Statistics and Its Interface Volume 7 (2014) 51–60

Genotype-based association models of complex diseases to detect gene-gene and gene-environment interactions

Iryna Lobach*†, Ruzong Fan‡, and Prashiela Manga†

A central problem in genetic epidemiology is to identify and rank genetic markers involved in a disease. Complex diseases, such as cancer, hypertension, diabetes, are thought to be caused by an interaction of a panel of genetic factors, that can be identified by markers, which modulate environmental factors. Moreover, the effect of each genetic marker may be small. Hence, the association signal may be missed unless a large sample is considered, or *a priori* biomedical data are used. Recent advances generated a vast variety of *a priori* information, including linkage maps and information about gene regulatory dependence assembled into curated pathway databases. We propose a genotype-based approach that takes into account linkage disequilibrium (LD) information between genetic markers that are in moderate LD while modeling gene-gene and gene-environment interactions. A major advantage of our method is that the observed genetic information enters a model directly thus eliminating the need to estimate haplotype-phase. Our approach results in an algorithm that is inexpensive computationally and does not suffer from bias induced by haplotype-phase ambiguity. We investigated our model in a series of simulation experiments and demonstrated that the proposed approach results in estimates that are nearly unbiased and have small variability. We applied our method to the analysis of data from a melanoma case-control study and investigated interaction between a set of pigmentation genes and environmental factors defined by age and gender. Furthermore, an application of our method is demonstrated using a study of Alcohol Dependence.

AMS 2000 subject classifications: 60K35.

1. INTRODUCTION

Case-control studies are widely used to investigate genetic markers involved in complex disease susceptibility. Complex diseases, such as cancer, hypertension, diabetes, are likely to be caused by multiple genetic markers working in concert. The effect of each individual marker on risk of developing a complex disease is likely to be small. Some markers may be missed with the standard main effect tests unless a large sample size is used to balance individual genetic variability and experimental noise. A suitably large sample size is not feasible in many studies. Hence, an alternative approach, i.e., to use *a priori* biomedical information, might be preferable. Recent developments in high-throughput technologies generated rich sources of data that can be used for detecting disease-gene associations in the presence of gene-gene and gene-environment interactions.

Our work is motivated by a melanoma case-control study. Melanoma is a highly morbid disease for which incidence has continued to rise sharply over the past few decades in the United States. Melanoma is one of the more frequent cancers in young adults and the second most common cancer among women with ages 20–29. Identification of individuals at increased risk of melanoma is the key to reducing the incidence of metastatic disease. Treatment of early melanomas is readily achievable through surgical excisions, while prognosis of patients with metastatic melanoma is extremely poor. Analysis of gene-environment interactions has the potential to improve the understanding of the genetic predisposition of melanoma and to yield insight into mechanism of action of the exposures under various settings of particular genetic backgrounds thus improving disease prevention strategies. We have conducted a case-control study consisting of 343 melanoma patients and 434 controls. The available genetic information consists of multiple genetic markers spanning well characterized pigmentation genes (PG). The goal of this work is to examine the role of PG-gender and PG-age interactions in melanoma susceptibility, as well as the effect of interaction between PG genes.

Functionally related genes work concordantly while being involved in disease susceptibility thus leading to gene-gene interactions. A number of bioinformatics databases have been developed and curated to provide information on functions and relatedness of genes and to classify genes into gene sets with common underlying features. These databases include the gene ontology (GO: http://geneontology.org/) database, the Kyoto Encyclopedia of Genes and

*Corresponding author.
†The research was supported by the National Cancer Institute grant 5P30CA16087-28 and New York University Langone Medical Center, Center of Excellence 'Cancers of the Skin' pilot project.
‡Supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH), Maryland, USA.

Genomes (KEGG: www.genome.jp/kegg), the Molecular Signatures Database (MSigDB: http://www.broadinstitute.org/gsea/msigdb), Meta Core Ingenuity Pathway Analysis (http://www.ingenuity.com/), and Ariadne Genomics Pathway Studio (http://www.ariadnegenomics.com/products/pathway-studio/) among many others.

Recent developments in genotyping technology have generated dense single nucleotide polymorphism (SNP) panels providing an unprecedented opportunity to take advantage of the dependence among these markers and hence to combine an association signal provided by a group of genetic markers that are located closely to each other. The dependence among test statistics representing genetic markers is moderate, unless genetic markers are designed to be sampled locally to form haplotype blocks. Furthermore, information available in pathway databases allows to focus the analysis on small groups of genetic markers that are known to serve a particular biological function that may be important for an outcome in the study. Hence, 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 disease traits [1, 2, 3, 4]. LD information in a small set of genetic markers reflects structure of the genome and hence provides a valuable opportunity for mapping genetic variants responsible for complex diseases.

We are interested in building statistical methods that relate genetic variation to complex phenotypes and permit detection of interaction between genetic variants and environmental factors. Our strategy consists of the following two steps. First, we bring attention to groups of genetic markers defined by LD structure in which many weaker signals, considered together, present strong evidence of association for the unit, without losing the ability to detect strong single-marker signals. Second, using *a priori* biomedical information about the interaction structure we form a multivariate regression model involving main effects and interaction of the genetic markers grouped using LD information.

The current state of the genotyping technology allows generating unphased genotypes. Because we are focusing on LD blocks or candidate gene regions, it is of interest to investigate an association signal produced by a group of SNPs. Moreover, we consider situations commonly encountered in practice when genetic markers are in moderate LD. The haplotype-based method offers an advantage of modeling LD through the construction of haplotype blocks. However, in our setting the LD is moderate and hence the uncertainty associated with haplotype phase and haplotype blocks may be large. And this uncertainty associated with the haplotype-based method can lead to loss of accuracy in parameter estimates [5, 6, 7, 8, 9, 10, 11].

