.[22,23] More recently, a number of powerful methods also have been developed for incorporating gene-gene interactions in association studies.[24,25] These methods, however, are mostly suitable for direct association studies involving candidate SNPs and cannot exploit the structure of indirect association studies involving groups of tightly linked SNPs that could be statistically correlated because of LD or functionally related because of underlying common biological mechanisms.

In this article, we propose a novel method for incorporating gene-gene and gene-environment interactions into association studies. When several SNPs are involved within a gene, the number of parameters required in standard statistical models of gene-gene and gene-environment interactions could easily become very large, potentially causing loss of power due to either the use of increased dfs or the need for multiple-testing adjustments. We consider use of Tukey's 1-df model of interaction.[26,27] We show that this parsimonious form of interaction can be motivated

From the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD (N.C.; R.M.; S.W.); Information Management System, Rockville, MD (Z.K.); and Fred Hutchinson Cancer Research Center, Seattle (U.P.)

Received July 17, 2006; accepted for publication September 22, 2006; electronically published October 20, 2006.

Address for correspondence and reprints: Dr. Nilanjan Chatterjee, 6120 Executive Boulevard, EPS 8038, Rockville, MD 20852. E-mail: chattern@mail.nih.gov

Am. J. Hum. Genet. 2006;79:1002–1016. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7906-0003$15.00

Table 1. Haplotype Frequencies Used for Simulating Genotype Data on Marker SNPs for Two Candidate Genes

Haplotype	Frequency
G_1:	
000000	.3211
001101	.1204
010000	.0909
000001	.0785
111001	.0722
110001	.0708
000010	.0610
011001	.0523
110000	.0468
100000	.0353
001000	.0279
010001	.0228
G_2:	
100010	.3506
010001	.2819
010100	.1274
100000	.0678
000000	.0407
101100	.0401
000010	.0307
010010	.0237
010000	.0226
100001	.0144

through a conceptual framework in which the observed SNPs within a gene affect the risk of the disease through an underlying common causal mechanism. Modern association studies in which tagging SNPs are selected as potential surrogates for underlying causal variants fit into this framework. Other examples where the framework is very natural are also discussed.

We propose a novel multilocus test of genetic association, based on Tukey's model, that can efficiently exploit the LD pattern among SNPs within a gene and can simultaneously account for their interactions with SNPs in another gene or with an environmental exposure. We simulate case-control data, in a way that mimics modern association study designs, to evaluate type I errors and powers of the proposed testing strategy. We also apply the proposed methodology to a case-control study designed to investigate the association between colorectal adenoma and DNA variants in *N*-acetyltransferase 2 (*NAT2* [MIM 243400]), a candidate gene that plays an important role in detoxification of aromatic amine carcinogens present in cigarette smoke. Both the simulated and real data examples demonstrate major power advantages for the proposed methodology over two alternative tests of association, one ignoring interactions and the other incorporating a saturated model of interactions.

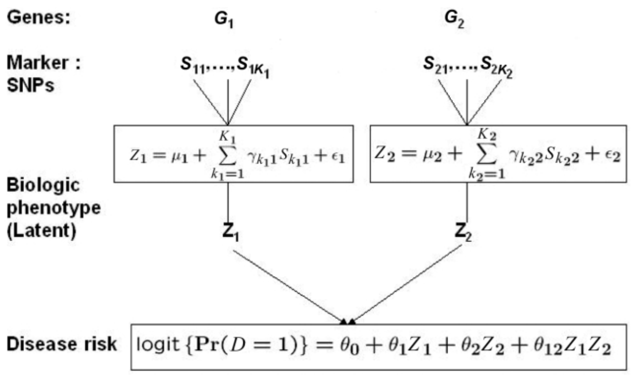
Genes:
G_1
G_2
Marker : SNPs
$S_{11},\ldots,S_{1K_1}$
$S_{21},\ldots,S_{2K_2}$
$Z_1 = \mu_1 + \sum_{k_1=1}^{K_1} \gamma_{k_1 1} S_{k_1 1} + \epsilon_1$
$Z_2 = \mu_2 + \sum_{k_2=1}^{K_2} \gamma_{k_2 2} S_{k_2 2} + \epsilon_2$
Biologic phenotype (Latent)
Z_1
Z_2
Disease risk
$\text{logit}\,\{\Pr(D=1)\} = \theta_0 + \theta_1 Z_1 + \theta_2 Z_2 + \theta_{12} Z_1 Z_2$

Figure 1. A conceptual framework for modeling gene-gene interactions in indirect-association studies.

Material and Methods

A Latent-Variable Model and Tukey's 1-df Form of Interaction

Suppose that G_1 and G_2 are two candidate genes of interest for which K_1 and K_2 marker SNPs, respectively, have been genotyped. Let $\mathbf{S}_1 = (S_{11},S_{21},\ldots,S_{K_1 1})$ and $\mathbf{S}_2 = (S_{12},S_{22},\ldots,S_{K_2 2})$ denote the genotype data for the corresponding sets of markers. In this article, we assume that each marker genotype S_{ij} is recorded as "0," "1," or "2," counting the number of copies of the minor or variant allele. Figure 1 shows a schematic diagram for a hypothesized model describing the relationship between the marker SNPs and the disease through an underlying causal mechanism. The model assumes that, for each gene G_i, the marker data $\mathbf{S}_i$ act as a surrogate for an underlying biological phenotype, Z_i, that is causally related to the disease. The associations between the markers and the biological phenotypes for the two genes are described by two separate linear models (fig. 1, upper two boxes), where the error terms ϵ_1 and ϵ_2 are assumed to be mean zero independent random variables. The risk of the disease, given the causal variables Z_1

Table 2. Approximate Relative-Risk Models Used for Simulating Disease End Points, Given the Genotypes for Two Causal Loci in Candidate Genes G_1 and G_2

	No. of Alleles			
Model	($S_1^* = 0$, $S_2^* = 0$)[a]	($S_1 \geqslant 1$, $S_2 = 0$)	($S_1 = 0$, $S_2 \geqslant 1$)	($S_1 \geqslant 1$, $S_2 \geqslant 1$)
General form	1	$\exp(\theta_1)$	$\exp(\theta_2)$	$\exp(\theta_1 + \theta_2 + \theta_{12})$
Purely epistatic	1	1	1	ϕ
Multiplicative	1	ϕ	ϕ	ϕ^2
Additive	1	ϕ_1	ϕ_2	$\phi_1 + \phi_2 - 1$
Crossover	1	ϕ_1 (<1)	1	ϕ_{12} (>1)

[a] S_1^* and S_2^* refer to the number of copies of the variant allele in the causal loci of G_1 and G_2, respectively.

Table 3. Empirical Significance Level for Test of Association with Region G_1

R^2_{geno},[a] f_2,[b] and Method	Relative Risk for Causal SNP in G_2 (θ_2)	
	1.0	2.0
90%:		
.04:		
Permutation	.008	.012
Asymptotic	.008	.011
.13:		
Permutation	.013	.011
Asymptotic	.012	.009
75%:		
.04:		
Permutation	.010	.009
Asymptotic	.009	.008
.13:		
Permutation	.009	.004
Asymptotic	.009	.004
60%:		
.04:		
Permutation	.011	.012
Asymptotic	.012	.012
.13:		
Permutation	.009	.009
Asymptotic	.009	.008

[a] Multiple R^2 between genotypes and causal and marker loci.

[b] Allele frequency for causal SNP in G_2.

and Z_2, is specified by a standard logistic model that involves both main and interaction effects (fig. 1, lower box). It is also implicitly assumed that, given the true biological exposures Z_1 and Z_2, the risk of the disease does not depend on the markers $\mathbf{S}_1$ and $\mathbf{S}_2$.

Before one proceeds further, it is useful to understand what the latent variables Z_1 and Z_2 may be in practice. If the gene G_i contains a single causal locus L_i, the variable Z_i could represent the genotype data for L_i itself. If, for example, one of the selected markers is the causal locus and Z_i denotes the count for the corresponding variant allele, then the assumed linear model describing the relationship between Z_i and $\mathbf{S}_i$ would fit perfectly—that is, the error term ϵ_i would vanish, by the setting of $\gamma_{ik} = 1$ for the causal locus and $\gamma_{ik} = 0$ for all the other markers. If the causal locus is not selected as a marker, then the error term will not generally disappear, but the magnitude of it could be expected to be small for modern association studies that aim to select the markers to be a panel of tagging SNPs that would have a very high degree of LD, as measured by the R^2 criterion, with all the genetic variations of the regions, including any possible causal ones. The validity of the proposed framework, however, does not depend on the existence of a single causal locus in each gene. The variable Z_i could, for example, represent a quantitative biological phenotype that may be governed by several different variants within the same gene G_i. In the study of colorectal adenoma (see the "Results" section), the underlying biological phenotype for the gene of interest, *NAT2,* is the *N*-acetyltransferase enzymatic activity level, which has been shown to be determined by several single–base-pair substitutions in the gene and the associated haplotypes/diplotypes.[28,29]

The logistic model shown in figure 1 (lower box) cannot be used directly for association testing because, typically, the variables Z_1 and Z_2 are not observable. However, in this model, expressing Z_1 and Z_2 in terms of $\mathbf{S}_1$ and $\mathbf{S}_2$ with use of the corresponding linear-regression models and assuming small variances for the error terms ϵ_1 and ϵ_2, a risk model for the disease (D), in terms of the observable SNPs, can be derived approximately in the formula

$$\operatorname{logit}[\Pr(D = 1|\mathbf{S}_1,\mathbf{S}_2)] = \alpha + \sum_{k_1=1}^{K_1} \beta_{k_1 1} S_{k_1 1} + \sum_{k_2=1}^{K_2} \beta_{k_2 1} S_{k_2 2} + \theta \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 1}\beta_{k_2 1} S_{k_1 1} S_{k_2 2} \tag{1}$$

(see appendix A for details). We observe that equation (1) resembles a traditional logistic-regression model, except that the SNPs across two genes have the parsimonious Tukey's 1-df form of interaction.[26,27] Thus, postulating the biological effect of the observed SNPs to be determined by a smaller set of casual variables leads to a very parsimonious model for gene-gene interactions.

