Published in final edited form as:

Biometrics. 2017 March ; 73(1): 344–355. doi:10.1111/biom.12567.

Detecting Rare and Common Haplotype–Environment Interaction under Uncertainty of Gene–Environment Independence Assumption

Yuan Zhang¹, Shili Lin², and Swati Biswas^{1,*}

¹Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75080, U.S.A.

²Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A.

Summary

Finding rare variants and gene-environment interactions (GXE) are critical in dissecting complex diseases. We consider the problem of detecting GXE where G is a rare haplotype and E is a non-genetic factor. Such methods typically assume G-E independence, which may not hold in many applications. A pertinent example is lung cancer — there is evidence that variants on Chromosome 15q25.1 interact with smoking to affect the risk. However, these variants are associated with smoking behavior rendering the assumption of G-E independence inappropriate. With the motivation of detecting GXE under G-E dependence, we extend an existing approach, logistic Bayesian LASSO, which assumes G-E independence (LBL-GXE-I) by modeling G-E dependence through a multinomial logistic regression (referred to as LBL-GXE-D). Unlike LBL-GXE-I, LBL-GXE-D controls type I error rates in all situations; however, it has reduced power when G-E independence holds. To control type I error without sacrificing power, we further propose a unified approach, LBL-GXE, to incorporate uncertainty in the G-E independence assumption by employing a reversible jump Markov chain Monte Carlo method. Our simulations show that LBL-GXE has power similar to that of LBL-GXE-I when G-E independence holds, yet has well-controlled type I errors in all situations. To illustrate the utility of LBL-GXE, we analyzed a lung cancer dataset and found several significant interactions in the 15q25.1 region, including one between a specific rare haplotype and smoking.

Keywords

GXE; G-E dependence; LBL; Missing Heritability; Rare variants; Reversible jump MCMC

1. Introduction

In the post-Genome-wide association studies (GWAS) era, the search for rare variants has attracted a major research focus in the quest for missing heritability. Unlike the so-called

* swati.biswas@utdallas.edu.

6. SUPPLEMENTARY MATERIALS

Web Appendix, Tables, and Figures referenced in Sections 2.1.3, 3, 4, and 5 are available at the *Biometrics* website on Wiley Online Library. An R package LBL implementing the method is also available here.

“collapsing” approaches for rare variants, haplotype-based methods can make use of common single nucleotide polymorphisms (SNP) from GWAS to detect associations of rare variants with common diseases. This is because rare haplotype variants (rHTVs) may result from the combination of common SNPs. Haplotype analyses are a natural next step subsequent to a single-SNP analysis. They are typically more powerful than their single-SNP counterparts when multiple causal variants act interactively on a disease or when causal variants are not genotyped (Clark, 2004). More recently, it has also been argued that rHTVs are more likely to tag rare Single Nucleotide Variants (rSNV), and haplotype methods can have greater power than collapsing methods in detecting these associations in the above-mentioned situations (Li, Byrnes, and Li, 2010; Lin et al., 2013; Wang and Lin, 2015; Datta et al., 2016; Datta and Biswas, 2015). Thus, it behooves us to dig into the amassed GWAS data for uncovering associations with rHTVs, especially because these data are “freely” available. To fully exploit the richness of the GWAS data, it is equally important to use them to investigate gene-environment interaction (GXE), another contributor to missing heritability. Indeed, there is a pressing need for methods to detect GXE where G is rHTV but this is a challenging task as one has to not only deal with infrequent events but also high dimensionality.

A recently proposed rHTV approach, logistic Bayesian LASSO (LBL), can detect GXE in case-control data by penalizing the regression coefficients through appropriate priors (Biswas and Lin, 2012; Biswas, Xia and Lin, 2014; Xia, 2014; Zhang and Biswas, 2015). As in most other GXE methods, a key assumption of the method is G-E independence in the control population. We refer to this method as LBL-GXE-I hereafter. Although LBL-GXE-I performs well under G-E independence, we show that it leads to inflated type I error rates when the assumption is violated. That is, inferences made with this method can be highly sensitive to the usually unverified assumption of G-E independence. In fact, often this assumption does not hold or there is an uncertainty about its validity (Mukherjee and Chatterjee, 2008). There are methods in the literature to handle these situations. For example, Mukherjee and Chatterjee (2008) combined a case-only and a case-control estimators but considering genetic factor that is directly observable. On the other hand, Chen, Chatterjee, and Carroll (2008) characterized haplotype–environment association using a semiparametric estimating equation methodology although their method is only applicable to common haplotypes.

An apt example of G-E dependence is lung cancer, the motivating example for this work. Tobacco smoking is the leading risk factor for lung cancer. Nonetheless, genetics plays an important role in interacting with smoking to affect the risk (Yokota, Shiraishi, and Kohno, 2010). Several variants on Chromosome 15q25.1 have been identified to affect lung cancer susceptibility (Thorgeirsson et al., 2008, Yokota et al., 2010; Yu et al., 2012; VanderWeele et al., 2012). At the same time, it has also been demonstrated in numerous studies that variants on Chromosome 15q25.1 are associated with smoking behavior (Spitz et al., 2008; Thorgeirsson et al., 2008; Liu et al., 2010). VanderWeele et al. (2012) studied two SNPs rs8034191 and rs1051730 in this region, which have been implicated in many single-SNP studies. They found that these variants interact with smoking to increase the risk. They noted that these variants may increase the amount of nicotine and toxins extracted from cigarettes and thus, such an effect would only be observed for smokers. In fact, there is a cluster of

nicotinic acetylcholine receptor genes in the 15q25.1 region. Thus, Yu et al. (2012) advocated a multi-SNP approach to study interactions. They studied the interaction of 15 tagging SNPs in this region with smoking intensity (number of packs per day) for the risk of lung cancer using the data from the Environment And Genetics in Lung cancer Etiology (EAGLE) study. They found that smokers with a specific genotype profile (obtained as principal components) had a higher lung cancer risk. From these studies, we can deduce two points of relevance to us — (1) it is unlikely that G-E independence assumption holds in this region and (2) due to the presence of multiple causal SNPs, a follow-up haplotype analysis of this region is warranted. Thus, there is a need for a haplotype association method that can take into account of uncertainty in the G-E independence assumption.

To address this gap, we propose two LBL-based methods: (1) LBL-GXE-D that models G-E dependence, and (2) LBL-GXE that unifies LBL-GXE-I and LBL-GXE-D to allow for uncertainty in the G-E independence assumption. We analyze a lung cancer dataset obtained from two studies — EAGLE and Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial and find a significant interaction between smoking and an rHTV on Chromosome 15q25.1 in the presence of G-E dependence. We compare LBL with two standard haplotype association methods through simulations. We also carry out extensive simulation studies to investigate and compare the performances of LBL-GXE-I, LBL-GXE-D, and LBL-GXE.