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. Hence, the haplotype phase need not to be estimated thus reducing computational burden and consequently reducing risk caused by potential bias due to haplotype-phase estimation. As the basis for estimation and inference, we will use the pseudo-likelihood function developed by [8]. The form of this pseudo-likelihood function offers several advantages. One is that it allows incorporating information about the probability of disease. In epidemiologic studies if the disease probability is unknown, a good bound can be specified. 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, body mass index (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, which is a reasonable assumption in many practical applications. For example, in situations when an individual cannot control exposure this assumption is valid. Alternatively, one can define strata and assume independence within a stratum. If the gene-environment independence is not valid in a setting, a distribution of genotype can be specified within strata defined by the environmental covariate.

2. PSEUDO-LIKELIHOOD ANALYSIS OF CASE-CONTROL STUDIES

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 denote all of the environmental (non-genetic) covariates of interest, such as age, gender, and exposure. The distribution of the environmental variable X is denoted as $f_X(x|\eta)$, where η is a column vector of unknown parameters of X.

Given the environmental covariates X 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

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

Here, $m_k(\cdot)$ is a function parameterizing the joint risk of the disease from genetic information $\mathbf{G}$ and environmental factors X in terms of the odds-ratio parameters β. Assume that all markers are di-allelic, e.g., SNPs. Let M_i and m_i be the major and minor alleles of a genetic marker i with frequencies P_{M_i} and P_{m_i}, respectively. Let $\Delta_{M_iM_j}$ be the measure of LD between markers i and j, i.e., $\Delta_{M_iM_j} = P(M_iM_j) - P(M_i)P(M_j)$. Let θ_i be frequency of the major allele at a genetic marker i, that is, $\theta_i = P(M_i)$. Denote

$\Theta = (\theta_1, \ldots, \theta_I)^{\mathrm{T}}$ and Δ the collection of all $\Delta_{M_iM_i}, 1 \le i < j \le I$ as a column vector. Further, define $\mathrm{P}(\mathbf{G}|\theta, \Delta)$ to be the distribution of genotype according to population genetic models, such as Hardy-Weinberg equilibrium.

Function $m_k(\mathbf{G}, X; \beta)$ parameterizes risk of developing disease explained by genetic information (G) and environmental factors (X). We model additive and dominance effects of a genetic marker i defined by the following variables

$$A_i = \begin{cases} 1 & \text{if } G_i = M_iM_i \\ 0 & \text{if } G_i = M_im_i \\ -1 & \text{if } G_i = m_im_i \end{cases}, \tag{1}$$

$$B_i = \begin{cases} -P_{m_i}^2 & \text{if } G_i = M_iM_i \\ P_{M_i}P_{m_i} & \text{if } G_i = M_im_i \\ -P_{M_i}^2 & \text{if } G_i = m_im_i \end{cases}. \tag{2}$$

The dummy variables A_i can be used to model additive effect and the dummy variables B_i can model the dominance effect of the phenotypic trait [8, 9, 10, 11]. In our previous work in Lobach et al. (2010), Fan and Xiong (2002), and Fan et al. (2006) [8, 9, 10, 11], we show that the genetic effect can be orthogonally decomposed into a summation of the additive effect of A_i and the dominance effect of B_i. This nice orthogonal decomposition is used in this paper without a repetition of justification. One may replace B_i by other definition such as $B_i = \begin{cases} 0 & \text{if } G_i = M_iM_i \\ 1 & \text{if } G_i = M_im_i \\ 0 & \text{if } G_i = m_im_i \end{cases}$, but the orthogonal decomposition will not be valid anymore.

Additive Effects Model (AEM) In the case when genetic markers are known to have additive effect, the risk function $m_k(\mathbf{G}, X; \beta)$ can be written in the following form

$$m_k(\mathcal{A}, X; \beta) = X\beta_{kX} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i \ne j} A_iA_j\beta_{kA_iA_j} + \sum_{i=1}^{I} XA_i\beta_{kAXi}. \tag{3}$$

In AEM (3), β_{kAi} is the additive effect of the dummy variable A_i, $\beta_{kA_iA_j}$ is the interaction between A_i and A_j, and β_{kAXi} is the interaction between X and A_i. Note that interaction terms define nonadditivity of effects. Specifically, the interaction term $\sum_{i \ne j} A_iA_j\beta_{kA_iA_j}$ defines epistasis, that is the nonadditivity of effects among the genetic marker loci. Further, dominance effect at a locus is defined to be the deviation of the observed genotypic value from the expectation based on the additive effects. Hence, dominance is a measure of nonadditivity of allelic effects within loci.

Genotype Effects Model (GEM) The following specification of the risk function $m_k(\mathbf{G}, X; \beta)$ incorporates both additive and dominance effects of genotype [8, 9], and gene-gene and gene-environment interactions

$$m_k(\mathcal{A}, \mathcal{B}, X; \beta) = X\beta_{kX} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i=1}^{I} B_i\beta_{kDi} + \sum_{i \ne j} A_iA_j\beta_{kA_iA_j} + \sum_{i=1}^{I} XA_i\beta_{kAXi} + \sum_{i \ne j} B_iB_j\beta_{kD_iD_j} + \sum_{i=1}^{I} XB_i\beta_{kDXi}. \tag{4}$$

In the GEM (4), β_{kAi}, $\beta_{kA_iA_j}$, and β_{kAXi} are the same as those of AEM (3). In addition, β_{kDi} is the dominance effect of the dummy variable B_i, $\beta_{kD_iD_j}$ is the interaction between B_i and B_j, and β_{kDXi} is the interaction between X and B_i.

In practice, the additive effect model (3) can be advantageous over the genotype effect model (4) because of the smaller number of parameters in (3). This situation may occur when the dominance effect is not significantly present or the dominance effect can not compensate for the increase of number of parameters in (4).

Pseudo-likelihood The following pseudo-sampling is along the lines of our previous work [8, 9]. Let n_0 be the number of control subjects; and for $k \ge 1$, denote by n_k the number of subjects in the sample with disease at 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 $\pi_k = \mathrm{P}(D = k), k = 0, 1, 2, \ldots, K$. 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 denote the indicator of whether a subject is selected in the sample. For the i-th subject, let us denote by $(D_i, \mathbf{G_i}, X_i, R_i)$ the observed values of variables $D, \mathbf{G}, X$ and R. Let us denote $\kappa_k = \beta_{k0} + \log(n_k/n_0) - \log(\pi_k/\pi_0)$ and $\widetilde{\kappa} = (\kappa_1, \ldots, \kappa_K)^{\mathrm{T}}$. In addition, let $\widetilde{\beta}_0 = (\beta_{10}, \ldots, \beta_{K0})^{\mathrm{T}}$, $\Omega = (\widetilde{\beta}_0^{\mathrm{T}}, \beta^{\mathrm{T}}, \Theta^{\mathrm{T}}, \widetilde{\kappa}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathcal{B} = (\Omega^{\mathrm{T}}, \eta^{\mathrm{T}})^{\mathrm{T}}$.

The development of this method relies on gene-environment independence assumption. Specifically, we assume that G and X are independently distributed in the underlying population. In many practical situations this assumption is reasonable, e.g., when an individual cannot control an environmental exposure. However, in some studies researchers may not be comfortable making this assumption. For example, in hormone-related diseases, certain genes may regulate a woman's age at menarche, menopause, or reproductive history. Another example is related to genes that

may regulate the degree of addiction, what needs to be accounted for in the analysis of lung cancer and nicotine dependence, alcohol dependence, etc.

To relax gene-environment independence assumption, genotype and environmental factors should be modeled conditionally on strata (if the independence assumption is reasonable within strata). Specifically, let S be the stratum. Hence, the environmental covariate consists of two sets of variables, S defining strata and Z is the set of environmental variables that are independent of G within strata S: $X = (S, Z)$. If an environmental variable is continuous, then these strata can be defined based on clinically-relevant cut-off values defined by the domain experts.

The distribution of the genotype within each strata for a pair of markers that are in LD can be written as follows

$$\mathrm{P}(\mathbf{G}|S,\theta,\Delta) = \begin{cases} P^s(M_1)P^s(M_2) + \Delta^s_{M1M2}, & \mathbf{G} = (M_1M_2) \\ \{1-P^s(M_1)\}P^s(M_2) - \Delta^s_{M1M2}, & \mathbf{G} = (m_1M_2) \\ P^s(M_1)\{1-P^s(M_2)\} - \Delta^s_{M1M2}, & \mathbf{G} = (M_1m_2) \\ \{1-P^s(M_1)\}\{1-P^s(M_2)\} + \Delta^s_{M1M2}, & \mathbf{G} = (m_1m_2), \end{cases}$$

where $P^s(M_i)$ is the allele frequency of allele M_i in the stratum s and Δ^s_{M1M2} is the corresponding linkage disequilibrium measure in the stratum. Define

(5)

$$S(k,\mathbf{g},z,s;\Omega,\Delta) = \frac{\exp\left[1_{(k\geq 1)}(k)\left\{\kappa_k + m_k(\mathbf{g},z,s;\beta)\right\}\right]}{1+\sum_{j=1}^K \exp\left\{\beta_{j0} + m_j(\mathbf{g},z,s;\beta)\right\}} \times \mathrm{P}(\mathbf{g}|s,\Theta,\Delta).$$

Similarly to [8] we propose to estimate parameters (Ω,η,Δ) based on a pseudo-likelihood function $L_{Pseudo}(k,\mathbf{g},z,s;\Omega,\Delta)$ defined as follows

$$\mathrm{P}(D=k,\mathbf{G}=\mathbf{g}|Z=z,S=s,R=1) = \frac{S(k,\mathbf{g},z,s;\Omega,\Delta)}{\sum_{k_1=0}^K \sum_{\mathbf{g}\in\mathcal{G}} S(k_1,\mathbf{g},z,s;\Omega,\Delta)},$$

where $\mathcal{G}$ is the set of all possible genotypes in the population. Observe that conditioning on $X = (Z, S)$ in L_{Pseudo} allows it to be free of the nonparametric density function $f_X(x|\eta)$, thus avoiding the difficulty of estimating potentially high-dimensional nuisance parameters.

Missing genetic data Missing genetic data arising in cases when genotype is not observed are handled by summing the likelihood function over all possible values that may be observed. Define $\mathcal{G}_o$ be the set of all possible genotypes that are consistent with the observed genotype. The pseudo-likelihood function $L_{Pseudo}(k,\mathbf{g},z,s;\Omega,\Delta)$ becomes

$$\mathrm{P}(D=k,\mathbf{G}=\mathbf{g}|Z=z,S=s,R=1) = \frac{\sum_{g^*\in\mathcal{G}_o} S(k,g^*,z,s;\Omega,\Delta)}{\sum_{k_1=0}^K \sum_{g^*\in\mathcal{G}} S(k_1,g^*,z,s;\Omega,\Delta)}.$$

In the same manner of Appendix in Lobach et al. (2010) [8], theoretical justification can show 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 can show 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 hence the regression does not depend on the markers [8].

Recall that in the case when genetic markers modeled in the risk function are in LD, the regression coefficients capture both the association signal and the LD information. In practice, we suggest to perform univariate analysis or use *a priori* knowledge to define a set of genetic markers to be analyzed for gene-gene interactions. 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).

Asymptotics Define $\Psi(k,g,z,s;\Omega,\Delta)$ to be the derivative of log pseudo-likelihood function $L_{Pseudo}(k,\mathbf{g},z,s;\Omega,\Delta)$ with respect to (Ω,Δ). Then define

$$\begin{aligned} \mathcal{L}_n(\Omega,\Delta) &= \sum_{i=1}^n \Psi(D_i,G_i,Z_i,S_i;\Omega,\Delta); \\ \mathcal{I} &= -n^{-1}\mathrm{E}[\partial\mathcal{L}_n(\Omega,\Delta)/\partial(\Omega,\Delta)^{\mathrm{T}}] \\ \Lambda &= \sum_k \frac{n_k}{n}\mathrm{E}\{\Psi(D,G,Z,S;\Omega,\Delta)|D=k\} \\ &\quad \times \mathrm{E}\{\Psi(D,G,Z,S;\Omega,\Delta)|D=k\}^{\mathrm{T}}, \end{aligned}$$

where all expectations are taken with respect to the case-control sampling design.

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

$$n^{-1/2}(\widehat{\mathcal{B}}-\mathcal{B}) \sim \mathrm{Normal}\{0,\mathcal{I}^{-1}(\mathcal{I}-\Lambda)\mathcal{I}^{-1}\}.$$

Remark 1. Matrices $\mathcal{I}$ and Λ can be consistently estimated as follows. $\mathrm{E}\{\Psi(D,G,Z,S;\Omega,\Delta)|D=k\}$ can be estimated by $n_k^{-1}\sum_{i=1}^n I(D_i=k)\Psi(k,G_i,Z_i,S_i,\widehat{\Omega},\widehat{\Delta})$. Similarly, $n^{-1}\partial\{\mathcal{L}_n(\widehat{\Omega},\widehat{\Delta})\}/\partial(\widehat{\Omega},\widehat{\Delta})$ estimates $\mathcal{I}$.

Remark 2. The intercept parameters β_{0j} are theoretically identifiable. When *a priori* estimate of probability of disease is available, it can be incorporated into an estimation scheme to potentially improve qualities of the estimate. Specifically, a probability of disease can be set into a grid spanning plausible values. For each of these values, supposing they are fixed, all other parameters can be estimated by maximizing the corresponding pseudo-likelihood function. Then, the estimate of probability of disease can be defined as a value that

maximizes the corresponding profile likelihood. In practical situations, likelihood function might be flat as a function of probability of disease and therefore intercept. In such situations, sensitivity analyses should be performed to examine differences in estimates of other parameters corresponding to various settings of probability of disease/intercept. In the case when a disease is rare, the likelihood function is expected to contain very little information about the intercept.

3. SIMULATION EXPERIMENTS

We performed a series of simulation experiments to investigate performance of the proposed procedure in various settings. We consider a case when the disease status D is binary. The genotype $\mathbf{G}$ was simulated under HWE for $I = 2, 3$ markers, respectively ($I = 3$ in Experiment 1, $I = 2$ in Experiments 2 and 3 below). Given the values of $(\mathbf{G}, X)$, we generated a binary disease outcome D using 2 logistic models, corresponding to the GEM and AEM. For the GEM, covariates are related to a disease via link function

$$\begin{aligned}&\text{logit}\{\mathrm{P}(D=1|\mathcal{A},\mathcal{B},X)\}\\&=\beta_0+X\beta_X+\sum_{i=1}^{I}A_i\beta_{Ai}+\sum_{i=1}^{I}XA_i\beta_{AXi}\\&\quad+\sum_{i=1}^{I}B_i\beta_{Di}+\sum_{i=1}^{I}XB_i\beta_{DXi}+A_1A_2\beta_{A_1A_2},\end{aligned}$$

where $I = 3$ in Experiment 1 and $I = 2$ in Experiment 2 and 3 below. The corresponding AEM was obtained by setting coefficients β_{Di}, β_{D12}, and β_{DXi} to be 0. Here, we omit the subscription k in the regression parameters βs since we have only one level disease cases and normal controls.

Experiment 1 We considered a situation when three SNPs are involved in a disease. The three SNPs have strong additive effect ($\beta_{A1} = \log(1.5)$; $\beta_{A2} = \log(2.2)$; $\beta_{A3} = \log(1.5)$), two of the SNPs have dominance effect ($\beta_{D1} = \log(3); \beta_{D3} = \log(2)$). The first two SNPs interact with each other $\beta_{A_1A_2} = \log(3)$ and, further, both additive and dominance effects interact with the environmental variable. Such a setting mimics complex interaction structure encountered in practice. The environmental variable (X) is binary with $\mathrm{P}(X = 1) = 0.5$.

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. To estimate the parameters, 500 samples are simulated and each sample contains of 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).

Simulation results are shown in Table 1. The results illustrate that the proposed methodology produced parameter estimates that are nearly unbiased and have small variability. Further, the root mean squared errors 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 situations when the data present strong evidence for the dominance effect. When $\mathrm{P}(D = 1)$ is known, parameter estimates have smaller variability compared to those in the case when the probability of disease is not known. This result illustrates the ability of the proposed method to incorporate information about the probability of disease, what cannot be done in the standard logistic regression model.

Table 1. Biases and Root Mean Squared Errors (RMSEs) of risk parameters in the case when $\mathrm{P}(D = 1)$ is known and when it is estimated. 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. The data is simulated and analyzed under the genotype effect model

		$\mathrm{P}(D=1)$ is known		$\mathrm{P}(D=1)$ is unknown	
Parameter	True value	Bias	RMSE	Bias	RMSE
κ	0.484	−0.017	0.037	−0.054	0.020
β_X	0.693	0.002	0.011	0.014	0.039
β_{A1}	0.406	−0.004	0.009	−0.012	0.016
β_{A2}	0.789	0.006	0.010	−0.003	0.015
β_{A3}	0.693	−0.001	0.008	−0.005	0.016
$\beta_{A_1A_2}$	1.099	0.007	0.009	0.009	0.012
β_{AX1}	0.916	0.006	0.014	0.039	0.046
β_{AX2}	0.693	−0.002	0.016	0.038	0.041
β_{AX3}	1.099	0.004	0.017	0.039	0.058
β_{D1}	0.262	0.022	0.064	0.026	0.152
β_{D2}	0.095	0.002	0.052	0.005	0.099
β_{D3}	0.693	0.007	0.046	0.018	0.128
β_{DX1}	1.099	−0.018	0.083	0.018	0.302
β_{DX2}	0.916	0.004	0.076	0.006	0.208
β_{DX3}	1.099	0.007	0.087	0.024	0.286
P_{M_i}	0.250	<0.001	<0.001	0.001	<0.001
$\mathrm{P}(D=1)$	0.005			0.003	<0.001

Experiment 2 To evaluate advantage offered by the proposed method in the case when genetic markers are in the LD, we performed the following simulation experiment. Genetic markers are simulated to have moderate LD ($\Delta_{M1M2} = 0.02$). Environmental factors and risk model are the same as in the first experiment.

Results presented in Table 2 illustrate that the proposed approach resulted in parameter estimates that are nearly unbiased and have small variability. The naive approach that ignores moderate LD resulted in parameter estimates that are biased and have elevated variability. For example, the gene-gene interaction parameter estimate had a bias that is 12.47% of the estimate thus masking the risk due to gene-gene interaction. Further, we simulated a case when the second marker is not involved in a disease and the naive anal-

Table 2. Biases and Root Mean Squared Errors (RMSEs) of risk parameters for the naive approach that ignores existence of the LD and the proposed method. 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. Probability of disease is* 0.0069 *and is assumed to be known in the population. The data is simulated and analyzed under the additive effect model and the LD measure* $\Delta_{M1M2} = 0.02$

		Naive approach		Proposed method	
Parameter	True value	Bias	RMSE	Bias	RMSE
κ	0.082	0.098	0.015	0.037	0.001
β_X	1.099	−0.007	0.010	−0.005	0.009
β_{A1}	0.693	−0.011	0.011	0.005	0.008
β_{A2}	0.000	−1.170	1.379	0.005	0.009
β_{AX1}	0.693	0.039	0.013	−0.004	0.013
β_{AX2}	0.693	−0.017	0.301	0.001	0.012
$\beta_{A_1A_2}$	1.099	0.137	0.301	0.002	0.007
P_{M_i}	0.250	0.004	<0.001	<0.001	<0.001
δ	0.500			−0.005	0.014

ysis results in largely negative bias hence producing a false positive result by announcing this marker to be protective (data not shown).

Experiment 3 To examine performance of inferences based on the proposed estimation procedure, we investigated variability of estimates obtained as a result of Experiment 2 and compared it to the estimated standard errors. Results presented in Table 3 illustrate that the proposed estimation procedure produced parameter estimates that have the expected variability.

4. MELANOMA DATA ANALYSIS

The melanoma case-control cohort consisted of 343 Caucasian melanoma patients recruited through the New York University Interdisciplinary Melanoma Cooperative Group and 434 obtained from the New York Cancer Project. Cases and controls are matched based on age and gender. SNPs spanning candidate melanoma susceptibility genes and 80 ancestry informative markers were genotyped. Because melanoma susceptibility is known to vary with ancestry, we selected participants of our study who (1) self-reported their ancestry as Northern European; and (2) principal combination of ancestry informative markers indicated their ancestry to be in the Northern European cluster. 126 cases and 160 controls reside in the Northern European cluster.

Using *a priori* biomedical knowledge, we selected two genes: Solute Carrier Family 45, member 2 gene (SLC45A2) and Oculocutaneous Albinism 2 gene (OCA2). The SLC45A2 gene is known to be associated with normal human skin variation [12]. It has been noted in the literature [12, 13, 14, 15, 16, 17] that the amounts of melanin

Table 3. Standard errors (SE) of risk parameters for the proposed approach. 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. Probability of disease is* 0.0069 *and is assumed to be known in the population. The data is simulated and analyzed under the additive effect model*

Parameter	True value	Standard error of estimates	Mean estimated standard error
β_X	1.099	0.007	0.008
β_{A1}	0.693	0.006	0.006
β_{AX1}	0.693	0.007	0.007
β_{A2}	0.000	0.013	0.015
β_{AX2}	0.693	0.010	0.013
$\beta_{A_1A_2}$	1.099	0.007	0.007

and cutaneous blood flow differs between men and women, yet the degree of this difference is not well established. The OCA2 gene is known to be involved in albinism [18] and plays a role in determining skin pigmentation [19]. Further, *a priori* knowledge collected in pathway databases suggests that OCA2 and SLC45A2 genes interact to determine skin pigmentation. Both OCA2 and SLC45A2 encode putative transmembrane bound transporters with unknown function. Mutations at these loci result in oculocutaneous albinism type 2 and 4 respectively, indicating that they are crucial for melanogenesis [18, 20]. We selected markers residing in SLC45A2 gene that are in moderate linkage disequilibrium. Our analysis suggested that haplotypes inferred based on these three markers are not significantly associated with melanoma. To examine gene-gender and gene-age interactions, we investigated the following risk model

$$\begin{aligned}(6)\qquad &\text{logit}\{\mathrm{P}(D=1|G, X_{Gender}, X_{Age})\}\\ &= \beta_0 + \beta_{Gender}X_{Gender} + \beta_{Age}X_{Age} + \beta_A A\\ &\quad +\beta_{A\times Gender}AX_{Gender} + \beta_{A\times Age}AX_{Age}.\end{aligned}$$

We considered a reduced additive effect model (6). Table 4 presents parameter estimates and the corresponding standard errors. These results suggest that for all three SNPs gene-gender and gene-age interactions are significant.