The motivation of Tukey's 1-df model for interaction through the above latent-variable framework also allows extension of the model in a number of different ways. For example, if some of the SNPs within a gene are known a priori to have functional significance, then it may be desirable to capture possible interactions between these functional SNPs of the same gene. Suppose S_{11} and S_{21} are two such SNPs for gene G_1. Then, the regression model for Z_1 could be extended to allow for interaction between S_{11} and S_{21}, as

$$Z_1 = \mu_1 + \sum_{k=1}^{K_1} \gamma_{k_1 1} S_{k_1 1} + \gamma_{(12)1} S_{11} S_{21} + \epsilon_1 \ . \tag{2}$$

With the assumption that the models for Z_2 and $\Pr(D = 1|Z_1,Z_2)$ remain the same as before, the model for the risk of the disease, in terms of the SNP data $\mathbf{S}_1$ and $\mathbf{S}_2$, can now be derived in the formula

$$\begin{aligned}\operatorname{logit}[\Pr(D = 1|\mathbf{S}_1,\mathbf{S}_2)] = {} & \alpha + \sum_{k_1=1}^{K_1} \beta_{k_1 1} S_{k_1 1} \\ & + \sum_{k_2=1}^{K_2} \beta_{k_2 2} S_{k_2 2} + \beta_{(12)1} S_{11} S_{21} \\ & + \theta \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 1}\beta_{k_2 2} S_{k_1 1} S_{k_2 2} \\ & + \tau \sum_{k_2=1}^{K_2} \beta_{(12)1}\beta_{k_2 2} S_{11} S_{21} S_{k_2 2} \ ,\end{aligned}$$

which includes both second- and third-order interactions. One could also account for SNP-SNP interactions within a gene by specifying the disease risk in terms of haplotypes instead of locus-specific genotypes.

The proposed modeling framework can be easily extended to incorporate gene-environment interactions. Suppose that the genomic region G_1 (e.g., *NAT2*) is believed to involve a biological pathway through which an environmental variable X (e.g., smoking) may act on the risk of a disease (e.g., colorectal adenoma). Again, on the basis of the latent-variable approach, a model for

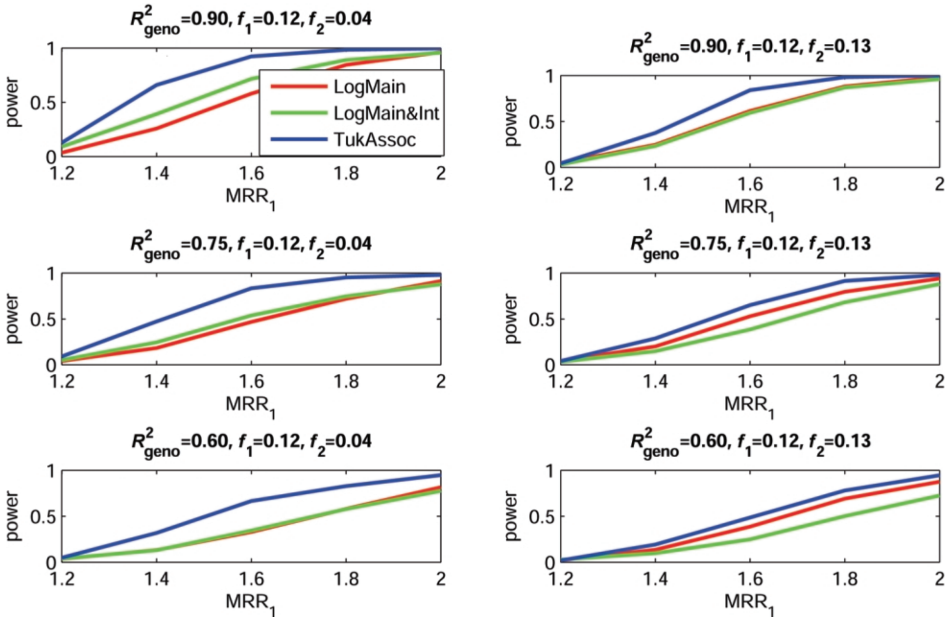

Figure 2. Empirical power, at $\alpha = 0.01$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the purely epistatic model (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

the disease risk in terms of the marker-SNPs $\mathbf{S}_1$ and the environmental variable X can be derived in the form

$$\operatorname{logit}[\Pr(D=1\,|\,\mathbf{S}_1,X)] = \alpha + \sum_{k_1=1}^{K_1}\beta_{k_11}S_{k_11} + \sum_{p=1}^{P}\gamma_pX_p + \theta\sum_{p=1}^{P}\sum_{k_1=1}^{K_1}\beta_{k_11}\gamma_pS_{k_11}X_p\ ,$$

where $\{X_1,X_2,\ldots X_P\}$ is a set of suitably chosen design variables, such as dummy variables for categorical exposures, for representing the effects of the exposure X.

Association Testing

In this section, we study methods for hypothesis testing based on the proposed model. When data on multiple putative risk factors, such as multiple candidate genes, are available, one could test a number of different types of hypotheses regarding the role of these factors in the risk of the disease. For association studies, the primary goal is to establish which of the factors, if any, are related to the risk of the disease. If multiple factors are found to be related to the disease, then a secondary hypothesis of interest could be generated to test for specific forms of interaction among the established risk factors. It is important, however, to realize that, although the test of interaction itself may be of only secondary interest, accounting for heterogeneity of genetic effects due to interactions can be vital for enhancing the power of the primary hypothesis of association testing.

In the following sections, we develop an association-testing framework involving two candidate genes, G_1 and G_2. The same framework can also be used to develop tests of associations involving a candidate gene and an environmental exposure. We assume a population-based case-control design of unrelated subjects. All of the methods, however, are easily extendable to alternative study designs, including family-based case-control and case-parent–trio designs. For possible strategies for using the methodology in general association studies that involve numerous candidate genes, see the "Discussion" section.

The general principle.—We focus on the test of association for G_1; the methods for G_2 are symmetric. In model (1), the null hypothesis of no association of disease with G_1 can be statistically stated as

$$H_0^{(1)}{:}\beta_{k_11} = 0,\quad \text{for all}\quad k_1 = 1,\ldots K_1\ ,$$

which implies conditional independence of D and G_1, given G_2. The parameter β_{k_11} not only appears in the model as the main effect for the marker S_{k_11} but also contributes to all K_2 interaction terms that could be defined involving S_{k_11} and the K_2 SNPs in G_2. Thus, it is best to describe β_{k_11}, $k_1 = 1,\ldots,K_1$ as a set of "generalized association parameters" instead of as traditional "main" or "interaction" effects.

A complication of association testing in model (1) is that, under the null hypothesis of $H_0^{(1)}$, the parameter θ disappears from the model and, hence, is not estimable from the data. Thus, standard statistical tests, such as score- or likelihood-ratio tests, which require estimation of all nuisance parameters of the model under the null hypothesis, are not applicable. However, for each fixed value of θ, irrespective of whether or not it is the true value for the population, model (1) gives a valid way of testing the null hypothesis $H_0^{(1)}$. In particular, for each fixed value of θ, the likelihood score function for the parameter vector $\beta_1 = (\beta_{11},\ldots,\beta_{K_11})$ can be shown to have zero expectation under the null hypothesis

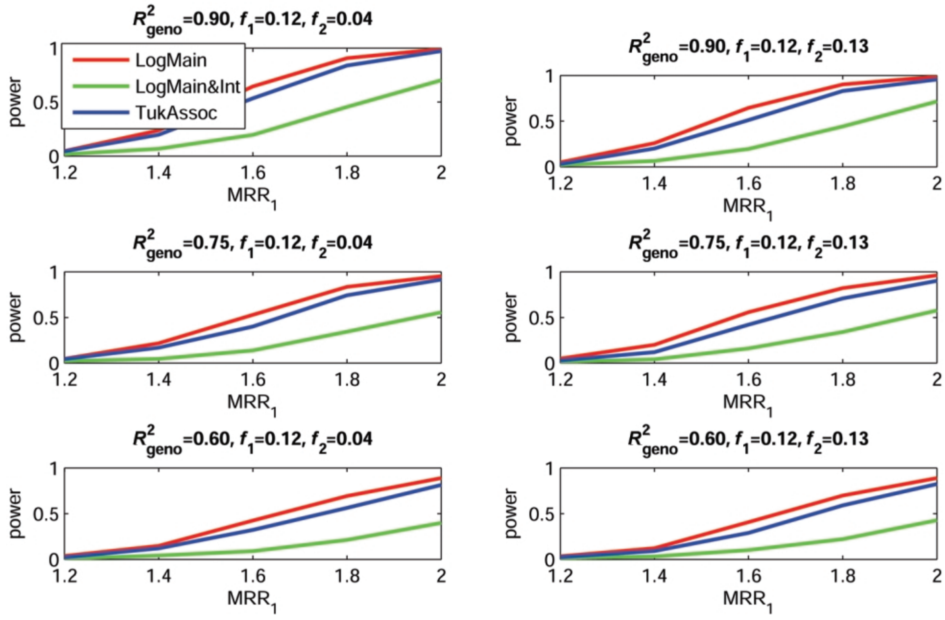

Figure 3. Empirical power, at $\alpha = 0.01$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the purely multiplicative model, with $\phi_1 = \phi_2$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

of $H_0^{(1)}$. Thus, for each fixed value of θ, an unbiased score statistic could be formed for testing $H_0^{(1)}$. Varying the value of θ, one can get a family of score statistics. We propose to use the maximum value of such score statistics over a suitable range of θ as the final test statistics to be used.

