The Annals of Applied Statistics
2013, Vol. 7, No. 1, 543–569
DOI: 10.1214/12-AOAS599

BAYESIAN SEMIPARAMETRIC ANALYSIS FOR TWO-PHASE STUDIES OF GENE-ENVIRONMENT INTERACTION[1]

BY JAEIL AHN[2], BHRAMAR MUKHERJEE[2,3,4], STEPHEN B. GRUBER[4] AND MALAY GHOSH[2]

University of Michigan, University of Michigan, University of Southern California and University of Florida

The two-phase sampling design is a cost-efficient way of collecting expensive covariate information on a judiciously selected subsample. It is natural to apply such a strategy for collecting genetic data in a subsample enriched for exposure to environmental factors for gene-environment interaction (G x E) analysis. In this paper, we consider two-phase studies of G x E interaction where phase I data are available on exposure, covariates and disease status. Stratified sampling is done to prioritize individuals for genotyping at phase II conditional on disease and exposure. We consider a Bayesian analysis based on the joint retrospective likelihood of phases I and II data. We address several important statistical issues: (i) we consider a model with multiple genes, environmental factors and their pairwise interactions. We employ a Bayesian variable selection algorithm to reduce the dimensionality of this potentially high-dimensional model; (ii) we use the assumption of gene–gene and gene-environment independence to trade off between bias and efficiency for estimating the interaction parameters through use of hierarchical priors reflecting this assumption; (iii) we posit a flexible model for the joint distribution of the phase I categorical variables using the nonparametric Bayes construction of Dunson and Xing [*J. Amer. Statist. Assoc.* **104** (2009) 1042–1051]. We carry out a small-scale simulation study to compare the proposed Bayesian method with weighted likelihood and pseudo-likelihood methods that are standard choices for analyzing two-phase data. The motivating example originates from an ongoing case-control study of colorectal cancer, where the goal is to explore the interaction between the use of statins (a drug used for lowering lipid levels) and 294 genetic markers in the lipid metabolism/cholesterol synthesis pathway. The subsample of cases and controls on which these genetic markers were measured is enriched in terms of statin users. The example and simulation results illustrate that the proposed Bayesian approach has a number of advantages for characterizing joint effects of genotype and exposure over existing alternatives and makes efficient use of all available data in both phases.

Received August 2011; revised June 2012.

[1]Genotyping and data collection were supported by R01 CA81488 and N01 CN43302.

[2]Supported in part by NSF Grant DMS-10-07494.

[3]Supported by R03 CA156608 and NIH/NIEHS Grant ES020811.

[4]Supported by NIH Grant U19 NCI-895700.

Key words and phrases. Biased sampling, colorectal cancer, Dirichlet prior, exposure enriched sampling, gene-environment independence, joint effects, multivariate categorical distribution, spike and slab prior.

1. Introduction. Case-control studies are popular analytical tools, particularly in cancer epidemiology, for assessing gene-disease association where the allele/genotype frequencies at a bi-allelic single nucleotide polymorphism (SNP) locus are compared between cases and controls. Recent genomewide case-control association studies (GWAS) have been remarkably successful in identifying susceptibility loci for many cancers [Yeager et al. (2007), Hunter et al. (2007), Amundadottir et al. (2009)]. A large fraction of variability in the different cancer traits still remain unexplained, with the identified SNPs contributing modestly to prediction of disease risk [Wacholder et al. (2010), Park et al. (2010)]. In search of the missing heritability, it is thus natural to study the genetic architecture of a cancer phenotype in conjunction with the known environmental risk factors (environmental toxins, dietary exposures, physical activity levels, medication use and other behavioral risk factors). In the post-GWAS era, more efficient statistical approaches to characterize such complex gene-environment (G x E) interactions, in terms of both design and analytic tools, have become a pressing need in cancer epidemiology research.

Variants of the case-control sampling design have been often employed in epidemiologic studies. Two-phase stratified sampling [Neyman (1938)] is an efficient alternative to the traditional cohort and case-control designs [Cochran (1963)] from cost and resource-saving perspectives. A typical application of two-phase sampling is for collecting expensive covariate information, for example, novel biomarkers or genotype data on a prioritized subsample of the initial study base. In particular, we will consider the following setup: the binary disease outcome or case-control status D, some relatively inexpensive covariates ($\mathbf{S}$) and environmental data (E) are collected at phase I (P_1). At phase II (P_2), genotype data (G) is collected on a subset selected from the phase I sample. To select this phase II subsample, stratified sampling with strata defined by phase I data (D, E and possibly $\mathbf{S}$) is implemented.

There is a large amount of literature on two-phase designs, using different likelihood based approaches [Horvitz and Thompson (1952), Flanders and Greenland (1991), Breslow and Cain (1988)] or estimating score approaches [Reilly and Pepe (1995), Chatterjee, Chen and Breslow (2003), Robins, Rotnitzky and Zhao (1994)]. Maximum likelihood inference for such problems was considered in the pioneering work of Scott and Wild (1997) and Breslow and Holubkov (1997a, 1997b). Lawless, Kalbfleisch and Wild (1999) and Breslow and Chatterjee (1999) compare and contrast several approaches for analyzing two-phase data. It has been noted that adding more phases can lead to further efficiency gains, consequently, the two-phase design has been generalized to multi-phase designs [Whittemore and Halpern (1998), Lee, Scott and Wild (2010)]. Haneuse and Chen (2011) propose an intermediate phase between phases I and II to reduce participation bias caused by differential participation.

The potential for such sampling designs for G x E studies has been indicated in Durt (2010). Many GWAS adopt this sampling at the design phase, but little atten-

tion is paid at the analysis stage to address the sampling design, thus potentially leading to biased estimates. To the best of our knowledge, literature on two-phase studies of G x E interaction is very limited. Chatterjee and Chen (2007) proposed maximum likelihood inference using a novel regression model for G x E interaction studies where second stage sampling was carried out based on disease outcome and family history. Asymptotic theories were established under the assumption of independence of the genetic and environmental factors in the population.

Multiple papers [Piegorsch, Weinberg and Taylor (1994), Umbach and Weinberg (1997), Chatterjee and Carroll (2005)] attest the phenomenon of gaining efficiency in studies of G x E by exploiting independence between the genetic and environmental factors under case-control sampling. Under such constraints, it is beneficial to use the retrospective likelihood for estimating interaction parameters instead of standard prospective logistic regression. However, with departures from these constraints, biases in estimating the interaction parameter can occur under retrospective methods. Several researchers have addressed this issue and proposed more robust strategies for testing G x E interaction [Mukherjee et al. (2008, 2010), Mukherjee and Chatterjee (2008), Vansteelandt, VanderWeele and Robins (2008), Li and Conti (2009), Murcray, Lewinger and Gauderman (2009)]. There is no standard multivariate tool for handling multiple genetic markers simultaneously for G x G and G x E studies that data-adaptively exploits gene–gene and gene-environment independence for gaining efficiency in estimating *multiple* SNP x E interaction parameters in a potentially high-dimensional model.

Bayesian literature on two-phase studies, even beyond the context of G x E studies, is also very limited. Haneuse and Wakefield (2007) presented the first hierarchical Bayesian work that closely relates to such data structure. The Bayesian framework presented in this paper appears to be a natural route to explore for multiple reasons. First, Bayesian estimation can lead to efficient computational algorithms, as the two-phase likelihood is naturally a missing data likelihood. Second, for G x E studies, Bayesian methods provide data-adaptive shrinkage to leverage the constraints of gene-environment independence by imposing informative priors around this assumption. Third, we incorporate Bayesian variable selection features which help us to handle a potentially high-dimensional disease risk model with main effects and interactions of multiple genes and environmental factors simultaneously. Fourth, we use the clever nonparametric Bayesian construction of Dunson and Xing (2009) as a substitute for profile likelihood in the frequentist setting to construct the retrospective likelihood under two-phase sampling. The current paper thus contributes to analysis of G x E studies with multiple markers/environmental exposures under an outcome-exposure stratified two-phase sampling design by offering a new Bayesian treatment of the problem. Our data analysis and simulation studies illustrate that for characterizing subgroup effects of the environmental exposure across genotype categories, our method provides gain in efficiency compared to other alternatives. Moreover, there are no comparable alternatives that can

offer the flexibility of our method in terms of multi-marker models and efficient G x E analysis under the two-phase design.

The paper is largely motivated by an example that originates from a population based case-control study of colorectal cancer (CRC) in Israel, namely, the Molecular Epidemiology of Colorectal Cancer (MECC) study. Statins (our environmental factor E) are a class of lipid-lowering drugs used by more than 25 million individuals worldwide for reducing cardiovascular disease risk. The MECC study was the first to establish a chemoprotective association of statins with risk of CRC [Poynter et al. (2005)]. Follow-up individual studies and a meta analysis of 18 studies have confirmed this association [Hachem et al. (2009)]. The benefit of statins for reducing CRC risk has been shown to vary with genetic variations in the HMGCR (3-Hydroxy-3-methylglutaryl coenzyme A reductase) gene, a gene involved in cholesterol synthesis [Lipkin et al. (2010)]. To understand the mechanism of effect modification further, investigators measured 294 SNPs in 40 genes, including HMGCR (our set of genetic factors $\mathbf{G}$), selected in the cholesterol synthesis/lipid metabolism pathway. The subsample selected for genotyping from the study population of all cases and controls was chosen by stratified sampling conditional on statin use (E) and case-control status (D) where statin users were purposefully oversampled. This sampling strategy was adopted due to limited budgetary resources and DNA samples. Complete statin use (E) data and other basic demographic covariates ($\mathbf{S}$) were available on the entire study base (phase I or P_1), and genetic data on these 294 SNPs were only available for the phase II subsample (P_2).

In addition, in the MECC study, due to experimental and laboratory logistics, genotype data were missing on a subset of individuals selected in P_2 on a group of genes ($\mathbf{G}_1$, say) and on a different subset of individuals on another group of genes ($\mathbf{G}_2$, say). This led to a nonmonotone missing data structure with some individuals in P_2 having observations on both $(\mathbf{G}_1, \mathbf{G}_2)$ [subset denoted by $P_2(\mathbf{G}_1, \mathbf{G}_2)$] and some only on $\mathbf{G}_1$ [subset denoted by $P_2(\mathbf{G}_1)$] and some only on $\mathbf{G}_2$ [subset denoted by $P_2(\mathbf{G}_2)$]. Figure 1 is a flow diagram of the sampling scheme and missingness pattern in the data.

The rest of the paper is organized as follows. In Section 2 we present the model ingredients: the likelihood, priors and posteriors. In Section 3 we discuss the analysis of statin x gene interaction in the MECC study. In Section 4 we conduct a simulation study to compare the various maximum likelihood and score based approaches with the Bayesian approach. Section 5 concludes with a discussion.