2. Methods

We first describe the model for LBL-GXE-D in its entirety for ease of exposition. Most of this methodology was developed in Biswas and Lin (2012), Biswas et al. (2014), and Zhang and Biswas (2014) except the part on modeling G-E dependence. Next we describe LBL-GXE-I and LBL-GXE. Finally, we describe how to test for association, which is common to all three approaches.

2.1 LBL-GXE-D

2.1.1 Retrospective likelihood—Consider a case-control sample consisting of n_1 cases and n_2 controls with $n_1 + n_2 = n$. Let $Y_i = 1/0$ denote the case/control status of the i th individual, $i = 1, \dots, n$, and $\mathbf{Y} = (Y_1, \dots, Y_n)$. Let G_i denote the observed genotype of the i th individual, and $\mathbf{G} = (G_1, \dots, G_n)$. As haplotype pair of a person may not be deduced unambiguously from the observed genotype data, we further let $\mathcal{S}(G_i)$ denote the set of haplotype pairs compatible with G_i and let Z_{ir} denote the r th element of $\mathcal{S}(G_i)$. Next we denote the vector of environmental covariates of the i th individual by $\mathbf{E}_i$ and $\hat{\mathbf{E}} = (\mathbf{E}_1, \dots, \mathbf{E}_n)$. Then the retrospective likelihood of the observed data is written as:

$$\begin{aligned}
L(\Psi^{(D)}) &\equiv L(\Psi^{(D)} | Y, \tilde{E}, G) \\
&= \prod_{i=1}^{n_1} \sum_{Z_{ir} \in S(G_i)} P(Z_{ir} | E_i, Y_i=1, \Psi^{(D)}) P(E_i | Y_i=1, \Psi^{(D)}) \\
&\times \prod_{i=n_1+1}^n \sum_{Z_{ir} \in S(G_i)} P(Z_{ir} | E_i, Y_i=0, \Psi^{(D)}) P(E_i | Y_i=0, \Psi^{(D)}), \quad (1)
\end{aligned}$$

where $\Psi^{(D)}$ consists of regression coefficients and parameters associated with haplotype pair frequencies, which we will define explicitly later in this sub-section.

In the following, we will individually model each term of the likelihood in (1). For simplicity, we suppress the subscripts i and r and the parameter vector $\Psi^{(D)}$ without loss of clarity. We start with modeling of $P(Z|E, Y=0) = a_{ZE}$, frequency of haplotype pair Z in the control population for a given E value. Suppose there are a total of m haplotypes in a haplotype block with frequencies in controls denoted by $f(E) = (f_1(E), \dots, f_m(E))$. Here we explicitly write f as a function of E to be able to model G-E dependence later.

First, we model a_{ZE} for a haplotype pair $Z = (z_k, z_{k'})$ as

$$a_{Z|E} = P(Z = (z_k, z_{k'}) | Y=0, f(E), d) = \delta_{kk'} f_k(E) d + (2 - \delta_{kk'}) f_k(E) f_{k'}(E) (1 - d),$$

where $\delta_{kk'} = 1(0)$ if $z_k = z_{k'} (z_k \neq z_{k'})$, f_k and $f_{k'}$ are frequencies of z_k and $z_{k'}$, and $d \in (-1, 1)$ is the within-population inbreeding coefficient which captures excess/reduction of homozygosity (Weir, 1996). For $d = 0$, the above expression is equivalent to assuming Hardy-Weinberg Equilibrium (HWE) while other values of d allow Hardy-Weinberg disequilibrium (HWD).

Secondly, we model $f(E)$ using a multinomial logistic regression model to allow G-E dependence (Mukherjee and Chatterjee, 2008). Let the m th haplotype be the baseline and consider C environmental covariate levels excluding baseline(s), $E = \{E_1, E_2, \dots, E_C\}$ corresponding to one or more categorical covariates. Thus,

$$f_k(E) = \frac{\exp(g_k(E))}{1 + \sum_{j=1}^{m-1} \exp(g_j(E))}, \quad k=1, \dots, m-1; \quad f_m(E) = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(g_j(E))}, \quad (2)$$

where $g_k(E) = \gamma_{k0} + \gamma_{k1}E_1 + \gamma_{k2}E_2 + \dots + \gamma_{kC}E_C$. Let γ be an $(m-1) \times (C+1)$ matrix whose (k, c) th element is γ_{kc} , $k=1, \dots, m-1$, $c=0, \dots, C$. Thus, $a_{ZE}(\gamma, d)$ is now fully specified.

Next let us consider $P(Z|E, Y=1) = b_{ZE}$. We express b_{ZE} in terms of a_{ZE} and the odds of disease for a given Z and E , $\theta_{ZE} = P(Y=1|Z, E)/P(Y=0|Z, E)$ as

$$b_{Z|E} = \theta_{Z,E} a_{Z|E} / \sum_{H \in \mathbf{H}} \theta_{H,E} a_{H|E}, \text{ where } \mathbf{H} \text{ is the set of all possible haplotype pairs. We model}$$

θ_{ZE} using a logistic regression model: $\theta_{ZE} = \exp(\mathbf{X}_{ZE}\boldsymbol{\beta})$, where $\mathbf{X}_{ZE} = (1, \mathbf{X}_Z, \mathbf{X}_E, \mathbf{X}_Z\mathbf{X}_E)$ and $\boldsymbol{\beta}$ is a vector of the corresponding regression coefficients. Specifically, $\mathbf{X}_Z = (x_1, x_2, \dots, x_{m-1})$, where x_k is the number of copies of haplotype z_k in haplotype pair Z with the m th haplotype assumed to be the baseline. $\mathbf{X}_E$ consist of usual dummy variables and $\mathbf{X}_Z\mathbf{X}_E$ is obtained by (scalar) multiplication of $\mathbf{X}_Z$ and $\mathbf{X}_E$. Therefore, there are in total $1+(m-1)+C+(m-1)C = mC + m$ regression coefficients including the intercept.

It remains to model $P(\mathbf{E}_j | Y_i = 0)$ and $P(\mathbf{E}_j | Y_i = 1)$ in (1). Assuming a saturated model for $P(\mathbf{E})$, $P(\mathbf{E} | Y) \propto P(Y | \mathbf{E})$ without loss of information (Prentice and Pyke, 1979; Kwee et al, 2007). Then using the Bayes rule, we get the following:

$$P(\mathbf{E} | Y=1) \propto \frac{\sum_H \theta_{H,E} a_{H|E}}{1 + \sum_H \theta_{H,E} a_{H|E}} \quad \text{and} \quad P(\mathbf{E} | Y=0) \propto \frac{1}{1 + \sum_H \theta_{H,E} a_{H|E}}.$$

We have now completely specified the likelihood $L(\boldsymbol{\Psi}^{(D)})$ in (1) where $\boldsymbol{\Psi}^{(D)} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, d)$.