Steps for deriving the test statistics.—We assume that N_1 cases and N_0 controls have been sampled in the study and that, for each subject, i, the SNP-genotype vectors $\mathbf{S}_{1i}$ and $\mathbf{S}_{2i}$ have been recorded. In the following list, we describe the four major steps for deriving the test statistics associated with G_1. The test statistics for G_2 could be derived by symmetry.

1. Obtain maximum-likelihood estimate α and $\beta_2 = (\beta_{12}, \dots, \beta_{K_2 2})$ under the local null hypothesis $H_0^{(1)}$. Under $H_0^{(1)}$, the model (1) becomes equivalent to a standard logistic-regression model involving the main effects of the SNPs in G_2. Thus, a standard logistic software package can be used to obtain $\hat{\psi} = (\hat{\alpha},\hat{\beta}_2)$. Let $\hat{P}_{H_0^{(1)}}(\mathbf{S}_2)$ denote $\Pr(D = 1|\mathbf{S}_1,\mathbf{S}_2) = \Pr(D = 1|\mathbf{S}_2)$ evaluated at $\beta_1 = 0$ and $\psi = \hat{\psi}$.
2. For a fixed value of θ, evaluate the score functions for the parameters $\beta_{k_1 1}$ and $k_1 = 1, \dots ,K_1$ at $\beta_1 = 0$ and $\psi = \hat{\psi}$, using the formula

$$S_{\beta_{k_1 1}}(\theta) = \sum_{i=1}^{N_0+N_1} \left(1 + \theta \sum_{k_2=1}^{K_2} S_{k_2 2i}\hat{\beta}_{k_2 2}\right) S_{k_1 1i}[D_i - \hat{P}_{H_0^{(1)}}(\mathbf{S}_{2i})] \ , \qquad (3)$$

which, in a vectorized form, can be written as

$$S_{\beta_1}(\theta) = \sum_{i=1}^{N_0+N_1} (1 + \theta\mathbf{S}_{2i}^T\hat{\beta}_2)\mathbf{S}_{1i}[D_i - \hat{P}_{H_0^{(1)}}(\mathbf{S}_{2i})] \ .$$

Interestingly, the score functions (eq. [3]) resemble those obtained from a standard logistic-regression model, except that the design vector $\mathbf{S}_{1i}$ has been replaced by $(1 + \theta\mathbf{S}_{2i}^T\hat{\beta}_2)\mathbf{S}_{1i}$, a quantity incorporating design variables for both the main and the interaction effects of $\mathbf{S}_1$.
3. Estimate the inverse of the variance-covariance matrix for $S_{\beta_1}(\theta)$, using the formula

$$I^{\beta_1\beta_1}(\theta) = [I_{\beta_1\beta_1}(\theta) - I_{\beta_1\psi}(\theta)I_{\psi\psi}^{-1}I_{\psi\beta_1}(\theta)]^{-1} \ , \qquad (4)$$

where the expressions for the component information matrices—$I_{\beta_1\beta_1}(\theta) = \partial L/\partial\beta_1\partial\beta_1^T, I_{\beta_1\psi}(\theta) = \partial L/\partial\beta_1\partial\psi^T$ and $I_{\psi\psi} = \partial L/\partial\psi\partial\psi^T$—evaluated at $\beta_1 = 0$ and $\psi = \hat{\psi}$ are given in the formulae (A2), (A3), and (A4) (in appendix A). All these quantities can be conveniently computed using standard logistic-regression software, by simply setting the design vector for each subject to be $X = [1,\mathbf{S}_{2i},(1 + \theta\mathbf{S}_{2i}^T\hat{\beta}_2)\mathbf{S}_{1i}]$.
4. For fixed value of θ, obtain the score statistics

$$T_1(\theta) = S_{\beta_1}(\theta)^T I^{\beta_1\beta_1}(\theta) S_{\beta_1}(\theta) \ .$$

Compute the final test statistics as $T_1^* = \max_{L \leqslant \theta \leqslant U} T(\theta)$, where L and U denote some prespecified values for lower and upper limits of θ, respectively.

Simulating the null distribution of the test statistics.—In appendix A, we show an asymptotic equivalent representation of the score statistics $T_1(\theta)$ as $U^T(\theta)V^{-1}(\theta)U(\theta)$, where

$$U(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_i(\theta)$$

denotes the efficient score function for β_1 for fixed θ (see formula [A5]) and $V(\theta)$ is the limit of $1/N\sum_{i=1}^{N} U_i(\theta)U_i^T(\theta)$. Further, under

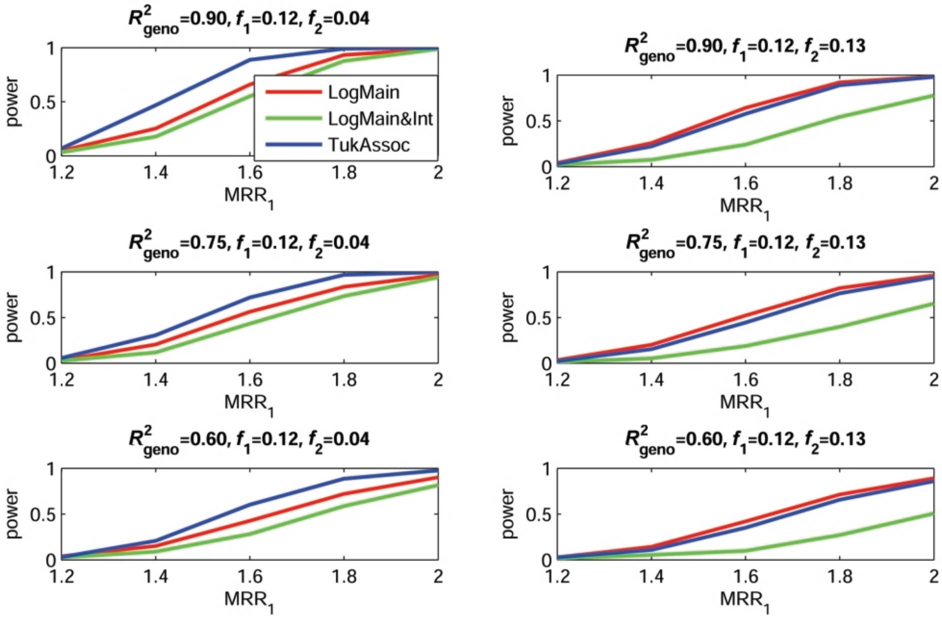

Figure 4. Empirical power, at $\alpha = 0.01$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the additive model, with ϕ_2 chosen so that MRR$_2$ = 2.0 when $f_2 = 0.12$ and MRR$_2$ = 5.0 when $f_2 = 0.04$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

$\beta_1 = 0$, we show that $U(\theta)$, as a stochastic process in θ, converges to a K_1-variate Gaussian process, $\mathcal{Z}(\theta)$, with mean zero and variance-covariance function

$$V(\theta_1,\theta_2) = \lim_{N\to\infty} 1/N \sum_{i=1}^{N} U_i(\theta_1)U_i^T(\theta_2) \ .$$

We propose to generate realization of the process $\mathcal{Z}(\theta)$ as

$$U_0(\theta) = \sum_{i=1}^{N} U_i(\theta)W_i \ ,$$

where W_i and $i = 1,\dots,N$ are independent standard normal random variables that are also independent of the data.[30] The null distribution of the test statistics T_1 is then simulated by repeatedly generating data as $T_1^0 = \max_{L\leqslant\theta\leqslant U} U_0^T(\theta)I^{\beta_1,\beta_1}(\theta)U_0^T(\theta)$, where, in each replication, a new realization of $U_0^T(\theta)$ is obtained by regenerating the random numbers $(W_1,\dots,W_N)$.

We also considered simulating the null distribution of T_1^*, using a permutation-based resampling method. We randomly permuted the value of the vector $\mathbf{S}_{1i}$ over different subjects $i = 1,\dots,N_0 + N_1$ while holding D_i and $\mathbf{S}_{2i}$ to be fixed at their observed values. This yields a valid way of generating null data under the assumption that $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent in the underlying population, because, in this case, the null hypothesis of $\beta_1 = 0$ corresponds to independence of $\mathbf{S}_{1i}$ and $(D_i,\mathbf{S}_{2i})$. By permuting all the components of $\mathbf{S}_{1i}$ simultaneously and keeping $(D_i,\mathbf{S}_{2i})$ fixed, the procedure allows within-gene LD patterns and marginal association structures of D_i and G_2 to be the same as the original data.

