Powerful multi-locus tests for genetic association in the presence of gene-gene and gene-environment interactions

Nilanjan Chatterjee[1], Zeynep Kalaylioglu[2], Roxana Moslehi[1], Ulrike Peters[3], Sholom Wacholder[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA

[2]Information Management System, Rockville, Maryland

[3]Fred Hutchinson Cancer Research Center, Seattle

Corresponding address:

Nilanjan Chatterjee

6120 Executive Blvd, EPS 8038

Rockville, MD 20852, USA

email: chattern@mail.nih.gov

Ph: 301-402-7933

Fax: 301-402-0081

Running title: Multi-locus tests and interactions

Summary

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 multi-locus 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 "one degree-of-freedom" model for interaction which is natural when individual SNPs within a gene are associated with disease through a common biologic mechanism; in contrast, many standard regression models are designed as if each SNP has unique functional significance. Based on Tukey's model, we propose a novel, but computationally simple, "generalized" test for 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 to two standard tests for association, one ignoring gene-gene/gene-environment interactions and the other based on a saturated model for interactions. We demonstrate major power advantages for our method in analysis of data from a case-control study for the association between colorectal adenoma and DNA variants in *NAT2* genomic region, which are well known to be related to a common biologic phenotype, and under different models for gene-gene interactions using simulated data.

Keywords: case-control study, efficient score, epistasis, multi-locus test, locus heterogeneity, omnibus test, Tukey's 1 d.f model for interaction;

Identification of large number of single nucleotide polymorphisms (SNP) across human genome has given rise to great opportunity for fine mapping of disease susceptibility loci (DSL) through population-based association studies[1–5]. An increasingly popular design for association studies has been the "indirect" approach, where 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 likely to be in linkage disequilibrium with the underlying causal variants, if any exists. The availability of linkage disequilibrium information across 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 for 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, either due to their direct causal effects or due to their linkage disequilibrium with the underlying causal variant(s) in the region. Two classes of multivariate tests, one based on multi-locus 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 for complex diseases is that the etiologic effect of the underlying causal variants are likely to be complex due to 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. Starting from the late 80's, a variety of "multi-point" methods were developed to account for gene-gene interaction in linkage analysis[17–21]. Methods for linkage scan 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 due to LD or functionally related due to underlying common biologic 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 for gene-gene and gene-environment interactions could easily become very large, potentially causing loss of power, either due to the use of increased degrees of freedom or due to the need of multiple testing adjustments. We consider use of the Tukey's 1 d.f model for interaction[26,27]. We show that this parsimonious form of interaction can be motivated through a conceptual framework where the observed SNPs within a gene affects the risk of the disease through an underlying common causal mechanism. Modern association studies where 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.

Based on Tukey's model, we propose a novel multi-locus test of genetic association that can efficiently exploit the LD pattern among SNPs within a gene and simultaneously can account for their interactions with SNPs in another gene or with an environmental exposure. We simulate case-control data mimicking 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 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 for interactions.

1 Materials and Methods

1.1 A latent variable model and Tukey's one-degree-of-freedom form of interaction

Suppose G_1 and G_2 are two candidate genes of interest for which K_1 and K_2 marker SNPs 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 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 "biologic phenotype" Z_i which is causally related to the disease. The associations between the markers and the "biologic phenotypes" for the two genes are described by two separate linear models (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 and Z_2 is specified by a standard logistic model involving both main- and interaction effects (lower box) . It is also implicitly assumed that given the true biologic 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 proceeding 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 setting $\gamma_{ik} = 1$ 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 which aim to select the markers to be a panel of "tagging SNPs" that would have a very high degree of linkage disequilibrium, as measured by the R^2 criterion, with all of 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 biologic phenotype that may be governed by several different variants within the same gene G_i. In the study of colorectal adenoma (see Results), the underlying biologic phenotype for the gene of interest *NAT2* is N-acetyltransferase enzymatic activity level which has been shown to be determined by several single based pair substitutions in the gene and the associated haplotypes/diplotypes[28,29].

The logistic model shown in Figure 1 (bottom 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 $\mathbf{S_2}$ using the corresponding linear regression models, and assuming small variances for the error terms ϵ_1 and ϵ_2, a risk-model for the disease in terms of the observable SNPs can be derived approximately in the form (see Appendix for details)

$$\text{logit}\,\{\Pr(D=1|\mathbf{S_1},\mathbf{S_2})\} = \alpha + \sum_{k_1=1}^{K_1}\beta_{k_11}S_{k_11} + \sum_{k_1=1}^{K_1}\beta_{k_21}S_{k_22} + \theta\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\beta_{k_11}\beta_{k_21}S_{k_11}S_{k_22}. \quad (1)$$

We observe that (1) resembles a traditional logistic regression model except that the SNPs across two genes have the parsimonious "Tukey's 1 d.f" form of interaction[26,27]. Thus, postulating the biologic 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 d.f 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

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

Assuming 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 be now derived in the form

$$\begin{aligned}
&\text{logit}\left\{\Pr(D = 1|\mathbf{S_1}, \mathbf{S_2})\right\} \\
&= \alpha + \sum_{k_1=1}^{K_1} \beta_{k_1 1} S_{k_1 1} + \sum_{k_2=1}^{K_2} \beta_{k_2 2} S_{k_2 2} + \beta_{(12)1} S_{11} S_{21} + \theta \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 1} \beta_{k_2 2} S_{k_1 1} S_{k_2 2} + \\
&\quad \tau \sum_{k_2=1}^{K_2} \beta_{(12)1} \beta_{k_2 2} S_{11} S_{21} S_{k_2 2},
\end{aligned}$$

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 modelling framework can be easily extended to incorporate gene-environment interactions. Suppose the genomic region G_1 (e.g NAT2) is believed to involve a biologic pathway through which an environmental variable X (e.g smoking) may act on the risk of a disease (e.g. colorectal adenoma). Again, based on the latent variable approach, a model for the risk of the disease in terms of the marker-SNPs $\mathbf{S_1}$ and the environmental variable X can be derived in the form

$$\text{logit}\left\{\Pr(D = 1|\mathbf{S_1}, X)\right\} = \alpha + \sum_{k_1=1}^{K_1} \beta_{k_1 1} S_{k_1 1} + \sum_{p=1}^{P} \gamma_p X_p + \theta \sum_{p=1}^{P} \sum_{k_1=1}^{K_1} \beta_{k_1 1} \gamma_p S_{k_1 1} X_p,$$

where X_1, X_2,..., X_P are a set of suitably chosen design variables, such as dummy variables for categorical exposures, for representing the effects of the exposure X.