2.1.2 Priors—We regularize the regression coefficients $\boldsymbol{\beta}$ using Bayesian LASSO by assigning each of them a double exponential prior with hyper-parameter λ :

$$\pi(\beta_j | \lambda) = \frac{\lambda}{2} \exp(-\lambda |\beta_j|), \quad -\infty < \beta_j < \infty, \quad j=0, \dots, (mC+m-1).$$

The above priors are centered at 0 to serve as a penalty for $\boldsymbol{\beta}$ s yet have heavy tails to allow for associated effects. This helps in weeding out the unassociated effects, making it possible for the associated ones, especially those involving the rare haplotypes, to stand out. The variance is $2/\lambda^2$, where λ controls the degree of penalty. We then let λ follow a Gamma(a, b) distribution whose mean is a/b . We set $a = b = 20$ following Biswas and Lin (2012) and Biswas et al. (2014). The marginal variance of $\boldsymbol{\beta}$ when $a = b$ is $2a^2/((a-1)(a-2))$, $a > 2$, which is a decreasing function of a . Following Biswas and Lin (2012), we use these values of a and b to cover a wide range of realistic scenarios as the true scenario is not known beforehand. However, one can adjust the values of these hyper-parameters to achieve more or less shrinkage, e.g., if one is interested in detecting an extremely rare haplotype with a very large odds ratio (OR) but possibly with a trade-off of reduced power for detecting common haplotypes.

For γ_{kc} , $k = 1, \dots, m-1$; $c = 0, \dots, C$, we use a double exponential prior with hyper-parameter ν . We set ν to be 0.5 in our simulation study and application. For d , we note that it is dependent on $f(\mathbf{E})$ as a_{ZE} should be nonnegative. Thus, $d > \{f_k(\mathbf{E})/(1 - f_k(\mathbf{E}))\}$ for all k . As $-1 < d < 1$, we get $\max_k \{-f_k(\mathbf{E})/(1 - f_k(\mathbf{E}))\} < d < 1$, $k = 1, \dots, m-1$. So we set the prior for d to be Uniform($\max_k \{-f_k(\mathbf{E})/(1 - f_k(\mathbf{E}))\}, 1$), where $f_k(\mathbf{E})$ is given by (2).

2.1.3 Posterior Distributions—These are estimated using Markov chain Monte Carlo (MCMC) methods. The detailed steps are provided in the Supplementary Materials.

2.2 LBL-GXE-I

LBL-GXE-I (Biswas et al., 2014) does not model haplotype frequencies $\mathbf{f} = (f_1, \dots, f_m)$ to be a function of $\mathbf{E}$. That is, $a_{ZE} = a_Z$ (independent of $\mathbf{E}$) and consequently there are no $\boldsymbol{\gamma}$ parameters in equation (2). The prior for $\mathbf{f}$ is Dirichlet $(1, \dots, 1)$, consisting of a total of m 1's for the m haplotypes. The rest of the modeling is the same as in LBL-GXE-D. We let $\boldsymbol{\Psi}^{(I)} = (\boldsymbol{\beta}, \mathbf{f}, d)$ and its retrospective likelihood as $L(\boldsymbol{\Psi}^{(I)})$.

2.3 LBL-GXE

To incorporate uncertainty about G-E independence, we introduce an additional parameter, M such that $M = I$ corresponds to the model under G-E independence and $M = D$ corresponds to the model incorporating G-E dependence. Denote parameters in Model I to be $\boldsymbol{\Theta}^{(I)} = (\boldsymbol{\beta}, \lambda, \mathbf{f}, d)$ and parameters in Model D to be $\boldsymbol{\Theta}^{(D)} = (\boldsymbol{\beta}, \lambda, \boldsymbol{\gamma}, d)$. As $\boldsymbol{\Theta}^{(I)}$ and $\boldsymbol{\Theta}^{(D)}$ are of different dimensions, a sampler that enables moves between models with different number of parameters is needed. Here we use the reversible-jump MCMC (RJMCMC) method (Green, 1995) to move between the model spaces of I and D , thereby proposing a unified method that incorporates the uncertainty in the G-E independence assumption.

At every iteration, we first determine whether a move from the current model to the other model will be proposed. Suppose the probability of staying in I at $(t+1)$ th iteration given that the current model at (t) th iteration is I , is $P(M^{(t+1)} = I | M^{(t)} = I) = p_1$. Similarly, let $P(M^{(t+1)} = D | M^{(t)} = D) = p_2$. Hence $P(M^{(t+1)} = D | M^{(t)} = I) = 1 - p_1$ and $P(M^{(t+1)} = I | M^{(t)} = D) = 1 - p_2$. If it is proposed to stay in the current model, we update the corresponding model parameters using the MCMC procedures for LBL-GXE-D or LBL-GXE-I depending on whether the current model is D or I because this type of move does not change the dimensionality of the parameter space. On the other hand, if a move to the other model is proposed, we have to ensure dimension matching, as we describe in details in the following.

As the parameters $\boldsymbol{\beta}$, λ , and d are common between the two models, their current values are retained in the moves across the models. Thus, dimension matching is only needed between the frequency parameters of the two models, namely, $\mathbf{f}$ and $\boldsymbol{\gamma}$. For the rest of this sub-section, we assume $C = 1$ to keep the exposition simple. That is, we consider only one environmental covariate with two levels denoted by $E = 0$ (baseline) and $E = 1$. Then there are two $\boldsymbol{\gamma}$ parameters in equation (2) — intercept γ_{k0} and slope γ_{k1} for each k ($k = 1, \dots, m-1$). If a move from I to D is proposed, we create γ_{k0} by setting $\gamma_{k0} = \log(f_k/f_m)$ while γ_{k1} is generated afresh by setting $\gamma_{k1} = u_k$, where $u_k \sim \text{Double exponential}(0, \zeta)$, for all k . Therefore, according to equation (2), the proposed $f_k(E=0)$ for model D is set equal to the current value of f_k coming from model I .

For generating γ_{k1} , we choose ζ to be 50 so that $\text{Var}(\gamma_{k1}) = \text{Var}(u_k) = 2/\zeta^2$ is small, resulting in γ_{k1} being close to 0. Consequently, as per equation (2), the proposed $f_k(E=1)$ for model D is set close to f_k coming from model I . Thus, when we move from model I to D (or vice versa), there is not much loss in information in the current values of the frequency parameters. The acceptance probability of the move from I to D is given by $\min\{1, A(I, D)\}$ where

$$A(I, D) = \frac{P(\mathbf{G}, \mathbf{E} | \mathbf{Y}, \Theta^{(D)}, M=D) P(\Theta^{(D)} | M=D) P(M=D) P(M=I | M=D)}{P(\mathbf{G}, \mathbf{E} | \mathbf{Y}, \Theta^{(I)}, M=I) P(\Theta^{(I)} | M=I) P(M=I) \pi(\mathbf{u}) P(M=D | M=I)} \left| \frac{d(\Theta^{(D)})}{d(\Theta^{(I)}, \mathbf{u})} \right|$$

$$= \frac{L(\Psi^{(D)}) \left(\prod_{k=1}^{m-1} \prod_{c=0}^1 \pi(\gamma_{kc}) \right) \pi(d | \gamma, \mathbf{E}) P(M=D) (1-p_2)}{L(\Psi^{(I)}) \pi(\mathbf{f}) \pi(d | \mathbf{f}) \pi(\mathbf{u}) P(M=I) (1-p_1)} \frac{1}{\prod_{i=1}^m f_i},$$

where $\mathbf{u} = (u_1, \dots, u_{m-1})$. $A(I, D)$ is a product of three ratios — likelihood, prior, and proposal distributions under model D to model I , multiplied by the Jacobian of the transformation.