Design for Simulation Study

We studied performance of the proposed test of association, using simulated case-control studies. We assumed that the true risk model involves two potentially interacting causal SNPs, S_1^* and S_2^*, residing on two separate candidate genes, G_1 and G_2, respectively. For each gene, we assumed that genotype data are available on six marker SNPs, none of which is the causal SNP. To simulate a realistic LD pattern among the markers, we used real haplotype data on glutathione peroxidase 3 (*GPX3* [MIM 138321]) and glutathione peroxidase 4 (*GPX4* [MIM 138322]), two candidate genes for prostate cancer that have been resequenced using a sample of 29 white subjects at the Core Genotyping Facility of the National Cancer Institute (NCI). In our simulation, we chose the marker SNPs for G_1 and G_2 to correspond to two sets of six tagging SNPs that have been respectively selected for *GPX3* and *GPX4* with use of the original resequencing data. Table 1 shows the distribution of the associated haplotypes.

To define haplotypes for each gene, including the causal locus, we allowed the major mass of the causal SNP to lie mainly on one marker haplotype: 001101 for G_1 and 010100 or 101100 for G_2, depending on whether a scenario with a common or rare variant, respectively, was considered. We fixed the marginal frequency for a causal SNP to be the same as that for the corresponding main haplotype: 12% for G_1 and 12.7% (common) or 4.1% (rare) for G_2. To allow for imperfect LD between the causal and the marker SNPs, we allowed for a small amount of recombination between the causal SNP and a set of other marker haplotypes: {000001,000010} for G_1 and {100000,101100} or {000010, 010010} for G_2, depending on whether a scenario with the common or rare variant was considered. We varied the recombination fraction (δ) at three different values to generate different degrees

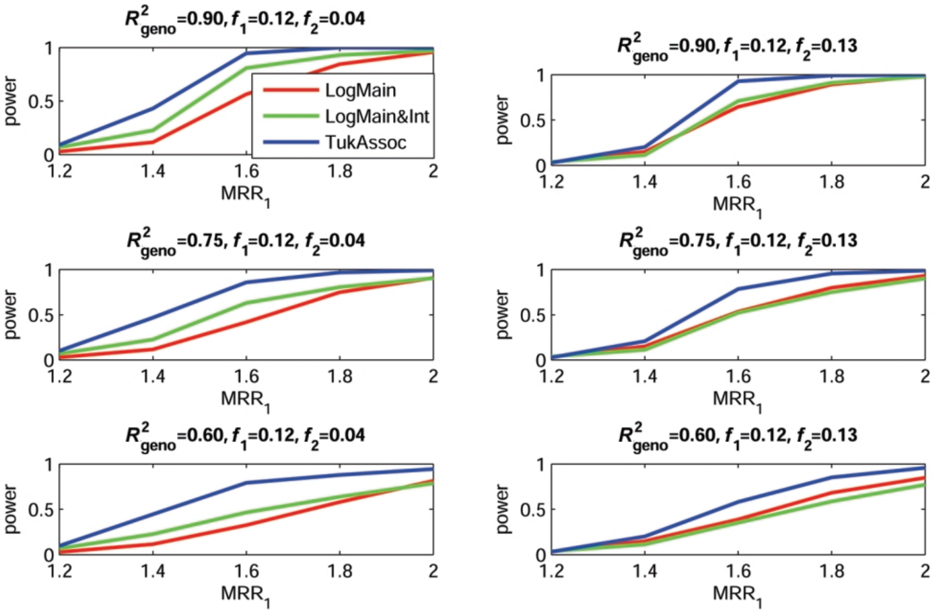

Figure 5. Empirical power, at $\alpha = 0.01$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the crossover model, with $\phi_1 = 0.90$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

of LD between the causal and marker SNPs. The values of R^2_{geno}, defined as the squared multiple correlation between the genotypes at the causal loci and those at the corresponding marker loci, were 90%, 75%, and 60% in these three settings.

Given the set of haplotype frequencies, in each simulation we first generated diplotype (haplotype-pair) data for a random sample of subjects, assuming random mating and no LD between genes. For each subject, we generated a binary disease end point, $D = 0$ or $D = 1$, assuming a general logistic-regression model of the formula

$$\Pr(D = 1) = \frac{\exp[\alpha + \theta_1 I(S_1^*) + \theta_2 I(S_2^*) + \theta_{12} I(S_1^*) I(S_2^*)]}{1 + \exp[\alpha + \theta_1 I(S_1^*) + \theta_2 I(S_2^*) + \theta_{12} I(S_1^*) I(S_2^*)]} , \tag{5}$$

where $I(S_1^*)$ and $I(S_2^*)$ are binary indicator variables for the presence of the variant allele at the respective causal loci. For each given set of parameter values θ_1, θ_2, and θ_{12}, the intercept parameter α was chosen in such a way that the marginal probability of the disease in the underlying population is fixed at 1%. In each replication, we first generated data for a large random sample of subjects, which we then treated as the "study base" to further select a case-control sample of given size. During analysis of each set of simulated data, we assumed that genotype data are variable for the marker SNPs but not for the causal SNPs.

We computed the empirical significance level of the proposed testing procedure, by simulating data under two different settings, both of which corresponded to the null hypothesis of no association of the disease with G_1. In the first setting, we assumed all the association parameters—θ_1, θ_2, and θ_{12}—to be zero, which implied that both G_1 and G_2 were not associated with the disease. In the second setting, we assumed θ_1 and θ_{12} to be null but allowed nonzero values for θ_2, so that G_2 could be associated with the disease even if G_1 is not. The significance thresholds for the test statistics T_1^* were obtained using two methods: (1) *permutation-based* resampling of the genotype data of SNPs in G_1 and (2) the *asymptotic-based* method, which requires generation of normal numbers.

To evaluate power, we simulated data using five different models for the joint effect of the two causal SNPs (see table 2). Assuming rare disease, these settings correspond to (1) the *purely epistatic* form, which assumes that the effect of one variant exists only in the presence of the other and vice versa; (2) the *multiplicative* form, which assumes that the joint effect of the two variants is given by the product of the main effects of the individual variants[31]; (3) the *purely additive* form, an approximation to the genetic heterogeneity model,[18] which assumes that the joint effect of the two variants is given by the sum of main effects of the individual variants; and (4) the *crossover* model, which assumes that the second variant has no effect by itself but that it reverses the effect of the first variant. For each model, we varied the value of the free risk parameter(s) in a way that the marginal relative risk (MRR)—the relative risk of the disease associated with one variant when the presence of the other is ignored—associated with S_1^* ranges in the set $\{1.2,1.4,1.6,1.8,2.0\}$. For the epistatic and multiplicative models, the MRR for S_2^* also varied in the same range. For the additive model, we fixed the MRR for S_2^* to be 2.0 (low penetrant) and 5.0 (high penetrant) in the common and rare variant scenarios, respectively. For the crossover model, we assumed $\phi_1 = 0.90(<1)$, which implies a modest protective effect of S_1^* in the absence of S_2^*.

We compared power for three different G_1-specific tests of association: (1) LogMain, an omnibus 6-df χ^2 test based on a logistic-regression model that involves the main effects of only the six marker SNPs in G_1[12]; (2) LogMain&Int, an omnibus 42-df χ^2 test

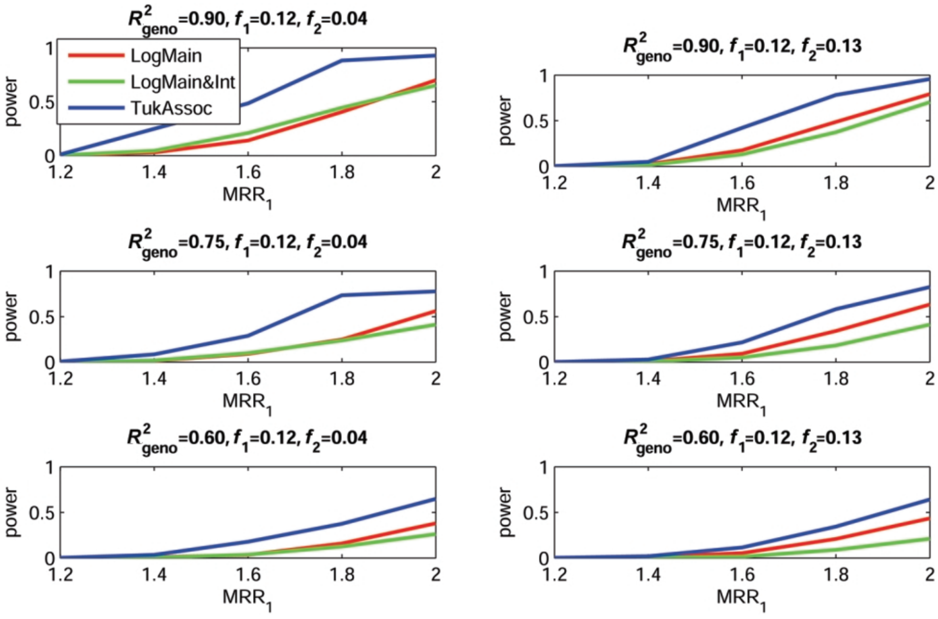

Figure 6. Empirical power, at $\alpha = 0.0001$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the purely epistatic model (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

based on a logistic-regression model that involves the main effects of all the SNPs in G_1 and G_2 and all pairwise interactions between SNPs across the two genes (the null model in this test involves only the main effects of the SNPs in G_2); and (3) TukAssoc, the proposed test of association based on Tukey's model of interaction. In each method, the genotype data for the marker SNPs were coded as continuous variables representing the count for the respective minor alleles. Asymptotic-based significance thresholds were used for all three test statistics. Both type I errors and powers were obtained empirically, on the basis of 1,000 simulated data sets.

