Genotype-Based Association Mapping of Complex Diseases: Gene-Environment Interactions with Multiple Genetic Markers and Measurement Error in Environmental Exposures

Iryna Lobach,[1] Ruzong Fan,[2–4]* and Raymond J. Carroll[2]

[1]*Division of Biostatistics, New York University, School of Medicine, New York, New York*
[2]*Department of Statistics, Texas A&M University, College Station, Texas*
[3]*Department of Epidemiology, MD Anderson Cancer Center, University of Texas, Houston, Texas*
[4]*Division of Cancer Control and Population Sciences, Surveillance Research Program, National Cancer Institute, Rockville, Maryland*

With the advent of dense single nucleotide polymorphism genotyping, population-based association studies have become the major tools for identifying human disease genes and for fine gene mapping of complex traits. We develop a genotype-based approach for association analysis of case-control studies of gene-environment interactions in the case when environmental factors are measured with error and genotype data are available on multiple genetic markers. To directly use the observed genotype data, we propose two genotype-based models: genotype effect and additive effect models. Our approach offers several advantages. First, the proposed risk functions can directly incorporate the observed genotype data while modeling the linkage disequilibrium information in the regression coefficients, thus eliminating the need to infer haplotype phase. Compared with the haplotype-based approach, an estimating procedure based on the proposed methods can be much simpler and significantly faster. In addition, there is no potential risk due to haplotype phase estimation. Further, by fitting the proposed models, it is possible to analyze the risk alleles/variants of complex diseases, including their dominant or additive effects. To model measurement error, we adopt the pseudo-likelihood method by Lobach et al. [2008]. Performance of the proposed method is examined using simulation experiments. An application of our method is illustrated using a population-based case-control study of association between calcium intake with the risk of colorectal adenoma development. *Genet. Epidemiol.* 34:792–802, 2010.

Key words: gene-environment interactions; EM-algorithm; errors in variables; linkage disequilibrium; pseudo-likelihood; semi-parametric methods

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: National Cancer Institute; Contract grant numbers: R01-CA133996; R37-CA057030; Contract grant sponsor: National Institutes of Health.
*Correspondence to: Ruzong Fan, Department of Statistics, Texas A&M University, 447 Blocker, College Station, TX 77843-3143.
E-mail: rfan@stat.tamu.edu
Received 21 September 2009; Revised 11 June 2010; Accepted 13 June 2010
Published online 28 October 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/gepi.20523

INTRODUCTION

Case-control studies are widely used to detect gene-environment and gene-gene interactions in the etiology of complex diseases, such as cancer, hypertension, and diabetes. Many variables of interest to biomedical researchers are very difficult to measure on the individual level and oftentimes uncertainty associated with the observed values cannot be avoided in practice. Measurement error causes bias in gene-environment parameter estimates, thus masking key features of data and leading to loss of power and spurious/masked associations [Lobach et al., 2008]. Loss of power prevents the ability to detect important relationships among variables [Carroll et al., 2006]. Nutrition—defined broadly to indicate diet, body size, physical activity—is likely to be causally related to cancer [Schatzkin et al., 2009]. Nevertheless, nutritional epidemiology of cancer remains problematic, largely because of persistent concerns that standard instruments measure diet and physical activity with too much error. For example, in large epidemiologic studies of impact of diet on development of a disease, nutrient intake is commonly measured using the food frequency questionnaire (FFQ). It is well known that the FFQ as a measure of long-term diet is subject both to biases and random errors [Subar et al., 2003].

With the advent of dense single nucleotide polymorphism (SNP) genotyping, population-based linkage disequilibrium (LD) mapping or case-control association studies have become the major tools for identifying human disease genes and for the fine gene mapping of complex traits [Hinds et al., 2005; The International HapMap Consortium, 2003, 2005, 2007; The International SNP Map Working Group, 2001]. LD information reflects structure of the genome and hence provides valuable opportunity for mapping genetic variants responsible for complex diseases. The availability of millions of SNPs greatly facilitates

detection of genetic variants of complex diseases using case-control association studies and provides a unique opportunity to increase resolution of the fine genotype mapping. In the meantime, association studies are challenging because of high false-positive rate and computational demands needed to handle massive number of SNPs. Moreover, it is believed that multiple genetic variants and environmental factors interactively cause complex traits [Chatterjee and Carroll, 2005; Lobach et al., 2008; Mukherjee and Chatterjee, 2008; Spinka et al., 2005]. We are interested in building statistical models that relate genetic variations to complex phenotypes and permit detection of interaction between genetic variants and environmental factors in the difficult but common situation when the environmental factors are measured with uncertainty.

The following important issues are likely to arise in the analysis of population-based case-control association studies: (1) Genetic markers are usually typed locus by locus and oftentimes moderate to high LD exists between the observed markers. The available genetic data are in the form of unphased genotypes and hence the haplotype-based approach requires haplotype-phase estimation; (2) Some of the environmental factors that are likely to interact with the genetic traits cannot be measured directly, and instead only surrogates are available; i.e. measurement error exists in the environmental factors. To be concrete, let us describe the association study of colorectal adenoma that motivated development of our models. The Colorectal Adenoma Study was designed to investigate the interaction between the dietary calcium intake and genetic variants in the calcium-sensing receptor (CaSR) region [Peters et al., 2004]. In this study, the dietary calcium intake is not measured directly. Instead, it is estimated from a baseline FFQ. In addition, genotype information is available on three non-synonymous SNPs in the CaSR region. To detect the interaction of dietary calcium intake and the CaSR genes, two challenges exist: (1) to utilize the FFQ to measure dietary calcium intake is prone to both bias and random error, what needs to be corrected by building appropriate models and (2) to effectively analyze the observed genotype data taking into account LD information between the genetic markers.

The genotype-based and haplotype-based models are the two major approaches widely used in association studies in the presence of LD. The haplotype-based method offers an advantage of modeling LD through the construction of haplotype blocks at the price of computational expense. Lobach et al. [2008] proposed a pseudo-likelihood method in the case when the environmental factors are prone to error. The haplotype-based approach can be computationally intensive in the case when the number of genetic markers is large and the LD is moderate. Hence, in these cases, the uncertainty associated with the haplotype-based methods can lead to loss of accuracy in risk coefficients estimation [Lin et al., 2002; Marchini et al., 2006; Qin et al., 2002; Stephens et al., 2001; Stephans and Donnelly, 2003].

We propose to develop a genotype-based approach for analysis of case-control studies of gene-environment interactions. A special feature of the proposed method is that the observed genetic information enters the model directly and the LD structure is captured in the regression coefficients. As the basis for estimation and inference, we will use the pseudo-likelihood function developed by Lobach et al. [2008]. The form of this pseudo-likelihood function offers several advantages. One is that it allows to incorporate information about the probability of disease. In epidemiologic studies, a good bound on the probability of disease in a population is generally available. Further, the formulation of the pseudo-likelihood function does not require specification of the distribution of environmental variables measured exactly. These variables include age, ethnicity, BMI and other demographic and clinical measurements. Thus gains in efficiency can be achieved by not having to model a distribution of a multivariate vector of these measurements. The pseudo-likelihood function exploits the gene-environment independence assumption. If the gene-environment independence is not valid in a setting, then a distribution of genotype can be specified within strata defined by the environmental covariate. We propose a risk model that be incorporated in this pseudo-likelihood function that captures the LD structure in the regression coefficients. Hence, the haplotype phase needs not to be estimated, thus reducing computational burden and consequently reducing risk caused by potential bias dueto haplotype-phase estimation.

The organization of the paper is as follows. First, we describe the problem, our methodology and related theoretical results. We then present the results of simulation studies, and we analyze the example discussed above. In Discussion Section, we give concluding remarks. All technical derivations are given in the Appendix A and in the Web Appendix.