We further included pair-wise gene-gene interactions between these three SNPs, however they are not significant. We examined interaction between a SNP residing in OCA2 gene (designated OCA2-SNP3) and the three SLC45A2 markers using the following risk model

$$\begin{aligned}(7)\qquad &\text{logit}\{\mathrm{P}(D=1|G, X_{gender}, X_{age})\}\\ &= \beta_0 + \beta_{Xgender}X_{gender} + \beta_{Xage}X_{age}\\ &\quad +\beta_{A1}A_1 + \beta_{A2}A_2 + \beta_{A_1A_2}A_1 * A_2,\end{aligned}$$

where A_1 denotes OCA2-SNP3, and A_2 corresponds to SNPs of SLC45A2 gene. Results presented in Table 5 il-

Table 4. Estimates and standard errors of risk parameter estimates in the melanoma study based on model (6)

SNP	Parameter	Estimate	Standard error
SLC45A2-SNP6	β_{Gender}	1.11	0.28
	β_{Age}	0.04	0.009
	$\beta_{A_{2.6}}$	0.45	0.13
	$\beta_{A_{2.6}\times Gender}$	−1.69	0.59
	$\beta_{A_{2.6}\times age}$	−0.02	0.007
SLC45A2-SNP7	β_{Gender}	0.60	0.32
	β_{Age}	0.04	0.009
	$\beta_{A_{2.7}}$	0.32	0.09
	$\beta_{A_{2.7}\times Gender}$	−1.13	0.48
	$\beta_{A_{2.7}\times Age}$	−0.01	0.006
SLC45A2-SNP8	β_{Gender}	0.63	0.31
	β_{Age}	0.04	0.009
	$\beta_{A_{2.8}}$	0.57	0.21
	$\beta_{A_{2.8}\times Gender}$	−1.19	0.49
	$\beta_{A_{2.8}\times Age}$	−0.02	0.005

Table 5. Estimates and standard errors of risk parameters in the melanoma study based on model (7)

SNP	Parameter	Estimate	Standard error
OCA23xSLC45A2-SNP6	$\beta_{A_1A_{2.6}}$	−0.74	0.32
OCA23xSLC45A2-SNP7	$\beta_{A_1A_{2.7}}$	−0.89	0.38
OCA23xSLC45A2-SNP8	$\beta_{A_1A_{2.8}}$	−0.74	0.37

lustrate that indeed these gene-gene interactions are significant. Hence, these genetic markers, both involved in pathways responsible for pigmentation, work together while being involved in melanoma.

Note that an estimate of probability of melanoma in a population (π_k) is well-established in literature (e.g., http://seer.cancer.gov/statfacts/html/melan.html), and overall it is a rare disease. We used this estimate to inform our estimation procedure, i.e., the disease probability was set on a conservative interval around this estimate and grid-search method was used to find a value that maximizes the likelihood function. Because melanoma is a rare disease (π_d is small), the likelihood function is expected to contain little information about the π_d and intercept.

5. ANALYSIS OF ALCOHOL DEPENDENCE

The goal of this analysis is to examine effect of an association between genetic and environmental factors and demonstrate analysis based on stratification.

The Collaborative Studies on the Genetics of Alcoholism (COGA) is a nine-center nationwide study that was initiated in 1989 and has had as its primary aim the identification of genes that contribute to alcoholism susceptibility and related characteristics [21, 22, 23]. COGA is funded through the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The focus of this study is a case-control design of unrelated individuals for a genetic association analysis of addiction. We focus the analysis on the role of gender. In contrast to alcoholism risk in men, evidence for a major genetic contribution to alcoholism risk in women from systematically ascertained adoptee and twin samples appears much weaker [24, 25, 26, 27]. The absence of strong evidence for a genetic influence on female alcoholism has been interpreted as supporting existence of a subtype of alcoholism that is predominant among women and that is only modestly heritable, contrasted with a more highly heritable male-limited subtype.

The genetic variable of our interest is a SNP rs2043602 residing in NCAM1 gene. This SNP is functionally linked to dopamine in the brain. Several association studies reported an association with alcohol dependence and drug dependence. However, the results have been inconsistent [27, 28]. We hypothesize that the effect of gender and an association between gender and a distribution of allele frequencies in the population may have contributed to the observed inconsistency of results.

The sample consists of 1,962 controls and 1,720 cases; 1,962 men and 1,720 women. Genotype and gender are associated in controls (the p-value of χ^2 test = 0.002). We note that within strata defined by the age when a person got drunk for the first time (AFGD), this association is not statistically significant (the p-values of χ^2 tests are 0.24 and 0.18 within a subset of participants whose AFGD is less than 21 and those whose AFGD is greater than 21, respectively). Note that AFGD represents an environmental exposure and may be an important factor in the context of the noted absence of strong evidence for a genetic influence on female alcoholism.

Hence, we consider two models. The first is based on the main effects of gender, the additive effect of genetics, and their interaction. The second model incorporates stratification on AFGD. We note that the first model is not technically correct because of the strong association between gender and genetics. Analysis of this model is useful for assessing effects of the association between genetic and environmental factors. Specifically, **Model 1**

$$\begin{aligned}(8)\qquad &\text{logit}\{\text{pr}(D=1|A,X_{Gender})\}\\ &=\beta_0+\beta_A A+\beta_{X_{Gender}}*X_{Gender}\\ &\quad+\beta_{AX_{Gender}}AX_{Gender}.\end{aligned}$$

and **Model 2** involves stratification on S_{AFGD} that is a binary variable defined using a cut-off age = 21.

Results presented in Table 6 demonstrate that the interaction term $\beta_{AX_{Gender}}$ is not statistically significant (p-value = 0.18). In the second model of Table 7, this interaction effect is significant (p-value = 0.045) and the main contributor to this effect is the effect of a genotype in women whose AFGD is less than 21 (log(OR) = −0.23, p-value =

Table 6. Estimates of risk parameter and corresponding p-values in the alcohol dependence study based on model (8)

Effect	Estimate	P-value
$\beta_{X_{Gender}}$	−1.06	< 0.001
β_A	−0.89	0.59
$\beta_{AX_{Gender}}$	−0.74	0.18

Table 7. Estimates of risk parameter and corresponding p-values in the alcohol dependence study based on model (8) with stratification on AFGD

Effect	Estimate	P-value
$\beta_{X_{Gender}}$	−1.06	<0.001
β_A	−0.14	0.008
$\beta_{AX_{Gender}}$	−0.20	0.045
$\beta_{S_{AFGD}}$	−1.03	<0.001

0.012). The directionality of the effect suggests that a minor allele may have a protective effect. Note that similarly to analysis of melanoma data, probability of alcohol dependence is estimated based on a grid-search algorithm around a well-established estimate of alcohol dependence in a population.

6. DISCUSSION

We proposed a genotype-based approach for the analysis of case-control studies of gene-gene and gene-environment interactions and applied it to the analysis of gene-gene, gene-gender, and gene-age interactions in the etiology of melanoma. The formulation of risk functions and estimation procedure are along the lines of previous work on the co-authors: genotype and additive effect models [10, 11] and pseudo-likelihood approach [8]. The risk model involves both the additive and dominance effect while taking into account possible interactions between genes expressed in terms of interaction between their additive and dominance components.

The proposed method has several unique aspects. First, the observed genetic information enters the model directly and pair-wise LD structure is captured in the regression coefficients. This aspect offers advantages from the practical point of view, the computational burden is less demanding because haplotype-phase need not to be estimated. In the cases when LD is moderate, which is the focus of our work, the computational demands can be substantial even with the current state of technology. Further, the risk due to uncertainty associated with the haplotype-phase estimation can be avoided. Similarly to the method investigated in [8], the estimating procedure is based on a pseudo-likelihood model that allows efficiently estimating 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.

We applied the proposed approach to the analysis of an association between melanoma and pigmentation genes, age, and gender while accounting for their interaction. The pigmentation genes were selected based on *a priori* information about their function; genetic markers within the genes were selected using information obtained from linkage maps. We found evidence of interaction between genetic markers and age/gender. The effect of age and gene-age interactions has been implicated in many complex diseases, including cancer. Genetic predisposition plays an important role at younger ages, while the exposure may become more important at older ages. In the etiology of melanoma, the environmental exposure is believed to play an important role and a small proportion (5–10%) of people who develop melanoma do so because of genetic susceptibility alone [30, 31]. The vast majority of melanoma is caused by ultraviolet light exposure often due to behavioral factors with some genetic contribution.

Melanoma rates and melanoma-related death rates differ between men and women [32]. The difference between behavioral factors plays an important role in gender disparity of melanoma. For example, a study conducted based on a 2005 Health Interview Survey [31] concluded that although men more often wear protective clothing and are less likely to use a tanning bed, women tend to avoid sun exposure and use sunscreen. This result suggested that preventive measures taken by women may have resulted in smaller incidence in melanoma compared to that of men. Further, a number of epidemiologic studies hypothesized that the disparity in melanoma predisposition is due to hormonal differences. For example, several studies reported epidemiologic evidence for a reduced melanoma risk with higher parity and a higher risk with higher age at first live birth and the use of oral contraceptives [33, 34, 35]. In summary, the gene-gender interactions that we found may be due to gender disparities in behavioral and hormonal factors. That is, gender is a surrogate of behavioral and hormonal factors that are different between men and women and that have an important role in melanoma etiology.

Based on our simulation experiments and application of the proposed method to the melanoma study, we found that the method offers advantages when the amount of LD is moderate and when genetic markers have moderate effect. However, when the number of markers is large and the LD is strong, the haplotype-based approach based on a pseudo-likelihood that we developed earlier [29] can be more useful. In the case when a disease is caused by one or two haplotypes and alleles forming the haplotype have small effects, the haplotype-based approach is superior to our proposed genotype-based modeling. Bayesian formulation provides a conceptually elegant way to incorporate *a priori* information. However, in our context, conventional Bayesian analysis may not be applied directly because the analysis is

based on a pseudo-likelihood function. Hence posterior credible intervals may have poor coverage. We have proposed a Bayesian model based on our pseudo-likelihood function and demonstrated advantages of this model in situations when an environmental variable is subject to measurement error or misclassification [36, 9]. Validation of pseudo-likelihood function is highly computationally intensive.

ACKNOWLEDGEMENT

R. Fan was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Maryland, USA. I. Lobach was partially supported by 1R21AA020356-01A1 grant from the National Institute on Alcohol Abuse and Alcoholism and P. Manga was partially supported by the National Cancer Institute grant 5P30CA16087-28 and New York University Langone Medical Center, Center of Excellence 'Cancers of the Skin' pilot project. We thank Iman Osman, the leader of the Interdisciplinary Melanoma Cooperative Group.

The authors thank Ivan Belousov for thoughtful comments. This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at New York University Langone Medical Center. Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). SAGE is one of the genomewide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning as well as with general study coordination was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1.

Received 24 January 2013

REFERENCES

[1] The International HapMap Consortium. (2003). The international HapMap project. *Nature* **426** 789–796.

[2] The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* **437** 1299–1320.

[3] The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449** 851–861.

[4] 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.

[5] Lin, D. Y. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101** 89–118. MR2268031

[6] Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z., Munro, H. M., Abecasis, G. R., and Donnelly, P. for the International HapMap Consortium (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* **78** 437–450.

[7] Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68** 978–989.

[8] Lobach, I., Fan, R., and Carroll, R. J. (2010). Genotype-based association mapping of complex diseases: Gene-environment interactions with multiple genetic markers and measurement error in environmental exposures. *Genetic Epidemiology* **34(8)** 792–802.

[9] Lobach, I. and Fan, R. (2012). Genotype-based Bayesian analysis of gene-environment interactions with multiple genetic markers and misclassification in environmental factors. *Journal of Statistics and Probability* **2012**, Article ID 151259, 15 pages, doi:10.1155/2012/151259. MR2959596

[10] Fan, R., Jung, J., and Jin, J. (2006). High resolution association mapping of quantitative trait loci, a population based approach. *Genetics* **172** 663–686.

[11] Fan, R. and Xiong, M. (2002) High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *European Journal of Human Genetics* **10** 607–615.

[12] Graf, J., Hodgson, R., and van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum. Mutat.* **25(3)** 278–284.

[13] Edwards, E. A. and Duntley, S. Q. (1939). The pigments and color of living human skin. *American Journal of Anatomy* **65** 1–33.

[14] Frost, P. (1988). Human skin color: A possible relationship between its sexual dimorphism and its social perception. *Perspectives in Biology and Medicine* **32** 38–58.