Results

Simulation Study

Table 3 shows the empirical type I error rates for the proposed testing procedure at a significance level of $\alpha =$ 0.01. Both methods performed well in maintaining the nominal significance level in all the different settings considered.

Figures 2–5 show the empirical power of different procedures for testing the association of the disease with G_1 at a significance level of 0.01 under different models for the joint effects of the underlying causal variants. Similar examples showing power at a significance level of 0.0001 are provided in figures 6–9.

When the true effects of the causal SNPs were purely epistatic (fig. 2), the proposed test of association (TukAssoc), which accounts for interactions, clearly outperformed the standard main-effect–based test (LogMain) in detecting the association of the disease with G_1. Given the same marginal-effect size for the causal SNP in G_1, the gain in power was larger when the causal SNP in the background gene, G_2, was rarer, because it corresponded to larger magnitude of the interaction parameter θ_{12}. In this rare-variant setting, the test based on the saturated model of interaction (LogMain&Int) also performed better than the main-effect–based test (LogMain) but lost major power compared with TukAssoc because of the use of large dfs. As the correlation between the causal and marker SNPs decreased, the absolute power of all of the different methods, as expected, decreased. Interestingly, the power of both interaction-based tests, LogMain&Int and TukAssoc, relative to LogMain, also decreased as R^2_{geno} decreased.

When the true effects of the causal SNPs were multiplicative (fig. 3), LogMain, which assumes no multiplicative interaction, as expected, had the highest power. The proposed test, TukAssoc, although not the best, remained a close second. In contrast, LogMain&Int, which used the saturated model for interaction, performed very poorly. When the true model was additive (fig. 4), the power of TukAssoc remained very close to that of LogMain when the causal SNP, S_2^*, in the background gene, G_2, was "common low penetrant." In contrast, under the same model, when S_2^* was "rare high penetrant," TukAssoc showed a major gain in power over LogMain. Finally, under the crossover model (fig. 5), in which the causal variant in G_2 reversed the effect of that in G_1, TukAssoc had much higher power than LogMain. Often, LogMain&Int also performed better than LogMain but remained far inferior to TukAssoc. As observed under the epistatic model, the

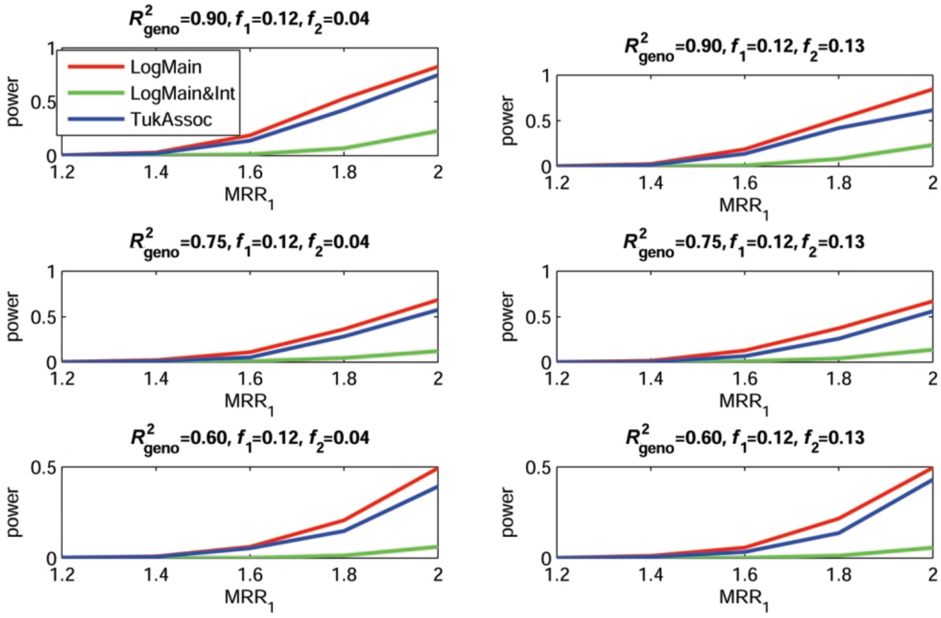

Figure 7. Empirical power, at $\alpha = 0.0001$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the purely multiplicative model, with $\phi_1 = \phi_2$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

power of both TukAssoc and LogMain&Int relative to LogMain decreased for lower values of R^2_{geno}.

Under each setting described above, the power advantage of TukAssoc compared with the other two procedures further increased when the significance level was chosen to be 0.0001 instead of 0.01 (see figs. 6–9).

A Study of NAT2 *Acetylation Activity, Smoking, and Risk of Colorectal Adenoma*

Cigarette smoking has been consistently associated with the risk of colorectal adenoma, a recognized precursor to colorectal cancer (MIM 114500). Thus, there is interest to study the risk of adenoma associated with candidate genes encoding *N*-acetyltransferase enzymes that are involved in the metabolism of aromatic amines derived from tobacco smoke. *N*-acetyltransferase 2 (*NAT2*), located at 8p21.3, is a candidate gene known to play an important role in the detoxification of certain aromatic carcinogens and, after *N*-hydroxylation, in the activation of other amine protocarcinogens to their ultimate carcinogenic form. We recently completed a report[32] on a case-control study of the association between *NAT2* genetic variants and colorectal adenoma, in relation to tobacco smoking, which used "case" individuals with left-sided prevalent advanced adenoma and sex- and age-matched "control" individuals selected from the screening arm of the large, ongoing Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial.[33,34] The study selected six SNPs (*C282T, T341C, C481T, G590A, A803G,* and *G857A*) for genotyping that are known to be informative for reconstructing diplotypes that have been described elsewhere and categorized in laboratory studies as having "slow," "intermediate," or "rapid" *N*-acetyltransferase enzymatic activity. On the basis of genotype data, 685 cases and 693 controls in the study were assigned diplotype and related phenotype status with use of an algorithm developed at the University of Louisville.[28,29] The frequency distribution of these diplotypes and associated phenotypes are shown in table 4. Questionnaire data on the smoking histories of these subjects were also available. We categorized subjects, on the basis of smoking history, as "current," "former," or "never."

Clearly, in the original study,[32] the availability of the prior data to group the numerous diplotypes into a smaller number of phenotypic categories provided us an opportunity to investigate the association between *NAT2* and adenoma in a very powerful way. In the current study, we compared the power of alternative tests that relied on the original diplotypes themselves, pretending as if the underlying phenotype variable was not observed. It is to be noted that, for most genomic regions, the phenotypic significance of the variants is not well understood and, thus, the opportunity for grouping the observed genetic variants into a smaller number of categories may not exist. Using the diplotype information shown in table 4, we performed three different tests of association between *NAT2* and adenoma: (D1) LogMain, an omnibus χ^2 test based on a logistic-regression model that involves a main-effect term for each of the 14 nonreferent diplotypes (df = 14);

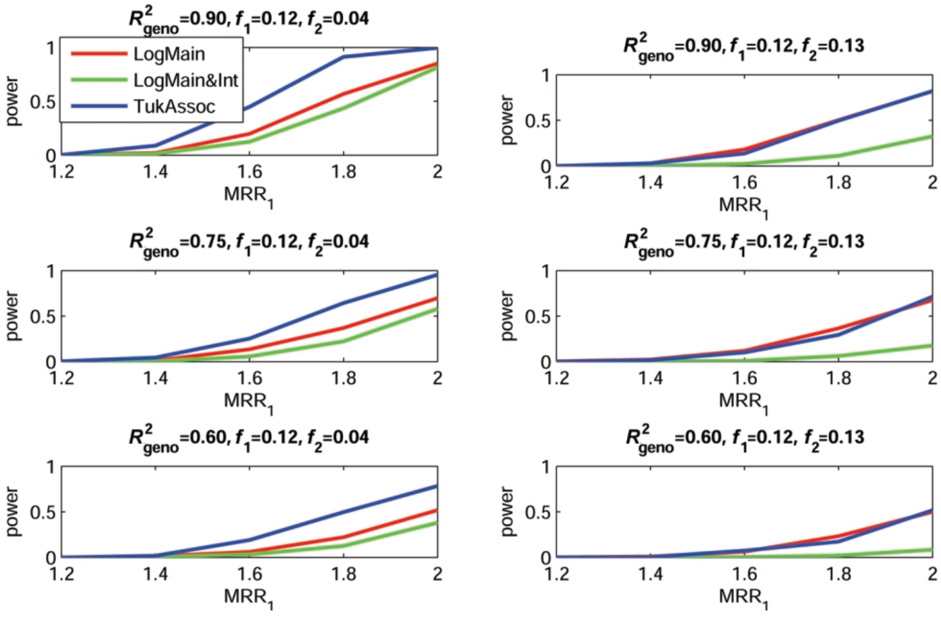

Figure 8. Empirical power, at $\alpha = 0.0001$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S_1^*. The joint effect of causal SNPs in G_1 and G_2 follows the additive model, with ϕ_2 chosen so that MRR$_2$ = 2.0 when $f_2 = 0.12$ and MRR$_2$ = 5.0 when $f_2 = 0.04$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