METHODOLOGY AND MAIN THEORETICAL RESULTS

MODEL AND NOTATION

Let D be the categorical indicator of disease status. We allow D to have $K+1$ levels with the possibility of $K \geq 1$ to accommodate different subtypes and stages of a disease. Let $D = 0$ denote the disease-free (control) subjects and $D = k$, $k \geq 1$ denote the diseased (case) subjects of the kth subtype. Suppose the genetic region of interest is spanned by I loci. Let (X,Z) denote all of the environmental (non-genetic) covariates of interest, where X are the factors susceptible to measurement error and Z are additional environmental factors measured without error. Given the environmental covariates X and Z and genotype data $\mathbf{G} = (G_1,G_2,\ldots,G_I)$, the risk of the disease in the underlying population is given by the polytomous logistic regression model

$$\mathrm{pr}(D = k \geq 1|\mathbf{G}, X, Z) = \frac{\exp\{\beta_{k0} + m_k(\mathbf{G}, X, Z; \beta)\}}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{G}, X, Z; \beta)\}}. \quad (1)$$

Here, $m_k(\cdot)$ is a known function parameterizing the joint risk of the disease from $\mathbf{G}$, X, and Z in terms of the odds-ratio parameters β. Assume that all markers are di-allelic, e.g. SNPs. Under the Hardy-Weinberg equilibrium (HWE) assumption, the distribution of the marker genotype can be specified in a parametric form $\mathrm{pr}(\mathbf{G}) = \mathrm{pr}(\mathbf{G}; \Theta)$, where $\Theta = (P_{M_i}, i = 1, 2, \ldots, I)$ are the allele frequencies. The formulation of our model is general enough to account for deviations from the HWE. For the ith marker, denote the two alleles by M_i and m_i, with frequencies P_{M_i} and P_{m_i}.

Define the following dummy variables

$$A_i = \begin{cases} 1 & \text{if } G_i = M_iM_i \\ 0 & \text{if } G_i = M_im_i \\ -1 & \text{if } G_i = m_im_i \end{cases}, \quad B_i = \begin{cases} -P^2_{m_i} & \text{if } G_i = M_iM_i \\ P_{M_i}P_{m_i} & \text{if } G_i = M_im_i \\ -P^2_{M_i} & \text{if } G_i = m_im_i \end{cases}. \tag{2}$$

Notice that A_i+1 is the number of allele M_i at the ith marker, and hence A_i can be used to model the allele or additive effect of M_i. In the following, we provide two examples of function $m_k(\cdot)$ using the genotype information $\mathbf{G} = (G_1, G_2, \ldots, G_I)$. Denote $\mathcal{A} = (A_1, \ldots, A_I)$ and $\mathcal{B} = (B_1, \ldots, B_I)$.

Genotype effect model (GEM). The following specification of the risk function incorporates both additive and dominance effects of genotype, as well as the multiplicative gene-environment.

$$\begin{aligned} m_k(\mathbf{G}, X, Z; \beta) &= m_k(\mathcal{A}, \mathcal{B}, X, Z, \beta) \\ &= X\beta_{kX} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i=1}^{I} XA_i\beta_{kAXi} \\ &\quad + \sum_{i=1}^{I} ZA_i\beta_{kAZi} + \sum_{i=1}^{I} B_i\beta_{kDi} + \sum_{i=1}^{I} XB_i\beta_{kDXi} \\ &\quad + \sum_{i=1}^{I} ZB_i\beta_{kDZi}. \end{aligned} \tag{3}$$

In this formulation, the regression coefficients β_{kA_i} and β_{kD_i} model risk due to the additive and dominance effect, respectively [Fan et al., 2006; Fan and Xiong, 2002]. The remaining terms capture the multiplicative gene-environmental interaction.

Additive effect model (AEM). Suppose that the dominance effect is not significantly present in the model (3). In this situation, the risk function takes the following form.

$$\begin{aligned} m_k(\mathbf{G}, X, Z; \beta) = m_k(\mathcal{A}, X, Z; \beta) &= X\beta_{kX} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} \\ &\quad + \sum_{i=1}^{I} XA_i\beta_{kAXi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi}. \end{aligned} \tag{4}$$

The difference between "genotype effect model" (3) and "additive effect model" (4) is that dominance effect and related gene-environmental interactions are not modeled in (4). The number of parameters in function (3) can be significantly larger than that of function (4). In practice, additive effect model (4) can be advantageous over the genotype effect model (3) because of the smaller number of parameters in (4). This situation may occur when the dominance effect is not significantly present or the dominance effect cannot compensate for the increase in the number of parameters in (3).

The model (1) cannot be used directly for analysis since the covariate X is measured with error. Let W denote the error-prone version of X. We assume a parametric model of the form $f_{\text{mem}}(w \mid D, \mathbf{G}, X, Z; \xi)$ for the conditional distribution of W given disease-status D, marker genotype $\mathbf{G}$, the true exposure X, and additional environmental factors Z. Measurement error can be modeled both as differential and non-differential. If measurement error can be assumed to be non-differential by disease status, then one can simplify the model as $f_{\text{mem}}(w \mid D, \mathbf{G}, X, Z; \xi) = f_{\text{mem}}(w \mid \mathbf{G}, X, Z; \xi)$, what does not depend on D. We assume that the joint distribution of the environmental factors in the underlying population can be specified according to a semi-parametric model of the form $f_{X,Z}(x, z) = f_X(x \mid z; \eta) f_Z(z)$, where $f_Z(z)$ is left completely unspecified.

Theoretical justification provided in the Appendix proves that the risk functions (3) and (4) are valid for analysis of case-control association studies in the case when genetic markers are in the LD. Briefly, we illustrated that (1) the LD is being modeled in the regression coefficients, and (2) if there is no association between observed genotype and trait locus, then all regression coefficients of A_i and B_i are zeros and so the regression does not depend on the markers.

SEMI-PARAMETRIC INFERENCE BASED ON A PSEUDO-LIKELIHOOD

Let n_0 be the number of control subjects; and for $k \geq 1$, denote by n_k the number of subjects in the sample with disease at a stage k. Let $n = n_0 + n_1 + \cdots + n_K$ be the total number of subjects in the sample. In addition, let us denote $\pi_k = \text{pr}(D = k)$, $k = 0, 1, 2, \ldots, K$. Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme, where the selection probability for a subject given his/her disease status $D = k$ is proportional to $\mu_k = n_k/\pi_k$. In addition, assume that the sampling only depends on the disease status, and so the selection of a subject is independent of the subject's marker information and environmental covariates.

Let $R = 1$ denote the indicator of whether a subject is selected in the sample. For the ith subject, let us denote by $(D_i, \mathbf{G_i}, X_i, W_i, Z_i, R_i)$ the observed values of variables D, $\mathbf{G}$, X, W, Z and R. Let us denote $\kappa_k = \beta_{k0} + \log(n_k/n_0) - \log(\pi_k/\pi_0)$ and $\tilde{\kappa} = (\kappa_1, \ldots, \kappa_K)^{\text{T}}$. In addition, let $\tilde{\beta}_0 = (\beta_{10}, \ldots, \beta_{K0})^{\text{T}}$, $\Omega = (\tilde{\beta}_0^{\text{T}}, \beta^{\text{T}}, \Theta^{\text{T}}, \tilde{\kappa}^{\text{T}})^{\text{T}}$, $\mathcal{B} = (\Omega^{\text{T}}, \eta^{\text{T}})^{\text{T}}$ and $v = (\eta^{\text{T}}, \xi^{\text{T}})^{\text{T}}$. Define

$$S(k, \mathbf{g}, x, z; \Omega) = \frac{\exp[1_{(k \geq 1)}(k)\{\kappa_k + m_k(\mathbf{g}, x, z; \beta)\}]}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{g}, x, z; \beta)\}} \text{pr}(g; \Theta).$$

We assume that G and (X, Z) are independently distributed in the underlying population. Only changes in notation are needed to model genotype and environment within strata thus relaxing gene-environment independence assumption. We suppose that the type of genetic covariate measured does not depend on the individual's true genetic covariate, given disease status, environmental covariates, and the measured genetic information. Further, we suppose that the observed genetic variable does not contain any additional information on disease status and true environmental covariate given the genetic variable of interest.

Similarly to Lobach et al. [2008], we propose to use the following pseudo-likelihood function in place of the likelihood function to estimate the parameters. In the Web Appendix A.1, the pseudo-probability is calculated as

$$\begin{aligned} &L_{\text{Pseudo}}(k, \mathbf{g}, w, z; \Omega, \eta, \xi) \\ &\equiv \text{pr}(D = k, \mathbf{G} = \mathbf{g}, W = w \mid Z = z, R = 1) \\ &= \frac{\int S(k, \mathbf{g}, x, z; \Omega) f_{\text{mem}}(w \mid k, \mathbf{g}, x, z; \xi) f_X(x \mid z; \eta)\, dx}{\sum_{k_1=0}^{K} \sum_{\mathbf{g} \in \mathcal{G}} \int S(k_1, \mathbf{g}, x, z; \Omega) f_X(x \mid z; \eta)\, dx}, \end{aligned} \tag{5}$$

where $\mathcal{G}$ is the set of all possible genotypes in the population. Lobach et al. [2008] proved that the maximization of

L_{Pseudo}, although not the actual retrospective-likelihood for case-control data, leads to consistent and asymptotically normal parameter estimates. Observe that conditioning on Z in L_{Pseudo} allows it to be free of the nonparametric density function $f_Z(z)$, thus avoiding the difficulty of estimating potentially high-dimensional nuisance parameters.

ASYMPTOTIC THEORY IN CASE WHEN GENETIC MARKERS ARE INDEPENDENT

In this section, we will provide asymptotic results along the lines of Lobach et al. [2008]. These results are readily applicable to the proposed model in the case when no LD is present between the genetic markers.

Estimation With Known Measurement Error Distribution. Assume that the parameter ξ controlling the distribution of the measurement error is known. We show that maximization of L_{Pseudo} leads to consistent and asymptotically normal parameter estimates. Let $\Psi(k, \mathbf{g}, w, z; \Omega, \eta, \xi)$ be the derivative of $\log\{L_{\text{Pseudo}}(k, \mathbf{g}, w, z; \Omega, \eta, \xi)\}$ with respect to $\mathcal{B}$. Then define

$$\mathcal{L}_n(\Omega, \eta, \xi) = \sum_{i=1}^{n} \Psi(D_i, G_i, W_i, Z_i; \Omega, \eta, \xi),$$

$$\mathcal{I} = -n^{-1} E[\partial\{\mathcal{L}_n(\Omega, \eta, \xi)\}/\partial\mathcal{B}^{\mathrm{T}}],$$

$$\Lambda = \sum_k \frac{n_k}{n} E\{\Psi(D, G, W, Z; \Omega, \eta, \xi)|D = k\} \times E\{\Psi(D, G, W, Z; \Omega, \eta, \xi)|D = k\}^{\mathrm{T}}$$

where all expectations are taken with respect to the case-control sampling design. The estimation $\hat{\mathcal{B}} = (\hat{\Omega}^{\mathrm{T}}, \hat{\eta}^{\mathrm{T}})^{\mathrm{T}}$ of $\mathcal{B}$ are the solution to

$$0 = \mathcal{L}_n(\Omega, \eta, \xi) = \mathcal{L}_n(\mathcal{B}, \xi). \tag{6}$$

In the Web Appendix A.2, we show the following limiting properties of $\hat{\mathcal{B}}$.

Theorem 1. The estimating function $\mathcal{L}_n(\Omega, \eta, \xi)$ is unbiased, i.e. has mean zero when evaluated at the true parameter values. In addition, under suitable regulatory conditions, there is a consistent sequence of solutions to (6), with the property that

$$n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}) \Rightarrow \text{Normal}\{0, \mathcal{I}^{-1}(\mathcal{I} - \Lambda)\mathcal{I}^{-1}\}.$$

Remark 1. An EM algorithm for the estimating the parameters, based along the lines of Lobach et al. [2008] and Spinka et al. [2005], is given in the Web Appendix A.3.

Estimated Measurement Error Distribution. In practice, the parameter ξ controlling the measurement error distribution will be unknown, and typically additional data are necessary to estimate it. Here, we consider the case of additive mean-zero measurement error with replications of W. Our convention is that there are at most J replications of the W for any individual. Let W_i denote this ensemble of the J replicates, and let t_i be the number of replicates we actually observe. Let $f_{\text{mem}}(w \mid k, \mathbf{g}, x, z, j; \xi)$ be the joint density of the first j replicates for $j = 1,\ldots,J$; $\Psi(D, \mathbf{G}, W, Z; \Omega, \eta, \xi, j)$, $\mathcal{I}_j$, and Λ_j be the matrices defined above for the case with exactly j replicates for each individual. Assume that j_i is independent of $(D_i, \mathbf{G_i}, X_i, W_i, Z_i)$ and that probability to observe j replicates is $p(j)$. Further, define $\mathcal{I} = \sum_{j=1}^{J} p(j)\mathcal{I}_j$. The estimating function for $\mathcal{B} = (\Omega^{\mathrm{T}}, \eta^{\mathrm{T}}, \xi)^{\mathrm{T}}$ can be written [Lobach et al., 2008] in the form

$$0 = \sum_{i=1}^{n} \sum_{j=1}^{J} 1_{(t_i=j)}(t_i)\Psi(D_i, \mathbf{G_i}, W_i, Z_i; \Omega, \eta, \xi, j). \tag{7}$$

The parameter estimates have the following asymptotic properties (Web Appendix A.4).

Theorem 2. The estimating function (7) is unbiased, i.e., has mean zero when evaluated at the true parameter values. In addition, under suitable regulatory conditions, there is a consistent sequence of solutions to (7), with the property that

$$n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}) \Rightarrow \text{Normal}\left[0, \mathcal{I}^{-1}\left\{\mathcal{I} - \sum_{j=1}^{J} p(j)\Lambda_j\right\}\mathcal{I}^{-1}\right]. \tag{8}$$

ASYMPTOTICS IN CASE WHEN GENETIC MARKERS ARE IN LINKAGE DISEQUILIBRIUM

Recall that in the case when genetic markers modeled in the risk function are in the LD, the regression coefficients capture both the association signal and the LD information. In practice, if the LD is present in the genetic material, we suggest to construct and investigate the model based on genetic markers that are known to be associated with disease. The association information can be obtained either based on a priori knowledge (e.g. previously reported studies, biological interpretation), or can be inferred using the observed data (e.g. model selection procedure).

The distribution of the genotype in the population for a pair of markers that are in LD can be written as follows:

$$\text{pr}(\mathbf{G}|\theta, \Delta) = \begin{cases} P(M_1)P(M_2) + \Delta_{M1M2}, & \mathbf{G} = (M_1M_2) \\ \{1 - P(M_1)\}P(M_2) - \Delta_{M1M2}, & \mathbf{G} = (m_1M_2) \\ P(M_1)\{1 - P(M_2)\} - \Delta_{M1M2}, & \mathbf{G} = (M_1m_2) \\ \{1 - P(M_1)\}\{1 - P(M_2)\} + \Delta_{M1M2}, & \mathbf{G} = (m_1m_2). \end{cases}$$

Define

$$S(k, \mathbf{g}, x, z; \Omega, \Delta) = \frac{\exp[1_{(k\geq 1)}(k)\{\kappa_k + m_k(\mathbf{g}, x, z; \beta)\}]}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{g}, x, z; \beta)\}} \text{pr}(\mathbf{g}; \Theta, \Delta). \tag{9}$$

We propose to estimate parameters $(\Omega, \eta, \xi, \Delta)$ based on the pseudo-likelihood function that is of the same form as (5), but is based on the $S(k, \mathbf{g}, x, z; \Omega, \Delta)$ function in the form (9), specifically,

$$L_{\text{Pseudo}}(k, \mathbf{g}, w, z; \Omega, \eta, \xi, \Delta) \equiv \text{pr}(D = k, \mathbf{G} = \mathbf{g}, W = w|Z = z, R = 1) = \frac{\int S(k, \mathbf{g}, x, z; \Omega, \Delta) f_{\text{mem}}(w|k, \mathbf{g}, x, z; \xi) f_X(x|z; \eta)\, dx}{\sum_{k_1=0}^{K} \sum_{\mathbf{g}\in\mathcal{G}} \int S(k_1, \mathbf{g}, x, z; \Omega, \Delta) f_X(x|z; \eta)\, dx}. \tag{10}$$

Justification of the risk model and regression coefficients presented in the first section of the Appendix A suggests that the LD information between the observed genetic markers and the trait locus is captured in the regression coefficients.

SIMULATION EXPERIMENTS

We performed a series of simulation experiments to investigate the performance of the proposed procedure in various settings. We consider a case when the disease status D is binary (i.e. $K = 1$). Note that

$$\begin{aligned}\mathrm{pr}(D=1) &= \int \sum_{\mathbf{g}\in\mathcal{G}} \mathrm{pr}(D=1|\mathbf{g},x)\mathrm{pr}(\mathbf{g};\Theta) f_X(x|\eta)\,\mathrm{d}x \\ &= \int \sum_{\mathbf{g}\in\mathcal{G}} \frac{\exp\{\beta_0+m_1(\mathbf{g},x;\beta)\}}{1+\exp\{\beta_0+m_1(\mathbf{g},x;\beta)\}}\mathrm{pr}(\mathbf{g};\Theta) f_X(x|\eta)\,\mathrm{d}x.\end{aligned} \tag{11}$$

Thus, the parameters ($\beta_0 = \beta_{10}$, β, Θ, η) are sufficient to identify $\mathrm{pr}(D = 1)$, i.e., $\kappa = \kappa_1$ is identified from (β_0, β, Θ, η). This means that simply using (5) as a likelihood function directly will be unstable because of over-parametrization. To overcome this, we may re-parametrize in terms of $\mathrm{pr}(D = 1)$ through (11). In addition, let κ be a function of $\mathrm{pr}(D = 1)$. This obvious solution can solve the over-parametrization problem.

The genotype $\mathbf{G}$ was simulated under HWE for $I = 1, 2, 3$. Given the values of ($\mathbf{G}$, X), we generated a binary disease outcome D using two logistic models, corresponding to the GEM and AEM. For the GEM, covariates are related to a disease via link function

$$\begin{aligned}\mathrm{logit}\{\mathrm{pr}(D=1|\mathcal{A},\mathcal{B},X)\} = \beta_0 + X\beta_X + \sum_{i=1}^{I} A_i\beta_{Ai} \\ + \sum_{i=1}^{I} XA_i\beta_{AXi} + \sum_{i=1}^{I} B_i\beta_{Di} \\ + \sum_{i=1}^{I} XB_i\beta_{DXi}, \quad I = 1,2,3,\end{aligned}$$

and the corresponding AEM was obtained by setting coefficients β_{Di} and β_{DXi} to be 0. Here we omit the subscription k in the regression parameters βs since we have one level disease cases and normal controls.

We considered three settings of the disease risk function. In the first setting, only one marker is involved in a disease. This marker has weak additive and dominance effects ($\beta_{A1} = \log(1.5)$ and $\beta_{D1} = \log(1.3)$), while the interaction effect with the environment is strong for both additive and dominance components ($\beta_{XA1} = \log(2.5)$ and $\beta_{XD1} = \log(3)$). In the second setting, in addition to the marker described above, we added one with stronger additive component ($\beta_{A2} = \log(2.2)$) and almost no dominance $\beta_{D2} = \log(1.1)$. Interaction effects of both additive and dominance components are strong ($\beta_{AX2} = \log(2.2)$ and $\beta_{DX2} = \log(2.5)$). Finally, in the third setting, we added one additional marker with strong additive and dominance components in addition to the strong interaction effects $\beta_{A3} = \log(2)$, $\beta_{AX3} = \log(3)$, $\beta_{D3} = \log(2)$, $\beta_{DX3} = \log(3)$.

THE DISCRETE CASE

Consider the case when disease status and environmental variables (X, W) are binary. Let $\mathrm{pr}(X = 1) = 0.5$. We simulated the observed environmental variable W using the following misclassification probabilities. $\mathrm{pr}(W = 0 \mid X = 1) = 0.25$ for the exposed participants and $\mathrm{pr}(W = 1 \mid X = 0) = 0.20$ for the non-exposed. We performed a simulation sub-study when probability of disease is not known and it is estimated via grid-search method. The values of π_d are set to be on interval (0.001, 0.04) with step 0.005 and the resulting estimate is a value that maximizes the pseudo-likelihood function. Parameter β_0 is estimated by solving equation (11). To estimate the parameters, 500 samples are simulated and each sample contains 1,000 cases and 1,000 controls. To illustrate performance and advantages of the proposed method, we presented biases and Root Mean-Squared Errors (RMSE).

Shown in the Table I are simulation results for the case when three independent genetic markers are observed. The results illustrate that the proposed methodology produced parameter estimates that are nearly unbiased and have small variability. The naive approach that ignores existence of the measurement error and pretends that the environment is observed exactly results in biased estimates with variability that is larger than that of the proposed approach. These simulation results illustrate that the RMSE of coefficients β_{Di} and β_{DXi} are generally larger than those of β_{Ai} and β_{AXi}, thus suggesting that the dominance effect should only be used in the situations when the data present strong evidence for the dominance effect. In Web Tables I and II, we present the results for two-marker case, additive and genotype effect models, respectively. Findings shown in the Web Tables I and II are similar to those of the Table I.

The setup described above simulates markers that are in linkage equilibrium. To evaluate performance of the proposed method in the case when genetic markers are in the LD, we considered the following simulation setup. We simulated the observed genotype according to the following frequencies:

$$\mathbf{G} = \begin{cases} M_1M_2, & P(M_1)P(M_2) + \Delta_{M1M2} \\ m_1M_2, & \{1-P(M_1)\}P(M_2) - \Delta_{M1M2} \\ M_1m_2, & P(M_1)\{1-P(M_2)\} - \Delta_{M1M2} \\ m_1m_2, & \{1-P(M_1)\}\{1-P(M_2)\} + \Delta_{M1M2}. \end{cases}$$

We further compared performance of the proposed approach and the one that ignores existence of the LD. We found that when the LD is small (e.g. 0.005), the parameter estimates based on the model (5) are nearly unbiased and have small variability (Web Table III), because the coefficients capture enough of the LD information. To test the performance of the proposed models in the case when moderate amount of LD is present ($\Delta = 0.01$, 0.02, 0.05) and compare it to the procedure that ignores existence of the LD, we performed the following simulation experiment. We simulated the observed data using a simulation setup described above with two markers that are in LD using AEM. Results presented in the Table II illustrate that the naive approach resulted in parameter estimates that are biased and are highly variable, while the proposed method eliminated bias and substantially reduced the variability of the estimates.

CONTINUOUS SIMULATIONS

In this simulation experiment, we considered a continuous environmental variables. We simulated the true

TABLE I. Biases and RMSEs of risk parameters for the naive approach that ignores existence of measurement error and the proposed method in the case when pr($D = 1$) is known and when it is estimated

		Naive analysis		Proposed approach			
		pr($D = 1$) is known		pr($D = 1$) is known		pr($D = 1$) is unknown	
Parameter	True value	Bias	RMSE	Bias	RMSE	Bias	RMSE
κ	0.484	0.481	0.231	−0.048	0.019	−0.054	0.020
β_X	0.693	−0.351	0.132	−0.005	0.036	0.014	0.039
β_{A1}	0.406	0.257	0.073	−0.006	0.016	−0.011	0.016
β_{A2}	0.789	0.194	0.046	0.002	0.015	−0.003	0.015
β_{A3}	0.693	0.283	0.089	−0.0002	0.015	−0.005	0.016
β_{AX1}	0.916	−0.425	0.193	0.012	0.040	0.039	0.046
β_{AX2}	0.693	−0.317	0.113	0.009	0.037	0.038	0.041
β_{AX3}	1.099	−0.515	0.282	0.011	0.053	0.039	0.058
β_{D1}	0.262	0.299	0.133	0.024	0.149	0.026	0.152
β_{D2}	0.095	0.258	0.105	−0.0003	0.097	0.005	0.099
β_{D3}	0.693	0.231	0.092	0.012	0.124	0.018	0.128
β_{DX1}	1.099	−0.495	0.326	−0.002	0.287	0.018	0.301
β_{DX2}	0.916	−0.413	0.235	0.019	0.197	0.006	0.208
β_{DX3}	1.099	−0.486	0.313	0.016	0.275	0.023	0.286
P_{M_i}	0.250	<0.001	<0.001	<0.001	<0.001	−0.001	<0.001
pr($X = 1$)	0.500			0.003	0.001	0.003	0.001
pr($D = 1$)	0.005					0.003	<0.001

The results are based on 500 samples of 1,000 cases and 1,000 controls. Genotype is simulated at the three marker loci with $P_{M_i} = 0.25$, $i = 1$, 2, 3. The environmental covariate (X) is binary and measured with error with misclassification probabilities being 0.20 for exposed and 0.25 for non-exposed subjects. The data are simulated and analyzed under the genotype effect model. RMSE, root mean-squared error.