1.2 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 on the risk of the disease. For association studies, the primary goal is to establish which of the factors, if any, 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 to test for specific forms of interaction among the established risk factors. It is, however, important to realize that although the test of interaction itself may only be of 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 what follows, we develop an association testing framework involving two candidate genes G_1 and G_2. The same framework can be also 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. Possible strategies for utilizing the methodology in general association studies that may involve numerous candidate genes are discussed later.

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 of hypothesis of "no association of disease with G_1", can be statistically stated as

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

which implies conditional independence of D and G_1 given G_2. The parameter $\beta_{k_1 1}$ not only appears in the model as the "main effect" for the marker $S_{k_1 1}$, but also it contributes to all K_2 "interaction terms" that could be defined involving $S_{k_1 1}$ and the K_2 SNPs in G_2. Thus it is best to describe $\beta_{k_1 1}$, $k_1 = 1, \ldots, K_1$ as a set of "generalized association parameters" instead of 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 it is the true value for the population or not, 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 $\beta_1 = (\beta_{11}, \ldots, \beta_{K_1 1})$ can be shown to have zero expectation under the null hypothesis of $H_0^{(1)}$. Thus, for each fixed value of θ, an unbiased score-statistics 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 N_1 cases and N_0 controls have been sampled into the study and for each subject, i, the SNP-genotype vectors $\mathbf{S_{1i}}$ and $\mathbf{S_{2i}}$ have been recorded. In the following, 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 $\beta_2 = (\beta_{12}, \ldots, \beta_{K_2 2})$ 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, 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 $\beta_1 = 0$, $\psi = \hat{\psi}$.

2. For fixed value of θ, evaluate the score-functions for the parameters $\beta_{k_1 1}$, $k_1 = 1, \ldots, K_1$ at $\beta_1 = 0$ and $\psi = \hat{\psi}$ using the formula

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

   which in a vectorized form can be written as

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

   Interestingly, the score-functions (3) resemble that obtained from a standard logistic regression model, except that the "design vector" $\mathbf{S_{1i}}$ has been replaced by $\left[1 + \theta \mathbf{S_{2i}}^T \hat{\beta}_{\mathbf{2}}\right] \mathbf{S_{1i}}$,

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

$$I^{\beta_1\beta_1}(\theta) = \left\{ I_{\beta_1\beta_1}(\theta) - I_{\beta_1\psi}(\theta) I_{\psi\psi}^{-1} I_{\psi\beta_1}(\theta) \right\}^{-1}, \tag{4}$$

where the expressions for the component information matrices $I_{\beta_1\beta_1} = \partial L/\partial\beta_1\partial\beta_1^T$,$I_{\beta_1\psi}(\theta) = \partial L/\partial\beta_1\partial\psi^T$ 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 (7),(8) and (9) in the Appendix. All of these quantities can be conveniently computed using standard logistic regression software by simply setting the "design vector" for each subject to be $X = \left[1, \mathbf{S_{2i}}, \left\{1 + \theta \mathbf{S_{2i}}^T \hat{\beta}_\mathbf{2}\right\} \mathbf{S_{1i}}\right]$

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 $T_1^* = max_{L\leq\theta\leq U} T(\theta)$, where U and L denote some pre-specified values for lower- and upper-limits of θ.

Simulating the null distribution of the test statistics

In the appendix, 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 10) and $V(\theta)$ is the limit of $1/N \sum_{i=1}^N U_i(\theta)U_i^T(\theta)$. Further, under $\beta_1 = 0$, 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

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

. Following[30], 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, $i = 1, \ldots, N$ are independent standard normal random variables that are also independent of the data. The null distribution of the test statistics T_1 is then simulated by

repeatedly generating data as $T_1^0 = max_{L \leq \theta \leq U} U_0^T(\theta) I^{\beta_1,\beta_1}(\theta) U_0^T(\theta)$, where in each replication a new realization of $U_0^T(\theta)$ is obtained by re-generating the random numbers $(W_1, \ldots, W_N)$.

We also considered simulating the null distribution of T_1^* using a permutation-based re-sampling method. We randomly permuted the value of the vector $\mathbf{S_{1i}}$ over different subjects $i = 1, \ldots, N_0 + N_1$, 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 the genomic regions $\mathbf{G_1}$ and $\mathbf{G_2}$ are unlinked 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 $(D_i, \mathbf{S_{2i}})$ to be fixed, the procedure allows within-gene LD patterns and marginal association structure of D_i and G_2 to be the same as the original data.

1.3 Design for simulation Study

We study 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. For each gene, we assumed genotype data are available on six marker SNPs, none of which is the causal SNP. To simulate realistic linkage disequilibrium pattern among the markers, we utilized 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 re-sequenced using a sample of 29 Caucasian subjects at the Core Genotyping Facility of the National Cancer Institute. 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* using the original re-sequencing data. Table 1 shows the distribution of the associated haplotypes.

To define haplotypes including the causal locus, for each gene, 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 a "common" vs "rare" variant scenario 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% or 4.1% 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 the "common" vs "rare" variant scenario considered. We varied the recombination fraction (δ) at three different values to generate different degrees of LD between the causal and the 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 linkage disequilibrium 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 form

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

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 value 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 simulated data, we assumed genotype data are variable for the marker SNPs, but not on 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, we assumed all of 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, we assumed θ_1 and θ_{12} to be null, but allowed non-zero value 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* re-sampling of the genotype data of SNPs in G_1 and (2) *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) *purely epistatic* form that assumes the effect of one variant exists only in the presence of the other and vice versa; (2) *multiplicative* form which assumes the joint effect of the two variants is given by the product of the main effects[†] of the individual variants[31]; (3) *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; (4) *Cross-over* model which assumes that the second variant has no effect by itself, but 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)[‡] associated with S_1^* ranges in the set $\{1.2, 1.4, 1.6, 1.8, 2.0\}$. For the *epistatic* and *multiplicative* model, 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 *cross-over* 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 d.f. chi-square test based on a logistic regression model that involves only the main effects of the 6 marker SNPs in G_1[12] (2) LogMain&Int: An omnibus 42 d.f chi-square test based on a logistic regression model that involves 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; (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 of the three test statistics. Both type-I errors and powers were obtained empirically based on 1000 simulated data sets.