(D2) LogMain&Int, an omnibus χ^2 test based on a logistic-regression model that involves the main effects of the diplotypes and all the interactions between the diplotypes and the two nonreferent categories of smoking; the null model in this test includes only the main effects of the smoking categories (df = 14 + 14 × 2 = 42); and (D3) TukAssoc, an omnibus test of association for *NAT2* diplotypes, based on the model

$$\operatorname{logit}\Pr(D=1) = \alpha + \sum_{j=1}^{14}\beta_j I(H=h_j) + \sum_{k=1}^{2}\gamma_k I(\mathrm{Smk}=k)$$
$$+\theta\sum_{j=1}^{14}\sum_{k=1}^{2}\beta_j\gamma_k I(H=h_j)I(\mathrm{Smk}=k)\ ,$$

where $I(H = h_j)$, $j = 1,\ldots,14$, and $I(\mathrm{Smk} = k)$,$k = 1,2$ denote the dummy variables for the diplotypes and the smoking categories. In addition, we also performed two phenotype-based tests: (P1), a 1-df test for the trend effect of the phenotype variable that codes it as a continuous variable—"0" for "slow," "1" for "intermediate," and "2" for "fast"; and (P2), an omnibus test for the main effect and interactions (with smoking categories) for the continuous phenotype variable (df = 1 + 2 = 3). All of the phenotype- and diplotype-based tests were adjusted for age and sex by including appropriate main-effect terms in the corresponding logistic-regression model. For computation of *P* values, we relied on permutation-based resampling, instead of on the asymptotic-based method, because of the small number of subjects in some of the diplotype categories.

From the results shown in table 5, it is clear that, in this example, the test that captures both the main and the interaction effects of the phenotype variable was the most sensitive in detecting the association of adenoma with *NAT2*. Among the diplotype-based methods, TukAssoc, although not significant at the traditional 5% level, provided more evidence of association than the other two methods considered. This example illustrates two important points. First, it shows how incorporating interaction can improve the power to discover genetic associations. Second, it shows that the most powerful test for a genetic association could be obtained when the phenotypic significance of the underlying variants are well understood a priori. If such prior data are not available but the variants within a genomic region are likely to be functionally related by a common biological mechanism, such as *NAT2* acetylation activity, then the proposed test of association based on Tukey's 1-df model of interaction could be a promising approach.

Discussion

In summary, we have proposed a powerful method for testing genetic association in case-control studies, by accounting for heterogeneity in disease risk due to gene-gene and gene-environment interactions. By considering a conceptual framework in which multiple SNPs within a gene are postulated to be related to a common causal mechanism, we motivate the use of a low-dimensional 1-df model for gene-gene and gene-environment interactions. On the basis of this model, we have developed an omnibus gene-specific test of association that can simultaneously account for the main effects of the variants within the region as well as for their interactions with the variants of another region or with an environmental exposure. We used both simulated and real data to study the efficiency of the proposed method relative to two standard logistic-regression–based tests, one ignoring interactions and the other incorporating a saturated model for interactions. These studies suggest that the proposed method can improve the power of genetic association tests in the presence of nonmultiplicative effects of the underlying causal variants. When the true effects are close to multiplicative, the proposed method, although it may not be the best, generally has robust power.

Table 4. Distribution of Cases and Controls by *NAT2* Diplotype and Acetylation Phenotype in the PLCO Adenoma Study

Diplotype	Acetylation Phenotype[a]	Cases	Controls
*5B/*6A	0	155	124
*5B/*5B	0	121	98
*6A/*6A	0	59	73
*5A/*5B	0	16	18
*5B/*7B	0	16	17
*5B/*5C	0	16	10
*6A/*7B	0	10	12
*5A/*6A	0	8	10
*5C/*6A	0	7	9
*4/*5B	1	109	138
*4/*6A	1	86	104
*4/*7B	1	17	8
*4/*5A	1	9	6
*4/*4	2	37	41
Rare	…	19	25

[a] For acetylation phenotype, 0 is slow, 1 is intermediate, and 2 is rapid.

Gene-gene and gene-environment interactions can cause the effect size of a genetic variant to be heterogeneous for different subgroups of the population. Tests of genetic association that ignore such heterogeneity may lack power, since the “marginal” effect of a variant when subgroups are ignored can be quite small, even though its effect can be quite large in specific subgroups. Under an extreme form of interaction in which the effect of a variant may be in opposite directions in different subgroups, there may be no marginal effects even if there are very strong subgroup effects. Accounting for interaction in association testing allows one to exploit the full variation in the effects of the causal variants, at the risk of increasing the number of parameters to be tested. Our applications involving the saturated model for interaction suggest that the power advantage of interaction-based tests may be negated if too many dfs are spent on model for interaction. The proposed test based on Tukey’s 1-df model for interaction provides a good compromise between detecting large genetic effects versus testing for many parameters.

When multiple SNPs are involved within a gene, one could attempt to reduce the dfs for related association tests

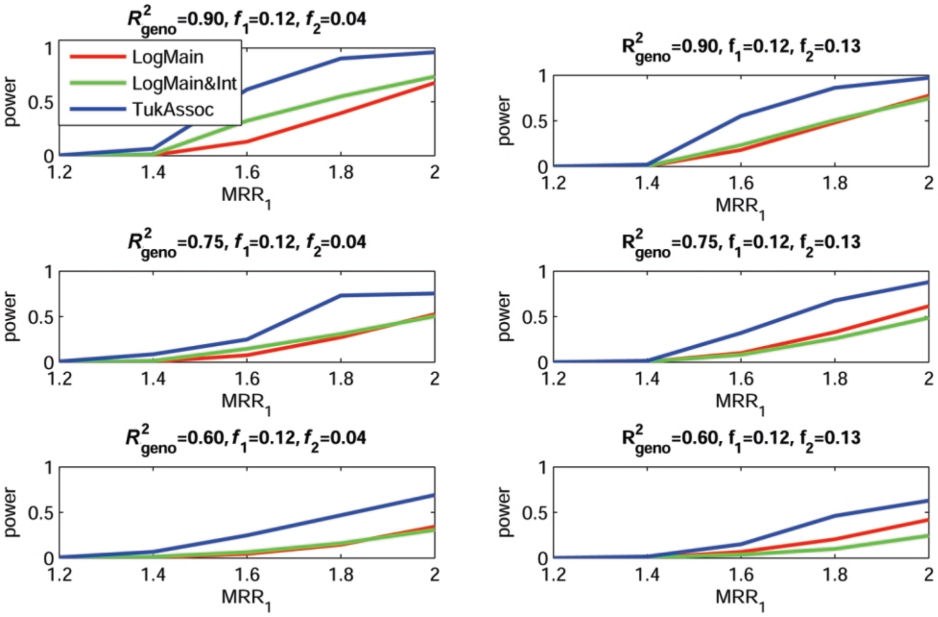

Figure 9. Empirical power, at $\alpha = 0.0001$, to detect the association of the disease with candidate gene G_1 as a function of the MRR of the underlying causal SNP, S^*_1. The joint effect of casual SNPs in G_1 and G_2 follows the crossover model, with $\phi_1 = 0.90$ (see table 2). f_1 and f_2 denote minor-allele frequencies for causal SNPs in G_1 and G_2, respectively, and R^2_{geno} denotes the value of multiple R^2 between the causal and marker loci within a gene.

Table 5. Test of Association of Adenoma with *NAT2,* with and without Accounting for *NAT2*-Smoking Interaction

Test	Test Statistic	df	*P*
Acetylation-based[a]:			
LogMain	3.30	1	.069
LogMain&Int	14.23	3	.003
Diplotype-based[b]:			
LogMain	18.25	14	.200
LogMain&Int	54.41	42	.156
TukAssoc	26.45	…	.071

[a] Uses continuous phenotype variable codes 0, 1, and 2.

[b] Uses diplotypes shown in table 4.

on the basis of a derived variable that can combine information across multiple SNPs by using prior knowledge about possible directionality of the effects of the variants.[35] The acetylation phenotype for the gene used in our data analysis, *NAT2,* is a derived variable defined on the basis of prior data. The scope of such analysis, however, is limited for contemporary association studies because of the lack of such prior data on the SNPs. The proposed method, which also uses derived variables—namely, the latent factors Z_1 and Z_2—does not require any explicit prior data on the directionality of the effects of the SNPs under study. In particular, the generalized association parameters (β) allow one to estimate the directionality as well as the strength of association from the data. Thus, the proposed method can use a low-df model for interaction without requiring explicit prior knowledge about the potential effects of the SNPs.

An alternative approach to reduce the df for association tests could be to follow a two-stage procedure in which SNPs are first tested for their main effects and, then, interaction-based tests involving only those SNPs for which main effects were found to be significant are considered. In general, obtaining the correct type I error rates for such sequential procedures is quite complex. A recent report suggested a conservative but simple approach for finding critical values for SNP-based two-stage tests.[36] In a limited simulation study, we found the power of such a procedure to be similar to the proposed gene-based one-stage test, TukAssoc, when each candidate gene under study involved only a single causal variant. In contrast, when the individual candidate genes involved multiple causal variants, TukAssoc was clearly superior. Further work is needed to develop more-efficient two-stage tests of association.