2. Proposed methods.

2.1. *The likelihood.* We refer to Figure 1 for understanding the data structure and construction of our likelihood. Let u and D denote the subject indicator and disease status, respectively. Here, E is environmental exposure and $\mathbf{S}$ are basic demographic covariates as described before. Let $W = (E, \mathbf{S})$. There are N indi-

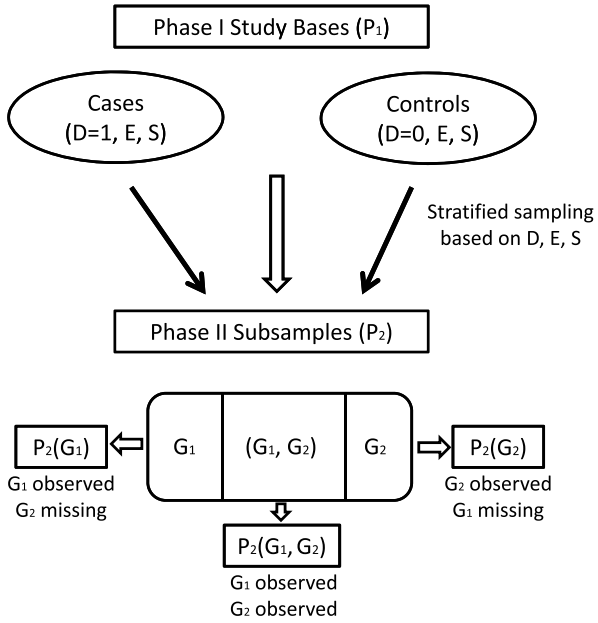

FIG. 1. *Data structure under two-phase sampling with partial missingness in phase II genetic covariates from the Molecular Epidemiology of Colorectal Cancer study.*

viduals in phase I and M individuals in phase II. To simplify notation, we write the retrospective likelihood corresponding to a two-gene model (G_1, G_2), with the understanding that the methods/notation can be directly extended to gene-sets $(\mathbf{G}_1, \mathbf{G}_2)$ where each contain multiple SNPs. The two-phase likelihood has the following form to capture the sampling phases and the missingness patterns in $\mathbf{G}$ (Figure 1):

$$\mathrm{L}^{\mathrm{TP}} = \prod_{u \in P_1 \setminus P_2} \mathrm{P}(W_u|D_u) \times \prod_{u \in P_2(G_1)} \mathrm{P}(G_{1u}, W_u|D_u) \times \prod_{u \in P_2(G_2)} \mathrm{P}(G_{2u}, W_u|D_u) \times \prod_{u \in P_2(G_1,G_2)} \mathrm{P}(G_{1u}, G_{2u}, W_u|D_u).$$

Each term in L^{TP} can be factorized by using $\mathrm{P}(G_1, G_2, W|D) = \{\mathrm{P}(D|G_1, G_2, W)\,\mathrm{P}(G_1, G_2|W)\mathrm{P}(W)\}/\mathrm{P}(D)$. This retrospective likelihood is then marginalized over the missing data in each term. We assume missing completely at random [Little and Rubin (2002)] for the genotype data collected at phase II. The likelihood is then expressed as

$$\mathrm{L}^{\mathrm{TP}} = \prod_{u \in P_1 \setminus P_2} \sum_{g_1, g_2} \mathrm{P}(D_u|g_1, g_2, W_u)\mathrm{P}(g_1, g_2|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u) \times \prod_{u \in P_2(G_1)} \sum_{g_2} \mathrm{P}(D_u|G_{1u}, g_2, W_u)\mathrm{P}(G_{1u}, g_2|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u) \tag{2.1}$$

$$\times \prod_{u\in P_2(G_2)} \sum_{g_1} \mathrm{P}(D_u|g_1, G_{2u}, W_u)\mathrm{P}(g_1, G_{2u}|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u)$$

$$\times \prod_{u\in P_2(G_1,G_2)} \mathrm{P}(D_u|G_{1u}, G_{2u}, W_u)\mathrm{P}(G_{1u}, G_{2u}|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u),$$

where $\mathrm{P}(D_u) = \sum_{g_1,g_2} \int_w \mathrm{P}(D_u|g_1, g_2, w)\mathrm{P}(g_1, g_2|w)\mathrm{P}(dw)$ with the integral replaced by the sum when components of W are discrete. Corresponding to this likelihood, there are three model ingredients:

1. A DISEASE RISK MODEL. We assume $\mathrm{P}(D = 1|G_1 = g_1, G_2 = g_2, W = w; \boldsymbol{\beta}) = H[\{\beta_0 + m(g_1, g_2, w; \boldsymbol{\beta})\}]$, where H is the logistic function $H(u) = \{1 + \exp(-u)\}^{-1}$. Typical choice of m involves, say, for two genes G_1 and G_2, $m(g_1, g_2, w; \boldsymbol{\beta}) = \beta_{G_1} g_1 + \beta_{G_2} g_2 + \beta_E e + \boldsymbol{\beta}_S^\top \mathbf{s} + \beta_{G_1G_2} g_1 g_2 + \beta_{G_1E} g_1 e + \beta_{G_2E} g_2 e$, noting that $w = (e, \mathbf{s})$.

2. A MODEL FOR $(G_1, G_2|W = (E, \mathbf{S}))$. For genotype data at a bi-allelic locus, G_j can take three possible values ("$g_0 =$ aa," "$g_1 =$ Aa" and "$g_2 =$ AA"). We assume, $\mathrm{P}(G_1 = g_j, G_2 = g'_j|W = w; \boldsymbol{\lambda}) = q_{jj'}(w; \boldsymbol{\lambda})$, $j, j' = 0, 1, 2$. This specification will require a joint model for multivariate categorical data (trinary for SNP data at a bi-allelic locus). Under gene–gene and gene-environment independence, the model can in general be factorized conditional on covariates $\mathbf{S}$, for $j, j' = 0, 1, 2$,

$$\underbrace{\mathrm{P}\big(G_1 = g_j, G_2 = g'_j|E = e, \mathbf{S} = \mathbf{s}; \boldsymbol{\lambda}\big) = \mathrm{P}(G_1 = g_j|\mathbf{S} = \mathbf{s}, \boldsymbol{\lambda}_1)\mathrm{P}\big(G_2 = g'_j|\mathbf{S} = \mathbf{s}, \boldsymbol{\lambda}_2\big)}_{\text{under G-G and G-E independence}}.$$

Instead of the above fully nonparametric model, we explore a parametric model for the joint distribution $\mathrm{P}(\mathbf{G}_1, \mathbf{G}_2|W)$. We consider a class of log-linear models with linear by linear structure [Agresti (2002)] for parsimonious modeling of the $(G_1, G_2|W)$ associations,

$$\begin{aligned}\log\big\{\mu\big(G_1 = g_j, G_2 = g'_j|E = e, \mathbf{S} = \mathbf{s}; \boldsymbol{\lambda}\big)\big\} \\ = \lambda_0 + \lambda_{G_1} g_j + \lambda_{G_2} g'_j + \lambda_E e + \boldsymbol{\lambda}_S^\top \mathbf{s} \\ + \lambda_{G_1G_2} g_j g_{j'} + \lambda_{G_1E} g_j e + \lambda_{G_2E} g_{j'} e + \boldsymbol{\lambda}_{G_1S}^\top g_j \mathbf{s} + \boldsymbol{\lambda}_{G_2S}^\top g_{j'} \mathbf{s},\end{aligned} \tag{2.2}$$

where g_j are chosen ordinal scores, typically 0, 1, 2 [Agresti (2002)]. This is the common allelic dosage coding under a log-additive genetic susceptibility model. Our method could easily be extended to a co-dominant coding of the genetic factor using two dummy variables. Since log-additivity is often assumed for screening interactions, and for simplicity of presentation in terms of one parameter estimate as opposed to two, we proceed with this additive coding. Ad-

ditionally, even if the true genetic susceptibility model is co-dominant with the disease-causing allele, for a tagging marker which is correlated to this causal allele, one would not a'priori know the direction of association of the marker allele and causal allele. Pfeiffer and Gail (2003) show that the additive scores are more robust to choice of marker allele and varying correlation scenarios. In case of high-dimensional $\mathbf{G}$, we can further reduce the dimensionality of the problem by assuming common association parameters λ_{GE} and λ_{GS} between similar functional groups of SNPs. As discussed in Agresti (2002), this Poisson log-linear model has a corresponding multinomial representation. Thus, the probability of $P_{G_1,G_2}(g_j, g'_j|\boldsymbol{\lambda}) = P(G_1 = g_j, G_2 = g'_j|E = e, \mathbf{S} = \mathbf{s})$ can be written in terms of the multinomial probabilities,

$$
\begin{aligned}
&P_{G_1,G_2}\big(g_j, g'_j|\boldsymbol{\lambda}\big) \\
&\quad = \exp\big(\lambda_{G_1} g_j + \lambda_{G_2} g'_j + \lambda_{G_1G_2} g_j g_{j'} \\
&\qquad + \lambda_{G_1E} g_j e + \lambda_{G_2E} g'_j e + \boldsymbol{\lambda}^{\top}_{G_1S} g_j \mathbf{s} + \boldsymbol{\lambda}^{\top}_{G_2S} g'_j \mathbf{s}\big) \\
&\qquad \times \Bigg(\sum_{l=0}^{2}\sum_{l'=0}^{2} \exp\big(\lambda_{G_1} g_l + \lambda_{G_2} g'_l + \lambda_{G_1G_2} g_l g'_{l'} \\
&\qquad\qquad + \lambda_{G_1E} g_l e + \lambda_{G_2E} g'_{l'} e + \boldsymbol{\lambda}^{\top}_{G_1S} g_l \mathbf{s} + \boldsymbol{\lambda}^{\top}_{G_2S} g'_{l'} \mathbf{s}\big)\Bigg)^{-1}.
\end{aligned}
$$

Note that gene–gene and gene-environment independence in the above model (2.2) will imply $\lambda_{G_1E} \equiv \lambda_{G_2E} \equiv \lambda_{G_1G_2} \equiv 0$.

3. A MODEL FOR $W = (E, \mathbf{S})$. A nonparametric and flexible model for the distribution of W is desired. Recall that W can be a mixed set of quantitative and categorical variables. For the MECC example W is a set of categorical covariates, which will be our primary focus in this paper. The approach for modeling the joint distribution of a set of categorical variables that we follow for W can also be applied to the the joint distribution of the trinary genotype variables $\mathbf{G}_1$ and $\mathbf{G}_2$ in (2.2) as well. However, reflecting prior faith on the gene–gene and gene-environment independence assumptions through direct priors on parameters $\lambda_{G_1E}, \lambda_{G_2E}, \lambda_{G_1G_2}$ in the log-linear model is more straightforward for a practitioner (2.2). This is the primary reason for using (2.2) for the second component $P(G_1, G_2|W = (E, \mathbf{S}))$.