If a move from D to I is proposed, we need to create $\mathbf{f}$ using the current values of γ_{k1} and γ_{k0} . To ensure reversible moves between D and I , we discard γ_{k1} and use γ_{k0} to get $\mathbf{f}$:

$$f_k = \frac{\exp(\gamma_{k0})}{1 + \sum_{j=1}^{m-1} \exp(\gamma_{j0})}, \quad k=1, \dots, m-1; \quad f_m = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(\gamma_{j0})}.$$

Acceptance probability of the move from D to I is $\min\{1, A(D, I)\}$, where $A(D, I) = A^{-1}(I, D)$.

In our simulation study and application, we set $P(M=I) = P(M=D) = 0.5$, $p_1 = 0.3$, and $p_2 = 0.7$. Such a choice helps in controlling type I error rate as this rate is high if model I is chosen too often in the presence of G-E dependence.

We monitor convergence of the chain using trace plots and $\hat{R}$ diagnostic statistic (Gelman et al., 2003). As the three methods vary in complexity, the total number of MCMC iterations required for convergence also vary. LBL-GXE-I, being the simplest, has the fastest convergence, and following Biswas and Lin (2012) and Biswas et al. (2014) we use a total number of 50,000 iterations with a burn-in period of 20,000 iterations. As LBL-GXE-D involves more parameters, a larger number of iterations is typically required while LBL-GXE, being the most complicated, needs to be run even longer to achieve satisfactory convergence. In our simulations, we run LBL-GXE-D and LBL-GXE for a total number of 70,000 and 100,000 iterations, respectively, with 20,000 burn-in. In the real data analysis, there are several haplotypes that are rarer than those in the simulated data and so to ensure convergence and obtain more precise results, we run LBL-GXE for 200,000 iterations from four different starting points (starting in models I as well as D), discard initial 100,000 as burn-in, and combine the four chains to obtain the posterior distributions (Gelman et al., 2003).

2.4 Association Testing

We test for the significance of each β coefficient by calculating its 95% credible set (CS) from its posterior distribution. If it excludes 0 (or equivalently the 95% CS of OR excludes 1), we conclude that the corresponding effect is significant. Alternatively, Bayes factor (BF) could be used for inference, e.g., $\text{BF} > 2$ was used to declare significance in Biswas and Lin

(2012) and Biswas et al. (2014). However, in our simulation studies, we found that when the BF method is used, LBL-GXE-D and LBL-GXE gives somewhat inflated type I error rates in a few G-E dependence situations. As such, we choose to use the CS method for calculating powers and type I error rates in our simulations, however, we report BF along with CS for the lung cancer data analysis.

3. Analysis of the Lung Cancer Data

We analyze lung cancer GWAS data collected in EAGLE and PLCO studies, having 2728 cases and 2821 controls in total. Following our motivation described in the Introduction section, we examine 15q25.1 region by combining the fifteen SNPs studied in Yu et al. (2012) and the two SNPs from VanderWeele et al. (2012). These seventeen SNPs (in chromosomal order) are rs1394371, rs12903150, rs12899131, rs2656069, rs13180, rs3743079, rs8034191, rs3885951, rs2036534, rs2292117, rs680244, rs578776, rs12914385, rs1051730, rs1948, rs11636753, and rs12441998. Of these, SNP number 7 (rs8034191) and 14 (rs1051730) correspond to VanderWeele et al. study while the rest are from Yu et al. First we use Haploview (Barrett et al., 2005) to find the haplotype blocks in this region (shown in Web Figure 1). The following five blocks are detected (in order): (1) rs1394371 to rs2656069 with 4 SNPs, (2) rs13180 to rs3743079 with 2 SNPs, (3) rs8034191 to rs2292117 with 4 SNPs, (4) rs12914385 to rs1948 with 3 SNPs, and (5) rs11636753 to rs12441998 with 2 SNPs. Note that rs680244 and rs578776 are not part of any haplotype block.

We consider (ever) smoking with two levels (Yes/No) as a covariate and apply LBL-GXE to these five haplotype blocks separately to study gene-environment interaction. As expected, smoking is a highly significant risk factor in all five analyses. Three of the five haplotype blocks reveal interaction of haplotypes with smoking; these are reported in Web Table 1. In particular, the third haplotype block shows an rHTV-smoking interaction. We report detailed results for this haplotype block in Table 1. There are a total of 10 haplotypes in this haplotype block. TTTA is used as the baseline as it is most common and has similar frequencies amongst cases and controls. Note that the effects involving rHTVs have wider CS as expected. The haplotype frequencies reported in the table were obtained using the software hapassoc (Burkett, Graham, and McNeney, 2006). These estimates were used as starting points for the frequency parameters in the MCMC procedures. The interaction effects of CTTG and TTTG with smoking are found to be significant (95% CS for OR does not cover 1); smokers with these haplotypes have increased risk of lung cancer. Specifically, the rHTV TTTG X smoking has an OR estimated to be 2.18, which combined with the main effect of TTTG, indicates that the carriers of TTTG have about two times higher odds of lung cancer than the carriers of baseline TTTA among smokers. Remarkably, the chains stay in D model in all iterations (after burn-in) for all starting points suggesting a strong G-E dependence. Thus, the results with LBL-GXE-D are the same as those obtained with LBL-GXE.

Interestingly, the two haplotypes CTTG and TTTG in the third block with significant interactions share TTG at the last three SNPs in this block. To investigate this further, we also analyzed the last three SNPs only in this third block as a haplotype block and found a significant interaction of haplotype –TTG with smoking, increasing the odds of lung cancer

("—" indicates the SNP at that position was not used in the analysis). Similarly, an analysis of a block with just the last two SNPs reveals —TG X smoking to be significant with $OR > 1$. On the other hand, in a haplotype block formed by the first 3 SNPs only, we found interactions of smoking with CTT— to be significant as well with $OR > 1$. Further, in a single-SNP analysis, the interaction effect of smoking with C allele of the first SNP is significant. Taken together, these results corroborate the earlier indication of multiple causal SNPs acting in cis in this block to increase the odds of lung cancer for smokers.