Computationally, the proposed score-test statistic is remarkably simple and can be implemented using standard logistic-regression software. We have described a simple and fast way of generating the asymptotic null distribution of the test statistics. The methodology can be easily generalized to alternative types of study designs and outcome traits by simply replacing the logistic model with a suitable alternative regression model. Moreover, the methods can be used to test for the collective effect of any group of functionally related SNPs, which need not be restricted to candidate genes.

The results from our simulation studies involving two candidate genes are quite intuitive. When the true effects of the causal loci across two genes were multiplicative, tests that were based on the marker SNPs of individual genes and that ignored gene-gene interactions were optimal. This result can be explained mathematically by the observation that, under the multiplicative model, the likelihood for case-control data can be factored into two pieces, each depending on the marker data from a single gene.[20] When the true effects of the causal loci were additive—a nonmultiplicative model that is often considered to be the default for specifying the joint effects of two exposures acting on nonoverlapping pathways[18,37]—the proposed test performed similarly to or substantially better than the main-effect–based test, depending on the strength of the main effects of the causal variants. When the main effects for both causal variants were modest, the additive model corresponded to only a small departure from multiplicative effects and, thus, TukAssoc performed similarly to LogMain. In contrast, when the main effect of the causal variant in one gene was large, the additive model corresponded to a large departure from multiplicative effects, and TukAssoc became superior. The largest gains in power for TukAssoc over LogMain were seen for the epistatic and crossover models, both of which corresponded to a large departure from multiplicative effects.

As expected, the absolute power of all the methods decreased as R^2_{geno}, the correlation between the causal and the marker SNPs, decreased. Interestingly, the power of both interaction-based tests, LogMain&Int and TukAssoc, relative to LogMain also decreased as R^2_{geno} decreased. When the markers have low correlation with the respective causal SNPs, the joint risk of the disease in terms of the markers may appear close to the multiplicative model (with nonnull main effects), even if the true effects of the causal variants are highly epistatic. Thus, for low values of R^2, models involving only the main effects of the markers may perform well, even if the true effects of the causal loci are highly interactive. In the context of association testing using single binary markers, a similar robustness property for the multiplicative model has been noted elsewhere.[38]

In this article, we focused on tests of association for one candidate gene by exploiting its interaction with another candidate gene or an environmental exposure. In practice, however, an association study may involve a variety of candidate genes and environmental exposures, each of which may interact with all the others. Clearly, if all the possible interactions are to be accounted for, the number of potential tests could be very large. To examine the effect of the associated multiple-testing problem, we performed a small simulation study. We used the same setting as shown in figure 2 but added eight null genes to the analysis. Similarly as we did for the two genes that contained the causal loci, for each of the null genes we assumed that genotype data are available on six marker SNPs. We used TukAssoc to assess the significance of a specific gene by pairing it with each of the other nine genes and then taking the maximum of the corresponding nine different test statistics. To evaluate the critical value of the final test statistic, we used permutation-based resampling that adjusts for multiple testing in an efficient way, by taking into account the correlation among the different test statistics. Alternatively, we used the standard main-effect–based test LogMain to test for each gene individually, ignoring interactions. For both TukAssoc and LogMain, the test for each specific gene was performed at the significance level of $0.01/10 = 0.001$, to maintain an overall significance level of 0.01. Even with multiple-testing adjustment, TukAssoc remained substantially more powerful than LogMain in a number of different settings. For example, in the setting of $R^2 = 90$, $f_2 = 0.13$, and $\mathrm{MRR} = 1.6$, the power for detecting the association of the disease with G_1 was 54% with use of TukAssoc and 34% with use of LogMain. With $f_2 = 0.04$ and R^2 and MRR remaining the same as the values above, the power for TukAssoc became 75%, whereas that for LogMain remained at 34%. In the context of a much-larger-scale study involving a whole-genome scan, a recent report has made a similar observation that tests that account for interactions among pairs of SNPs could be substantially more powerful than those based only on the main effects of the SNPs, even though the former class of tests may require a much higher level of multiple-testing adjustment.[36]

Nevertheless, we believe that the power advantage of interaction-based tests would be best realized if the number of interactions to be considered could be reduced a priori, on the basis of biological knowledge, previous data, or/and some prescreening methods. Biological knowledge of a pathway, for example, may help investigators choose a few “high-prior” candidate genes that are likely to have central roles in mediating the biological effects of various genetic and environmental exposures. In such a setting, the power of association for the other candidate genes in that pathway can be improved by accounting for their interactions with the selected high-prior genes. Data from previous linkage and association studies could also guide the selection of such high-prior candidates.

A prescreening method could also reduce the number of interactions to be tested. For case-control studies involving candidate SNPs, Millstein et al.[25] described a method that first screens for potential interactions by testing for the significance of the correlations among pairs of SNPs in the pooled case-control sample. If, for a pair of SNPs, no LD is expected in the population but correlation is evident in the pooled case-control data, it indicates non-multiplicative effects of the variants on the risk of the disease. Moreover, because the screening is done only on the basis of the genotype data of the subjects, without regard to their case-control status, subsequent tests of association do not require adjustment. Similar ideas can be adopted to reduce the number of gene-gene interactions in our setting. For example, one may use a global test of independence between two sets of SNPs to decide whether the corresponding gene-gene interaction should be included in the subsequent association analysis.

In conclusion, the proposed method, given its efficiency, computational simplicity, and broad applicability, seems a promising approach for testing genetic association in the presence of gene-gene and gene-environment interactions. Future work is needed to develop and evaluate practical strategies for the applications of the methodology in large-scale association studies involving specific biologic pathways or the whole genome.

Acknowledgments

We thank Drs. Glen Satten and Alice Whittemore and two anonymous reviewers for their positive comments on an earlier version of this article. Resequencing of the *GPX3* and *GPX4* genes and genotyping assays for the *NAT2* gene were performed at the Core Genotyping Facility at the NCI Advanced Technology Center, in Gaithersburg, MD.[39] This research was supported by the Intramural Program of the National Institutes of Health.

Appendix A
Heuristic Derivation of Tukey’s 1-df Model of Interaction

Let $X_j = \mu_j + \sum_{k_j=1}^{K_j} \gamma_{k_j j} S_{k_j j}$ for $j = 1{,}2$, and define the function

$$f(x_1,x_2) = \frac{\exp(\theta_0 + x_1 + x_2)}{1 + \exp(\theta_0 + x_1 + x_2)} \ .$$

By substituting the regression formula for Z_1 and Z_2 (fig. 1, upper boxes) into the disease-risk model (lower box) and taking a Taylor’s series expansion, with respect to ϵ_1 and ϵ_2, one can write

$$\begin{aligned}\Pr_{\epsilon_1,\epsilon_2}(D = 1 \mid \mathbf{S}_1,\mathbf{S}_2) = {} & f(X_1,X_2) \\ & + \epsilon_1 f_1(X_1,X_2) + \epsilon_2 f_2(X_1,X_2) \\ & + O(\epsilon_1^2 + \epsilon_1^2) \ ,\end{aligned}$$

where $f_j(x_1,x_2) = \partial f(x_1,x_2)/\partial x_j$, $j = 1{,}2$, and $O(\epsilon_1^2 + \epsilon_1^2)$ denotes a term that can be bounded above by $K(\epsilon_1^2 + \epsilon_1^2)$ for a suitable positive constant K. Noting that ϵ_1 and ϵ_2 are mean zero random variables (conditional on $\mathbf{S}_1$ and $\mathbf{S}_2$), we can write

$$\begin{aligned}\Pr(D = 1 \mid S_1,S_2) &= E_{\epsilon_1,\epsilon_2} \Pr_{\epsilon_1,\epsilon_2}(D = 1 \mid S_1,S_2) \\ &= f(X_1,X_2) + O(\sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2) \ ,\end{aligned}$$

where $\sigma_{\epsilon_j}^2$ denotes the variance of ϵ_j and $j = 1{,}2$. Thus, if $\sigma_{\epsilon_j}^2$ and $j = 1{,}2$ are small, then $\Pr(D = 1 \mid S_1,S_2) \approx f(X_1,X_2)$, which is precisely the model shown in equation (1), with $\alpha = \theta_0 + \theta_1^* \mu_1 + \theta_2^* \mu_2 + \theta_{12}\mu_1\mu_2$, $\beta_{k_j j} = \theta_j^* \gamma_{k_j j}$, $k_j = 1, \ldots ,K_j; j = 1{,}2$, and $\theta = \theta_{12}/(\theta_1^* \times \theta_2^*)$, where $\theta_j^* = \theta_j + \theta_{12}\mu_{3-j}, j = 1{,}2$.