Let $\mathbf{W}_u = (E_u, \mathbf{S}_u)$ denote the W data corresponding to subject u, $u = 1, \ldots, N$. Here W_u is $p \times 1$ vector of p categorical variables, that is, $W_u = (w_{u1}, \ldots, w_{up})$ for a subject u. Assume that the jth component of W can have d_j values $j = 1, \ldots, p$. In order to parsimoniously model this $(d_1 \times d_2 \times \cdots \times d_p)$ joint distribution, DX first note that the joint distribution of two categorical variables can always be expressed as a finite mixture of product-multinomial distributions. Ex-

tending this idea, DX introduce a latent class index variable $z_u \in \{1, \ldots, k\}$, such that $w_{ur}, w_{ut}, r, t \in \{1, \ldots, p\}, r \neq t$, are conditionally independent given z_u. Then the joint distribution for $\mathbf{w}_u$ has this finite mixture representation,

$$\begin{aligned} &P_W(w_{u1} = c_1, \ldots, w_{up} = c_p) \\ &\quad = \sum_{h=1}^{k} P(w_{u1} = c_1, \ldots, w_{up} = c_p | z_u = h) P(z_u = h) \\ &\quad = \sum_{h=1}^{k} P(z_u = h) \prod_{j=1}^{p} P(w_{uj} = c_j | z_u = h). \end{aligned} \tag{2.3}$$

For notational convenience, we rewrite (2.3) as

$$\begin{aligned} &P_W(w_{u1} = c_1, \ldots, w_{up} = c_p) = \pi_{c_1 \cdots c_p} = \sum_{h=1}^{k} \nu_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}, \\ &\sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \pi_{c_1 \cdots c_p} = 1, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^\top$ is a probability vector with $\nu_h = P(z_u = h)$ and $\psi_{hc_j}^{(j)} = P(w_{uj} = c_j | z_u = h)$ is a $d_j \times 1$ probability vector, that is, the conditional probability of $w_{uj} = c_j$, given that subject u is in latent class h for $j = 1, \ldots, p$. We will discuss the choice of k through a Dirichlet process prior structure on this latent class probability model in the next section.

REMARK 1. While Chatterjee and Chen (2007) and Chatterjee and Carroll (2005) use profile likelihood for handling the distribution of W nonparametrically, it has been a challenging task in the Bayesian framework to posit a flexible model for $\mathbf{W} = (E, \mathbf{S})$ which could be a mixture of categorical and continuous covariates. In this mixed case, Müller et al. (1999) model the joint distribution of the continuous covariates through a Dirichlet process mixture of normals. Then, conditional on the continuous covariates, the categorical variables have a joint multivariate probit distribution. A recent paper by Bhattacharya and Dunson (2012) extends the above DX construction for categorical data to handle joint distribution modeling of more complex data, including continuous and discrete data. They extend the conditional independence idea and replace the product-multinomial structure in (2.4) by a product of various kernels, such as Gaussian, Poisson and more complex univariate or multivariate distributional kernels. The MECC example does not require going beyond the original DX construction, but with continuous E, this is what we would adopt.

REMARK 2. If the phase I sample is a cohort study, with disease endpoint D, then the corresponding likelihood is proportional to

$$
\begin{aligned}
\mathrm{L}^{\mathrm{cohort,TP}} \propto & \prod_{u\in P_1\setminus P_2} \sum_{g_1,g_2} \mathrm{P}(D_u|g_1,g_2,W_u)\mathrm{P}(g_1,g_2|W_u) \\
& \times \prod_{u\in P_2(G_1)} \sum_{g_2} \mathrm{P}(D_u|G_{1u},g_2,W_u)\mathrm{P}(G_{1u},g_2|W_u) \\
& \times \prod_{u\in P_2(G_2)} \sum_{g_1} \mathrm{P}(D_u|g_1,G_{2u},W_u)\mathrm{P}(g_1,G_{2u}|W_u) \\
& \times \prod_{u\in P_2(G_1,G_2)} \mathrm{P}(D_u|G_{1u},G_{2u},W_u)\mathrm{P}(G_{1u},G_{2u}|W_u).
\end{aligned} \tag{2.5}
$$

Similarly, if environmental data E is collected in phase II as well, the first term representing the phase I cohort likelihood can also involve an integral over the missing E data with respect to a probability distribution $dF(E)$, exactly as in equation (3) of Chatterjee and Chen (2007). A surrogate measure of E, namely, E^*, may be available in phase I and a measurement error model relating E and E^* can also be used to construct a joint likelihood of phases I and II data.

2.2. *Priors*. As mentioned before, for this complex retrospective likelihood formulation, we have three sets of parameters from the above three ingredients of the likelihood. For $\boldsymbol{\beta}$ in the disease risk model, we use a spike and slab type mixture prior to handle variable selection in a high-dimensional disease risk model with multiple markers. For $\boldsymbol{\lambda}$ in the multivariate gene model, the Bayesian hierarchical approach provides a flexible way to allow for uncertainty around the assumption of gene–gene and gene-environment independence, through prior on $\lambda_{G_1G_2}$, λ_{G_1E} and λ_{G_2E}. When sparsity occurs in a certain configuration of $(\mathbf{G}_1,\mathbf{G}_2,\mathbf{W})$ or dimension of $(\mathbf{G}_1,\mathbf{G}_2,\mathbf{W})$ grows, the frequentist profile likelihood estimation may become unstable and the log-linear model with shared parameters across gene-sets and the DX latent mixture construction aid with such situations. We follow the same sequence as in the previous section to describe the prior structure on the parameters.

1. In the presence of multiple genes in $\mathbf{G}_1$ and $\mathbf{G}_2$, the logistic disease risk model can potentially have many pairwise and higher order interaction terms. We implement a scalable variable selection framework via spike and slab type priors [Mitchell and Beauchamp (1988), George and Mcculloch (1993)] on the parameters $\boldsymbol{\beta}$ in the disease risk model $P(D|\mathbf{G}_1,\mathbf{G}_2,W;\boldsymbol{\beta})$. We impose mixture prior distributions on each component of $\boldsymbol{\beta}$, say, $(\beta_0,\beta_{G_1},\beta_{G_2},\beta_E,\beta_S,\beta_{G_1G_2},\beta_{G_1E},\beta_{G_2E})$ for a two-gene model. In general, we denote this vector by $\boldsymbol{\beta}_{n_\beta\times 1}=\{\beta_r, r=1,\ldots,n_\beta\}$. Given a latent variable p_0 representing the mixture weight on the "not informative" regression coefficients, we describe the hierarchical prior structure as

follows:

$$\begin{aligned}
\beta_r | f_r, \tau_r &\overset{\text{ind}}{\sim} N(0, f_r \tau_r^2), \qquad r = 1, \ldots, n_\beta, \\
f_r | v_0, p_0 &\overset{\text{i.i.d.}}{\sim} p_0 \delta_{v_0}(\cdot) + (1 - p_0)\delta_1(\cdot), \\
\tau_r^{-2} | a_1, a_2 &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_1, a_2), \\
p_0 &\overset{\text{i.i.d.}}{\sim} \text{Beta}(a, b).
\end{aligned}$$

As discussed in Ishwaran and Rao (2003), v_0 in the above specification is assumed to be a small positive value near 0. Note that f_r can assume two values v_0 or 1. At each iteration of posterior sampling, f_r takes value 1 if sampled β_r is significantly away from zero, implying that the rth covariate is potentially informative. Note that a key feature of this prior specification is that the marginal prior variance of β_r is calibrated as $\text{var}(\beta_r) = f_r \tau_r^2$ and has a bimodal distribution. Large $\text{var}(\beta_r)$ can occur when $f_r = 1$ and τ_r^2 is large, inducing large values of β_r, identifying potentially informative covariates. Small values of $\text{var}(\beta_r)$ occur when f_r assumes value v_0, leading to values of β_r that are near zero, suggesting that β_r is potentially uninformative. The value of p_0 controls how likely it is for f_r to be v_0 or 1, thus controlling how many β_r are nonzero or the complexity of the model. The Gamma parameters (a, b) control the degree of parsimony through the prior on p_0. We set $(a, b) = (1, 1)$, that is, a uniform prior on p_0, for the analysis we present in the main text. Note that (a_1, a_2) determines the prior on τ_r^2 and thus the variance of β_r. We fix (a_1, a_2) at $(5, 50)$ to allow the possibility of large prior variances on β. The values used for the hyperparameters in the hierarchy are exactly as recommended in Ishwaran and Rao (2003).

2. In the joint log-linear model (2.2), we typically assume vague normal priors with large variance on the parameters $(\lambda_{G_1}, \lambda_{G_2}, \lambda_{G_1 S}, \lambda_{G_2 S})$. In our data example, we have used a $N(0, 10^4)$ prior. On the other hand, for the G-E pairwise association parameters $(\lambda_{G_1 G_2}, \lambda_{G_1 E}, \lambda_{G_2 E})$, we reflect a priori information on G-G or G-E independence via a normal prior centered at zero but with two different choices for the prior variance. In the first set of priors we reflect the belief that with 95% probability the association parameter lies between $\log(0.8)$ and $\log(1.2)$. This leads to an approximate $\text{SD} = 0.1$ under a normal distribution and, thus, we assume an informative prior of $N(0, 10^{-2})$. In the second choice, following the empirical Bayes estimation of Mukherjee and Chatterjee (2008), we compute association parameters for G_1-G_2, G_1-E, and G_2-E in the control subjects in the data, say, $\hat{\theta}$, and use a data-driven prior $N(0, \hat{\theta}^2)$ on $\lambda_{G_1 G_2}$, $\lambda_{G_1 E}$ and $\lambda_{G_2 E}$.

3. The mixture representation in (2.4) requires determining the number of latent classes k. Following DX, instead of selecting a fixed k, a Bayesian nonparametric approach is carried out through the Dirichlet process prior specification on $\boldsymbol{\nu}$:

$$\boldsymbol{\pi} = \sum_{h=1}^{\infty} \nu_h \boldsymbol{\psi}_h, \qquad \boldsymbol{\psi}_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \qquad h = 1, \ldots, \infty,$$

$$\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}) \qquad \text{independently for } j = 1, \ldots, p,$$

$$\nu_h = \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l), \qquad V_h \sim \text{Beta}(1, \alpha),$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha),$$

where $\otimes$ is the outer product. The parameter α is a hyper-parameter that controls the rate of decrease from the stick-breaking process [Sethuraman (1994)]. For example, in the case of small values of α, ν_h decreases toward zero quickly with increasing h, thus putting most of the weight on the first few components, leading to a sparse representation. The hyperprior on α allows one to data-adaptively determine the degree of sparseness or the number of components needed. As discussed in Dunson and Xing (2009), we set $(a_\alpha, b_\alpha) = (1/4, 1/4)$ for a vague prior which implies the probability of independence across components of w in the product multinomial model to be 0.5. We set uniform priors for each category probability $\boldsymbol{\psi}$ with $a_{j1} = \cdots = a_{jd_j} = 1$, for $j = 1, \ldots, p$ and let the data dominate over priors. To minimize large numbers of mixture components instead of using infinite mixtures, we truncate the maximum of the number of mixture components k at 30 in the real data example [Ahn et al. (2013)]. We study sensitivity with respect to this truncation threshold in Table 1.

2.3. *Posterior sampling.* In the full likelihood (2.1), we would like to point out that the three components are linked with each other through the sum over each component in the expression for $P(D)$ in the denominator. We denote the two-phase likelihood in (2.1) by L_{TP} which involves the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \mathbf{V}, \alpha)$. The full conditionals are not reducible to a simpler closed form and are best represented by the following proportionality relations:

$$\beta_r|\cdot \propto \text{L}^{\text{TP}} \times \exp\left(-\frac{\beta_r^2}{2 f_r \tau_r^2}\right), \qquad r = 1, \ldots, n_\beta,$$

$$\tau_r^{-2}|\cdot \propto \text{Gamma}\left(a_1 + 0.5, a_2 + \frac{\beta_r^2}{2 f_r}\right),$$

$$f_r|\cdot \propto \{I(f_r = v_0) p_0 + I(f_r = 1)(1 - p_0)\} \times \exp\left(-\frac{1}{2 f_r \tau_r^2} \beta_r^2\right) \times f_r^{-0.5},$$

$$p_0|\cdot \propto \text{Beta}\left(a + \sum_{r=1}^{n_\beta} I(f_r = v_0), b + \sum_{r=1}^{n_\beta} I(f_r = 1)\right),$$

$$\lambda_l|\cdot \propto \text{L}^{\text{TP}} \times \exp\left(-\frac{\lambda_l^2}{2\sigma^2}\right), \qquad l = 1, \ldots, n_\lambda,$$

where n_β and n_λ again represent the number of parameters in $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, respectively.

Posterior sampling corresponding to $P(\mathbf{W})$: Let us recapitulate the model structure for $\mathbf{W}$ which is essentially a Dirichlet process mixture of discrete Dirichlet kernels. For $u = 1, \ldots, N$ and $j = 1, \ldots, p$,

$$w_{uj} \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi^j_{z_u,1}, \ldots, \psi^j_{z_u,d_j}),$$

$$z_u \sim V_h \prod_{l<h} (1 - V_l)\delta_h, \qquad V_h \sim \text{Beta}(1, \alpha), \qquad \alpha \sim \text{Gamma}(a_\alpha, b_\alpha).$$

DX present an efficient data-augmented Gibbs sampling algorithm by augmenting the likelihood with latent constructs following Walker (2007). The details of the updating steps are described in the supplemental article [Ahn et al. (2013)].

Note that while the entire likelihood in DX is constituted of W data only, in our problem, $P(\mathbf{W})$ is embedded as a component in the joint retrospective likelihood L^{TP} in (2.1). Thus, for updating the parameters involved in $\text{P}(\mathbf{W})$, say, $\boldsymbol{\theta}(=\{\boldsymbol{\psi}, \mathbf{V}, \alpha\})$, we use the Metropolis Hastings algorithm. Only the terms $\prod_u \text{P}(\mathbf{W}_u)/\text{P}(D_u)$ from the full likelihood (2.1) involve $\boldsymbol{\theta}$, where $\text{P}(D_u) = \sum_{g_1,g_2} \sum_{\mathbf{w}} \text{P}(D_u|g_1, g_2, \mathbf{w})\ \text{P}(g_1, g_2|\mathbf{w})\text{P}(\mathbf{w})$. We draw $\boldsymbol{\theta}$ following the DX algorithm and for the proposal density of $\boldsymbol{\theta}$ we consider the implied full conditional $q(\boldsymbol{\theta}^{\text{new}}|\mathbf{W})$ as determined by this algorithm. Then given $\boldsymbol{\lambda}, \boldsymbol{\beta}$, we repeat the following updates of $\boldsymbol{\theta}$:

- At iteration l, sample a vector $\boldsymbol{\theta}^{\text{new}}$ from $q(\boldsymbol{\theta}^{\text{new}}|\mathbf{W})$ as described in the Dunson and Xing (2009) algorithm.
- Compute the acceptance ratio

$$r(\boldsymbol{\theta}^{\text{new}}, \boldsymbol{\theta}_l) = \min\left[1, \frac{\prod_u \text{P}(D_u|\boldsymbol{\theta}_l, \boldsymbol{\lambda}, \boldsymbol{\beta})}{\prod_u \text{P}(D_u|\boldsymbol{\theta}^{\text{new}}, \boldsymbol{\lambda}, \boldsymbol{\beta})}\right].$$

  In calculating the acceptance ratio, we note that the numerator and denominator $\prod_u\{\text{P}(W_u|\boldsymbol{\theta}^{\text{new}})\}p(\boldsymbol{\theta}^{\text{new}})\ q(\boldsymbol{\theta}_l|\mathbf{W})/\prod_u\{\text{P}(W_u|\boldsymbol{\theta}_l)\}p(\boldsymbol{\theta}_l)q(\boldsymbol{\theta}^{\text{new}}|\mathbf{W})$ are canceled out where $p(\boldsymbol{\theta})$ is a prior for $\boldsymbol{\theta}$.
- If $r(\boldsymbol{\theta}^{\text{new}}, \boldsymbol{\theta}_l) < U$ where $U \sim \text{unif}(0, 1)$, we set $\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}^{\text{new}}$. Otherwise, the candidate vector θ^{new} is rejected and $\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l$.
- Repeat the steps until the posterior chains converge.

Given the full conditionals, we implement the Gibbs sampler [Geman and Geman (1984)] with Metropolis Hastings updates to sample from respective full conditional distributions. For each parameter, we iterate 50,000 times and discard the first 40,000 iterations as "burn-in." We check convergence of the chains using trace plots and the numerical diagnostic statistic "potential scale reduction factor" [David (1992)] using the R package CODA [Plummer et al. (2009)]. Auto and cross-correlation checks are performed and a thinning of every tenth observation is carried out. Remaining posterior samples are used to construct estimated posterior summaries needed for Bayesian inference.

3. The Molecular Epidemiology of Colorectal Cancer study. In this section we describe the motivating example from the MECC study in detail and present analysis results. We use data on 1746 cases and 1853 controls with completely observed response to the question whether statins were used for more than 5 years. The binary variable "statin use of at least 5 years" (E) is the environmental factor of interest with 91% "NO" and 9% "YES."

We adjust for completely observed confounders and precision variables ($\mathbf{S}$): age (S_1), gender (S_2), ethnicity (S_3), physical activity (S_4), family history of CRC (S_5), vegetable consumption (S_6), NSAID usage within 3 year (S_7) and Aspirin usage within 3 year (S_8). Age and ethnicity variables were dichotomized as Age $\geq$ or $<$ 50 (94% and 6%, resp.), and "Ashkenazi" and "Non-Ashkenazi" (68% and 32%, resp.). Gender (S_3) was coded as 1 (50%) for male and 0 (50%) for female. The remaining binary factors (S_4, S_5, S_6, S_7, S_8) are classified to 1 or "YES" with the proportions of (0.36, 0.09, 0.31, 0.02, 0.20), respectively.

For genotyping at phase II, stratified-sampling based on the disease status (D) and statin use (E) was carried out. All case-control subjects with statin use ("YES") were included at the phase II sample. We have 1200 cases and 1200 controls at phase II with data available on 294 trinary SNPs $\mathbf{G} = (G_1, \ldots, G_{294})$. Genotype data are not completely observed even at phase II due to technical genotyping failures for a limited number of SNPs. Among 2400 case-control subjects at phase II, 56 subjects and 20 had partial genotype information on two subsets of SNPs. We did not have a dense set of markers typed across the genome to successfully impute these missing genotypes, thus we consider a marginalized likelihood as in (2.1).

Among 294 SNPs, we first illustrate our methods with two SNPs on two genes, $RS1800775$ on CETP (G_1) and $RS1056836$ on CYP1B1 (G_2), where both SNPs exhibit significant interactions with statin use in an initial single marker interaction analysis. We compare our methods for this simple two SNP model to some of the alternative methods that can only handle single marker interaction analysis. The raw frequencies of the cross-classification of case-control status (D), statins (E), genotypes G_1 and G_2 are shown in online supplementary Table 1 [Ahn et al. (2013)]. Simple logistic regression analysis was carried out to examine G_1-E and G_2-E association among control subjects and yielded odds ratios of 1.11 and 1.01 and corresponding p-values of 0.30 and 0.91, respectively. Based on a chi-squared test for independence, G_1-G_2 reveals no association (p-value of 0.90) These tests suggest that the data support G_1-E, G_2-E and G_1-G_2 independence assumption.

We report the results of the multivariate analysis in Table 1. Along with the two-phase full Bayes approach (TPFB), we consider five alternative methods. Unfortunately, none of these competing methods use the data in both phases and make use of the independence constraints. The first three use phase II data only (i) unconstrained maximum likelihood (UML), a retrospective analysis that does not specify any constraints on $P(G_1, G_2|E, \mathbf{S})$, (ii) constrained maximum likelihood (CML),

TABLE 1

(a) Analysis results for the MECC study data with statins (E), G_1 $RS1800775$ on CETP and G_2 $RS1056836$ on CYP1B1. The model adjusts age (S_1, ">50" = 1, "≤50" = 0), gender (S_2, male = 1, female = 0), ethnicity (S_3, Ashkenazi = 1, Non-Ashkenazi = 0), sports activity (S_4, Yes = 1, No = 0), vegetable consumption (S_5, High = 1, Low = 0), family history of CRC (S_6, Yes = 1, No = 0), the use or nonuse of NSAID within 3 years (S_7, Yes = 1, No = 0), the use or nonuse of Aspirin within 3 years (S_8, Yes = 1, No = 0). Under the TPFB method the "est." corresponds to the posterior mean, whereas PSD corresponds to posterior standard deviation. The methods that yield the smallest PSD are in bold font in each row

	TPFB est.(PSD)	**TPFB_{emp} est.(PSD)**	**WL est.(se)**	**PL est.(se)**	**UML est.(se)**	**CML est.(se)**	**EB est.(se)**
Exposure variables							
G_1	0.04 (0.09)	0.01 (0.09)	0.00 (0.09)	0.00 (0.09)	0.00 (0.09)	**0.00 (0.08)**	**−0.07 (0.08)**
G_2	−0.04 (0.10)	−0.06 (0.10)	−0.13 (0.10)	−0.13 (0.10)	−0.13 (0.10)	**−0.12 (0.08)**	**−0.12 (0.08)**
Statin use	−1.29 (0.30)	**−1.32 (0.27)**	−1.30 (0.30)	−1.30 (0.30)	−1.40 (0.30)	−1.54 (0.28)	−1.51 (0.29)
G_1 x G_2	0.01 (0.07)	**0.03 (0.05)**	0.05 (0.08)	0.06 (0.08)	0.06 (0.08)	0.06 (0.06)	0.06 (0.06)
G_1 x statin use	0.34 (0.17)	**0.34 (0.15)**	0.25 (0.18)	0.25 (0.18)	0.25 (0.18)	**0.38 (0.15)**	0.34 (0.17)
G_2 x statin use	**0.33 (0.16)**	**0.33 (0.16)**	0.38 (0.18)	0.38 (0.19)	0.38 (0.20)	0.38 (0.18)	0.38 (0.19)
Gene-statin and gene–gene association parameters from $P(G_1, G_2 \mid E, \mathbf{S})$							
$\lambda_{G_1G_2}$	0.02 (0.05)	0.00 (0.05)					
λ_{G_1E}	0.05 (0.07)	0.08 (0.06)					
λ_{G_2E}	0.01 (0.07)	0.01 (0.07)					

(b) Sensitivity analysis with respect to the maximum number of allowable mixture components $k_{\max}$, and the prior on G-G and G-E association parameters $\boldsymbol{\lambda}$

		G_1	**G_2**	**Statin use**	**G_1 x G_2**	**G_1 x statin use**	**G_2 x statin use**
$k_{\max} = 10$	TPFB	0.05 (0.10)	−0.03 (0.10)	−1.29 (0.30)	0.01 (0.07)	0.36 (0.16)	0.32 (0.17)
	TPFB_{emp}	0.01 (0.09)	−0.06 (0.10)	−1.32 (0.29)	0.03 (0.07)	0.31 (0.15)	0.32 (0.16)
$k_{\max} = 30$	TPFB_{non}	0.05 (0.11)	−0.03 (0.11)	−1.29 (0.31)	0.01 (0.07)	0.34 (0.19)	0.34 (0.21)

TPFB, TPFB_{emp}, TPFB_{non}: Two-phase full Bayes [with informative prior $N(0, 10^{-2})$, using empirical estimates for prior variances, with noninformative prior $N(0, 10^4)$] on G-E association parameters; UML: unconstrained maximum likelihood, CML: constrained maximum likelihood, EB: empirical-Bayes, WL: weighted likelihood and PL: pseudo-likelihood.

that imposes the Hardy–Weinberg equilibrium as well as G_1-E/G_1-G_2 independence, (iii) empirical-Bayes (EB), using data-adaptive "shrinkage estimation" between the constrained and unconstrained ML estimates. Since methods (ii) and (iii) are developed for single marker analysis, G_2-E independence cannot be enforced in existing software [we used the "CGEN" package by Bhattacharjee, Chatterjee and Wheeler (2011)]. These three methods completely ignore biased sampling at phase II and may thus lead to biased estimation of the main effect of E, particularly if the exposure sampling rates were the differential among cases and controls. The next two approaches use information from both phases under a prospective likelihood framework: (iv) a Horvitz–Thompson estimator, typically known as a weighted likelihood (WL) approach [Manski and Lerman (1977), Breslow and Chatterjee (1999)]. This approach uses sampling fractions n_{ij}/N_{ij}, where n_{ij} and N_{ij} are the number of subjects corresponding to $D = i$, $E = j$ at phases II and I, respectively. The sampling fraction serves as weights in the likelihood to adjust for biased sampling [we used the *svyglm* function in the "survey" package in R by Lumley (2011)]. Finally, (v) a pseudo-likelihood (PL) approach which also adjusts for biased sampling probabilities in a likelihood framework [Schill et al. (1993)]. Briefly, if we denote $P_{ij} = P(D = i|E = j) = \exp(i\alpha_j)/\{1 + \exp(\alpha_j)\}$ where α_j is the log-odds for $D = 1$ when $E = j$, then the pseudo-likelihood is defined as $\prod_{i,j} P_{ij}^{N_{ij}} \prod_{i,j,k} p_{ijk}$. Here,

$$p_{ijk} = \frac{n_{ij} \exp\{i(\beta_0 - \alpha_j + s_{ijk}\beta)\}}{n_{0j} + n_{1j} \exp(\beta_0 - \alpha_j + s_{ijk}\beta)},$$

where s_{ijk} denotes covariate values for a subject with $D = i$ and $E = j$.

Note that all of these five methods use completely observed phase II data on G_1 and G_2 as opposed to our proposed method that includes partially observed data by marginalization of the likelihood in terms of G_1 and G_2 when needed.

As previously explained, we present our method (TPFB) corresponding to two different priors on the G-E and G-G association parameters in model (2.2). First, we consider informative prior $N(0, 10^{-2})$ that enforces fixed prior belief around G-E and G-G independence; we denote this by TPFB. The analysis using an alternative prior where the prior variances on λ_{GG} and λ_{GE} are estimated based on observed association in the data is denoted by TPFB_{emp}. In Table 1, the variable selection scheme is excluded in the TPFB and TPFB_{emp} by assuming all $f_r = 1$, $r = 1, \ldots, n_\beta$, so that all covariates are included across all methods. This is done so that the method can be fairly compared to other alternatives which do not have the variable selection feature.

Under all methods, note in Table 1 that the estimated coefficients corresponding to statin-use suggest strong negative association with CRC status. The estimated effect size varies depending on whether the method accounts for biased sampling and/or gene-environment independence. In the presence of interactions, we cannot really interpret the main effect estimates and need to combine the model results

to present estimated subgroup effects. Recall that G_1-E and G_2-E independence does appear to be plausible in light of this data. Note that while G_2 x E interaction is detected by all methods, G_1 x E interaction can only be detected by CML, EB, TPFB and TPFB_{emp}, that is, methods that use the independence assumption. The TPFB estimates of terms involving E are slightly different in effect sizes with smaller standard errors when compared to the other methods. Smaller standard errors corresponding to interaction parameters are noted in all retrospective methods that explicitly model (G_1, G_2, E) dependence structure.

We also carried out a sensitivity analyses with respect to the choice of threshold to truncate the maximum value of k in the DX construction and the prior on G-E and G-G association. As can be seen from Table 1(b), the results are almost identical with a smaller number ($k_{\max} = 10$) of components in the mixture distribution for W. This suggests further computational efficiency gain is possible by imposing more parsimonious constraint on k. In another sensitivity analysis, when the prior on G-E association is noninformative $N(0, 10^4)$, we notice TPFB estimates slightly drift toward the estimates from PL and WL while losing some efficiency on the G_1 x E and G_2 x E terms.

To reflect our main interest in subgroup effects of statin across genotype configurations, we report effects of statin across genotype subgroups of one SNP, holding the other SNP fixed at the common genotype category for that second SNP (coded as 0) in Table 2. It seems that statin effect is strongly modified by G_1 and G_2. Ac-

TABLE 2

Odds ratio estimates (*confidence interval or credible interval*) *for CRC corresponding to statin users vs nonusers across genotype subgroups. Under all five methods, a model with main effect of* G_1, G_2, E *controlling for S was fit as in Table* 1. *Common alleles in* G_1 (*RS*1800775 *on CETP*) *and* G_2 (*RS*1056836 *on CYP*1*B*1) *are A and C, respectively, and minor alleles in* G_1 *and* G_2 *are C and G, respectively*

	Statins				
G_1 G_2	**A/A C/C**	**A/C C/C**	**C/C C/C**	**A/A G/C**	**A/A G/G**
TPFB	0.27 (0.15, 0.39)	0.35 (0.22, 0.44)	0.48 (0.26, 0.65)	0.38 (0.24, 0.51)	0.48 (0.30, 0.77)
TPFB_{emp}	0.26 (0.15, 0.42)	0.36 (0.23, 0.45)	0.49 (0.31, 0.69)	0.37 (0.23, 0.50)	0.50 (0.31, 0.79)
WL	0.27 (0.15, 0.49)	0.35 (0.22, 0.55)	0.45 (0.26, 0.77)	0.40 (0.26, 0.62)	0.59 (0.34, 1.02)
PL	0.27 (0.15, 0.49)	0.35 (0.22, 0.55)	0.45 (0.26, 0.79)	0.40 (0.25, 0.63)	0.59 (0.33, 1.05)
UML	0.25 (0.14, 0.44)	0.32 (0.20, 0.50)	0.41 (0.23, 0.72)	0.36 (0.23, 0.57)	0.53 (0.29, 0.95)
CML	0.22 (0.12, 0.37)	0.33 (0.21, 0.51)	0.49 (0.29, 0.83)	0.34 (0.20, 0.48)	0.46 (0.26, 0.80)
EB	0.22 (0.13, 0.39)	0.31 (0.20, 0.49)	0.43 (0.24, 0.79)	0.32 (0.21, 0.49)	0.47 (0.27, 0.82)

TPFB, TPFB_{emp}: Two-phase full Bayes (with empirical estimates for prior variances), UML: unconstrained maximum likelihood, CML: constrained maximum likelihood, EB: empirical-Bayes, WL: weighted likelihood, and PL: pseudo-likelihood.

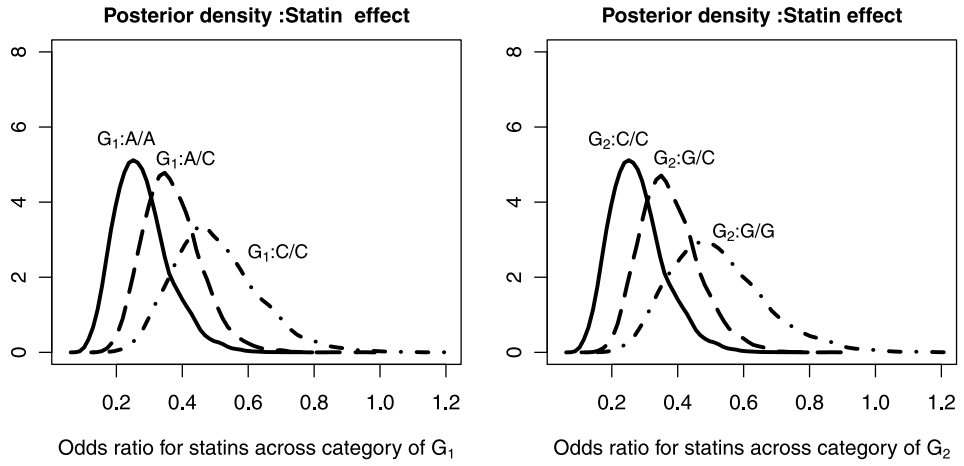

FIG. 2. *The left figure shows the posterior densities of the odds ratio estimates of CRC corresponding to statin users versus nonusers across three genotypes in RS*1800775 *on CETP*(G_1), *holding the genotype in RS*1056836 *on CYP*1*B*1 *at the most frequent category, that is,* $(G_2) = (C/C)$. *Similarly, the right figure shows the posterior densities of the odds ratio estimates corresponding to statin users versus nonusers across three genotypes in RS*1056836 *of CYP*1*B*1(G_2), *holding the genotype in RS*1800775 *of CETP fixed at the most frequent category, that is,* $(G_1) = (A/A)$.

cording to TPFB_{emp} estimates, keeping the G_2 genotype fixed at C/C, the benefit of taking statins to reduce the risk of CRC is maximum in the A/A genotype of G_1 with the posterior estimate (and 95% HPD) of the odds-ratios (corresponding to statin users versus nonusers) being 0.26 (0.15, 0.42). The corresponding ORs in genotype category A/C and C/C are 0.36 (0.23, 0.45) and 0.49 (0.31, 0.69), respectively. Figure 2 illustrates estimated posterior densities of the odds ratios corresponding to statin-use across each genotype of G_1 (left) or G_2 (right), respectively, while holding the other SNP fixed at the most common category. This figure indicates that the protective effect of statin in CRC is diminishing as the allelic dosage for the minor allele increases in both G_1 and G_2. Overall, the TPFB approaches provide much narrower credible intervals compared to PL and WL by exploiting G_1-E and G_2-E independence. The estimates from methods that use phase II data only, like CML, UML and EB, are numerically slightly different.

VARIABLE SELECTION: We explore how variable selection feature performs in this example for the TPFB method. Previous research by Ishwaran and Rao (2003) discussed the performance of spike and slab prior for general variable selection. We introduce three SNPs (RS5925224, RS10174721, RS10077453) and all possible pairwise G x G and G x E interactions to the previous two SNP model as fit in Table 1. The dimension of the disease risk model is now 34. None of the main effects and interactions corresponding to these three additional SNPs were found significant in an initial single marker analysis.

We set $f_r = 1$ for S_1 through S_8 to always keep the confounders and precision variables in the model. The tuning parameters v_0 are fixed at 0.0001 for this application with sensitivity analysis results presented for $v_0 = 0.001$ in Table 3. We would like to see if the variable selection can still detect a significant G_1 x E and G_2 x E interaction. Moreover, we would like to assess if the three additional SNPs and the corresponding interactions we added (with null effects as observed in our initial analysis) are also identified to be not informative by this process. We tabulate the posterior distribution of $f_r = 1$ among $\mathbf{f} = (f_1, \ldots, f_{n_\beta})$ which indicate "in-and-out" frequencies of the corresponding parameters. These posterior frequencies of $f_r = 1$ can be used to define a ranking of important predictors. An alternative is to rank the top models (not just the predictors individually). Before implementing the TPFB, we reduced the dimensionality of parameters in the model $P(\mathbf{G}|\mathbf{W})$ where $\mathbf{G} = (G_1, G_2, G_3, G_4, G_5)$ by assuming common λ_{GG} and λ_{GE} association parameters across all SNPs. We use $N(0, 0.1^2)$ prior on this common parameter. In addition, we further assume a single common parameter λ_{GS} for all G-S associations with a vague normal prior $N(0, 10^4)$. These are assumptions that may be stringent in certain situations, but to reduce estimation burden in the log-linear model, we do need to make these assumptions for the TPFB methods. For SNPs on a same functional pathway like in our example, it may not be too unrealistic to assume a shared association parameter across SNPs.

In Table 3, we present numerical results on model and predictor ranking as well as the Bayesian Information Criterion (BIC) corresponding to each model. We only present the top 10 models. According to the result, the model with main effects of E and G_2 x E interactions seems to be the most preferred model (posterior probability 13.1%) followed by the model with E and both G_1 x E and G_2 x E interactions (posterior probability 9.2%). With $v_0 = 0.001$, the ranking of predictors is slightly different, as the main effects of G_1 through G_5 are now selected more often. The bottom panel of Table 3 shows the frequency of retaining a predictor in the model according to the posterior distribution of $\mathbf{f}$. The main effect of E appears most of the times (100%) with large selection probabilities for G_1 x E and G_2 x E interactions (36.8% and 64.0%), respectively. Overall, nonsignificant interactions/main effects are well filtered under this variable selection scheme.

4. Simulation study. In this section we assess the performance of the proposed method by conducting a simulation study. We mainly consider two aspects: (i) varying gene-gene/gene-environment association structure and (ii) when phase II sampling is the differential between cases and controls. We compare our method with the five alternative methods mentioned before: WL, PL, UML, CML and EB in terms of the average bias and mean squared errors (MSE), based on 1000 simulated data sets.

TABLE 3

(a) The top 10 promising models in terms of estimated posterior probabilities of the models. All **S** adjustment variables are retained in the model by default and variable selection is performed only on the five genetic and environmental factors and all possible pairwise interactions. Bayesian Information Criterion (BIC) is provided for each model. Results in parentheses represent the sensitivity analysis carried out with $v_0 = 0.001$

Model	**Posterior probability %**	**BIC**
$[E][\text{All } S][G_2 \text{ x } E]$	13.1% (12.5%)	43,992 (43,967)
$[E][\text{All } S][G_1 \text{ x } E][G_2 \text{ x } E]$	9.2% (6.3%)	43,993 (43,978)
$[E][\text{All } S][G_1 \text{ x } E]$	7.7% (5.1%)	43,994 (43,967)
$[E][\text{All } S]$	7.5% (10.6%)	43,997 (43,977)
$[E][\text{All } S][G_2 \text{ x } E][G_3 \text{ x } E]$	3.9% (4.6%)	44,004 (43,977)
$[E][\text{All } S][G_2 \text{ x } E][G_4 \text{ x } E]$	2.4% (2.2%)	44,002 (43,974)
$[E][\text{All } S][G_2 \text{ x } E][G_3 \text{ x } G_4]$	2.1% (1.5%)	43,998 (43,971)
$[E][\text{All } S][G_1 \text{ x } E][G_3 \text{ x } E]$	2.1% (2.2%)	44,005 (43,974)
$[E][\text{All } S][G_1][G_2 \text{ x } E]$	1.8% (0.7%)	43,996 (43,976)
$[E][\text{All } S][G_1 \text{ x } E][G_2 \text{ x } E][G_5 \text{ x } E]$	1.6% (2.0%)	44,010 (43,974)

BIC represents Bayesian Information Criterion.

(b) The estimated posterior probabilities of appearance corresponding to **G** and **E** main effects and their interactions are shown under the identical setting as in Table 3(a). Results in parentheses represent the sensitivity analysis carried out with $v_0 = 0.001$

G_1	G_2	G_3	G_4	G_5	E	E x G_1	E x G_2	E x G_3	E x G_4	E x G_5
6.7	5.0	4.3	4.1	7.6	100.0	36.8	64.0	18.4	10.1	13.9
(5.6)	(6.8)	(10.6)	(7.3)	(9.4)	100.0	(29.5)	(55.4)	(19.9)	(9.7)	(9.5)

We first describe the data generation procedure. We consider two genes G_1 and G_2, and one environment factor E, with disease status D, all binary. We generate data from the following log-linear model [Li and Conti (2009)]:

$$\begin{aligned}\log(\mu|D, G_1, G_2, E) = {} & \gamma_0 + \gamma_{G_1} G_1 + \gamma_{G_2} G_2 + \gamma_E E + \gamma_D D \\ & + \lambda_{G_1E} G_1 E + \lambda_{G_2E} G_2 E + \lambda_{G_1G_2} G_1 G_2 \\ & + \beta_{G_1} G_1 D + \beta_{G_2} G_2 D + \beta_E E D \\ & + \beta_{G_1E} G_1 E D + \beta_{G_2E} G_2 E D + \beta_{G_1G_2} G_1 G_2 D,\end{aligned} \tag{4.1}$$

where μ denotes expected cell counts corresponding to the (D, G_1, G_2, E) configuration. Under this model, we are capable of fixing G_1-E, G_2-E and G_1-G_2 association under controls by setting values of λ_{G_1E}, λ_{G_2E} and $\lambda_{G_1G_2}$, respectively. These parameters are approximately equivalent to those in model $P(G_1, G_2|W)$ (2.2) when the disease is rare. Similarly, we can set β_{G_1E}, β_{G_2E} or $\beta_{G_1G_2}$, corresponding to the G x E or G x G interactions in the disease risk model. The

parameters $(\gamma_0, \gamma_{G_1}, \gamma_{G_2}, \gamma_E)$ control the marginal frequencies of G_1, G_2 and E in controls. A large negative value of γ_D ensures that the disease is rare.

For the model parameters in (4.1), we fixed $(\gamma_0, \gamma_{G_1}, \gamma_{G_2}, \gamma_E, \gamma_D) = (-6, -0.5, -0.5, -2.0, -4.5)$ that produces approximately 2.5% of the cases, frequency of $G_1 = 1$ and $G_2 = 1$ both at 45% while the prevalence of $E = 1$ is 15%. We assign $(\beta_{G_1E}, \beta_{G_2E}, \beta_{G_1G_2}) = (0, \log(2), \log(2))$ in (4.1). For setting parameters corresponding to G-E/G-G association, we set $(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$ to reflect G_1-G_2 and G_2-E dependence, and $(0, 0, 0)$ for the independence scenario.

Now we turn our attention to the sampling design. We randomly generate 1000 cases and 1000 controls with complete (D, G_1, G_2, E) data. We then carry out (D, E)-stratified sampling as follows. We select 600 cases and 600 controls in phase II. We consider two scenarios regarding this the stratified sampling strategy: (a) all subjects with a positive $E(=1)$, in cases and controls, are automatically included in phase II; (b) all subjects with a positive $E(=1)$ in cases are included in phase II, however, 600 controls for phase II are randomly selected regardless of E status. Finally, information on G_1 and G_2 from phase I subjects, that is, 400 cases and 400 controls, is treated as missing by design. We iterate this step to generate 1000 replicate data sets under each sampling scheme.

Tables 4 and 5 display the simulation results based on two different sampling schemes (a) and (b), respectively. We follow the convention that $\perp$ and $\sim$ represent independence and dependence between two variables, respectively. Under $G_1 \perp E$, $G_2 \perp E$ and $G_1 \perp G_2$ the CML method yields the smallest MSE with respect to G_1 x E and G_1 x G_2 interaction followed by TPFB, TPFB_{emp} and EB, while WL, PL and UML present relatively larger MSE. Here we need to note that the current implementation of CML and EB can only use G_1-E and G_1-G_2 independence, but not G_2-E independence. As phase II sampling becomes differential between cases and controls from scenario (a) (Table 4) to (b) (Table 5), we notice the substantial increase in the bias for estimating the main effect of E from CML, UML and EB as expected, while WL, PL, TPFB and TPFB_{emp} provide relatively less biased estimates. This trend remains present in the case where $G_1 \perp E$, $G_2 \sim E$ and $G_1 \sim G_2$. Beyond the bias in β_E from CML, UML and EB, we note that under the departure from the independence assumption, namely, $G_1 \perp G_2$, there is a dramatic increase in the bias corresponding to the G_1 x G_2 interaction under CML and to some extent in TPFB. TPFB_{emp} and EB are more robust to this assumption. Both TPFB show gain in efficiency for interaction estimation compared to PL and WL. Overall, our proposed methods, especially TPFB_{emp}, yield obvious gain in efficiency compared to PL and WL in terms of the G x E or G x G interactions in the presence of independence. On the other hand, TPFB_{emp} provides less biased estimates of the E effect compared to UML, CML and EB which use only phase II data. When the subsampling ratio is 80%, the pattern remains the same as seen in online supplemental Table 2 [Ahn et al. (2013)]. We also provide the sum of the MSEs across all parameters in order to capture the

TABLE 4

Simulation results under exposure enriched sampling with all $E = 1$ in phase I data selected in phase II for both cases and controls. We consider two association scenarios: (1) $G_1 \perp E$, $G_1 \perp G_2$ *and* $G_2 \perp E$ *association*, (2) $G_1 \perp E$, $G_1 \sim G_2$ *and* $G_2 \sim E$. *The results are based on* 1000 *replicated data sets*, *each with* 1000 *cases and* 1000 *controls in phase I and* 600 *cases and* 600 *controls in phase II. The approaches listed*, *TPFB*, *TPFB*$_{\text{emp}}$, *WL*, *PL*, *UML*, *CML and EB*, *each represent two-phase full Bayes* (*with empirically obtained prior variance*), *weighted likelihood*, *pseudo-likeliohod*, *unconstrained maximum likelihood*, *constrained maximum likelihood*, *and empirical-Bayes*, *respectively. The CML imposes* G_1-E *and* G_1-G_2 *independence*, *however*, *no constraints on* G_2-E *association. We set* $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ *for all scenarios. The rows with the smallest two sum* (*MSE*) *are in bold*

		$G_1 \perp E, G_1 \perp G_2, G_2 \perp E$					$G_1 \perp E, G_1 \sim G_2, G_2 \sim E$				
Stratified sampling (a)[†]		E	G_1 x G_2	G_1 x E	G_2 x E	**Sum (MSE)**[*]	E	G_1 x G_2	G_1 x E	G_2 x E	**Sum (MSE)**[*]
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$					$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$				
TPFB	Bias	−0.024	−0.017	0.020	−0.017		−0.056	0.166	−0.022	0.119	
	(MSE)	(0.093)	(0.044)	(0.120)	(0.135)	(0.392)	(0.117)	(0.081)	(0.091)	(0.122)	(0.411)
TPFB$_{\text{emp}}$	Bias	0.007	−0.019	−0.021	−0.062		−0.033	0.043	−0.029	0.026	
	(MSE)	(0.089)	(0.025)	(0.111)	(0.126)	(**0.351**)	(0.113)	(0.064)	(0.091)	(0.120)	(**0.388**)
WL	Bias	−0.038	−0.025	0.043	0.009		−0.038	0.011	0.011	0.006	
	(MSE)	(0.099)	(0.058)	(0.144)	(0.157)	(0.458)	(0.105)	(0.057)	(0.101)	(0.121)	(0.384)
PL	Bias	−0.038	−0.026	0.043	0.009		−0.038	0.011	0.011	0.006	
	(MSE)	(0.098)	(0.056)	(0.144)	(0.157)	(0.455)	(0.105)	(0.056)	(0.101)	(0.121)	(**0.383**)
UML	Bias	−0.093	−0.026	0.043	0.009		−0.096	0.011	0.011	0.006	
	(MSE)	(0.110)	(0.056)	(0.144)	(0.157)	(0.467)	(0.116)	(0.056)	(0.101)	(0.121)	(0.394)
CML	Bias	−0.085	−0.020	0.026	0.003		−0.100	0.700	0.011	0.009	
	(MSE)	(0.099)	(0.025)	(0.083)	(0.155)	(**0.362**)	(0.112)	(0.520)	(0.070)	(0.116)	(0.818)
EB	Bias	−0.087	−0.025	0.036	0.004		−0.099	0.089	0.010	0.008	
	(MSE)	(0.099)	(0.036)	(0.099)	(0.155)	(0.389)	(0.112)	(0.069)	(0.075)	(0.116)	(0.392)

[†]All subjects with $E = 1$ in case and control are subsampled for phase II.

[*]The combined MSEs as summed over all four parameters.

TPFB uses the informative prior $N(0, 10^{-2})$ on G-G and G-E associations in the model (2.2). TPFB$_{\text{emp}}$ uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model (2.2) where $\hat{\theta}^2$ is empirically estimated as the G-G or G-E association parameter under the controls.

TABLE 5

Simulation results under exposure enriched sampling with all $E = 1$ in phase I data selected in phase II for cases but a random sample of controls are selected in phase II. We consider two association scenarios: (1) $G_1 \perp E$, $G_1 \perp G_2$, *and* $G_2 \perp E$ *association*, (2) $G_1 \perp E$, $G_1 \sim G_2$, *and* $G_2 \sim E$. *The results are based on* 1000 *replicated data sets*, *each with* 1000 *cases and* 1000 *controls in phase I and* 600 *cases and* 600 *controls in phase II. The approaches listed*, *TPFB*, $TPFB_{\text{emp}}$, *WL*, *PL*, *UML*, *CML*, *and EB*, *each represent two-phase full Bayes* (*with empirically obtained prior variance*), *weighted likelihood*, *pseudo-likeliohod*, *unconstrained maximum likelihood*, *constrained maximum likelihood*, *and empirical-Bayes*, *respectively. The CML imposes* G_1-E *and* G_1-G_2 *independence*, *however, no constraints on the* G_2-E *association. We set* $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ *for all scenarios. The rows with the smallest two sum* (*MSE*) *are in bold*

		$G_1 \perp E, G_1 \perp G_2, G_2 \perp E$					$G_1 \perp E, G_1 \sim G_2, G_2 \sim E$				
Stratified sampling (b)†		E	G_1 x G_2	G_1 x E	G_2 x E	**Sum (MSE)***	E	G_1 x G_2	G_1 x E	G_2 x E	**Sum (MSE)***
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$					$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$				
TPFB	Bias	0.007	0.022	−0.007	−0.022		−0.105	0.160	0.032	0.128	
	(MSE)	(0.081)	(0.040)	(0.113)	(0.124)	(**0.358**)	(0.125)	(0.073)	(0.106)	(0.128)	(**0.432**)
$TPFB_{\text{emp}}$	Bias	0.039	0.015	−0.054	−0.073		−0.027	0.036	0.024	0.000	
	(MSE)	(0.086)	(0.031)	(0.127)	(0.122)	(**0.366**)	(0.121)	(0.058)	(0.115)	(0.135)	(**0.429**)
WL	Bias	−0.012	0.016	0.021	0.024		−0.044	0.004	0.076	−0.014	
	(MSE)	(0.098)	(0.059)	(0.165)	(0.167)	(0.489)	(0.133)	(0.056)	(0.150)	(0.147)	(0.486)
PL	Bias	−0.012	0.015	0.020	0.025		−0.046	0.002	0.077	−0.013	
	(MSE)	(0.097)	(0.059)	(0.164)	(0.166)	(0.486)	(0.132)	(0.055)	(0.149)	(0.146)	(0.482)
UML	Bias	0.538	0.015	0.020	0.025		0.520	0.002	0.077	−0.013	
	(MSE)	(0.395)	(0.059)	(0.164)	(0.166)	(0.784)	(0.407)	(0.055)	(0.149)	(0.146)	(0.757)
CML	Bias	0.544	0.017	−0.002	0.018		0.530	0.699	0.046	−0.013	
	(MSE)	(0.384)	(0.030)	(0.088)	(0.161)	(0.663)	(0.399)	(0.515)	(0.073)	(0.141)	(1.128)
EB	Bias	0.543	0.016	0.008	0.018		0.528	0.078	0.059	−0.013	
	(MSE)	(0.385)	(0.039)	(0.112)	(0.161)	(0.697)	(0.401)	(0.066)	(0.097)	(0.141)	(0.705)

†All cases with $E = 1$ are included in phase II, however, controls are randomly selected for phase II.

*The combined MSEs over all four parameters.

TPFB uses the informative prior $N(0, 10^{-2})$ on the G-G and G-E associations in the model (2.2). TPFB_{emp} uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model (2.2) where $\hat{\theta}^2$ is empirically estimated as the G-G or G-E association parameter under the controls.

accuracy of estimating subgroup effects defined by different G-E configurations. This summary measure in the last columns of Tables 4 and 5 clearly suggests that our methods yield more efficient characterization of the joint effect of exposure and genetic factors.

Table 3 in the supplemental article [Ahn et al. (2013)] presents simulation results under the traditional or unstratified case-control design when a random sample of cases and controls are taken irrespective of E status. We can note clear efficiency gains from stratified sampling when comparing Table 4 to Table 3 in the supplemental article for estimating the interaction parameters.

5. Discussion. We presented a flexible Bayesian approach to estimate gene–gene (G x G) and/or gene-environment (G x E) interactions under two-phase sampling with multiple markers. The proposed approach can handle multiple genetic and environmental factors. The method can trade off between bias and efficiency by incorporating uncertainty around gene-environment independence through the hierarchical structure in a data-adaptive way. The underlying ingredients of this hierarchy are the disease risk model, the multivariate gene model and the joint model for the environment factors/covariates, respectively. Our method can also handle potential missingness in genetic information due to technical inconsistency or due to merging different studies or cohorts, leading to nonmonotone missing data structure at the phase II subsample. This paper is the first Bayesian paper with retrospective modeling for G x E studies under two-phase sampling that can handle multiple markers.

We compared our method to simpler alternatives such as UML, CML and EB that use gene-environment independence but only based on phase II data, ignoring biased sampling. We also considered methods that account for biased sampling at phase II: weighted likelihood and pseudo-likelihood, but do not leverage the independence assumption. Our method provides a framework that integrates both of these features. In a clinical study like the MECC example, where interest lies in estimating the differential effect of statin use across genetic subgroups for devising targeted prevention strategies, estimates of main effects as well as gene-environment interaction are equally important, thus both estimates need to be assessed. In terms of aggregate MSE, our method has superior performance across a wide range of scenarios over the competing method.

There are some limitations of the current paper that need to be expanded and explored in future studies. First, we do not fully address the performance of our method in the presence of a truly high-dimensional gene model through simulation studies. The method is scalable to handle up to 294 SNPs and pairwise interactions in our data example, but we have not carried out a simulation study due to computation time. We also need to deal with the exponentially increasing number of G x E and G x G interactions in the disease risk model as well as G-E/G-G/G-S associations in the multivariate gene model, as we add more G-variables in the model. We address this by Bayesian variable selection and assuming a common parameter

for G-E/G-G/G-S association on genes in the same pathway in the multivariate gene model. The latter is a rather ad-hoc strategy for reducing the dimension and is a limitation of our method. Bias in parameter estimates is expected to arise under departures from this assumption. Calculation of $P(D)$ in the denominator of the likelihood could also pose challenges with truly high-dimensional data. Second, we have not tested the Bhattacharya and Dunson (2012) algorithm for the mixed set of discrete and continuous covariates in W. Future research will focus on the higher-dimensional G and E settings, more general structure of the W vector as well as the possibility of capturing higher order interactions, not just pairwise interactions.

For practitioners who want to choose a design strategy to enhance the power of screening G x E effects with a relatively rare exposure, exposure enrichment of cases and controls for collecting genotype data is a better strategy than random sampling. The tools we developed in the paper provides a way to account for the biased sampling. The approach also allows one to explore a multivariate model with multiple SNPs and environmental exposure and identify potentially informative predictors. If the interest lies in characterizing subgroup effects of E across different subgroups defined by $\mathbf{G}$, this design and analysis strategy is particularly powerful. We recommend the use of default prior choices in the codes available at http://www.umich.edu/~jaeil/tp.zip and recommend using TPFB_{emp} as the analysis to be reported. For prescribing a preventive medicine prophylactically, like use of statins for colorectal cancer, identifying genetic subgroups that will receive the most benefit from such a therapy is particularly helpful. Characterizing G x E effects furthers our understanding of such subgroup effects for tailoring targeted prevention strategies.

SUPPLEMENTARY MATERIAL

Bayesian semiparametric analysis for two-phase studies of gene-environment interaction (DOI: 10.1214/12-AOAS599SUPP; .pdf). We consider two-phase studies of G x E interaction where phase I data is available on exposure, covariates and disease status and stratified sampling is done to prioritize individuals for genotyping at phase II. We consider a Bayesian analysis based on the joint retrospective likelihood of phases I and II data that handles multiple genetic and environmental factors, data adaptive use of gene-environment independence.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

AHN, J., MUKHERJEE, B., GRUBER, S. B. and GHOSH, M. (2013). Supplement to "Bayesian semiparametric analysis for two-phase studies of gene-environment interaction." DOI:10.1214/12-AOAS599SUPP.

AMUNDADOTTIR, L., KRAFT, P., STOLZENBERG-SOLOMON, R. Z., FUCHS, C. S., PETERSEN, G. M., ARSLAN, A. A. et al. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* **41** 986–990.

BHATTACHARJEE, S., CHATTERJEE, N. and WHEELER, W. (2011). An R package for analysis of case-control studies in genetic epidemiology. Package CGEN, Version 1.0.0. Available at http://dceg.cancer.gov/bb/tools/genetanalcasecontdata.

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377.

BRESLOW, N. E. and CAIN, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75** 11–20. MR0932812

BRESLOW, N. E. and CHATTERJEE, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J. Appl. Stat.* **48** 457–468.

BRESLOW, N. E. and HOLUBKOV, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 447–461. MR1440590

BRESLOW, N. E. and HOLUBKOV, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Stat. Med.* **16** 103–116.

CHATTERJEE, N. and CARROLL, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92** 399–418. MR2201367

CHATTERJEE, N., CHEN, Y.-H. and BRESLOW, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.* **98** 158–168. MR1965682

CHATTERJEE, N. and CHEN, Y.-H. (2007). Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 123–142. MR2325268

COCHRAN, W. G. (1963). *Sampling Techniques*. Wiley, New York.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

DURT, T. (2010). Experimental proposal for testing the emergence of environment induced (EIN) classical selection rules with biological systems. *Studia Logica* **95** 259–277. MR2729664

FLANDERS, W. D. and GREENLAND, S. (1991). Analytic methods for 2-stage case-control studies and other stratified designs. *Stat. Med.* **10** 739–747.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–472.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

HACHEM, C., MORGAN, R., JOHNSON, M., MUEBELER, M. and EL-SERAG, H. (2009). Statins and the risk of colorectal carcinoma: A nested case-control study in veterans with diabetes. *Am. J. Gastroenterol.* **104** 1241–1248.

HANEUSE, S. and CHEN, J. (2011). A multiphase design strategy for dealing with participation bias. *Biometrics* **67** 309–318. MR2898843

HANEUSE, S. J.-P. A. and WAKEFIELD, J. C. (2007). Hierarchical models for combining ecological and case-control data. *Biometrics* **63** 128–136, 312. MR2345582

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M. et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39** 870–874.

ISHWARAN, H. and RAO, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.* **98** 438–455. MR1995720

LAWLESS, J. F., KALBFLEISCH, J. D. and WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 413–438. MR1680310

LEE, A. J., SCOTT, A. J. and WILD, C. J. (2010). Efficient estimation in multi-phase case-control studies. *Biometrika* **97** 361–374. MR2650744

LI, D. and CONTI, D. V. (2009). Interactions using a combined case-only and case-control approach. *Am. J. Epidemiol.* **169** 497–504.

LIPKIN, S. M. et al. (2010). Genetic variation in 3-hydroxy-3-methylglutaryl CoA reductase modifies the chemopreventive activity of statins for colorectal cancer. *Cancer Prev. Res.* (*Phila*) **3** 597–603.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LUMLEY, T. (2011). R for analyzing data from complex surveys. Package Survey, Version 3.2.4. Available at http://cran.r-project.org/web/packages/survey/index.html.

MANSKI, C. F. and LERMAN, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45** 1977–1988. MR0501708

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

MUKHERJEE, B. and CHATTERJEE, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64** 685–694. MR2526617

MUKHERJEE, B., AHN, J., STEPHEN, B. G., RENNERT, G., VICTOR, M. and CHATTERJEE, N. (2008). Testing gene-environment interaction from case-control data: A novel study of type-1 error, power and designs. *Gen. Epid.* **32** 615–626.

MUKHERJEE, B., AHN, J., GRUBER, S. B., GHOSH, M. and CHATTERJEE, N. (2010). Bayesian sample size determination for case-control studies of gene-environment interaction. *Biometrics* **66** 934–948. MR2758230

MÜLLER, P., PARMIGIANI, G., SHILDKRAUT, J. and TARDELLA, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55** 858–866.

MURCRAY, C. E., LEWINGER, J. P. and GAUDERMAN, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169** 219–226.

NEYMAN, J. (1938). Contribution to the theory of sampling from human populations. *J. Amer. Statist. Assoc.* **33** 101–116.

PARK, J. H., WACHOLDER, S., GAIL, M. H., PETERS, U., JACOBS, K. B., CHANOCK, S. J. and CHATTERJEE, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42** 570–575.

PFEIFFER, R. M. and GAIL, M. H. (2003). Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet. Epidemiol.* **25** 136–148.

PIEGORSCH, W. W., WEINBERG, C. R. and TAYLOR, J. (1994). Non hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13** 153–162.

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2009). Output analysis and diagnostics for MCMC. Package CODA, Version 0.13-4. Available at http://cran.r-project.org/web/packages/coda/.

POYNTER, J. N., GRUBER, S. B., HIGGINS, P. D. R., ALMOG, R., BONNER, J. D., RENNERT, H. S., LOW, M., GREENSON, J. K. and RENNERT, G. (2005). Statins and the risk of colorectal cancer. *N. Engl. J. Med.* **352** 2184–2192.

REILLY, M. and PEPE, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82** 299–314. MR1354230

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

SCHILL, W., JÖCKEL, K. H., DRESCHER, K. and TIMM, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* **80** 339–352. MR1243509

SCOTT, A. J. and WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71. MR1450191

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

UMBACH, D. M. and WEINBERG, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat. Med.* **11** 259–272.

VANSTEELANDT, S., VANDERWEELE, T. J. and ROBINS, J. M. (2008). Multiply robust inference for statistical interactions. *J. Amer. Statist. Assoc.* **103** 1693–1704. MR2510295

WACHOLDER, S., HARTGE, P., PRENTICE, R., GARCIA-CLOSAS, M. et al. (2010). Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362** 986–993.

WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36** 45–54. MR2370888

WHITTEMORE, A. S. and HALPERN, J. (1998). Multi-stage sampling in genetic epidemiology. *Stat. Med.* **16** 153–167.

YEAGER, M., ORR, N., HAYES, R. B., JACOBS, K. B., KRAFT, P., WACHOLDER, S. et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39** 645–649.

J. AHN
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF MICHIGAN
M4153 SPH II
1415 WASHINGTON HEIGHTS
ANN ARBOR, MICHIGAN 48109-2029
USA
E-MAIL: jaeil@umich.edu

B. MUKHERJEE
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF MICHIGAN
M4166 SPH II
1415 WASHINGTON HEIGHTS
ANN ARBOR, MICHIGAN 48109-2029
USA
E-MAIL: bhramar@umich.edu

S. B. GRUBER
USC NORRIS COMPREHENSIVE CANCER CENTER
UNIVERSITY OF SOUTHERN CALIFORNIA
1441 EASTLAKE AVE, NOR 8302L
LOS ANGELES, CALIFORNIA 90089-9181
USA
E-MAIL: sgruber@usc.edu

M. GHOSH
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
P.O. BOX 118545
GAINESVILLE, FLORIDA 32611-8545
USA
E-MAIL: ghoshm@stat.ufl.edu