4. Simulation Study

4.1 Simulations based on the Lung Cancer Data

We consider three simulation settings based on the above lung cancer data analysis. In the first setting, we use the estimated ORs (for associated effects) and haplotype frequencies (in smoking and non-smoking groups). The second setting has somewhat different ORs while the third one has an interaction effect of a haplotype that is rarer than what we found in the above analysis. These simulations serve a two-fold goal — to examine LBL-GXE under realistic linkage disequilibrium patterns and to compare it with two standard haplotype association approaches, hapassoc (Burkett et al., 2006) and haplo.glm (Lake et al., 2003). Both are GLM based methods and use Expectation Maximization (EM) algorithm. EM algorithm may not converge in the presence of rHTVs, and thus, by default, these approaches pool rHTVs into a single group. Hapassoc allows pooling tolerance to be set to 0 (i.e., no haplotypes are grouped together) while haplo.glm does not allow pooling tolerance to be set below 0.001. We used their respective minimum pooling tolerance as higher values will limit the possibility of detecting rHTVs and their interactions.

Similar to the real data, the sample size is 5000 with equal number of cases and controls. The procedure for generating a sample is the same in this and the next sub-section, and is as follows. For each individual, first we simulate a binary covariate value, say E using a specific value of prevalence p_E (0.83 in this case from Table 1). Then we generate a phased haplotype pair, say Z , using frequencies for the two E categories (from Table 2) and assuming HWE. Next, the individual is assigned to be a case or control using a logistic regression model: $\log(p/(1-p)) = X_{Z,E}\beta$, where p is the probability that the individual is case, and $X_{Z,E}$ is the design vector coded according to the generated haplotype pair, covariate, and their interactions. The intercept is calculated using a baseline prevalence of 0.1, i.e., $\beta_0 = \log(0.1/0.9)$. For other β coefficients, we use the corresponding ORs (from Table 2). We set the most frequent haplotype as the baseline in the regression model. After the case/control status is assigned, the phase information is removed and only genotypes are retained. The process is repeated until the required numbers of cases and controls in a sample are generated. This procedure is replicated 2000 times. For each coefficient, we report the percentage of times that it is significant (95% CS excludes 0) to study the power or type I error rate.

Hapassoc did not converge for 98.15% of the sampled replicates due to rHTVs; their results are not meaningful, and thus are not reported. The non-convergence rate for haplo.glm is 2.92%. In Table 2, we report the results for haplo.glm and LBL-GXE. Note that the causal haplotypes are uncommon (U), common (C), and rare (R). We see that haplo.glm has

inflated type I error rates for several main and interaction effects, especially those involving rHTVs while these rates are well-controlled for LBL-GXE. These results hold even if data are simulated under G-E independence (see Web Table 2), consistent with Datta and Biswas (2015) and Datta et al. (2016).

4.2 General Simulations

4.2.1 Settings and Data Generation—Next we carry out extensive simulation studies with a slightly different goal. We focus on comparing the performance of LBL-GXE with that of LBL-GXE-I and LBL-GXE-D in a wide range of scenarios involving rare (and common) haplotype-environment interactions as well as the main effects. As `haplo.glm` and `hapassoc` did not perform satisfactorily in the above simulations, we do not consider them in these simulations. We consider three haplotype settings with 6, 9, and 12 haplotypes in a haplotype block as listed in Table 3. Each haplotype block is formed by five SNPs with alleles labeled as 0 or 1. There is a binary environmental covariate $E (= 0/1)$ with prevalence p_E varied to be 0.1, 0.25, and 0.5. We explore three G-E dependence levels: strong G-E dependence, moderate/weak G-E dependence, and G-E independence. Under strong G-E dependence, frequencies of all haplotypes are set to be different between the $E = 0$ and $E = 1$ categories. For moderate/weak G-E dependence, only 3 (out of 6, 9, and 12) haplotype frequencies are set to be different, while frequencies of all haplotypes are the same in both categories of E under G-E independence.

For generating association with disease, we use various combinations of effects for the following three haplotypes — a rare, an uncommon, and a common haplotypes denoted as R, U, and C, respectively in Table 3. In each association scenario, one of these three haplotypes has a main effect and another has an interaction effect, as listed in Table 3. We also simulate a completely null model with all ORs set to be 1 (scenario 8). For a specific combination of association scenario, G-E dependence level, haplotype setting, and p_E value, we generate a sample of 1000 cases and 1000 controls in the same manner as described in the previous sub-section.

4.2.2 Results—Figures 1 to 3 show the comparison of powers and type I error rates between LBL-GXE-I, LBL-GXE-D, and LBL-GXE for association scenarios 1 – 3 for $p_E = 0.5$. Web Figures 2 – 5 show results for association scenarios 4 – 7. We see that LBL-GXE-I leads to seriously inflated type I error rates under G-E dependence while LBL-GXE-D controls type I error. However, LBL-GXE-D has lower power than LBL-GXE-I when G-E independence holds. In contrast, LBL-GXE provides a good balance between the two by having power similar to that of LBL-GXE-I in the case of G-E independence yet retaining the advantage of well-controlled type I errors in the case of G-E dependence. In other words, LBL-GXE is able to optimize power without sacrificing type I error under both G-E independence and dependence. This is also confirmed in the type I error rates for the Null Model (see Web Figure 6). The inflation or deflation of the type I error rate for a given haplotype does not seem to depend on its rarity. We also note that the powers to detect interactions (but not necessarily main effects) is typically higher under G-E independence.

Results of association scenario 4 with $p_E = 0.25$ and 0.1 are shown in Web Figures 7 and 8. The general trend of the results is similar to what we found for $p_E = 0.5$ but the powers decrease with decreasing p_E as expected. Also, we ran simulations with rarer causal haplotypes (listed in Web Table 3) and obtained similar results albeit lower powers (see Web Figure 9). Finally, we also generated data without assuming HWE ($d = 0$) and found the results to be similar, confirming that the methods can easily accommodate HWD (as reported in more details in Biswas and Lin, 2012).

5. Discussion

There is a pressing need for methods to detect rHTV interactions without relying on G-E independence assumption. To this end, we have proposed two approaches based on LBL framework. Our first proposed approach, LBL-GXE-D, can be used in situations where it is known apriori that there is G-E dependence. We showed that it controls type I error rates in the presence of G-E dependence. Nevertheless, it has lower powers than LBL-GXE-I when G-E independence holds.

Our second approach LBL-GXE optimizes this trade-off between power and type I error rate. It is able to retain increased power of LBL-GXE-I in presence of G-E independence while keeping the false positive rates under control in all situations, a desirable feature of LBL-GXE-D. Thus, LBL-GXE is a bridge between LBL-GXE-I and LBL-GXE-D, retaining the advantages of both and incorporating uncertainty about the G-E independence assumption. LBL-GXE is a general approach and can be applied in all situations. Nonetheless, if it is known beforehand that G-E independence or dependence holds, then LBL-GXE-I or LBL-GXE-D can be used accordingly as they may provide slightly higher power with potential gain in computational efficiency.

When applied to the lung cancer data, LBL-GXE detected interactions of several haplotypes in the 15q25.1 region with smoking. This includes a common haplotype CTTG and a rare haplotype TTTG in the haplotype block formed by four SNPs (numbers 7 - 10 out of 17 SNPs considered). VanderWeele et al. (2012) found allele C of SNP 7 (rs8034191) to increase risk while Yu et al. (2012) found alleles C and T to be of risk type for SNPs 8 (rs3885951) and 9 (rs2036534), respectively. Consistent with them, the significant common haplotype CTTG interaction has risk alleles C and T at SNP 7 and 9 but, different from Yu et al, has allele T at SNP 8. Combining this with the novel finding of the rare haplotype TTTG, which has only one previously implicated risk allele (T at SNP 9), it appears that these SNPs act in cis to interact with smoking to increase lung cancer risk. The presence of such haplotype effect is further strengthened by the fact that these two haplotypes share TTG at SNPs 8, 9, and 10, and as discussed in the Results section we found significant interaction with TTG also. This application clearly demonstrates how the richness of GWAS data can be mined for uncovering GXE with rare variants under G-E dependence.

For the lung cancer data analysis, a relevant issue is multiple testing. It is a non-trivial issue due to dependency across SNPs and blocks, and thus we simply used 95% CS for inference. Nonetheless, we also permuted the affection status to simulate a null scenario. We replicated this process 1000 times and estimated the type I error rates of LBL-GXE for all the effects in

all the five haplotype blocks. We found them to be in the range of 0 to 0.04. Thus, it appears that the significant effects that we reported are likely to be true positive.

LBL-GXE-D is computationally slower than LBL-GXE-I as it has more parameters and consequently also requires more iterations for convergence. For example, LBL-GXE-I takes 127, 200, and 308 seconds to finish 50,000 iterations under Setting 1, 2, and 3, respectively for a sample of size 2000 while the corresponding times for LBL-GXE-D with 70,000 iterations are 294, 488 and 766 seconds. LBL-GXE requires the largest number of iterations (100,000 for these same settings) but its computation time depends on the level of G-E dependence. It is fastest when there is G-E independence (as the chain tends to stay in model λ) and becomes slower with increasing strength of G-E dependence. For data generated under G-E independence, LBL-GXE takes 161, 175, and 211 seconds for the three settings while the respective times are 300, 505, and 780 seconds when there is strong G-E dependence. These computing times are for a 3.60 GHz Xeon processor under Linux operating system with 15.55 GB RAM.

With regard to the choice of values of hyper-parameters (for priors) and tuning parameters (for proposal distributions), we used the same values as those used in Biswas and Lin (2012), where it was shown that the MCMC procedure is not overly sensitive to the chosen values and that these choices provide well-calibrated results. The same holds for the newly introduced hyper-parameter ν and tuning parameter ζ . However, if any large scale changes are made to their values, the convergence of the chain as well as the acceptance rates should be monitored to ensure the results are still well calibrated (Gelman et al., 2003). The speed of convergence also depends on the number of rare haplotypes and their extent of rarity. The lung cancer dataset has several very rare haplotypes but we were able to achieve reasonable convergence with 200,000 iterations (few trace plots and their respective $\hat{R}$ statistics are shown in Web Figure 10). However, if there are many extremely rare haplotypes, more iterations may be needed for convergence.

As a simpler alternative to LBL-GXE, we have explored a two-stage approach whose first stage is a test for G-E independence at 5% level (using a chi-square or Fisher's exact test), and depending on the results, the second stage involves application of either LBL-GXE-I or LBL-GXE-D. We found that it gives inflated type I error rates sometimes when there is G-E dependence (e.g., see Web Table 4). In another two-stage approach, we applied LBL-GXE-D in the first stage. If G-E independence seems plausible (from CS for γ_1), then LBL-GXE-I is run in the second stage. This approach is less powerful than LBL-GXE when G-E independence holds (e.g., see Web Table 5) and is more computationally intensive when both stages have to be run.

Due to the computational intensity of LBL-GXE, we do not recommend it for GWAS. Instead, LBL-GXE seems most useful for zooming in on a candidate region of interest as we demonstrated in the lung cancer application. Another limitation is that LBL-GXE may not be powerful enough for detecting interactions with extremely rare haplotypes unless the effect sizes and/or sample sizes are large. Although the method seems to perform satisfactory for the haplotype frequencies that we considered in our study, we strongly recommend checking convergence to ensure validity of results, especially if some

haplotypes are very rare (as in the lung cancer data). The width of CS should also be taken into consideration in this regard. At the other end of spectrum of frequencies, if the interest is only in common haplotypes, then the existing methods such as haplo.glm are preferable (with pooling of rHTVs) to LBL as they are simpler and computationally faster.

Notwithstanding the limitations, we believe LBL is a useful tool for detecting GXE especially in light of the fact that the existing methods such as hapassoc and haplo.glm run into the issues of non-convergence, even lower power, and/or inflated type I error rate in the presence of rHTVs. Also, LBL has been extended to handle longitudinal data (Xia and Lin, 2014) and case-parent triad data (Wang and Lin, 2014). Thus, we believe LBL can serve as a useful toolkit in the ongoing hunt for rare variants and their interactions in a wide range of data. Our current and future directions of research include extending the methods to analyze data generated from complex sampling schemes, quantitative traits, and general pedigrees.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was partially supported by the grants R03CA171011 from the National Cancer Institute, NIH and DMS-1208968 from the National Science Foundation, and by allocations of computing times from the Texas Advanced Computing Center at the University of Texas at Austin. The lung cancer dataset was obtained through dbGaP accession number phs000093.v2.p2. We are thankful to the editor, associate editor, and two referees for their constructive comments and suggestions.

References

- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005; 21:263–265. [PubMed: 15297300]
- Biswas S, Lin S. Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics*. 2012; 68:587–597. [PubMed: 21955118]
- Biswas S, Xia S, Lin S. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. *Genetic Epidemiology*. 2014; 38:31–41. [PubMed: 24272913]
- Burkett K, Graham J, McNeney B. Hapassoc: software for likelihood inference of trait associations with SNP haplotypes and other attributes. *Journal of Statistical Software*. 2006; 16:1–19.
- Chen YH, Chatterjee N, Carroll RJ. Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*. 2008; 9:81–99. [PubMed: 17490987]
- Clark AG. The role of haplotypes in candidate gene studies. *Genetic Epidemiology*. 2004; 27:321–333. [PubMed: 15368617]
- Datta A, Zhang Y, Zhang L, Biswas S. Association of rare haplotypes on ULK4 and MAP4 genes with hypertension. *BMC Proceedings*. 2016 in press.
- Datta A, Biswas S. Comparison of Haplotype-based Statistical Tests for Disease Association with Rare and Common Variants. *Briefings in Bioinformatics*. 2015 doi: 10.1093/bib/bbv072.
- Gelman, A., Carlin, JB., Stern, HS., Rubin, DB. *Bayesian data analysis*. Chapman and Hall/CRC; Boca Raton, FL: 2003.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995; 82:711–732.

- Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA. Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology*. 2007; 31:75–90. [PubMed: 17123302]
- Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity*. 2003; 55:56–65. [PubMed: 12890927]
- Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *American Journal of Human Genetics*. 2010; 87:728–735. [PubMed: 21055717]
- Lin WY, Yi N, Lou XY, Zhi D, Zhang K, Gao G, et al. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genetic Epidemiology*. 2013; 37:560–570. [PubMed: 23740760]
- Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature Genetics*. 2010; 42:436–440. [PubMed: 20418889]
- Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. *Biometrics*. 2008; 64:685–694. [PubMed: 18162111]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Spitz MR, Amos CI, Dong Q, Lin J, Wu X. The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *Journal of the National Cancer Institute*. 2008; 100:1552–1556. [PubMed: 18957677]
- Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008; 452:638–642. [PubMed: 18385739]
- VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, Han Y, Spitz MR, Shete S, et al. Genetic variants on 15q25.1, smoking, and lung cancer: An assessment of mediation and interaction. *American Journal of Human Genetics*. 2012; 175:1013–1020.
- Wang M, Lin S. FamLBL: detecting rare haplotype disease association based on common SNPs using case-parent triads. *Bioinformatics*. 2014; 30:2611–2618. [PubMed: 24849576]
- Wang M, Lin S. Detecting associations of rare variants with common diseases: collapsing or haplotyping? *Briefings in Bioinformatics*. 2015; 16:759–768. [PubMed: 25596401]
- Weir, BS. *Genetic data analysis II*. Sinauer Associates Inc.; Sunderland, MA: 1996.
- Xia S, Lin S. Detecting longitudinal effects of haplotypes and smoking on hypertension using B-Splines and Bayesian LASSO. *BMC Proceedings*. 2013; 8:S85.
- Xia, S. Ph.D. Thesis. The Ohio State University; 2014. Detecting Rare Haplotype-Environment Interaction and Dynamic Effects of Rare Haplotypes using Logistic Bayesian LASSO..
- Yokota J, Shiraishi K, Kohno T. Genetic basis for susceptibility to lung cancer: Recent progress and future directions. *Advances in Cancer Research*. 2010; 109:51–72. [PubMed: 21070914]
- Yu K, Wacholder S, Wheeler W, Wang Z, Caporaso N, Teresa M, et al. A flexible Bayesian model for studying gene-environment interaction. *PLoS Genetics*. 2012; 8:e1002482. [PubMed: 22291610]
- Zhang Y, Biswas S. An improved version of logistic Bayesian LASSO for detecting rare haplotype-environment interactions with application to lung cancer. *Cancer Informatics*. 2015; 14:11–16.

Figure 1.

Powers (in gray shadow) and type I error rates of LBL-GXE-I, LBL-GXE-D, and LBL-GXE when $p_E = 0.5$, $OR.R = 3$, $OR.UXE = 3$, and all other $ORs = 1$. Each plot has two panels for main effects (bottom) and interactions of the corresponding haplotypes with E (top). 5% is marked by a gray horizontal dashed line. The overall frequency of each haplotype is shown in the parenthesis in the x-axis; the frequencies under $E = 0$ and $E = 1$ are listed in Table 3. The total number of replications in each simulation is 1000. This figure appears in color in the electronic version of this article.

Figure 2.

Powers (in gray shadow) and type I error rates of LBL-GXE-I, LBL-GXE-D, and LBL-GXE when $p_E = 0.5$, $OR.U = 3$, $OR.RXE = 5$, and all other $ORs = 1$. Each plot has two panels for main effects (bottom) and interactions of the corresponding haplotypes with E (top). 5% is marked by a gray horizontal dashed line. The overall frequency of each haplotype is shown in the parenthesis in the x-axis; the frequencies under $E = 0$ and $E = 1$ are listed in Table 3. The total number of replications in each simulation is 1000. This figure appears in color in the electronic version of this article.

Figure 3.

Powers (in gray shadow) and type I error rates of LBL-GXE-I, LBL-GXE-D, and LBL-GXE when $p_E = 0.5$, $OR.C = 1.5$, $OR.RXE = 5$, and all other $ORs = 1$. Each plot has two panels for main effects (bottom) and interactions of the corresponding haplotypes with E (top). 5% is marked by a gray horizontal dashed line. The overall frequency of each haplotype is shown in the parenthesis in the x-axis; the frequencies under $E = 0$ and $E = 1$ are listed in

Table 3. The total number of replications in each simulation is 1000. This figure appears in color in the electronic version of this article.

Table 1

Analysis of the lung cancer data. The four SNPs in this haplotype block are rs8034191, rs3885951, rs2036534 and rs2292117.

Effect	Overall Freq	Case Freq	Control Freq	OR	95% CS for OR	BF
CCCG	0.0001	0.0000	0.0002	0.78	(0.07, 5.19)	0.73
CCTG	0.1250	0.1368	0.1135	1.01	(0.74, 1.38)	0.10
CTCG	0.0013	0.0014	0.0012	0.97	(0.25, 3.37)	0.50
CTTA	0.0029	0.0030	0.0030	0.86	(0.26, 2.11)	0.43
CTTG	0.2772	0.2979	0.2571	0.92	(0.71, 1.19)	0.11
TCTG	0.0019	0.0028	0.0010	1.56	(0.50, 5.74)	0.71
TTCA	0.0002	0.0003	0.0000	1.12	(0.20, 7.38)	0.65
TTCG	0.2113	0.1884	0.2335	0.84	(0.62, 1.10)	0.28
TTTG	0.0149	0.0190	0.0109	0.95	(0.41, 1.93)	0.33
TTTA	0.3653	0.3505	0.3797	NA	NA	NA
smoking	0.8300	0.9200	0.7500	2.76	(2.00, 3.86) [*]	>100
CCCG × smoking				0.92	(0.09, 7.32)	0.73
CCTG × smoking				1.36	(0.98, 1.91)	0.94
CTCG × smoking				1.19	(0.35, 5.12)	0.53
CTTA × smoking				1.75	(0.64, 6.91)	0.75
CTTG × smoking				1.43	(1.09, 1.89) [*]	3.03
TCTG × smoking				2.00	(0.58, 10.06)	0.95
TTCA × smoking				1.39	(0.25, 12.14)	0.70
TTCG × smoking				1.06	(0.79, 1.45)	0.11
TTTG × smoking				2.18	(1.02, 5.27) [*]	2.96

Freq: Frequency

BF: Bayes Factor

^{*} 95% CS for OR excludes 1

Table 2

Proportion of replicates showing significance for three simulations based on the lung cancer data.