Log-Likelihood, Score Function, and Information Matrices

Let $P_{\alpha,\beta_1,\beta_2;\theta}(\mathbf{S}_1,\mathbf{S}_2)$ denote $\Pr(D=1|\mathbf{S}_1,\mathbf{S}_2)$, as defined by the proposed model in equation (1). The log-likelihood of the data under case-control design can be written as

$$L=\sum_{i=1}^{N_1+N_0} D_i \log P_{\alpha,\beta_1,\beta_2;\theta}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})$$

$$+(1-\mathbf{D}_{\mathbf{i}})\log[1-\mathbf{P}_{\alpha,\beta_1,\beta_2;\theta}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})] \ . \qquad \text{(A1)}$$

Under the null hypothesis that $\beta_{k_11}=0$ for all k_1, the maximum-likelihood estimates of the parameters $\psi=(\alpha,\beta_2)$, for each fixed value of θ, can be obtained by solving the score equation $S_\psi(\psi;\theta)=0$, where

$$S_\psi(\psi;\theta)=\sum_{i=1}^{N_0+N_1}\mathbf{Z}_{\mathbf{2}\mathrm{i}}[D_i-P_{\alpha,\beta_1=0,\beta_2;\theta}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})]$$

and $\mathbf{Z}_{\mathbf{2}\mathrm{i}}=(1,\mathbf{S}_{\mathbf{2}\mathrm{i}}^T)^T$. Note that the quantity $P_{\alpha,\beta_1,\beta_2;\theta}$ does not depend on θ when $\beta_1=0$, and $S_\psi(\psi;\theta)\equiv S_\psi(\psi)$ corresponds to standard logistic-regression score function that involves only main-effect terms for the marker SNPs in G_2. Let the maximum-likelihood estimates of ψ under $\beta_1=0$ be denoted by $\hat{\psi}=(\hat{\alpha},\hat{\beta}_2)$. Further, let $\hat{P}_{\mathrm{NULL}}(\mathbf{S}_1,\mathbf{S}_2)$ denote $P_{\hat{\alpha},\beta_1=0,\hat{\beta}_2;\theta}(\mathbf{S}_1,\mathbf{S}_2)$. Now, for a fixed value of θ, the score function for the association parameters β_{k_11} and $k_1=1,\ldots,K_1$, evaluated under the null hypothesis that $\beta_{k_11}=0$ for all k_1, can be written in the form of equation (3).

Define $Z_2=(1,\mathbf{S}_2^T)$ to be a design matrix associated with the standard logistic-regression analysis of the data that allows for the constant intercept term α and a main-effect term for each of the markers in G_2. Ignoring terms with expectation zero, the formulae for the information matrices in equation (4), evaluated at $\beta_1=0$ and $\psi=\hat{\psi}$, can be written as

$$I_{\beta_1\beta_1}(\theta)=\sum_{i=1}^{N_0+N_1}\left[1+\theta\hat{\beta}_2^T\mathbf{S}_{\mathbf{2}\mathrm{i}}\right]^2\hat{P}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}}) \qquad \text{(A2)}$$

$$\times[1-\hat{\mathbf{P}}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})]\mathbf{S}_{\mathbf{1}\mathrm{i}}\mathbf{S}_{\mathbf{1}\mathrm{i}}^{\mathbf{T}} \ ,$$

$$I_{\beta_1,\psi}(\theta)=\sum_{i=1}^{N_0+N_1}\left[1+\theta\hat{\beta}_2^T\mathbf{S}_{\mathbf{2}\mathrm{i}}\right]\hat{P}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}}) \qquad \text{(A3)}$$

$$\times[1-\hat{\mathbf{P}}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})]\mathbf{S}_{\mathbf{1}\mathrm{i}}\mathbf{Z}_{2\mathrm{i}}^{\mathbf{T}} \ ,$$

and

$$I_{\psi,\psi}=\sum_{i=1}^{N_0+N_1}\hat{P}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})[1-\hat{\mathbf{P}}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}\mathrm{i}})]\mathbf{Z}_{\mathbf{2}\mathrm{i}}\mathbf{Z}_{\mathbf{2}\mathrm{i}}^{\mathbf{T}} \ . \qquad \text{(A4)}$$

Efficient Score Function and Asymptotic Theory

Let $S_{\beta_1,i}(\theta)$ be the contribution of the i-th subject in the score vector $S_{\beta_1}(\theta)$ defined in equation (3). Similarly, let $S_{\psi,i}=\mathbf{Z}_{\mathbf{2}\mathrm{i}}[D_i-\hat{P}_{\mathrm{NULL}}(\mathbf{S}_{\mathbf{1}\mathrm{i}},\mathbf{S}_{\mathbf{2}}\mathrm{i})]$ be the contribution of the i-th subject to the score vector $S_\psi(\psi)$, evaluated at $\psi=\hat{\psi}$. Define $i_{\beta_1,\beta_1}(\theta)$, $i_{\beta_1,\psi}(\theta)$, and $i_{\psi,\psi}$ to be the asymptotic limits of the scaled information matrices $N^{-1}I_{\beta_1\beta_1}(\theta)$, $N^{-1}I_{\beta_1\psi}(\theta)$, and $N^{-1}I_{\psi,\psi}$. By using a standard Taylor's series argument, one can represent the score vector $S_{\beta_1}(\theta)$ in its asymptotic form

$$N^{-1/2}S_{\beta_1}(\theta)=N^{-1/2}\sum_{i=1}^{N}U_i(\theta)+o_p(1) \ ,$$

where $U_i(\theta)$ denotes the efficient influence function defined by

$$U_i(\theta)=S_{\beta_1,i}(\theta)-i_{\beta_1,\psi}(\theta)i_{\psi,\psi}^{-1}S_{\psi,i} \qquad \text{(A5)}$$

and $o_p(1)$ represents the term that converges to zero in probability. On the basis of the standard central-limit theorem, one can then show that, for any fixed value of θ and under the null hypothesis of $\beta_1=0$, $N^{-1/2}S_{\beta_1}(\theta)$ converges to K_1-variate normal distribution with zero mean and variance-covariance matrix given by $i_{\beta_1,\beta_1}(\theta)-i_{\beta_1,\psi}(\theta)i_{\psi,\psi}^{-1}i_{\beta_1,\psi}(\theta)^T$. Moreover, on any given compact interval $\overline{\Theta}$ for θ, the convergence of the score vector $N^{-1/2}S_{\beta_1}(\theta)$ to the corresponding normal distribution can be shown to be uniform over θ. Thus, it follows that $N^{-1/2}S_{\beta_1}(\theta)$, as a K_1-dimensional stochastic process in θ, converges to a zero mean Gaussian process for which the covariance function for the pair of value (θ_1,θ_2) is given by the asymptotic limit of $N^{-1}\sum_{i=1}^{N}U_i(\theta_1)U_i^T(\theta_2)$.

Web Resources

The URLs for data presented herein are as follows:

NCI Advanced Technology Center, http://cgf.nci.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *NAT2, GPX3, GPX4,* and colorectal cancer)

References

1. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
2. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856
3. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99
4. Carlson CS, Eberle MA, Kruglyak LA, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
5. Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118
6. International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

7. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
8. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933
9. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
10. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Genova GD, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237
11. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27–36
12. Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31
13. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120
14. Clayton DG, Chapman JM, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415–428
15. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151
16. Schaid D, Rowland C, Tines D, Jacobson R, Poland G (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434
17. Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 83:7353–7357
18. Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228
19. Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127–1136
20. Dupuis J, Brown PO, Siegmund D (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 140:843–856
21. Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 66:1273–1286
22. Gauderman WJ, Siegmund KD (2001) Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 52:34–46
23. Peng J, Tang HK, Siegmund D (2005) Genome scans with gene-covariate interaction. Genet Epidemiol 29:173–184
24. Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157
25. Millstein J, Conti DV, Gilliland FD, Gauderman WJ (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 78:15–27
26. Tukey JW (1949) One degree of freedom for non-additivity. Biometrics 5:232–242
27. Scheffe H (1959) The analysis of variance. John Wiley and Sons, New York, pp 129–134
28. Hein D, Ferguson R, Doll M, Rustan T, Gray K (1994) Molecular genetics of human polymorphic N-acetyltransferase: enzymatic analysis of 15 recombinant wild-type, mutant, and chimeric NAT2 allozymes. Hum Mol Genet 3:729–734
29. Hein DW, Doll MA, Ferguson RJ (1995) Metabolic activation of carcinogenic arylamines by rapid acetylator, slow acetylator, and chimeric recombinant Syrian hamster NAT2 allozymes. Proc West Pharmacol Soc 38:59–62
30. Lin DY, Zou F (2004) Assessing genomewide statistical significance in linkage studies. Genet Epidemiol 27:202–214
31. Hodge SE (1981) Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. Am J Hum Genet 33:381–395
32. Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB (2006) Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. Pharmacogenomics 7:819–829
33. Hayes RB, Reding D, Kopp W, Subar AF, Bhat N, Rothman N, Caporaso N, Ziegler RG, Johnson CC, Weissfeld, Hoover RN, P PH, Palace C, Gohagan JK, Prostate Lung Colorectal and Ovarian Cancer Screening Trial Project Team (2000) Etiologic and early marker studies in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials 21:349S–355S
34. Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P, Reding D, Gelmann EP, Rothman N, Pfeiffer RM, Hoover RN, Berg CD, for the PLCO Trial Team (2005) Methods for etiologic and early marker investigations in the PLCO trial. Mutat Res 592:147–154
35. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76:780–793
36. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417
37. Thompson WD (1991) Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232
38. Pfeiffer RM, Gail MH (2003) Sample size calculations for population- and family-based case-control association studies on marker genotypes. Genet Epidemiol 25:136–148
39. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, Eckert A, Puri V, Gerhard D, Chanock SJ (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res 34:D617–D621