TABLE II. Biases and RMSEs of risk parameters for the naive approach that ignores existence of the LD and the proposed method

		Naive approach		Proposed method	
Parameter	True value	Bias	RMSE	Bias	RMSE
κ	−5.000	−0.060	0.024	−0.024	0.017
β_X	1.099	0.063	0.045	0.005	0.038
β_{A1}	0.693	−0.043	0.019	−0.007	0.013
β_{A2}	0.000	−0.046	0.022	0.002	0.014
β_{AX1}	0.693	0.169	0.064	0.005	0.028
β_{AX2}	0.693	0.113	0.049	0.005	0.026
P_{M_i}	0.250	<0.001	<0.001	<0.001	0.001
pr($X = 1$)	0.500	0.002	0.001	<0.001	0.001

The results are based on 500 samples of 1,000 cases and 1,000 controls. Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1,2$. The environmental covariate (X) is binary and measured with error with misclassification probabilities with misclassification probabilities being 0.20 for exposed and 0.25 for non-exposed subjects. Probability of disease is 0.0069 and is assumed to be known in the population. The data are simulated and analyzed under the additive effect model and the LD measure $\Delta_{m_1 m_2} = 0.02$. RMSE, root mean-squared error.

environmental covariate X from a Normal distribution with zero mean and variance 0.1. To simulate observed environmental variables, we used additive model of the form $W = X + U$, where U is generated from the normal distribution with zero mean and variance $\xi = 0.25$. Note that we are simulating a case of large measurement error, to mimic a situation that occurs in practice while measuring diet. To estimate the probability of disease, we used grid-search method on the interval (0.001, 0.051) with step 0.005 by maximizing the pseudo-likelihood function for values of probability of disease fixed on a grid and then performing a grid-search to identify the value of probability of disease that maximized the likelihood.

Within this simulation setup, we suppose that the measurement error distribution is known. Table III presents the results of three-marker case under the additive effect model. We found that for our method there is no noticeable bias in parameter estimates, whereas the naive approach that ignores existence of the measurement error results in substantial bias (Table III). The RMSEs of coefficients β_{Ai} in Table III are reasonable. However, the RMSEs of β_{AXi} in Table III are generally larger because they are based on the continuous covariate with noise that is 2.5 times more variable than the signal. We are giving a very stringent test to our method in the case when the environmental covariate is continuous because in practice the measurement error is massive. In the Web Tables IV and V, we present the results of one marker case for additive effect model and genotype model, respectively; in Web Table VI, we report the results of two-marker case for the additive effect model. The three Web Tables provide similar results as those of Table III.

To investigate accuracy of the proposed variance estimator, we performed an experiment and reported results in Table IV. The results suggest that the mean estimated standard error is nearly unbiased. However, the variability of the parameter estimates is elevated. This phenomena is well known in the measurement error

TABLE III. Biases and RMSEs of risk parameters for the naive approach that ignores existence of measurement error and the proposed method

		Naive approach		Proposed method	
Parameter	True value	Bias	RMSE	Bias	RMSE
κ	−5.000	0.001	0.050	−0.069	0.090
β_X	1.099	−0.775	0.780	0.018	0.319
β_{A1}	0.693	0.019	0.056	−0.003	0.057
β_{A2}	0.000	0.029	0.064	<0.001	0.061
β_{A3}	0.693	0.038	0.067	−0.001	0.060
β_{AX1}	0.693	−0.483	0.489	0.010	0.251
β_{AX2}	0.693	−0.485	0.493	0.005	0.277
β_{AX3}	1.099	−0.775	0.778	0.001	0.247
P_{M_i}	0.250	<0.001	0.005	<0.001	0.005
μ_x	0.000			0.002	0.090
σ_X^2	0.100			<0.001	0.011

The results are based on 500 samples of 1,000 cases and 1,000 controls. Genotype is simulated at the three marker loci with $P_{M_i} = 0.25$, $i = 1, 2, 3$. The environmental covariate (X) is Normal(μ_x, σ_x^2). The measurement error is additive with variance of noise that is 0.25. The data are simulated and analyzed under the additive effect model. RMSE, root mean-squared error.

TABLE IV. Standard errors (SE) of risk parameters for the proposed approach

Parameter	True value	SE of parameter estimates	Mean estimated SE	True SE
β_X	1.099	0.308*	0.284	0.293
β_{A1}	0.693	0.073	0.092	0.092
β_{AX1}	0.693	0.266*	0.255	0.252
β_{A2}	0.000	0.076	0.081	0.082
β_{AX2}	0.693	0.308*	0.278	0.278

Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1,2$. The environmental covariate (X) is Normal (0, 0.1). The measurement error is additive with variance of noise that is 0.25. The data are simulated and analyzed under the additive effect model. 5%-trimmed values are marked by *. The results are based on 500 samples of 1,000 cases and 1,000 controls.

literature and noted in our previous work [Lobach et al., 2008]. When the measurement noise is large, which is the case in our situation, the sampling distribution of the parameter estimates can be skewed. Hence, we reported 5%-timmed standard errors of the parameter estimates that are close to the true and mean estimated standard errors.

COLORECTAL ADENOMA STUDY DATA ANALYSIS

SINGLE MARKER ANALYSIS

Model. To illustrate the application of the proposed method we analyzed the colorectal adenoma study described in the introduction. To recap, there were 772 cases and 778 controls, the response D was colorectal adenoma status, the genetic data observed were three SNPs in the calcium receptor gene CaSR, the environmental variable X measured with error was log(1+calcium intake), which was measured by W, the result of a FFQ. The variables Z measured without error were age, sex, and race, which are not significant and are not included in the final model. The two alleles at the first SNP are A and G, the two alleles at the second SNP are C and G, and the two alleles at the third SNP are G and T. Let us denote $M_1 = A$, $m_1 = G$, $M_2 = C$, $m_2 = G$, $M_3 = G$ and $m_3 = T$; and then define dummy variables A_i accordingly by (2). Based on the observed genetic data only, we estimated the LD measures as follows: $\Delta_{M_1M_3} = 0.011$ for the first and third markers; $\Delta_{M_2M_3} = 0.012$ for second and third markers; and $\Delta_{M_1M_2}$ is approximately zero. Given calcium intake X and genotype information $\mathbf{G} = (G_1, G_2, G_3)$, we considered several risk model based on the following strategy. We first analyzed the three observed markers using the risk models based on one marker at a time in the following form

$$\text{logit}\{P(D = 1|\mathbf{G}, X)\} = \beta_0 + X\beta_X + A_i\beta_{Ai} + XA_i\beta_{AXi}, \quad i = 1, 2, 3,$$

and the model is denoted as AEM1, AEM2, and AEM3, respectively.

Measurement Error Modeling. Unfortunately, there is no direct information in the study to assess the measurement error properties of calcium intake W. We used a combination of outside data and sensitivity analysis instead. The outside data came from the WISH Study [Potischman et al., 2002]. There were ≈400 women in this study, which used the same FFQ as in the colorectal adenoma study and also included the results of six 24-hr recall measurements, which we denoted by T_{ij} for the ith individual and jth replicate. The model for these data were that

$$W_i = \alpha_0 + \alpha_i X_i + U_i,$$
$$T_{ij} = X_i + V_{ij},$$

where $U_i = \text{Normal}(0, \sigma_u^2)$ and $V_{ij} = \text{Normal}(0, \sigma_v^2)$. Using variance components analysis, we estimated $(\alpha_0, \alpha_1, \sigma_u^2)$, and took these as fixed and known in the colorectal adenoma study, although we also varied σ_u^2. The distribution of X was taken to be Gaussian with mean linear in Z and variance ξ. We used the method of Fuller 1987, Chapter 2, 5 and found estimates $\hat{\alpha}_0 = 0.22$, $\hat{\alpha}_1 = 0.75$, $\hat{\sigma}_u^2 = \hat{\xi} = 0.65$. To assess sensitivity to the measurement error model specification, we considered several scenarios by imposing measurement error structure estimated using WISH data and varying it through σ_u^2.

Results. After fitting the three models, AEM1, AEM2, and AEM3, we found significant results for AEM1 and AEM2 based on analysis of parameter estimates and 95% Wald confidence intervals (Table V). For the AEM1, each of the three regression coefficients β_{A1}, β_X, and β_{AX1} was significantly different from 0; and so was each of the three regression coefficients β_{A2}, β_X and β_{AX2} for AEM2 (Table V). Since the estimate −0.478 of parameter β_{AX1} was negative in AEM1 and so the estimate −0.771 of β_{AX2} in AEM2, the results suggested protective effect of an interaction between the calcium intake and additive effect of allele $M_1 = A$ of the first marker and allele $M_2 = C$ of the second marker.

For AEM3, none of the three regression coefficients β_{A3}, β_X and β_{AX3} was significant (data not shown). In addition,

TABLE V. Estimates and 95% Wald confidence intervals of parameters for the colorectal adenoma study

Model	Parameter	Estimate	Confidence interval
AEM1	β_{A1}	−0.310	(−0.524, −0.096)
	β_X	−0.687	(−0.977, −0.400)
	β_{AX1}	−0.478	(−0.699, −0.256)
AEM2	β_{A2}	−0.636	(−0.891, −0.381)
	β_X	−0.986	(−1.321, −0.651)
	β_{AX2}	−0.771	(−1.015, −0.527)
AEM12	β_{A1}	−0.273	(−0.581, 0.035)
	β_{A2}	−0.568	(−0.901, −0.235)
	β_X	−1.253	(−1.522, −0.985)
	β_{AX1}	−0.430	(−0.757, −0.103)
	β_{AX2}	−0.664	(−0.983, −0.344)

we added dominance effect and the related gene-environmental interaction terms to the AEM1, AEM2, and AEM3 to fit the data. No significant result was found for the dominance effect and the related gene-environmental interactions (data not shown). Thus, the additive effect models (AEM1 and AEM2) are enough to fit the data.

MULTIPLE MARKER ANALYSIS

Motivated by the results of single marker analysis, we considered the following risk model that models the effect of a pair of markers 1 and 2

$$\text{logit}\{P(D=1|\mathbf{G},X)\} = \beta_0 + X\beta_X + A_1\beta_{A1} + A_2\beta_{A2} + XA_1\beta_{AX1} + XA_2\beta_{AX2},$$

which we denoted as AEM12. The results are presented in the Table V. All regression coefficients except β_{A1} in AEM12 were significantly different from 0. The results of Table V confirmed the protective effects of the allele $M_1 = A$ at the first marker and the allele $M_2 = C$ at the second marker.

In addition to AEM12 using markers 1 and 2 in analysis, we fitted two more models: (1) AEM13: markers 1 and 3; and (2) AEM23: markers 2 and 3, respectively. The results, however, suggested that the third marker does not provide significant effect on the models AEM13 and AEM23 (data not shown). Hence, markers 1 and 2 are enough to model the association with the disease trait, and marker 3 contributed no significant information in addition to marker 1 or 2.

COMPARISON WITH THE RESULTS OF HAPLOTYPE-BASED APPROACHES

In Lobach et al. [2008], two haplotypes h_4 and h_5 have protective effects against colorectal tumor development and a haplotype h_2 has no significant effect. h_4 is haplotype *GCG*, h_5 is a combination of a common haplotype *AGG* plus three rare haplotypes *AGT*, *GGG*, and *GCT*. Compared with our results that the allele $M_1 = A$ at the first marker and the allele $M_2 = C$ at the second marker have protective effects, the protective effect of $h_4 = GCG$ may be due to the allele $M_2 = C$ at the second marker and the protective effect of h_5 may be due to the allele $M_1 = A$ at the first marker. The alleles at the third marker make no significant contribution to the colorectal tumor development.

DISCUSSION

In this article, we proposed a genotype-based approach for the analysis of case-control studies of gene-environment interactions in the presence of measurement error in the environmental factors. Two types of risk functions are proposed along the lines of the previous work of the second author: genotype and additive effect models [Fan et al., 2006; Fan and Xiong, 2002]. The genotype effect models capture both the additive and dominance effects, while the additive effect model takes into account the additive effect of genetic markers. The proposed method has several unique aspects. First, the observed genetic information enters the model directly and the pairwise LD structure is captured in the regression coefficients. Compared with the haplotype-based approaches, the proposed models are simpler and the theoretical justification/inference/asymptotics can be simpler too. Subsequently, in practice, the computational burden is less demanding since there is no need to estimate the haplotype phase. Moreover, there is no risk of potential bias due to haplotype phase estimations. The second unique aspect of the proposed method efficiently models the environmental covariates are measured with substantial error. So far, there is no genotype-based methods in literature to deal with the issue and the proposed method can fill the gaps. Similary to the method investigated in Lobach et al. [2008], the estimating procedure is based on a pseudo-likelihood model that allows to efficiently estimate parameters, model environmental covariates completely non-parametrically, and incorporate information about the probability of disease. In epidemiologic studies, the vector of environmental covariates measured exactly is oftentimes high dimensional and a good estimate about probability of disease in a population is known. Thus, the use of the proposed pseudo-likelihood function offers advantages.

Simulation experiments illustrated that the proposed methods can lead to nearly unbiased parameter estimates and the variability is generally low for the additive effect terms. In the genotype effect models, the variability of the dominance effect terms can be slightly elevated. Hence, in the case when dominance effect is moderate or is not significant, the additive effect models are superior to the genotype effect models as we noted in our previous work [Fan et al., 2006; Fan and Xiong, 2002]. In comparison, the naive estimation that ignores existence of measurement error and/or LD results in parameter estimates that are largely biased and variable.

The proposed methods will prove useful when the amount of LD is small or moderate. In this case, the number of possible haplotypes that are consistent with the observed genotype is large and hence the estimating procedure can be computationally intensive. Simulation experiments illustrated that the proposed method resulted in parameter estimates that are nearly unbiased and have small variability in the cases: (1) the LD is small and the LD is only modeled in the regression coefficients; (2) the LD is moderate and the proposed method estimates that LD and risk parameters simultaneously. When the number of markers is much larger than two and the strong LD is present, haplotype-based approaches developed in Lobach et al. [2008] can be more useful. In particular, if a disease is mainly due to one or two haplotypes and the haplotype consists of multiple alleles at different markers and none of the alleles has big

effect on the disease, the proposed genotype-based approach can be less powerful while the haplotype-based approach can be more powerful. On the other hand, the proposed models can be powerful when some alleles have strong effect on the disease.

As a result of application of the proposed method to the analysis of colorectal adenoma study, we found that the protective effects of the allele $M_1 = A$ at the first marker and the allele $M_2 = C$ at the second marker. In Lobach et al. [2008], two haplotypes $h_4 = GCG$ and h_5 (mainly AGG) have protective effects against colorectal tumor development. Compared with our results, the protective effect of h_4 may be due to the allele $M_2 = C$ at the second marker and the protective effect of h_5 may be due to the allele $M_1 = A$ at the first marker. Therefore, the proposed approach has the ability to identify important alleles and markers, which have significant contribution to the colorectal tumor development.

The proposed risk model is readily expendable for the analysis of gene-gene interactions in the case when the genetic markers involved in an interaction are independent. For the colorectal adenoma study data, we fit the model by adding the product terms A_1A_2 of dummy variables A_1 and A_2 but no significant gene-gene interaction effect was found.

ACKNOWLEDGMENTS

The research of Fan was supported by the National Cancer Institute grant R01-CA133996 from MD Anderson Cancer Center (P.I. C. Amos), where Fan spent his sabbatical during the 2008–2009 academic year; and a Research and Travel Support from the Intergovernmental Personnel Act (IPA), National Institutes of Health, in 2010. The research of Carroll was supported by a grant from the National Cancer Institute (R37-CA057030).

REFERENCES

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. Measurement Error in Nonlinear Models, 2nd edition. London, Boca Raton: Chapman & Hall, CRC Press.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation in case-control studies of gene-environmental interactions. Biometrika 92:399–418.

Fan R, Xiong M. 2002. High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. Eur J Hum Genet 10:607–615.

Fan R, Jung J, Jin J. 2006. High resolution association mapping of quantitative trait loci, a population based approach. Genetics 172: 663–686.

Fuller WA. 1987. Measurement Error Models. New York: Wiley.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

Lin S, Cutler DJ, Zwick ME, Chakravarti A. 2002. Haplotype inference in random population samples. Am J Hum Genet 71:1129–1137.

Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N. 2008. Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. Biometrics 64:673–684.

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin Z, Munro HM, Abecasis GR, Donnelly P, for the International HapMap Consortium. 2006. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78: 437–450.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of casecontrol studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694.

Peters U, Chatterjee N, Yeager M, Chanock SJ, Schoen RE, McGlynn KA, Church TR, Weissfeld JL, Schatzkin A, Hayes RB. 2004. Association of genetic variants in the calcium-sensing receptor with risk of colorectal adenoma. Cancer Epidemiol Biomarkers Prev 13:2181–2186.

Potischman N, Coates RJ, Swanson CA, Carroll RJ, Daling JR, Brogan DR, Gammon MD, Midthune D, Curtin J, Brinton LA. 2002. Increased risk of early stage breast cancer related to consumption of sweet foods among women less than age 45. Cancer Causes Control 13: 937–946.

Schatzkin A, Subar AF, Moore S, Park Y, Potischman N, Thompson FE, Leitzmann M, Hollenbeck A, Morrissey KG, Kipnis V. 2009. Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). Cancer Epidemiol Biomarkers Prev 18:1026.

Spinka C, Carroll RJ, Chatterjee N. 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 29:108–127.

Stephans M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballaard-Barbash R, Sunshine J, Schatzkin A. 2003. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the Observing Protein and Energy Nutrition (OPEN) study. Am J Epidemiol 158:1–13.

Qin ZS, Niu T, Liu JS. 2002. Partial-ligation-expectation-maximization for haplotype inference with single nucleotide polymorphisms. Am J Hum Genet 71:1242–1247.

APPENDIX A

JUSTIFICATION OF THE MODEL AND REGRESSION COEFFICIENTS

To justify that the proposed risk functions (3) and (4) are valid for analysis of case-control association studies in the case when genetic markers are in the LD. Consider a situation when genetic markers are in the HWE. Suppose that only one trait locus Q affects the disease status, and there are two alleles Q_1 and Q_2 at the trait locus. Let q_i be the frequency of allele Q_i, $i = 1,2$. Let $\mu_{ij}^{(k)} = \log\{P(D = k|Q_iQ_j, X, Z)/P(D = 0|Q_iQ_j, X, Z)\}$ be the log ratio of the disease given genotype Q_iQ_j and environmental covariates (X, Z), $i,j = 1,2$. Denote $a_k = \mu_{11}^{(k)} - \{\mu_{11}^{(k)} + \mu_{22}^{(k)}\}/2$ and $d_k = \mu_{12}^{(k)} - \{\mu_{11}^{(k)} + \mu_{22}^{(k)}\}/2$. As traditional quantitative genetics, the average effect of gene substitution is $\alpha_{Qk} = a_k + (q_2 - q_1)\ d_k$, i.e. the difference between the average

effects of the trait locus alleles, and dominance deviation is $\delta_{Qk} = 2d_k$ [Falconer and Mackay, 1996]. Let $\mu^{(k)} = \mu_{11}^{(k)} q_1^2 + 2\mu_{12}^{(k)} q_1 q_2 + \mu_{22}^{(k)} q_2^2$ be the population mean effect. It can be shown that

$$\begin{aligned}&\mu_{11}^{(k)} 1_{(Q_1Q_1)}(G_Q) + \mu_{12}^{(k)} 1_{(Q_1Q_2)}(G_Q) + \mu_{22}^{(k)} 1_{(Q_2Q_2)}(G_Q)\\ &\quad = \mu^{(k)} + A_Q \alpha_{Qk} + B_Q \delta_{Qk},\end{aligned} \tag{A.1}$$

where

$$A_Q = \begin{cases} 2q_2 & \text{if } G_Q = Q_1Q_1 \\ q_2 - q_1 & \text{if } G_Q = Q_1Q_2 \\ -2q_1 & \text{if } G_Q = Q_2Q_2 \end{cases}, \quad B_Q = \begin{cases} -q_2^2 & \text{if } G_Q = Q_1Q_1 \\ q_1q_2 & \text{if } G_Q = Q_1Q_2 \\ -q_2^2 & \text{if } G_Q = Q_2Q_2 \end{cases}.$$

Hence, the genotypic value $\mu_{ij}^{(k)}$ can be expressed by a linear combination of $\mu^{(k)}$, A_Q, and B_Q. Suppose that the marker M_1 coincides with the trait locus Q, and marker allele M_i is the trait allele Q_1 and marker allele m_i is the trait allele Q_2. Then after a linear transformation, the relation (A.1) can be re-expressed by a linear combination of $\mu^{(k)}$, $A_1 = A_Q$, and $B_1 = B_Q$ (actually, it can be seen that $A_Q + 2q_1 = A_1 + 1$).

First, denote for $k = 1,\ldots,K$

$$\begin{aligned}y_k(\mathbf{g}, x, z) &= \log\left\{\frac{\text{pr}(D = k|\mathbf{G} = \mathbf{g}, X = x, Z = z)}{\text{pr}(D = 0)|\mathbf{G} = \mathbf{g}, X = x, Z = z}\right\}\\ &= \log\left\{\frac{\sum_{l,j} \text{pr}(D = k, Q_lQ_j|\mathbf{G} = \mathbf{g}, X = x, Z = z)}{\sum_{l,j} \text{pr}(D = 0, Q_lQ_j|\mathbf{G} = \mathbf{g}, X = x, Z = z)}\right\}.\end{aligned} \tag{A.2}$$

If no covariates are considered, it can be shown that the LD information between trait locus Q and markers M_i is contained in the probabilities $\text{pr}(D = k, Q_lQ_j, \mathbf{G} = \mathbf{g})$. To see this, let $h^{\text{dip}} = (h_1, h_2)$ denote a phased haplotype of an individual at the markers. Notice

$$\begin{aligned}\text{pr}(D = k, Q_lQ_j, \mathbf{G}) &= \text{pr}(D = k|Q_lQ_j, \mathbf{G}) \times \text{pr}(Q_lQ_j, \mathbf{G})\\ &= \text{pr}(D = k|Q_lQ_j) \times \text{pr}(Q_lQ_j, \mathbf{G}),\end{aligned}$$

and the probability $\text{pr}(Q_lQ_j, \mathbf{G}) = \sum_{h^{\text{dip}} \in \mathbf{G}} \text{pr}(Q_lQ_j, h^{\text{dip}})$ contains the LD information between trait locus Q and markers M_i. Here, the subscript $h^{\text{dip}} \in \mathbf{G}$ denote all haplotype phases, which are consistent with the genotype $\mathbf{G}$. If the covariates are considered, then

$$\begin{aligned}&\text{pr}(D = k, Q_lQ_j, \mathbf{G}, X, Z)\\ &\quad = \text{pr}(D = k|Q_lQ_j, \mathbf{G}, X, Z) \times \text{pr}(Q_lQ_j, \mathbf{G}, X, Z)\\ &\quad = \text{pr}(D = k|Q_lQ_j, X, Z) \times \text{pr}(Q_l, Q_j, \mathbf{G}, X, Z),\end{aligned}$$

and the probability $\text{pr}(Q_lQ_j, \mathbf{G}, X, Z) = \sum_{h^{\text{dip}} \in \mathbf{G}} \text{pr}(Q_lQ_j, h^{\text{dip}}, X, Z)$ contains the LD information between trait locus Q and markers M_i. Thus, $Y_k(\mathbf{G}, X, Z)$ is a function of the LD information between trait locus Q and markers M_i, and it contains association information between the trait locus and the markers.

Denote the pairwise measure of LD between marker M_i and marker M_j by $\Delta_{M_iM_j} = P(M_iM_j) - P_{M_i}P_{M_j}, i<j, i, j = 1,\ldots,I$ Let the additive and dominance variance-covariance matrices of the indicator variables defined in (2) be (second section of the Appendix A)

$$\mathbf{V}_A = 2\begin{pmatrix} P_{M_1}P_{m_1} & \Delta_{M_1M_2} & \cdots & \Delta_{M_1M_I} \\ \Delta_{M_1M_2} & P_{M_2}P_{m_2} & \cdots & \Delta_{M_2M_I} \\ \vdots & \vdots & \ldots & \vdots \\ \Delta_{M_1M_I} & \Delta_{M_2M_I} & \cdots & P_{M_I}P_{m_I} \end{pmatrix},$$

$$\mathbf{V}_D = \begin{pmatrix} P_{M_1}^2P_{m_1}^2 & \Delta_{M_1M_2}^2 & \cdots & \Delta_{M_1M_I}^2 \\ \Delta_{M_1M_2}^2 & P_{M_2}^2P_{m_2}^2 & \cdots & \Delta_{M_2M_I}^2 \\ \vdots & \vdots & \ldots & \vdots \\ \Delta_{M_1M_I}^2 & \Delta_{M_2M_I}^2 & \cdots & P_{M_I}^2P_{m_I}^2 \end{pmatrix}. \tag{A.3}$$

Define

$$\mathcal{B}_A = \begin{pmatrix} \beta_{kA1} + X\beta_{kAX1} + Z\beta_{kAZ1} \\ \vdots \\ \beta_{kAI} + X\beta_{kAXI} + Z\beta_{kAZI} \end{pmatrix} \quad \text{and}$$

$$\mathcal{B}_D = \begin{pmatrix} \beta_{kD1} + X\beta_{kDX1} + Z\beta_{kDZ1} \\ \vdots \\ \beta_{kDI} + X\beta_{kDXI} + Z\beta_{kDZI} \end{pmatrix}.$$

Given X and Z, the coefficients of model (A.2) are derived as (third section of the Appendix A)

$$\begin{aligned}\mathcal{B}_A &= \mathbf{V}_\mathbf{A}^{-1} \times \text{Cov}(Y_k, A|X, Z),\\ \mathcal{B}_D &= \mathbf{V}_\mathbf{D}^{-1} \times \text{Cov}(Y_k, B|X, Z).\end{aligned} \tag{A.4}$$

The form of $\text{Cov}(Y_k, A\,|\,X, Z)$ and $\text{Cov}(Y_k, B\,|\,X, Z)$ is given in the Appendix A.3.