Effect	f(E=0)	f(E=1)	Simulation 1				Simulation 2				Simulation 3			
			OR	LBL-GXE	haplo.glm	OR	LBL-GXE	haplo.glm	OR	LBL-GXE	haplo.glm	OR	LBL-GXE	haplo.glm
CCCG	0.0005	0.0000	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00
CCTG	0.1260	0.1247	1	0.02	0.05	1	0.02	0.05	1	0.02	0.05	1	0.03	0.05
CTCG	0.0000	0.0016	1	0.00	0.64	1	0.00	0.61	1	0.00	0.61	1	0.00	0.67
CTTA (R)	0.0054	0.0025	1	0.00	0.26	1	0.00	0.19	1	0.00	0.19	1	0.02	0.20
CTTG (C)	0.2694	0.2788	1	0.04	0.05	1	0.04	0.05	1	0.04	0.05	1	0.04	0.04
TCTG	0.0032	0.0017	1	0.00	0.42	1	0.00	0.38	1	0.00	0.38	1	0.00	0.40
TTCA	0.0000	0.0002	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00
TTCG	0.2277	0.2080	1	0.03	0.05	1	0.03	0.04	1	0.03	0.04	1	0.03	0.05
TTTG (U)	0.0119	0.0154	1	0.01	0.05	1	0.02	0.04	1	0.02	0.04	1	0.00	0.04
smoking		2.76	1.00	0.97	0.99	2	0.99	0.93	2	0.99	0.93	2	0.99	0.93
CCCG × smoking		1	0.00	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00
CCTG × smoking		1	0.02	0.05	0.05	1	0.02	0.05	1	0.02	0.05	1	0.03	0.05
CTCG × smoking		1	0.00	0.64	0.64	1	0.00	0.62	1	0.00	0.62	1	0.00	0.66
CTTA × smoking		1	0.00	0.27	0.27	1	0.00	0.21	5	0.28	0.21	5	0.28	0.38
CTTG × smoking	1.43		0.67	0.52	0.52	1.5	0.84	0.69	1.5	0.83	0.69	1.5	0.83	0.68
TCTG × smoking		1	0.00	0.43	0.43	1	0.00	0.40	1	0.00	0.40	1	0.00	0.41
TTCA × smoking		1	0.00	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00
TTCG × smoking		1	0.03	0.06	0.06	1	0.03	0.05	1	0.03	0.05	1	0.03	0.05
TTTG × smoking		2.18	0.26	0.20	0.20	3	0.57	0.46	1	0.00	0.46	1	0.00	0.04

f(E=0) and f(E=1) are haplotype frequencies for E = 0 and E = 1 groups. The total number of replications in each simulation is 2000. For haplo.glm, an effect with p-value < 0.05 is considered as significant.

Table 3

Simulation Setup: OR under association scenarios 1 – 8 and haplotype frequencies under three G-E dependence levels.

Setting	Haplotype	Association Scenarios (OR)								Strong G-E Dep			Moderate/Weak G-E Dep			G-E Indep	
		1	2	3	4	5	6	7	8	E=0	E=1	E=0	E=0	E=1	E=0	E=1	All
1	h1: 01100	–	–	–	–	–	–	–	–	0.35	0.25	0.2375	0.3625	0.005	0.0075	0.3	0.0075
	h2: 10100 (R)	3	5*	5*	3	–	–	–	–	0.01	0.005	0.01	0.005	0.015	0.015	0.015	0.015
	h3: 11011 (U)	3*	3	–	5*	3*	3	5*	–	0.01	0.02	0.015	0.015	0.015	0.155	0.11	0.4125
	h4: 11100	–	–	–	–	–	–	–	–	0.03	0.28	0.155	0.05	0.05	0.05	0.05	0.05
	h5: 11111 (C)	–	–	1.5	–	1.5	2*	1.5	–	0.05	0.17	0.17	0.4125	0.4125	0.4125	0.4125	0.4125
	h6: 10011	–	–	–	–	–	–	–	–	0.55	0.275	0.4125	0.4125	0.4125	0.4125	0.4125	0.4125
2	h1: 01010	–	–	–	–	–	–	–	–	0.02	0.1	0.06	0.06	0.06	0.06	0.06	0.06
	h2: 01100	–	–	–	–	–	–	–	–	0.18	0.32	0.1775	0.3225	0.08	0.08	0.08	0.08
	h3: 10000	–	–	–	–	–	–	–	–	0.13	0.03	0.08	0.005	0.015	0.015	0.015	0.015
	h4: 10100 (R)	3	5*	5*	3	–	–	–	–	0.01	0.005	0.01	0.005	0.015	0.015	0.015	0.015
	h5: 11011 (U)	3*	3	–	5*	3*	3	5*	–	0.01	0.02	0.015	0.015	0.015	0.015	0.015	0.015
	h6: 11100	–	–	–	–	–	–	–	–	0.15	0.03	0.09	0.09	0.09	0.09	0.09	0.09
	h7: 11101	–	–	–	–	–	–	–	–	0.06	0.11	0.085	0.085	0.085	0.085	0.085	0.085
	h8: 11111 (C)	–	–	1.5	–	1.5	2*	1.5	–	0.05	0.15	0.17	0.3125	0.3125	0.3125	0.3125	0.3125
	h9: 10011	–	–	–	–	–	–	–	–	0.39	0.235	0.3125	0.3125	0.3125	0.3125	0.3125	0.3125
3	h1: 00111	–	–	–	–	–	–	–	–	0.03	0.11	0.07	0.07	0.07	0.07	0.07	0.07
	h2: 01000	–	–	–	–	–	–	–	–	0.01	0.03	0.02	0.02	0.02	0.02	0.02	0.02
	h3: 01011	–	–	–	–	–	–	–	–	0.03	0.07	0.05	0.05	0.05	0.05	0.05	0.05
	h4: 01101	–	–	–	–	–	–	–	–	0.03	0.09	0.06	0.06	0.06	0.06	0.06	0.06
	h5: 01110	–	–	–	–	–	–	–	–	0.22	0.06	0.14	0.14	0.14	0.14	0.14	0.14
	h6: 10010	–	–	–	–	–	–	–	–	0.11	0.05	0.08	0.08	0.08	0.08	0.08	0.08
	h7: 10100 (R)	3	5*	5*	3	–	–	–	–	0.01	0.005	0.01	0.005	0.015	0.015	0.015	0.015
	h8: 11011 (U)	3*	3	–	5*	3*	3	5*	–	0.01	0.02	0.015	0.015	0.015	0.015	0.015	0.015
	h9: 11101	–	–	–	–	–	–	–	–	0.13	0.05	0.09	0.09	0.09	0.09	0.09	0.09
	h10: 11110	–	–	–	–	–	–	–	–	0.18	0.08	0.13	0.13	0.13	0.13	0.13	0.13
	h11: 11111 (C)	–	–	1.5	–	1.5	2*	1.5	–	0.05	0.15	0.1575	0.0425	0.0425	0.0425	0.0425	0.1

Setting	Haplotype	Association Scenarios (OR)								Strong G-E Dep		Moderate/Weak G-E Dep		G-E Indep
		1	2	3	4	5	6	7	8	E=0	E=1	E=0	E=1	All
h12: 10001		-	-	-	-	-	-	-	-	0.19	0.285	0.1775	0.2975	0.2375

An OR indicated by * is an interaction effect between that haplotype and covariate otherwise it denotes the main effect of that haplotype. An OR of “-” denotes null effect (OR = 1). Covariate has no main effect in all scenarios. Haplotype frequencies are different for E = 0 and E = 1 groups under the two G-E dependence models. Dep: Dependence; Indep: Independence