[†]Relative-risk of the disease associated with one variant in the absence of the other

[‡]Relative risk of the disease associated with one variant ignoring the presence of the other

2 Results

2.1 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 of the different settings considered.

Figures 2-5 shows 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 figures 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 to *TukAssoc* due to the use of large degrees of freedom. 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 the 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 was 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 (Figure 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* gained major power

over *LogMain*. Finally, under the cross-over model (Figure 5), where 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 it remained far inferior compared to *TukAssoc*. As observed under the *epistatic* model, the 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 to the other two procedures further increased when the significance level was chosen to be 0.0001 instead of 0.01 (see figures 6-9).

2.2 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 pre-cursor of 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 that is known to play important role in detoxification of certain aromatic carcinogens and, following N-hydroxylation, the activation of other amine-proto carcinogens to their ultimate carcinogenic form. We have recently completed a report[32] studying the association between NAT2 genetic variants and colorectal adenoma in relationship to tobacco smoking using left-sided prevalent advanced adenoma cases and gender and age-matched controls 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 which are known to be informative for reconstructing diplotypes that have been previously described and categorized in laboratory studies as having "slow", "intermediate", or "rapid" N-acetyltransferase enzymatic activity. Based on the genotype data, 685 cases and 693 controls in the study were assigned diplotype and related phenotype status using an algorithm developed in University of Louisville[28,29]. The

frequency distribution of these diplotypes and associated phenotypes are shown in Table 4. Questionnaire data on smoking history of these subjects were also available. We categorized each subject based on their smoking history as "current", "former" or "never".

Clearly, in the original study, availability of the prior data to group the numerous diplotypes into 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 are not well understood and thus the opportunity of grouping the observed genetic variants into 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 chi-square test based on a logistic regression model that involves a main effect term for each of the 14 non-referent diplotypes (df=14) (D2) *LogMain&Int*: An omnibus chi-square test based on a logistic regression model that involves the main effects of the diplotypes and all of the interactions between the diplotypes and the two non-referent categories of smoking. The null model in this test includes only the main effects of the smoking categories(df=14+14*2=42); (D3) *TukAssoc*: An omnibus test of association for *NAT2* diplotypes based on the model:

$$\text{logitPr}(D=1) = \alpha + \sum_{j=1}^{14} \beta_j I(H=h_j) + \sum_{k=1}^{2} \gamma_k I(Smk=k) + \theta \sum_{j=1}^{14}\sum_{k=1}^{2} \beta_j \gamma_k I(H=h_j) I(Smk=k)$$

where $I(H=h_j), j=1,\ldots,14$ and $I(Smk=k), k=1,2$ denote the dummy variables for the diplotypes and the smoking categories. In addition, we also performed two phenotype-based tests: (P1) a 1-d.f test for the trend-effect of the phenotype variable by coding it as a continuous variable: 0 for "slow", 1 for "medium" and 2 for "fast" and (P2) an omnibus test for the main-effect and interactions (with smoking categories) for the continuous phenotype variable (d.f=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 re-sampling, instead of the asymptotic-based method, because of 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 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 for the association than the other two methods considered. This example illustrates several 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 biologic mechanism, such as *NAT2* acetylation activity, then the proposed test of association based on Tukey's 1 d.f model for interaction could be a promising approach.

3 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 where multiple SNPs within a gene are postulated to be related to a common causal mechanism, we motivate the use of a low-dimensional 1 d.f model for gene-gene and gene-environment interactions. Based on 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 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 proposed method can improve power of genetic association tests in the presence of non-multiplicative 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.

Gene-gene and gene-environment interactions can cause the effect-size of a genetic variant to be heterogeneous for different sub-groups of the population. Tests of genetic association that ignores such heterogeneity may lack power as the "marginal" effect of a variant, ignoring sub-groups, can be quite small even though its effect can be quite large in specific subgroups. Under an extreme form of interaction, where 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 the power advantage of interaction-based tests may be negated if too many degrees of freedoms are spent to model interaction. The proposed test based on Tukey's 1 d.f model for interaction provides a good compromise between detecting large genetic-effects vs testing for many parameters.

When multiple SNPs are involved within a gene, one could attempt to reduce the degrees

of freedom for related association tests based on a "derived variable" that can combine information across multiple SNPs by utilizing prior knowledge about possible directionality of the effects of the variants[35]. The acetylation phenotype for the gene *NAT2*, utilized in our data analysis, is a "derived variable" defined based on prior data. The scope of such analysis, however, is limited for contemporary association studies due to lack of such prior data on the SNPs. The proposed method, which also utilizes "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 utilize a low degree-of-model for interaction without requiring explicit prior knowledge about the potential effects of the SNPs.

An alternative approach to reduce the degree-of-freedoms for association tests could be to follow a "two-stage" procedure where SNPs are first tested for their main effects and then interaction-based tests are considered only involving those SNPs for which main effects were found to be significant. In general, obtaining the correct type-I error rates for such sequential procedures is quite complex. A recent report has suggested a conservative but simple approach of 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 statistics 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 based on the maker SNPs of individual genes, ignoring possible gene-gene interactions, were optimal. This result can be explained mathematically by observing 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 non-multiplicative model that is often considered to be the "default" for specifying the joint effects of two exposures acting on non-overlapping 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 the causal variants were modest, the additive model corresponded to only small departure from multiplicative effects and thus *TukAssoc* performed similar to *LogMain*. In contrast, when the main effect of the causal variant in one gene was large, the additive model corresponded to large departure from multiplicative effects and *TukAssoc* became superior. The largest gains in power for *TukAssoc* over *LogMain* were seen for the *epistatic* and *cross-over* models, both of which corresponded to 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 the 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 non-null 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 before[38]

In this article, we focussed 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 potentially interact with all the others. Clearly, if all of 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 carried out a small simulation study. We used the same setting of Figure 2, but added eight null genes to the analysis. Similar to the two genes that contained the causal loci, for each of the null genes we assumed 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 re-sampling 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 carried out 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 $R^2 = 90$, $f_2 = 0.12$, and $MRR = 1.6$, the power for detecting the association of the disease with G_1 was 54% using *TukAssoc* and 34% using *LogMain*. With $f_2 = 0.04$, R^2 and MRR remaining the same, the power for *TukAssoc* became 75% while that for *LogMain* remained at 34%. In the context of a much larger scale study involving whole genome scan, a recent report has made a similar observation in that tests that account for interactions among pairs of SNPs could substantially be 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 can be reduced a priori, based on biologic knowledge, previous data or/and some pre-screening methods. Biologic knowledge of a pathway, for example, may help investigators to chose few "high-prior" candidate genes which are likely to have central roles in mediating the biologic effects of various different genetic and environmental exposures. In such 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 selection of such "high-prior" candidates.