Since $Y_k(\mathbf{G}, X, Z)$ contains association information between the trait locus and the markers, Equations (A.4) show that the measures of LD are contained in the mean coefficients. Model (3) simultaneously captures the LD and the effects of the trait locus.

Given X and Z, assume that Q_lQ_j and G are independent or, equivalently, there is no association between the trait and markers. Then Y_k does not depend on $\mathbf{G}$ since

$$\begin{aligned}y_k(\mathbf{g}, x, z) &= \log\left\{\frac{\sum_{l,j} \text{pr}(D = k|Q_lQ_j, \mathbf{G} = \mathbf{g}, X = x, Z = z) \times \text{pr}(Q_lQ_j|\mathbf{G} = \mathbf{g}, X = x, Z = z)}{\sum_{l,j} \text{pr}(D = 0|Q_lQ_j, \mathbf{G} = \mathbf{g}, X = x, Z = z) \times \text{pr}(Q_lQ_j, |\mathbf{G} = \mathbf{g}, X = x, Z = z)}\right\}\\ &= \log\left\{\frac{\sum_{l,j} \text{pr}(D = k|Q_lQ_j, X = x, Z = z) \times \text{pr}(Q_lQ_j|X = x, Z = z)}{\sum_{l,j} \text{pr}(D = 0|Q_lQ_j, X = x, Z = z) \times \text{pr}(Q_lQ_j|X = x, Z = z)}\right\}\\ &= \log\left\{\frac{\text{pr}(D = k|X = x, Z = z)}{\text{pr}(D = 0|X = x, Z = z)}\right\}.\end{aligned}$$

Therefore, $\text{Cov}(Y_k, A_i\,|\,X, Z) = 0$ and $\text{Cov}(Y_k, B_i\,|\,X, Z) = 0$ for all markers $i = 1,\ldots,I$. Hence, the regression coefficients of A_i and B_i are all zero, and the function

$m_k(\cdot)$ does not depend on the marker genotype data.

In summary, we illustrated that (1) the LD is being modeled in the regression coefficients, and (2) if there is no association between observed genotype and trait locus, then all regression coefficients of A_i and B_i are zeros and so the regression does not depend on the markers.

DERIVATION OF VARIANCE-COVARIANCE MATRICES (A.3)

Similarly to the Appendix A, Fan and Xiong [2002], the following expectations, variance and covariances can be derived accordingly: $E(A_i) = P_{M_i} - P_{m_i}$, $E(B_i) = 0$, $\mathrm{Var}(A_i) = 2P_{M_i}P_{m_i}$, $\mathrm{Var}(B_i) = P^2_{M_i}P^2_{m_i}$, $\mathrm{Cov}(A_i, A_j) = 2\Delta_{M_iM_j}$, $\mathrm{Cov}(B_i, B_j) = \Delta^2_{M_jM_l}$, $i,j = 1,\ldots,I$, $I \neq j$; in addition, $\mathrm{Cov}(A_i, B_j) = 0$ for all i and j.

Define $\mathcal{A} = (A_1, \ldots, A_I)$ and $\mathcal{B} = (B_1, \ldots, B_I)$. Further, let $\mathbf{V_A}$ be the $I \times I$ matrix with diagonal elements $2P_{M_im_i}$ and off-diagonals $2\Delta_{M_iM_j}$. Similarly, let $\mathbf{V_D}$ be the $I \times I$ matrix with diagonal elements $P^2_{M_i}P^2_{m_i}$ and off-diagonals $\Delta^2_{M_iM_j}$. Note that $\mathbf{V_A}$ and $\mathbf{V_D}$ are the covariance matrices of the additive and dominance effects, respectively. Let $\mathbf{O_I}$ be a $I \times I$ matrix with zero elements. Then based on the expectations and covariances described above,

$$\mathrm{Cov}(\mathcal{A}, \mathcal{B}) = \begin{pmatrix} \mathbf{V_A} & \mathbf{O_I} \\ \mathbf{O_I} & \mathbf{V_D} \end{pmatrix}.$$

PROOF OF (A.4)

Based on the definition of Y_k and the form of the covariance matrices (A.3), it can be easily seen that for $j = 1,\ldots,I$

$$\begin{aligned}\mathrm{Cov}(Y_k, A_j|X,Z) &= \sum_{i=1}^{I}\{\mathrm{Cov}(A_i,A_j)\times\mathcal{B}_{Ai} + \mathrm{Cov}(B_i,A_i)\times\mathcal{B}_{D_i}\} \\ &= \mathbf{V_A}\times\mathcal{B}_{A_j}; \\ \mathrm{Cov}(Y_k, B_j|X,Z) &= \sum_{i=1}^{I}\{\mathrm{Cov}(A_i,A_j)\times\mathcal{B}_{Ai} + \mathrm{Cov}(B_i,A_i)\times\mathcal{B}_{D_i}\} \\ &= \mathbf{V_D}\times\mathcal{B}_{D_j}.\end{aligned}$$

Equation (A.4) readily follows. To calculate the covariance between Y_k and A_i, B_i, let us denote $\mathcal{G}(M_iM_i) = \{(G_1, \ldots, G_{i-1}, M_iM_i, G_{i+1}, \ldots, G_I): G_j \in (M_jM_j, M_jm_j, m_jm_j), j \neq i\}$, and similarly we may define $\mathcal{G}(M_im_i)$ and $\mathcal{G}(m_im_i)$. Using these notations, we may calculate

$$E(Y_k|X,Z) = \sum_{g\in\mathcal{G}} P(\mathbf{G})\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D{=}k, Q_lQ_j|\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D{=}0, Q_lQ_j|\mathbf{G},X,Z)}\right\},$$

$$\begin{aligned}E(Y_kA_i|X,Z) = &\sum_{g\in\mathcal{G}(M_iM_i)} P(\mathbf{G}) \\ &\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D=k, Q_lQ_j|\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D=0, Q_lQ_j|\mathbf{G},X,Z)}\right\} \\ &- \sum_{g\in\mathcal{G}(m_im_i)} P(\mathbf{G}) \\ &\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D=k, Q_lQ_j|\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D=0, Q_lQ_j|\mathbf{G},X,Z)}\right\},\end{aligned}$$

$$\begin{aligned}E(Y_kB_i|X,Z) = &-P^2_{m_i}\sum_{g\in\mathcal{G}(M_iM_i)} P(\mathbf{G}) \\ &\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D=k, Q_lQ_j|\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D=0, Q_lQ_j|\mathbf{G},X,Z)}\right\} \\ &+ P_{M_i}P_{m_i}\sum_{g\in\mathcal{G}(M_im_i)} P(\mathbf{G}) \\ &\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D=k, Q_lQ_j,\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D=0, Q_lQ_j,\mathbf{G},X,Z)}\right\} \\ &- P^2_{M_i}\sum_{g\in\mathcal{G}(m_im_i)} P(\mathbf{G}) \\ &\times\log\left\{\frac{\sum_{l,j}\mathrm{pr}(D=k, Q_lQ_j,\mathbf{G},X,Z)}{\sum_{l,j}\mathrm{pr}(D=0, Q_lQ_j,\mathbf{G},X,Z)}\right\}.\end{aligned}$$