[15] Frost, P. (2005). *Fair Women, Dark Men. The Forgotten Roots of Color Prejudice.* Cybereditions: Christchurch (New Zealand).

[16] Hulse, F. S. (1967). Selection for skin color among the Japanese. *American Journal of Physical Anthropology* **27** 143–156.

[17] Jablonski, N. G. and Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution* **39** 57–106.

[18] Ramsay, M., Colman, M. A., Stevens, G., Zwane, E., Kromberg, J., Farrall, M., and Jenkins T. (1992). The tyrosinase-positive oculocutaneous albinism locus maps to chromosome 15q11.2-q12. *American Journal of Human Genetics* **51(4)** 879–84.

[19] Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Palsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics* **39(12)** 1443–52.

[20] Newton, J. M., Cohen-Barak, O., Hagiwara, N., Gardner, J. M., Davisson, M. T., King, R. A., and Brilliant, M. H. (2001). Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4. *American Journal of Human Genetics* **69(5)** 981–988.

[21] Edenberg, H. J. (2002). The collaborative study on the genetics of alcoholism: An update. *Alcohol Research and Health* **26(3)** 214–218.

[22] Bierut, L. J., Saccone, N. L., Rice, J. P., Goate, A., Foroud, T., Edenberg, H., Almasy, L., Conneally, P. M., Crowe, R., Hesselbrock, Li, V. T. K., Nurnberger, J., Porjesz, B., Schuckit, M. A., Tischfield, J., Begleiter, H., and Reich, T. (2002). Defining alcohol-related phenotypes in humans: The collaborative study on the genetics of alcoholism. *Alcohol Research and Health* **26(3)** 208–213.

[23] Edenberg, H. J. and Foroud, T. (2006). The genetics of alcoholism: identifying specific genes through family studies. *Addiction Biology* **11(3–4)** 386–396.

[24] Heath, A. G. (1995). Genetic influences on alcoholism risk: A review of adoption and twin studies. *Alcohol Health and Research World* **19** 166–171.

[25] Heath, A. C., Bucholz, K. K., Madden, P. A., Dinwiddie, S. H., Slutske, W. S., Bierut, L. J., Statham, D. J., Dunne, M. P., Whitfield, J. B., and Martin, N. G. (1997). Genetic and environmental contributions to alcohol dependence risk in a national twin sample: Consistency of findings in women and men. *Psychological Medicine* **27** 1381–1396.

[26] Heath, A. C., Jardine, R., and Martin, N. G. (1989). Interactive effects of genotype and social environment on alcohol consumption in female twins. *Journal of Studies on Alcohol* **50** 38–48.

[27] Yang, B. Z., Kranzler, H. R., Zhao, H., Gruen, J. R., Luo, X., and Gelernter, J. (2008). Haplotypic variants in DRD2, ANKK1, TTC12, and NCAM1 are associated with comorbid alcohol and drug dependence. *Alcohol. Clin. Exp. Res.* **32(12)** 2117–2127.

[28] Nelson, E. C., Lynskey, M. T., Heath, A. C., Wray, N., Agrawal, A., Shand, F. L., Henders, A. K., Wallace, L., Todorov, A. A., Schrage, A. J., Saccone, N. L., Madden, P. A., Degenhardt, L., Martin, N. G., and Montgomery, G. W. (2013). ANKK1, TTC12, and NCAM1 polymorphisms and heroin dependence: Importance of considering drug exposure. *Journal of The American Medical Association Psychiatry* **70(3)** 325–333.

[29] Lobach, I., Carroll, R. J., Spinka, C., Gail, M. H., and Chatterjee, N. (2008). Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics* **64** 673–684. MR2526616

[30] Hansson, J. (2010). Familial cutaneous melanoma. *Adv. Exp. Med. Biol.* **685** 134–145.

[31] Jemal, A., Siegel, R., Ward, E. et al. (2009). Cancer statistics. *CA Cancer J. Clin.* **54** 225–249.

[32] Reuter, N. P., Bower, M., Scoggins, C. R., Martin, R. C., McMasters, K. M., and Chagpar, A. B. (2010). The lower incidence of melanoma in women may be related to increased preventative behaviors. *Am. J. Surg.* **200(6)** 765–768.

[33] Feskanich, D., Hunter, D. J., Willett, W. C, Spiegelman, D., Stampfer, M. J., Speizer, F. E., and Colditz, G. A. (1999). Oral contraceptive use and risk of melanoma in premenapausal women. *British Journal of Cancer* **81(5)** 918–923.

[34] Karagas, M. R., Zens, M. S., Stukel, T. A., Swerdlow, A. J., Rosso, S., Osterlind, A., Mack, T., Kirkpatrick, C., Holly, E. A., Green, A., Gallagher, R., Elwood, J. M., and Armstrong, B. K. (2006). Pregnancy history and incidence of melanoma in women: A pooled analysis. *Cancer Causes Control* **17(1)** 11–19.

[35] Kvale, G., Heuch, I., and Nilssen, S. (1994). Parity in relation to mortality and cancer incidence: A prospective study of Norwegian women. *Int. J. Epidemiol.* **23(4)** 691–699.

[36] Lobach, I., Mallick, B. K., and Carroll, R. J. (2011). Semiparameteic Bayesian analysis of gene-envrionment interactions with error in measurement of environmental covariates and missing genetic data. *Statistics and Its Interface* **4(3)** 305–316. MR2834965

Iryna Lobach
Department of Neurology
School of Medicine
University of California, San Francisco, CA 94185
USA
E-mail address: ilobach@memory.ucsf.edu

Ruzong Fan
Biostatistics and Bioinformatics Branch
Division of Intramural Population Health Research
Eunice Kennedy Shriver National Institute of Child Health and Human Development
National Institutes of Health, Rockville, MD 20852
USA
E-mail address: ruzong.fan@nih.gov

Prashiela Manga
Ronald O. Perelman Department of Dermatology
School of Medicine
New York University, New York, NY 10016
USA
E-mail address: prashiela.manga@nyumc.org