A pre-screening 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 linkage disequilibrium 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 based on 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 of 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 for large scale association studies, involving specific biologic pathways or the whole genome.

ACKNOWLEDGEMENTS

The authors would like to thank Drs. Glen Satten, Alice Whittemore and two anonymous reviewers for their positive comments on an earlier version of this paper. Re-sequencing of the GPX3 and GPX4 gene and genotyping assays for the NAT2 gene were performed at Core Genotyping Facility at the NCI Advanced Technology Center, Gaithersburg, MD (http://cgf.nci.nih.gov)[39]. This research was supported by the Intramural Program of the National Institute of Health, USA.

4 APPENDICES

Heuristic Derivation of Tukey's 1 d.f Model of Interaction

Let $X_j = \mu_j + \sum_{k_j=1}^{K_j} \gamma_{k_j j} S_{k_j j}$, $j = 1, 2$. 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 (top boxes in figure 1) into disease-risk model (bottom box) and taking a Taylor's series expansion with respect to ϵ_1 and ϵ_2, one can write

$$\Pr_{\epsilon_1,\epsilon_2}(D = 1|\mathbf{S_1}, \mathbf{S_2}) = f(X_1, X_2) + \epsilon_1 f_1(X_1, X_2) + \epsilon_2 f_2(X_1, X_2) + O(\epsilon_1^2 + \epsilon_1^2),$$

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

$$\Pr(D = 1|S_1, S_2) \;=\; E_{\epsilon_1,\epsilon_2}\Pr_{\epsilon_1,\epsilon_2}(D = 1|S_1, S_2) = f(X_1, X_2) + O(\sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2)$$

where $\sigma_{\epsilon_j}^2$ denotes the variance of ϵ_j, $j = 1, 2$. Thus, if $\sigma_{\epsilon_j}^2$, $j = 1, 2$ are small, then $\Pr(D = 1|S_1, S_2) \approx f(X_1, X_2)$ which is precisely the model shown in (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

$$L = \sum_{i=1}^{N_1+N_0} D_i \log P_{\alpha,\beta_1,\beta_2;\theta}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) + (\mathbf{1} - \mathbf{D_i}) \log\{\mathbf{1} - \mathbf{P}_{\alpha,\beta_1,\beta_2;\theta}(\mathbf{S_{1i}}, \mathbf{S_{2i}})\} \qquad (6)$$

Under the null hypothesis that $\beta_{k_1 1} = 0$ for all k_1, the maximum-likelihood (ML) 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

$$S_\psi(\psi;\theta) = \sum_{i=1}^{N_0+N_1} \mathbf{Z_{2i}} \left\{ D_i - P_{\alpha,\beta_1=0,\beta_2;\theta}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \right\},$$

and $\mathbf{Z_{2i}} = (1, \mathbf{S_{2i}^T})^T$. 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 ML maximum likelihood estimates of ψ under $\beta_1 = 0$ be denoted by $\hat{\psi} = (\hat{\alpha}, \hat{\beta}_2)$. Further, let $\hat{P}_{Null}(\mathbf{S_1}, \mathbf{S_2})$ denote $P_{\hat{\alpha},\beta_1=0,\hat{\beta}_2;\theta}(\mathbf{S_1}, \mathbf{S_2})$. Now, for fixed value of θ, the score-function for the association parameters $\beta_{k_1 1}$, $k_1 = 1, \ldots, K_1$, evaluated under the null hypothesis that $\beta_{k_1 1} = 0$ for all k_1, can be written in the form of equation (3).

Define $Z_2 = (1, \mathbf{S_2}^T)$ to be 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

$$I_{\beta_1\beta_1}(\theta) = \sum_{i=1}^{N_0+N_1} \left[1 + \theta\hat{\beta}_2^T \mathbf{S_{2i}}\right]^2 \hat{P}_{NULL}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \left\{ \mathbf{1} - \mathbf{\hat{P}_{NULL}}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \right\} \mathbf{S_{1i}} \mathbf{S_{1i}}^{\mathbf{T}}, \quad (7)$$

$$I_{\beta_1,\psi}(\theta) = \sum_{i=1}^{N_0+N_1} \left[1 + \theta\hat{\beta}_2^T \mathbf{S_{2i}}\right] \hat{P}_{NULL}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \left\{ \mathbf{1} - \mathbf{\hat{P}_{NULL}}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \right\} \mathbf{S_{1i}} \mathbf{Z_{2i}^T} \quad (8)$$

and

$$I_{\psi,\psi} = \sum_{i=1}^{N_0+N_1} \hat{P}_{NULL}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \left\{ \mathbf{1} - \mathbf{\hat{P}_{NULL}}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \right\} \mathbf{Z_{2i}} \mathbf{Z_{2i}^T} \quad (9)$$

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 $S_{\psi,i} = \mathbf{Z_{2i}} \left\{ D_i - \hat{P}_{NULL}(\mathbf{S_{1i}}, \mathbf{S_{2i}}) \right\}$ 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 $i_{\psi,\psi}$ 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 standard Taylor 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 $U_i(\theta)$ denote the efficient influence function defined by

$$U_i(\theta) = S_{\beta_1,i}(\theta) - i_{\beta_1,\psi}(\theta)i_{\psi,\psi}^{-1}S_{\psi,i} \tag{10}$$

and $o_p(1)$ represents term that converges to zero in probability. Based on 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 $N^{-1/2}S_{\beta_1}(\theta)$, 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

Web resources: Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim.

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 U S A 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 Inc

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

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 Database issue:D617-D621.

Table 1: Haplotype frequencies used for simulating genotype data on marker SNPs for two candidate genes

G_1		G_2	
Haplotypes	Freq	Haplotypes	Freq
000000	0.3211	100010	0.3506
001101	0.1204	010001	0.2819
010000	0.0909	010100	0.1274
000001	0.0785	100000	0.0678
111001	0.0722	000000	0.0407
110001	0.0708	101100	0.0401
000010	0.0610	000010	0.0307
011001	0.0523	010010	0.0237
110000	0.0468	010000	0.0226
100000	0.0353	100001	0.0144
001000	0.0279		
010001	0.0228		

Table 2: Approximate relative-risk models used for simulating disease end points given the genotypes for two causal loci in two candidate genes G_1 and G_2

Model	$(S_1^* = 0, S_2^* = 0)$	$(S_1 \geq 1, S_2 = 0)$	$(S_1 = 0, S_2 \geq 1)$	$(S_1 \geq 1,\ S_2 \geq 1)$
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	$\phi_1 + \phi_2 - 1$
Cross-over	1	$\phi_1 (< 1)$	1	$\phi_{12} (> 1)$

*Number of copies of variant allele in the causal loci of G_1 and G_2

Table 3: Empirical significance level for test of association with region G_1

			Relative-risk for causal SNP in G_2	
R^{2*}_{geno}	f_2^{**}	Method	1.0	2.0
90%	0.04	Permutation	0.008	0.012
		Asymptotic	0.008	0.011
	0.13	Permutation	0.013	0.011
		Asymptotic	0.012	0.009
75%	0.04	Permutation	0.010	0.009
		Asymptotic	0.009	0.008
	0.13	Permutation	0.009	0.004
		Asymptotic	0.009	0.004
60%	0.04	Permutation	0.011	0.012
		Asymptotic	0.012	0.012
	0.13	Permutation	0.009	0.009
		Asymptotic	0.009	0.008

*Multilple R^2 between genotypes at causal and marker loci

*Allele frequency for causal SNP in G_2

Table 4: Distribution of cases and controls by NAT2 diplotypes and acetylation status in the PLCO adenoma study

Diplotypes	Acetylation Phenotype	Cases	Controls
*5B/*6A	0 (slow)	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 (Medium)	109	138
*4/*6A	1	86	104
*4/*7B	1	17	8
*4/*5A	1	9	6
*4/*4	2 (Rapid)	37	41
Rare		19	25

Table 5: Test of association for adenoma with *NAT2* with and without accounting for *NAT2*-smoking interaction

	Test-stat	d.f	p-value
Acetylation-based*			
LogMain	3.30	1	0.069
LogMain&Int	14.23	3	0.003
Diplotype-based**			
LogMain	18.25	14	0.200
LogMain&Int	54.41	42	0.156
TukAssoc	26.45	-	0.071

*Uses continuous phenotype variable codes as 0,1,2

**Uses diplotypes shown in Table 4

Figure legends

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

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 marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual SNPs in G_1 and G_2 follows the purely *epistatic* model (See Table 2). In Figure 2-9, f_1 and f_2 denote minor allele frequencies for causal SNP 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.

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 marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual SNPs in G_1 and G_2 follows the purely *multiplicative* model with $\phi_1 = \phi_2$(See Table 2).

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 marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual 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).

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 marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual SNPs in G_1 and G_2 follows the *cross-over* model with $\phi_1 = 0.90$ (See Table 2).

Figure 6: Empirical power at $\underline{\alpha = 0.0001}$ to detect the association of the disease with candidate gene G_1 as a function of the marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual SNPs in G_1 and G_2 follows the *purely epistatic* model

(See Table 2).

Figure 7: Empirical power at $\underline{\alpha = 0.0001}$ to detect the association of the disease with candidate gene G_1 as a function of the marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual SNPs in G_1 and G_2 follows the *purely multiplicative* model with $\phi_1 = \phi_2$ (See Table 2).

Figure 8: Empirical power at $\underline{\alpha = 0.0001}$ to detect the association of the disease with candidate gene G_1 as a function of the marginal relative-risk (MRR) of the underlying causal SNP S_1^*: The joint effect of casual 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).

Figure 9: Empirical power at $\underline{\alpha = 0.0001}$ to detect the association of the disease with candidate gene G_1 as a function of the marginal relative-risk (MRR) of the underlying causal SNP S_1^* when the joint effect of casual SNPs in G_1 and G_2 follows the *cross-over* model with $\phi_1 = 0.90$ (See Table 2).

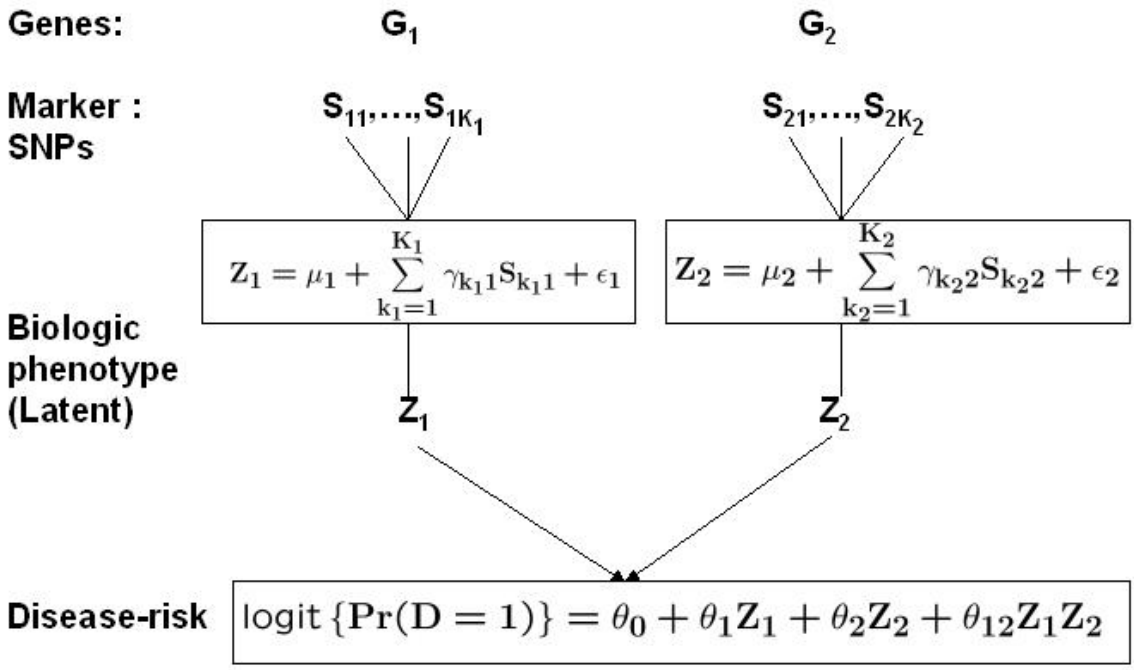

Figure 1:

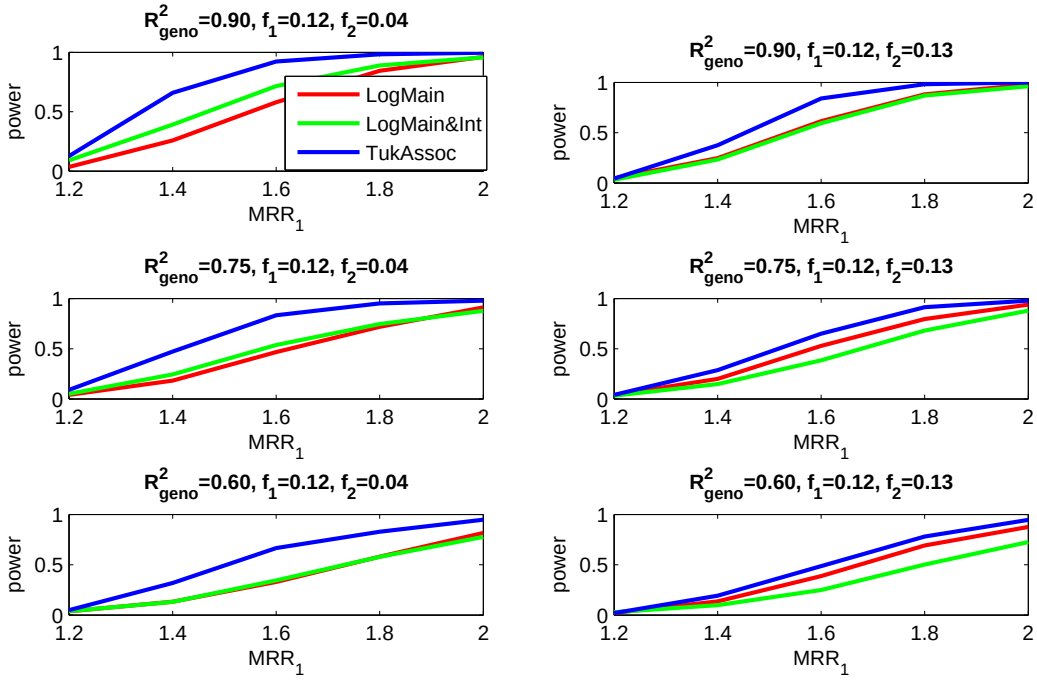

Figure 2:

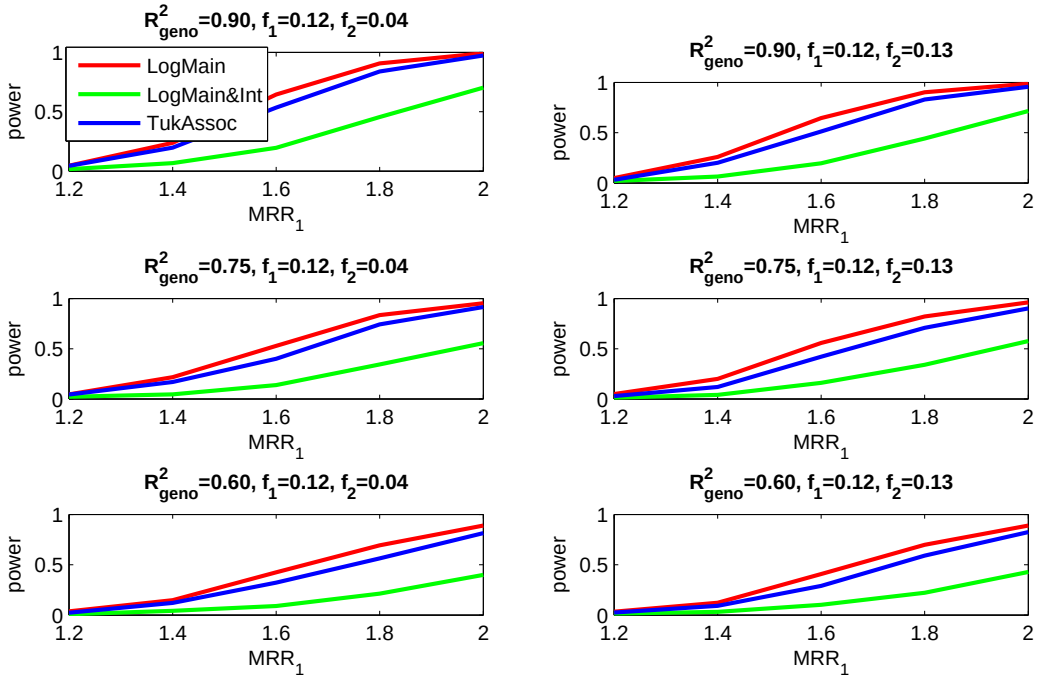

Figure 3:

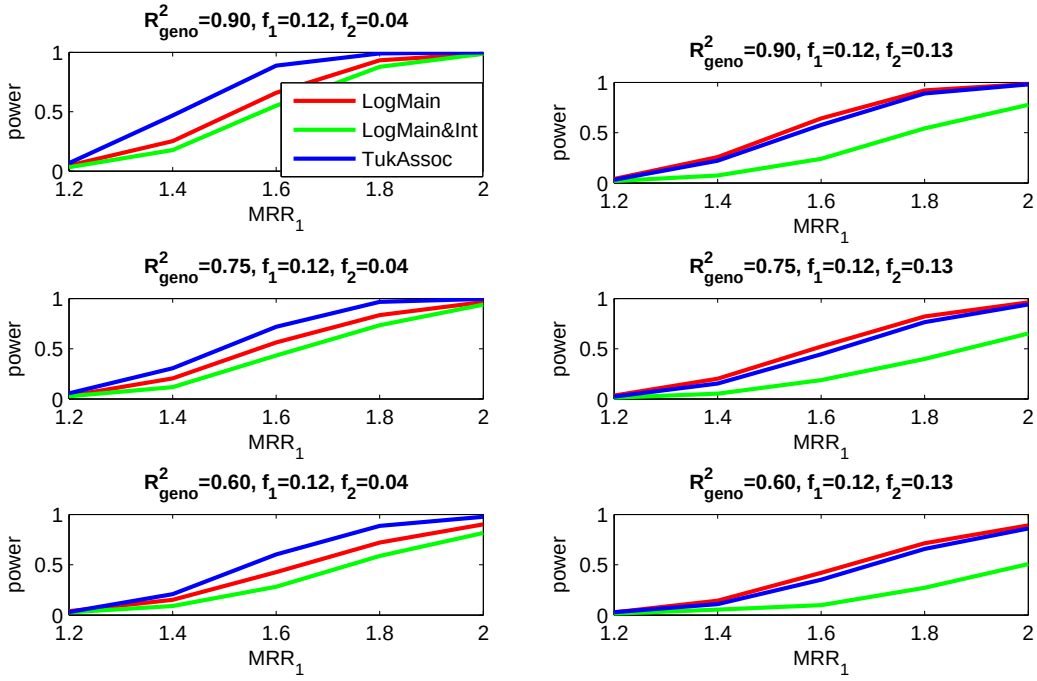

Figure 4:

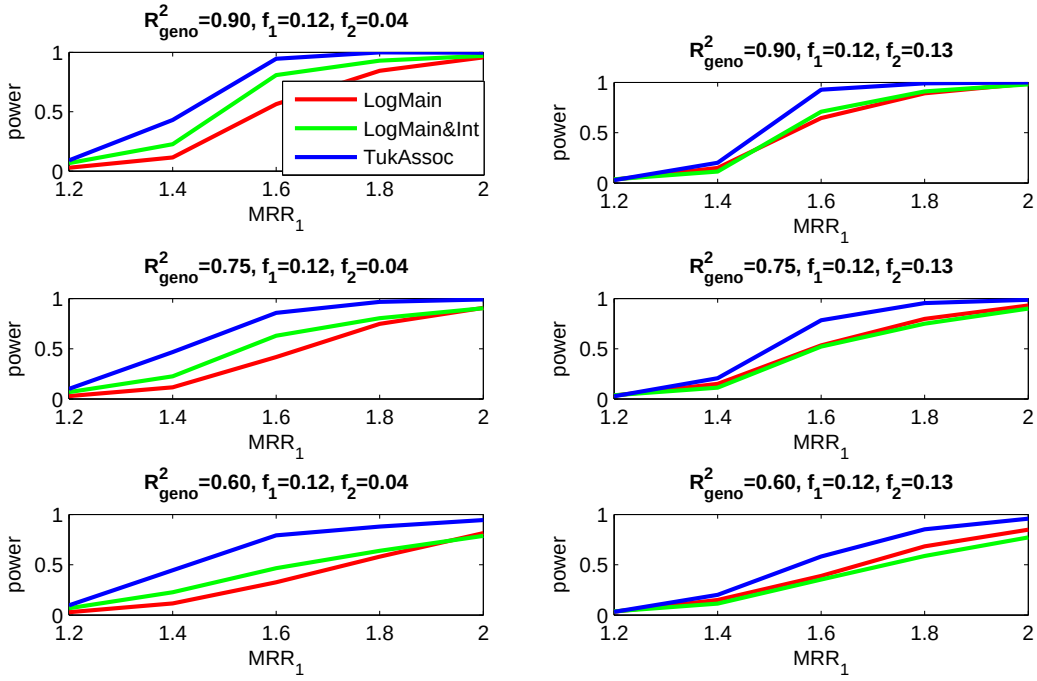

Figure 5:

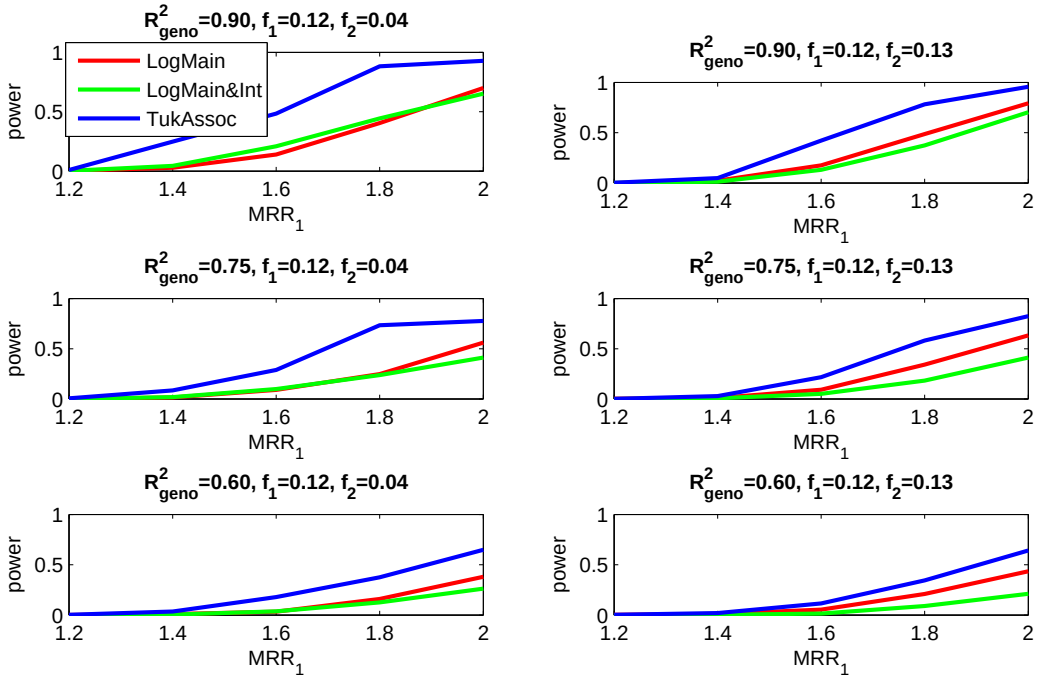

Figure 6:

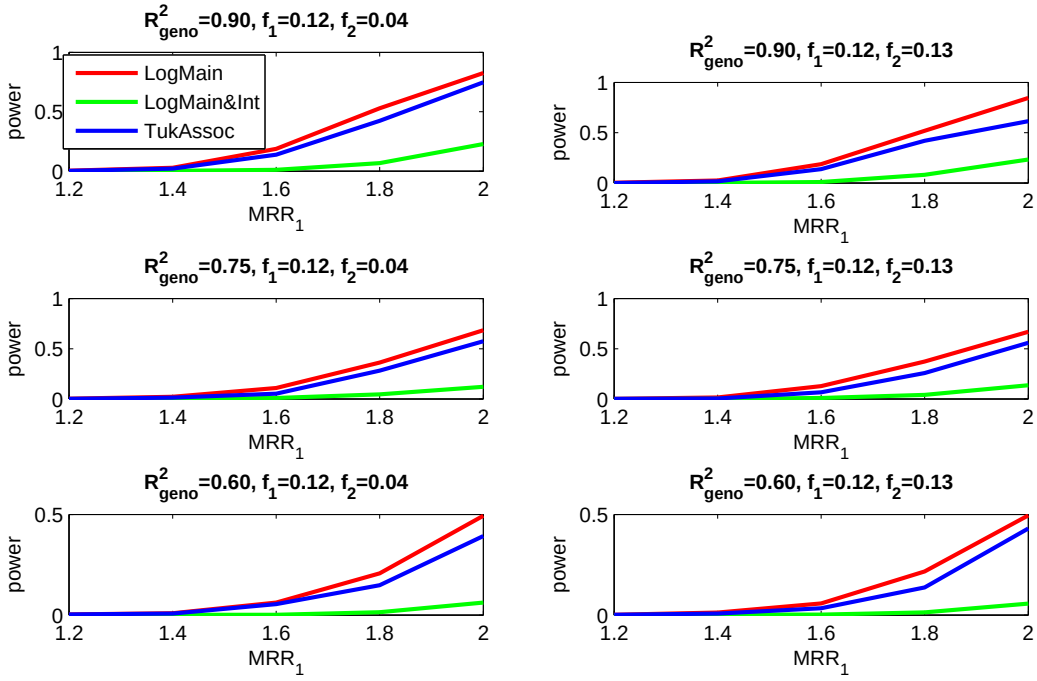

Figure 7:

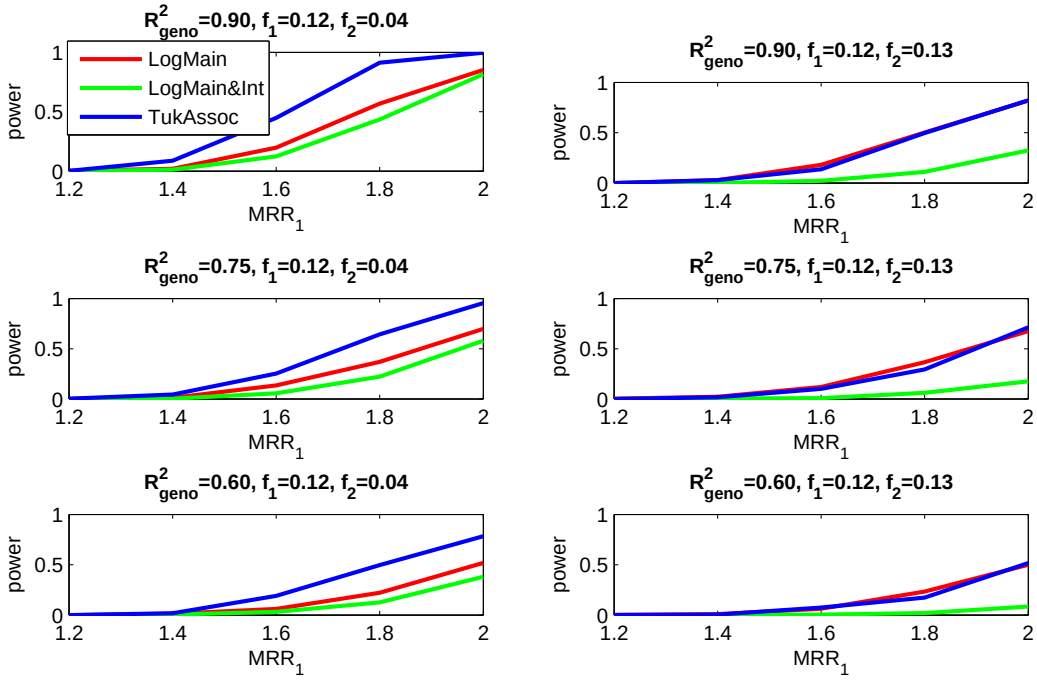

Figure 8:

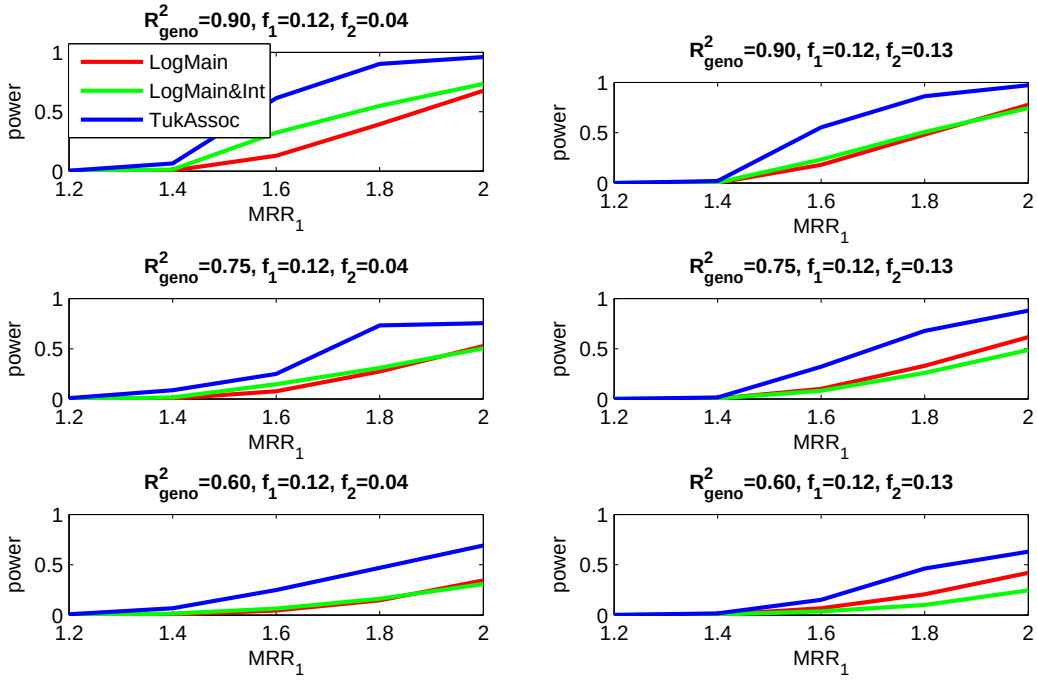

Figure 9: