Published in final edited form as:

*J Hum Genet*. 2017 September ; 62(9): 819–829. doi:10.1038/jhg.2017.43.

# Logistic Bayesian LASSO for Genetic Association Analysis of Data from Complex Sampling Designs

**Yuan Zhang**[1], **Jonathan N. Hofmann**[2], **Mark P. Purdue**[2], **Shili Lin**[3,*], and **Swati Biswas**[1,*]

[1]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080

[2]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892

[3]Department of Statistics, The Ohio State University, Columbus, OH 43210

## Abstract

Detecting gene-environment interactions (GXE) with rare variants is critical in dissecting the etiology of common diseases. Interactions with rare haplotype variants (rHTV) are of particular interest. At the same time, complex sampling designs, such as stratified random sampling, are becoming increasingly popular for designing case-control studies, especially for recruiting controls. The US Kidney Cancer Study (KCS) is an example, wherein all available cases were included while the controls at each site were randomly selected from the population by frequency matching with cases based on age, sex, and race. There is currently no rHTV association method that can account for such a complex sampling design. To fill this gap, we consider logistic Bayesian LASSO (LBL), an existing rHTV approach for case-control data, and show that its model can easily accommodate the complex sampling design. We study two extensions that include stratifying variables either as main effects only or with additional modeling of their interactions with haplotypes. We conduct extensive simulation studies to compare the complex sampling methods with the original LBL methods. We find that when there is no interaction between haplotype and stratifying variables, both extensions perform well while the original LBL methods lead to inflated type I error rates. However, when such an interaction exists, it is necessary to include the interaction effect in the model to control the type I error. Finally, we analyze the KCS data and find a significant interaction between (current) smoking and a specific rHTV in the N-acetyltransferase 2 (NAT2) gene.



*Address for correspondence: Swati Biswas, PhD, Department of Mathematical Sciences, University of Texas at Dallas, 800 W Campbell Rd., FO35 Richardson, TX 75080-3021, Tel: (972) 883-6686, swati.biswas@utdallas.edu; Shili Lin, PhD, Department of Statistics, The Ohio State University, 1958 Neil Avenue Columbus, OH 43210, Tel: (614) 292-7404, shili@stat.osu.edu.

**Software**

The methods have been implemented in an R package LBL available at http://www.utdallas.edu/~swati.biswas http://www.stat.osu.edu/~statgen/SOFTWARE/LBL

**Conflict of Interest**

The authors have no conflict of interest to declare.

**Keywords**

GXE; G-E dependence; Missing Heritability; Rare variants; NAT2 gene

## Introduction

Rare variants and gene-environment interactions (GXE) have been suggested in the literature as potential causes of "missing heritability" in common diseases. We consider these problems by focusing on G being a rare haplotype variant (rHTV), which may reflect a combination of common single nucleotide polymorphisms (SNP). Thus, rHTVs can be studied even in existing genome-wide association studies (GWAS) data without the need to sequence any additional data. Recently, we have proposed an approach for rHTV association for case-control data called logistic Bayesian LASSO (LBL).[1] We have extended it to handle GXE under the assumption of G-E independence as well as when this assumption is relaxed or there is an uncertainty about it.[2–4] LBL shrinks the effects of unassociated haplotypes or their interactions with environmental covariates towards zero, so that the associated effects can be identified with considerable power.[5–7] In fact, LBL is one of the most powerful rHTV methods.[8]

Complex sampling designs are being utilized with increasing frequency in case-control studies, especially for sampling of the controls. Typically, all available cases are included while controls are selected by stratified sampling using frequency matching with cases. Strata are usually formed based on known risk factors such as race, age, and sex. Often one or more strata, especially those containing minorities, are oversampled to obtain more controls. To account for different sampling rates arising from unequal sampling among strata, population weights are calculated, which indicate the number of population members represented by each sample subject. It is important to use these weights in the analysis to avoid bias in the results. However, at the same time, the use of weights also eliminates the power and efficiency in case-control studies due to the fact that population weights for controls are usually much larger than those for cases, leading to large variability in weights.[9] To regain some of the lost efficiency, rescaling of population weights has been suggested.[10] For example, one way of rescaling is such that the sum of the case (control) weights is equal to case (control) sample size. Another type of rescaling is to have the sum of weights of controls be equal to the sum of weights of cases.

The US Kidney Cancer Study (KCS) was designed using a complex sampling scheme through stratified random sampling for recruiting subjects.[11,12] It was conducted at two sites — Chicago and Detroit. Cases identified from the Metropolitan Detroit Cancer Surveillance System and Cook County hospitals were recruited. At each site, the controls were frequency matched to cases based on age, sex, and race. The matching rate of controls to cases was 2:1 in blacks and 1:1 in whites. Age groups were formed at 5-year intervals starting from 20 to 79 years. For age groups ≥ 65 years, controls were chosen from the database of Medicare beneficiaries, which has information on age, sex, and race. For age groups < 65 years, controls were chosen from a listing of the Department of Motor Vehicles (DMV), which contains information on age and sex but not on race. As a proxy for race, strata of low and

high black densities were formed based on Census data. Thus, the overall strata were formed by cross-classification of age, sex, and race (or black density). In addition to these stratifying variables, KCS collected covariates such as smoking status, high blood pressure, education level, and body mass index. As described in Colt et al.,[12] to account for features related to the complex sampling design (differential sampling rates for controls and cases, survey nonresponse, and deficiencies in coverage of the population-at-risk in the DMV and Medicare files), population weights were formed for each sampled individual.

Several authors have analyzed the KCS data and reported risk factors for kidney cancer such as smoking, obesity, and hypertension.[12–14] Besides, genetic susceptibility and its interaction with environmental factors have been reported to affect the risk as reported in the KCS and other studies.[15–19] In particular, the N-acetyltransferase 2 (NAT2) gene is known to code for an enzyme involved in tobacco-carcinogen mechanism. Semenza et al.[15] found that smoking-related risk of kidney cancer is higher among those carrying a polymorphic variant of NAT2 called slow acetylator genotype than rapid acetylators. Longuemaux et al.[20] observed a higher risk of kidney cancer for subjects with NAT2 slow acetylators combined with CYP1A1 variants, however, they did not study gene-environment interactions.

To the best of our knowledge, there is currently no rHTV association method that can account for complex sampling design such as that adopted in the KCS. To fill this gap, we adapt the LBL model to analyze this type of data. We show that stratified sampling with frequency matching can be easily accounted for in the framework of LBL without any additional modeling. We conduct simulation studies to investigate the properties of the extensions and compare with the original LBL method. Finally, we also analyze the KCS data to study the NAT2–smoking interaction.

## Materials and Methods

The method mostly follows from Zhang et al.[4] with necessary adaptation to include stratifying variables and population weights. Suppose we have 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, $i = 1, \ldots, n$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Let $G_i$ denote the observed genotype of the $i$th individual, and $\boldsymbol{G} = (G_1, \ldots, G_n)$. We then let $\mathscr{S}(G_i)$ be the set of haplotype pairs compatible with $G_i$ as the haplotype pair of a person may not be completely determined from the observed genotypes. Further we denote the $r$th haplotype pair in $\mathscr{S}(G_i)$ by $Z_{ir}$. Next we denote the vector of environmental covariates of the $i$th individual by $\boldsymbol{E}_i$. For a complex sampling design, the stratifying variables play a key role, and they are denoted collectively as $\boldsymbol{S}_i$ for individual $i$. In this paper, we consider both $\boldsymbol{E}$ and $\boldsymbol{S}$ to be categorical.

### Complex Sampling Design Structure and Analysis

For the type of complex sampling considered in this paper, the sampling mechanism leads to known (rescaled) population weights, $w_i$, for the $i$th individual. In simple terms, $w_i$ is the number of individuals in the population that the $i$th sampled person represents. It is essentially the ratio of the number of individuals available to be sampled (population size) to the number of individuals actually sampled (sample size) in the stratum to which the $i$th individual belongs. In surveys, non-response and post-stratification adjustments are further

made to these weights, and they are made available along with the rest of the sample data.[9] The weights are typically rescaled to increase efficiency as mentioned in the Introduction section. Further details on calculation of weights will be provided in the simulation study section.

The basic principle that we follow for incorporating complex sampling design in the Bayesian framework is to write the analysis model conditional on the information and variables that describe the data collection process.[21] That is, for writing the likelihood, we condition on the fact that the frequencies of cases and controls were matched (in some way that will become apparent below) in each stratum, and on the values of the variables used for matching (in this case the stratifying variables).

## Retrospective Likelihood

Conditional on $\{w_i\}$, $i = 1, \ldots, n$, and $\boldsymbol{S}$, the retrospective likelihood of the observed data is written as:

$$\begin{aligned} L(\boldsymbol{\Psi}) &= \prod_{i=1}^{n_1} (P(G_i, \boldsymbol{E}_i | Y_i=1, \boldsymbol{S}_i, \boldsymbol{\Psi}))^{w_i} \prod_{i=n_1+1}^{n} (P(G_i, \boldsymbol{E}_i | Y_i=0, \boldsymbol{S}_i, \boldsymbol{\Psi}))^{w_i} \\ &= \prod_{i=1}^{n_1} \left( \sum_{Z_{ir} \in \mathscr{S}(G_i)} P(Z_{ir} | Y_i=1, \boldsymbol{E}_i, \boldsymbol{S}_i, \boldsymbol{\Psi}) P(\boldsymbol{E}_i | Y_i=1, \boldsymbol{S}_i, \boldsymbol{\Psi}) \right)^{w_i} \\ &\quad \prod_{i=n_1+1}^{n} \left( \sum_{Z_{ir} \in \mathscr{S}(G_i)} P(Z_{ir} | Y_i=0, \boldsymbol{E}_i, \boldsymbol{S}_i, \boldsymbol{\Psi}) P(\boldsymbol{E}_i | Y_i=0, \boldsymbol{S}_i, \boldsymbol{\Psi}) \right)^{w_i} \end{aligned} \quad (1)$$

where $\boldsymbol{\Psi}$ consists of the regression coefficients and the parameters associated with the haplotype pair frequencies, which will be specified more explicitly later. Note that conditioning on the case/control status ($Y$), stratifying variable information, and the weight for each person automatically takes care of matching frequencies of cases and controls in all strata in the retrospective (in contrast to prospective) likelihood formulation. Now we will specify the model for each component of the likelihood. In the following, we suppress the subscripts $i$ and $r$ for simplicity without causing ambiguity.

## Modeling of $P(Z|\boldsymbol{E}, \boldsymbol{S}, Y = 0)$

We start with modeling $P(Z|\boldsymbol{E}, \boldsymbol{S}, Y = 0) = a_{Z|\boldsymbol{E},\boldsymbol{S}}$, the frequency of haplotype pair $Z$ in the control population for a given $\boldsymbol{E}$ and $\boldsymbol{S}$. Suppose there are a total of $m$ haplotypes and assume gene–environment (G–E) dependence is only due to some of the stratifying variables and/or covariates, defined as $\boldsymbol{C}_{dep}$, a subset of $\{\boldsymbol{E}, \boldsymbol{S}\}$. That is, conditional on $\boldsymbol{C}_{dep}$, G and E are independent.[22,23] Then we denote the haplotype frequencies in the control population by $f(\boldsymbol{C}_{dep}) = (f_1(\boldsymbol{C}_{dep}), \ldots, f_m(\boldsymbol{C}_{dep}))$. We model $a_{Z|\boldsymbol{E},\boldsymbol{S}}$ for a haplotype pair $Z = (z_k, z_{k'})$ as follows:

$$\begin{aligned} a_{Z|\boldsymbol{E},\boldsymbol{S}} &= P(Z=(z_k, z_{k'})|Y=0, \boldsymbol{E}, \boldsymbol{S}) \\ &= P(Z=(z_k, z_{k'})|Y=0, \boldsymbol{C}_{dep}) \\ &= \delta_{kk'} d f_k(\boldsymbol{C}_{dep}) + (2-\delta_{kk'})(1-d) f_k(\boldsymbol{C}_{dep}) f_{k'}(\boldsymbol{C}_{dep}), \end{aligned} \quad (2)$$

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 that captures excess/reduction of homozygosity.[24] For $d = 0$, the above expression is equivalent to assuming Hardy-Weinberg Equilibrium (HWE) while other values of $d$ allow Hardy-Weinberg Disequilibrium (HWD).

We then model $\boldsymbol{f}(\boldsymbol{C}_{dep})$ using a multinomial logistic regression model to allow G-E dependence.[25] Let the $m$th haplotype be the baseline and assume $\boldsymbol{C}_{dep}$ has $L$ levels excluding baseline(s):$\boldsymbol{C}_{dep} = \{C_1, C_2, \ldots, C_L\}$. For example, if $\boldsymbol{C}_{dep}$ consists of two binary variables, then $L = 2$ with exclusion of baseline category of each variable. Then we have

$$\log\left(\frac{f_k(\boldsymbol{C}_{dep})}{f_m(\boldsymbol{C}_{dep})}\right) = \gamma_{k0} + \gamma_{k1}C_1 + \gamma_{k2}C_2 + \ldots + \gamma_{kL}C_L = g_k(\boldsymbol{C}_{dep}), k=1, 2, \ldots, m-1. \quad (3)$$

Thus,

$$f_k(\boldsymbol{C}_{dep}) = \frac{exp(g_k(\boldsymbol{C}_{dep}))}{1+\sum_{j=1}^{m-1} exp(g_j(\boldsymbol{C}_{dep}))}, k=1, \ldots, m-1; f_m(\boldsymbol{C}_{dep}) = \frac{1}{1+\sum_{j=1}^{m-1} exp(g_j(\boldsymbol{C}_{dep}))}. \quad (4)$$

Let $\boldsymbol{\gamma}$ denote an $(m-1)\times(L+1)$ matrix with the $(k, l)$th element being $\gamma_{kl}$, $k = 1, \ldots, m-1$, $l = 0, \ldots, L$. Combining (2) and (4), we have now fully specified $a_{Z|\boldsymbol{E},\boldsymbol{S}}(\boldsymbol{\gamma}, d)$.

## Modeling of $P(Z|\boldsymbol{E}, \boldsymbol{S}, Y = 1)$

Next let us consider $P(Z|\boldsymbol{E}, \boldsymbol{S}, Y = 1) = b_{Z|\boldsymbol{E},\boldsymbol{S}}$, the frequency of haplotype pair $Z$ in the case population for a given value of $\boldsymbol{E}$ and $\boldsymbol{S}$. We express $b_{Z|\boldsymbol{E},\boldsymbol{S}}$ in terms of $a_{Z|\boldsymbol{E},\boldsymbol{S}}$ and the odds of disease for a given $Z$, $\boldsymbol{E}$, and $\boldsymbol{S}$, $\theta_{Z,\boldsymbol{E},\boldsymbol{S}}(= P(Y = 1|Z, \boldsymbol{E}, \boldsymbol{S})/P(Y = 0|Z, \boldsymbol{E}, \boldsymbol{S}))$:

$$b_{Z|\boldsymbol{E},\boldsymbol{S}} = P(Z|\boldsymbol{E}, \boldsymbol{S}, Y=1) = \frac{P(Y=1|Z, \boldsymbol{E}, \boldsymbol{S})P(Z, \boldsymbol{E}, \boldsymbol{S})}{\sum_H P(Y=1|H, \boldsymbol{E}, \boldsymbol{S})P(H, \boldsymbol{E}, \boldsymbol{S})} = \frac{\theta_{Z,\boldsymbol{E},\boldsymbol{S}} a_{Z|\boldsymbol{E},\boldsymbol{S}}}{\sum_H \theta_{H,\boldsymbol{E},\boldsymbol{S}} a_{H|\boldsymbol{E},\boldsymbol{S}}}, \quad (5)$$

where $H$ is the set of all possible haplotype pairs and $\theta_{Z,\boldsymbol{E},\boldsymbol{S}}$ is modeled using logistic regression. We consider two different ways of modeling $\theta_{Z,\boldsymbol{E},\boldsymbol{S}} = \exp(\boldsymbol{X}\boldsymbol{\beta})$ with respect to the stratifying variables. They are included as covariates either just as main effects (LBLc-GXE) or with additional modeling of interaction effects of $\boldsymbol{S}$ with haplotypes (LBLc-GXE-GXS); "c" in LBLc represents complex sampling. More specifically, $\boldsymbol{X}$ is (1, $\boldsymbol{X}_{\boldsymbol{S}}$, $\boldsymbol{X}_{\boldsymbol{E}}$, $\boldsymbol{X}_{Z}$,

$\boldsymbol{X}_Z\boldsymbol{X}_{\boldsymbol{E}}$) in LBLc-GXE and (1, $\boldsymbol{X}_{\boldsymbol{S}}$, $\boldsymbol{X}_{\boldsymbol{E}}$, $\boldsymbol{X}_Z$, $\boldsymbol{X}_Z\boldsymbol{X}_{\boldsymbol{S}}$, $\boldsymbol{X}_Z\boldsymbol{X}_{\boldsymbol{E}}$) in LBLc-GXE-GXS. For each model, $\beta$ is the vector comprising the corresponding regression coefficients. Here $\boldsymbol{X}_Z = (x_1, x_2, \ldots, x_{m-1})$, 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. $\boldsymbol{X}_{\boldsymbol{E}}$ and $\boldsymbol{X}_{\boldsymbol{S}}$ consist of the usual dummy variables corresponding to $\boldsymbol{E}$ and $\boldsymbol{S}$, respectively. $\boldsymbol{X}_Z\boldsymbol{X}_{\boldsymbol{E}}$ and $\boldsymbol{X}_Z\boldsymbol{X}_{\boldsymbol{S}}$ are obtained by (scalar) multiplication of $\boldsymbol{X}_Z$ and $\boldsymbol{X}_{\boldsymbol{E}}$ and $\boldsymbol{X}_Z$ and $\boldsymbol{X}_{\boldsymbol{S}}$, respectively.

## Modeling of $P(\boldsymbol{E}_i|Y_i = 0, \boldsymbol{S}_i)$ and $P(\boldsymbol{E}_i|Y_i = 1, \boldsymbol{S}_i)$

It remains to model $P(\boldsymbol{E}_i/Y_i = 0, \boldsymbol{S}_i)$ and $P(\boldsymbol{E}_i/Y_i = 1, \boldsymbol{S}_i)$ in (1). Assuming a saturated model for $P(\boldsymbol{E}, \boldsymbol{S})$, $P(\boldsymbol{E}|Y, \boldsymbol{S}) \propto P(Y|\boldsymbol{E}, \boldsymbol{S})$ without loss of information.[26,27] Then using the Bayes rule, we get the following:

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

Thus, we can write the observed data retrospective likelihood in (1) as:

$$L(\Psi) \propto \prod_{i=1}^{n_1} \left( \frac{\sum_{Z_{ir} \in \mathcal{S}(G_i)} \theta_{Z_{ir},\boldsymbol{E}_i,\boldsymbol{S}_i} a_{Z_{ir}|\boldsymbol{E}_i,\boldsymbol{S}_i}}{1+\sum_H \theta_{H,\boldsymbol{E}_i,\boldsymbol{S}_i} a_{H|\boldsymbol{E}_i,\boldsymbol{S}_i}} \right)^{w_i} \prod_{i=n_1+1}^{n} \left( \frac{\sum_{Z_{ir} \in \mathcal{S}(G_i)} a_{Z_{ir}|\boldsymbol{E}_i,\boldsymbol{S}_i}}{1+\sum_H \theta_{H,\boldsymbol{E}_i,\boldsymbol{S}_i} a_{H|\boldsymbol{E}_i,\boldsymbol{S}_i}} \right)^{w_i},$$

where $\Psi = (\beta, \gamma, d)$.

## Priors, Posterior Distributions, and Inference on Association

These follow closely from LBL-GXE[4] as elucidated briefly in the following. Bayesian LASSO is used to regularize the regression coefficients $\beta$s by assigning each of them a double exponential prior centered at 0 and variance $2/\lambda^2$: $\pi(\beta|\lambda) = \frac{\lambda}{2}\exp(-\lambda|\beta|)$, $-\infty < \beta < \infty$. Such regularization helps in weeding out the unassociated effects, making it possible for the associated ones, especially those involving rHTVs, to stand out. The parameter $\lambda$ controls the degree of penalty. It is assigned a Gamma($a$, $b$) hyper-prior with parametrization such that its mean is $a/b$. When $a = b = 20$, we obtain SD($\beta$) = 1.53, which corresponds to a realistic variability in odds ratios. For $\gamma$ parameters, we use a double exponential prior with hyper-parameter $\nu$ set to be 0.5, which provides well-calibrated results as seen in our simulation study. For $d$, we note that it is dependent on $\boldsymbol{f}(\boldsymbol{C}_{dep})$ as $a_{Z|\boldsymbol{E},\boldsymbol{S}}$ should be nonnegative. Thus, $d > \{f_k(\boldsymbol{C}_{dep})/(1 - f_k(\boldsymbol{C}_{dep}))\}$, $k = 1, \ldots, m-1$. As $-1 < d < 1$, we get $\max_k \{-f_k(\boldsymbol{C}_{dep})/(1 - f_k(\boldsymbol{C}_{dep}))\} < d < 1$. Therefore, we set the prior for $d$ to be uniformly distributed in that range.

The posterior distributions of all parameters in $\Psi$ are estimated using Markov chain Monte Carlo (MCMC) methods. Finally, we test for significance of each $\beta$ coefficient by computing its 95% credible set (CS) using MCMC samples from its posterior distribution. A 95% CS not covering 0 is considered as an evidence for significance. Alternatively, Bayes factor (BF) > 2 can be also used to declare significance.[1] For the KCS data analysis, we report both 95% CS and BF.

## Results

### Simulation Study

**One Stratifying Variable—**We carry out simulation studies to investigate the performance of LBL for complex sampling data. In this subsection, we consider one binary stratifying variable S (= 0/1) with prevalence $p_S = P(S = 1) = 0.3$. There is also a binary environmental covariate E (= 0/1) with prevalence $p_{E/S=0} = P(E = 1|S = 0) = 0.3$ and $p_{E/S=1} = P(E = 1|S = 1) = 0.7$. There are three haplotype settings with 6, 9, and 12 haplotypes in a haplotype block as listed in Table 1. Each haplotype block is formed by five SNPs with alleles labeled as 0 or 1. There are two rHTVs, denoted as R1 and R2, in each block. Note that there is G-S dependence as frequencies of haplotypes differ in the two strata. This, in turn, induces G-E dependence as prevalence of E differs across strata.

For creating association scenarios, we use various combinations of the following effects: R1, R2XS, R2XE, S, and E, as listed in Table 1. We also simulate a completely null model with all ORs set to be 1 (scenario 6).

To mimic a complex sampling design for generating data, we first generate a population of cases and controls, and then sample from it using matching based on the stratifying variable. For a specific combination of association scenario and haplotype setting, we generate a population of 10,000 subjects in the following manner. For each individual, first we simulate a stratifying variable value, say $S$ using the $p_S$ value. Then we generate an environmental covariate value, $E$, using the $p_{E/S}$ value. Then we generate a phased haplotype pair, say $Z$, using the frequencies given in Table 1 and assuming HWE ($d = 0$). Next, the individual is assigned to be a case or control using a logistic regression model: $\log(p/(1-p)) = \boldsymbol{X\beta}$, where $p$ is the probability that the individual is case, and $\boldsymbol{X} = (1, \boldsymbol{X_S}, \boldsymbol{X_E}, \boldsymbol{X_Z}, \boldsymbol{X_Z X_S}, \boldsymbol{X_Z X_E})$. The intercept is calculated using a baseline prevalence of 0.1, i.e., $\beta_0 = \log(0.1/0.9)$. For the other $\beta$ coefficients, we use the corresponding ORs as listed in Table 1. 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. Once a population of 10,000 subjects is generated in this manner, we obtain a sample from it as described next.

Suppose the numbers of cases and controls in the population of Stratum $h$ ($h = 0, 1$; $h = 0$ corresponds to S = 0 and $h = 1$ corresponds to S = 1) are $N_h^{Ca}$ and $N_h^{Co}$. Correspondingly, let the number of cases and controls in the sample of the Stratum $h$ be $n_h^{Ca}$ and $n_h^{Co}$. First, we select all the cases in the population to be included in the sample for each of the strata, i.e., $n_h^{Ca} = N_h^{Ca}$, $h = 0, 1$.[19] For selecting controls, to mimic the KCS data, we use differential sampling rates in the two strata. In Stratum 0, the number of controls is set to be the same as

the number of cases, that is, $n_0^{Co}=n_0^{Ca}(=N_0^{Ca})$. While in Stratum 1, we select a simple random sample of size $2n_1^{Ca}$ controls, i.e., $n_1^{Co}=2n_1^{Ca}(=2N_1^{Ca})$. In most situations, out of a population of size 10,000, we get a sample of size of 2200 – 3800 with the number of cases varying between 1000 – 1500 (700 – 900 in Stratum 0 and 150 – 800 in Stratum 1) depending on the scenario.

Next we calculate the population weights for sampled cases and control in each stratum and rescale them. The rescaling is such that the sum of weights for cases is the same as the sum of weights for controls, as in the analysis of the KCS data reported by Hofmann et al.[14] Denote the rescaled weights of sampled cases and controls in stratum $h$ by $w_h^{Ca}$ and $w_h^{Co}$. As all cases are sampled, the weight for a case is 1, i.e., $w_h^{Ca}=1$, $h$ = 0, 1. Thus, the sum of weights for cases in the sample is the sample size of the cases $(N_0^{Ca}+N_1^{Ca})$. The population weights of controls in stratum $h$ is $\frac{N_h^{Co}}{n_h^{Co}}$. For rescaling, we divide these population weights by their sum, i.e., $(N_0^{Co}/n_0^{Co})*n_0^{Co}+(N_1^{Co}/n_1^{Co})*n_1^{Co}=N_0^{Co}+N_1^{Co}$, and then multiply by case sample size, i.e., $N_0^{Ca}+N_1^{Ca}$. Thus, $w_0^{Co}=\frac{N_0^{Co}(N_0^{Ca}+N_1^{Ca})}{n_0^{Co}(N_0^{Co}+N_1^{Co})}=\frac{N_0^{Co}(N_0^{Ca}+N_1^{Ca})}{N_0^{Ca}(N_0^{Co}+N_1^{Co})}$ and $w_1^{Co}=\frac{N_1^{Co}(N_0^{Ca}+N_1^{Ca})}{n_1^{Co}(N_0^{Co}+N_1^{Co})}=\frac{N_1^{Co}(N_0^{Ca}+N_1^{Ca})}{2N_1^{Ca}(N_0^{Co}+N_1^{Co})}$. Therefore, we can see that if we oversample the controls for one stratum, their weights will be reduced. This can be clearly seen from the above expressions of weights if the control to case ratio in the population is constant across different strata. Note that all persons in a stratum have the same weight, original as well as rescaled, and these are computed only once for a given sample.

We analyze each sample using LBLc-GXE and LBLc-GXE-GXS. For comparison, we also apply LBL-GXE from Zhang et al.,[4] which models G-E dependence but ignores the stratifying variables (Note that in Zhang et al.,[4] this method was referred as LBL-GXE-D, however, for the sake of simplicity here we refer to it as LBL-GXE). Additionally, we also analyze the data using a variation of LBL-GXE, referred to as LBL-GXE-GXS, which includes the stratifying variables as covariates but does not use sampling weights, i.e., ignores the complex sampling scheme. For each of these four methods, we use a total number of 120,000 iterations with a burn-in period of 20,000 iterations to ensure satisfactory convergence.[21] The total number of replications in each simulation is 500. For each $\beta$ coefficient, we calculate the percentage of times (out of 500) that its 95% credible sets (CS) does not cover 0 to study the power or type I error rate.

Figures 1 to 3 and Supplementary Figures 1 to 3 show the powers and type I error rates for LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS, and LBL-GXE for association scenarios 1 through 6 (null model), respectively. In scenario 1 (Figure 1), the performance of LBLc-GXE and LBLc-GXE-GXS are comparable, detecting the main haplotype and interaction effects with E with similar powers and keeping the type I error rates under control, while LBL-GXE-GXS and LBL-GXE have inflated type I error rates. In scenario 2 (Figure 2) where an interaction effect with S is present in the data, LBLc-GXE-GXS continues to performs well, while the other three methods, including LBLc-GXE, have inflated type I

error rates. In scenario 3 (Figure 3) where the main effect of S is included in the data, LBLc-GXE, LBLc-GXE-GXS, and LBL-GXE-GXS control the type I error rates successfully while LBL-GXE leads to inflated type I error rates. However, we should note that the main effect of S detected by LBL-GXE-GXS here is not really an indication of its power because this method detects the main effect of S to be significant always irrespective of whether S has a true main effect or not, as seen in Figures 1, 2 and Supplementary Figures 1 to 3. In summary, LBLc-GXE controls type I error rates in situations where there is no interaction between haplotype and stratifying variable, while LBLc-GXE-GXS performs well in all scenarios. The Supplementary Figure 3 for the null model (scenario 6) shows that LBL-GXE-GXS and LBL-GXE lead to seriously inflated type I error rates while LBLc-GXE and LBLc-GXE-GXS control these rates well.

We also explore scenarios 2 and 6 with $p_S = 0.15$ and $p_{E/S=0} = p_{E/S=1} = 0.19$ to mimic S and E to be race and smoking. We use the fact that the prevalence of blacks in the US is about 15% and the prevalence of smoking among whites or blacks in the US is about 19%. Supplementary Figures 4 and 5 show the corresponding results. The methods perform similarly as before except that with lower prevalences of S and E, LBLc-GXE and LBLc-GXE-GXS have reduced power, as expected.

For scenario 2 and setting 1, we also analyzed the data using a standard haplotype association method haplo.glm.[28] Haplo.glm is based on the generalized linear model and uses maximum likelihood methods for inference. The results are reported in Figure 4 and Supplementary Table 1, which show that haplo.glm has inflated type I error rates.

Additionally, we investigate a different rescaling of the weights such that the sum of the case (control) weights is equal to case (control) sample size. We compare the two types of rescaling by applying LBLc-GXE and LBLc-GXE-GXS to the data generated under setting 1 of scenario 2. The results of these two types of rescaling are comparable as shown in Supplementary Table 2. We also examine the methods for data generated under HWD by setting $d = 0.1$ in the data simulation procedure for setting 1 of scenario 2. The relative performances of the methods are similar to what we found earlier under HWE. The detailed results are shown in Supplementary Figure 6.

**Two Stratifying Variables—**We next conduct simulation studies using two stratifying variables $S_1$ (0/1) and $S_2$ (0/1) to mimic race and sex. We set the prevalence $p_{S_1}=P(S_1=1)=0.15$ and $p_{S_2}=P(S_2=1)=0.5$. These two stratifying variables form 4 strata: Stratum 1 ($S_1 = 0$, $S_2 = 0$), Stratum 2 ($S_1 = 0$, $S_2 = 1$), Stratum 3 ($S_1 = 1$, $S_2 = 0$), and Stratum 4 ($S_1 = 1$, $S_2 = 1$). The binary environmental covariate E has prevalence $p_{E|S_2=0}=P(E=1|S_2=0)=0.15$ and $p_{E|S_2=1}=P(E=1|S_2=1)=0.2$, which mimics that prevalence of smoking among females and males are about 15% and 20%, respectively (http://kff.org/other/state-indicator/smoking-adults-by-gender/). We consider 6 haplotypes and two types of G-S dependence — dependence on $S_1$ only (G-$S_1$ dependence) or on both $S_1$ and $S_2$ (G-$S_1$-$S_2$ dependence), as listed in Table 2.

The sample generation and weights calculation procedure is similar to that in the One Stratifying Variable subsection. Specifically, we generate a population of size 10,000 and select all cases in the population. In Strata 1 and 2, we select a simple random sample of controls of the same size as the number of cases in the corresponding stratum. In Strata 3 and 4, we select a simple random sample of controls with size double of that of the cases in the corresponding stratum. The total sample sizes range from 2000 – 2500 with roughly 1000 cases (about 400 each in Strata 1 and 2 and 100 each in Strata 3 and 4).

Figure 5 shows the results for both G-$S_1$ dependence and G-$S_1$-$S_2$ dependence. The relative performances of the methods are comparable to what we observe in the case of one stratifying variable. That is, LBLc-GXE-GXS has type I error rates well controlled while the other three methods, including LBLc-GXE, have inflated type I error rates as the simulation model includes non-null effects of both GXE and GXS. The powers are lower under G-$S_1$-$S_2$ dependence compared to G-$S_1$ dependence since the former involves additional modeling.

**Application to the KCS data—**Following our motivation described in the Introduction section, we study the NAT2 gene and its interaction with smoking. Deitz et al.[29] report that seven SNPs (rs1801279, rs1041983, rs1801280, rs1799929, rs1799930, rs1208, and rs1799931) explain 100% of the alleles detected in NAT2. Out of these seven SNPs, six are available in the KCS data. From them, a haplotype block consisting of the following five SNPs is detected by Haploview:[30] rs1041983, rs1801280, rs1799929, rs1799930, and rs1208. We focus on analyzing this 5-SNP haplotype block.

The KCS data include rescaled population weights; the rescaling is such that the sum of the weights for the cases is the same as the sum of the weights for the controls. We used these weights in our analyses to account for complex sampling design. We consider smoking status as a covariate with three levels: never smoking, former smoking, and current smoking (consisting of occasional and regular current smokers). Further, we adjust for all four stratifying variables: site (Detroit, Chicago), age (<45, 45–54, 55–64, 65–74, ≥75 year), race (white, black), and sex following Li and Graubard[19] and Hofmann et al.[14] Note that at each site (city), both cases and controls were recruited, and so using site as a stratifying variable along with race can address population stratification due to geographical location to some extent.

After removing subjects with missing genotype or smoking status, there are 909 cases and 936 controls in the KCS data. Table 3 shows some characteristics of these data. There is a higher proportion of current smokers amongst cases than in controls for both whites and blacks. More details about these data can be found in Hofmann et al.[14] Haplotype frequencies as estimated using the hapassoc software[31] based on maximum likelihood estimation are shown in Table 4. They vary substantially between the two races as well as between cases and controls. These estimates are used as starting values of the frequency ($\boldsymbol{\gamma}$) parameters in the MCMC procedures.

In our analysis, we set haplotype TTCAA as the baseline as it has similar frequencies in the cases and controls among whites as well as blacks. In addition, we assume that G-E dependence can be captured through the dependence of haplotypes on race, that is, $\boldsymbol{C}_{dep}$ =

{Race}. As there are several haplotypes that are extremely rare, we run LBL for a large number of iterations to ensure convergence and accurate results. In particular, to monitor convergence, we run three chains from three different starting points and make diagnostic plots and calculate the $R^2$ statistics.[21] We run each chain for 300,000 iterations, discard initial 100,000 as burn-in, and combine the three chains to obtain the posterior distributions.

The results are reported in Table 5. Both LBLc-GXE and LBLc-GXE-GXS find an interaction effect of a rare haplotype CTCGG and current smoking to be highly significant with BF > 100. LBLc-GXE also detects the main effects of CTCGG and current smoking to be significant while LBLc-GXE-GXS finds only the latter to be significant. Specifically, LBLc-GXE-GXS estimates the OR of the interaction to be 0.37 and the main effect of CTCGG to be null. Therefore, among current smokers, the carriers of CTCGG have reduced odds of kidney cancer compared to the carriers of the baseline haplotype TTCAA. The two methods also detect a few other effects with their 95% CS excluding 1; however, their corresponding BF values are small.

On the other hand, if the complex sampling design is ignored in the analysis (i.e., LBL-GXE-GXS or LBL-GXE are used for analysis), we fail to detect the main effect of former or current smoking. Besides, LBL-GXE-GXS, which models main and interaction effects of stratification variables, even detects a protective effect of the black race, which contradicts the fact that blacks are at an increased risk of kidney cancer than whites.[17] These contradictory results illustrate the importance of accounting for complex sampling design in the analysis.

## Discussion

Complex sampling schemes such as stratified sampling with frequency matching are now increasingly used in practice. At the same time, in the quest to dissect the etiology of common diseases, tremendous efforts are being directed towards detecting rare variants and their interactions with environmental covariates. Yet most of the current genetic association methods do not take the design of data collection into account, which can lead to biased results. Thus, there is a pressing need for methods, especially for rare variants, that can properly account for complex sampling design. Here we adapted the LBL framework to analyze data originating from complex sampling schemes. As LBL is based on retrospective likelihood, it automatically conditions on the matched frequencies of cases and controls in each stratum once we condition on the stratifying variables. The differential sampling rates across strata are accounted for using the (rescaled) population weights.

When there is no interaction between stratifying variable and haplotype, we found that LBLc-GXE provides considerable powers and controlled type I error rates. However, it has increased type I error rates when such type of interaction is present. In such situations, the method that additionally models the interaction term, LBLc-GXE-GXS, performs well. On the other hand, the originally proposed LBL method has high type I error rates even when stratifying variables are included as covariates in the model. In addition to inference on association, which is our main focus, we also report in Supplementary Table 3 bias, standard errors (SE), and mean squared errors (MSE) of the point estimates of the regression

coefficients whose true OR > 1. For the null effects (OR = 1), these values are smaller than the ones reported in the table and thus omitted for brevity. As we can see from the table, these are all small for LBLc-GXE-GXS. The same is true for LBLc-GXE except for the bias and the MSE of the R2XE effect when there are two stratifying variables and there is also R2XS effect. In this case, LBLc-GXE is not the correct model and thus gives inflated type I errors, as already noted above.

To examine the methods under realistic linkage disequilibrium patterns and potential cryptic relatedness amongst subjects, we also carried out simulations based on the haplotypes and results from the KCS data analysis. We use the haplotype frequencies from Table 4 (separately for whites and blacks) and use race as the stratifying variable (S) and smoking as a binary environmental covariate (E). To mimic the prevalences of blacks in the US and smoking amongst the two races, we set $p_S = 0.15$ and $p_{E/S=0} = p_{E/S=1} = 0.19$, as used earlier in some simulations. The data are generated in the same manner as described in the Simulation Study section. We consider two scenarios — (1) Null with all ORs set to 1 and (2) Non-null with OR = 1.4 for E and OR = 0.3 for interaction of haplotype CTCGG with E, which are similar to those estimated in the KCS data analysis. The results, presented in Supplementary Figure 7, are consistent with our earlier simulation study results.

When applied to the KCS data, our method found current smokers to be at an increased risk for kidney cancer, consistent with the literature. Further, our finding of interaction between smoking and NAT2 gene has been also reported in the literature. However, this is the first time, to the best of our knowledge, that an interaction with a specific rHTV has been implicated. Moreover, we found that the current smokers carrying the rHTV CTCGG have reduced odds of the disease compared to those with baseline haplotype. Semenza et al.[15] and Chow et al.[17] state that kidney cancer risk is higher for NAT2 slow acetylators than rapid acetylators among smokers. The haplotype CTCGG appears to be of a rapid acetylator type as per http://www.snpedia.com/index.php/NAT2, which might explain its protective effect for current smokers. However, the finding of this significant interaction effect appears to be novel and should be investigated in future studies. Moreover, the population stratification issue, in general, might need to be handled more carefully because genetic background can sometimes vary even within the same race and site.

As an alternative to LBLc-GXE, which models stratifying variables as covariates, we also explored including stratifying variables in the model by assigning to each stratum its own intercept,[32] denoted by LBLc-GXE(I). We compared LBLc-GXE and LBLc-GXE(I) for a few simulation settings when there is one binary stratifying variable and they perform similarly. This is expected as the two models are actually equivalent in this case. When there are two or more stratifying variables and their effects are not additive, LBLc-GXE(I) may perform better than LBLc-GXE, however, its power will suffer if the model is additive given that it has a large number of intercept parameters.

The LBL methods are computationally intensive and hence are more suited for zooming into genes/regions of interest implicated previously by fast, typically single-SNP-based and genome-wide, algorithms. LBLc-GXE-GXS is computationally slower than LBLc-GXE as it has more parameters. For example, when there is one stratifying variable, LBLc-GXE takes

915, 1379, and 1993 seconds to finish 120,000 iterations under settings 1, 2, and 3 of scenario 2, respectively, while the corresponding times for LBLc-GXE-GXS are 1095, 1694, and 2435 seconds. These computing times are for a 3.60 GHz Xeon processor under Linux operating system with 15.55 GB RAM.

To summarize, we have extended the original LBL method to incorporate complex sampling schemes, in particular, stratified random sampling. Its main advantage stems from the fact that none of the current haplotype association methods can handle both rare variants and complex sampling design in the model. Another complex sampling scheme that is gaining popularity is matching controls to cases individually rather than with frequency matching (typically referred as matched case-control). Although we focus on stratified sampling design for a more concise discussion, the model for an individually matched case-control design would be similar because the retrospective likelihood will take care of conditioning on individual level matching, similar to frequency matching. LBL has been also extended to handle longitudinal data[33] and case-parent triad data.[34] Thus, LBL is now a comprehensive suite of rHTV methods, which can be used for various types of data. We plan to extend the methods to quantitative traits and extended family data as well as other sampling designs such as nested case-control and case-cohort to further increase LBL's capability.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was partially supported by the grant R03CA171011 from the National Cancer Institute, NIH and by allocations of computing times from the Texas Advanced Computing Center at the University of Texas at Austin. The US Kidney Cancer Study was supported by the Intramural Research Program of the NIH, National Cancer Institute. We are thankful to the two anonymous referees for their constructive comments and suggestions.

## References

1. 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]
2. Biswas S, Xia S, Lin S. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. Genet Epidemiol. 2014; 38:31–41. [PubMed: 24272913]
3. Zhang Y, Biswas S. An improved version of logistic Bayesian LASSO for detecting rare haplotype-environment interactions with application to lung cancer. Cancer Inform. 2015; 14:11–16.
4. Zhang Y, Lin S, Biswas S. Detecting rare haplotype-environment interaction under uncertainty of gene-environment independence assumption. Biometrics. 2017; 73:344–355. [PubMed: 27478935]
5. Biswas S, Papachristou C. Evaluation of logistic Bayesian LASSO for identifying association with rare haplotypes. BMC Proceedings. 2014; 8:S54. [PubMed: 25519334]
6. Datta AS, Zhang Y, Zhang L, Biswas S. Association of rare haplotypes on ULK4 and MAP4 genes with hypertension. BMC Proceedings. 10(Suppl 7):44.
7. Wang M, Lin S. Detecting associations of rare variants with common diseases: collapsing or haplotyping? Brief in Bioinform. 2015; 16:759–768.
8. Datta AS, Biswas S. Comparison of haplotype-based statistical tests for disease association with rare and common variants. Brief Bioinform. 2016; 17:657–671. [PubMed: 26338417]
9. Korn, EL., Graubard, B. Analysis of health surveys. Wiley; New York, NY, USA: 1999.

10. Scott AJ, Wild CJ. Case-control studies with complex sampling. Applied Statistics. 2001; 50:389–401.
11. Digaetano, R., Graubard, B., Rao, S., Severynse, J., Wacholder, S. Sampling racially matched population controls for case-control studies: using DMV lists and oversampling minorities. 2003. https://fcsm.sites.usa.gov/files/2014/05/2003FCSM_DiGaetano.pdf. Accessed Aug 8, 2016
12. Colt JS, Schwartz K, Graubard BI, Davis F, Ruterbusch J, DiGaetano R, Purdue M, Rothman N, Wacholder S, Chow W-H. Hypertension and risk of renal cell carcinoma among white and black Americans. Epidemiology. 2011; 22:797–804. [PubMed: 21881515]
13. Purdue MP, Moore LE, Merino MJ, Boffetta P, Colt JS, Schwartz KL, Bencko V, Davis FG, Graubard BI, Janout V, et al. An investigation of risk factors for renal cell carcinoma by histologic subtype in two case-control studies. Int J Cancer. 2013; 132:2640–2647. [PubMed: 23150424]
14. Hofmann JN, Schwartz K, Chow WH, Ruterbusch JJ, Shuch BM, Karami S, Rothman N, Wacholder S, Graubard BI, Colt JS, et al. The association between chronic renal failure and renal cell carcinoma may differ between black and white Americans. Cancer Causes Control. 2013; 24:167–174. [PubMed: 23179659]
15. Semenza JC, Ziogas A, Largent J, Peel D, Anton-Culver H. Gene-environment interactions in renal cell carcinoma. Am J Epidemiol. 2001; 153:851–859. [PubMed: 11323315]
16. Moore LE, Brennan P, Karami S, Menashe I, Berndt SI, Dong L, Meisner A, Yeager M, Chanock S, Colt J, et al. Apolipoprotein E/C1 Locus Variants Modify Renal Cell Carcinoma Risk. Cancer Res. 2009; 69:8001–8008. [PubMed: 19808960]
17. Chow W, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. Nat Rev Urol. 2010; 7:245–257. [PubMed: 20448658]
18. Purdue MP, Johansson M, Zelenika D, Toro JR, Scelo G, Moore LE, Prokhortchouk E, Wu X, Kiemeney LA, Gaborieau V, et al. Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. Nat Genet. 2011; 43:60–65. [PubMed: 21131975]
19. Li Y, Graubard BI. Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples. Biostatistics. 2012; 13:711–723. [PubMed: 22522235]
20. Longuemaux S, Delomenie C, Gallou C, Mejean A, Vincent-Viry M, Bouvier R, Droz D, Krishnamoorthy R, Galteau MM, Junien C, et al. Candidate genetic modifiers of individual susceptibility to renal cell carcinoma: a study of polymorphic human xenobiotic-metabolizing enzymes. Cancer Res. 1999; 59:2903–2908. [PubMed: 10383153]
21. Gelman, A., Carlin, JB., Stern, HS., Rubin, DB. Bayesian data analysis. Chapman and Hall/CRC; Boca Raton, FL, USA: 2003.
22. Chatterjee N, Carroll R. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92:399–418.
23. Mukherjee B, Zhang L, Ghosh M, Sinha S. Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. Biometrics. 2007; 63:834–844. [PubMed: 17489972]
24. Weir, BS. Genetic data analysis II. Sinauer Associates Inc; Sunderland, MA, USA: 1996.
25. Mukherjee B, Chatterjee N. An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. Biometrics. 2008; 64:685–694. [PubMed: 18162111]
26. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.
27. 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. Genet Epidemiol. 2007; 31:75–90. [PubMed: 17123302]
28. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered. 2003; 55:56–65. [PubMed: 12890927]
29. Deitz AC, Rothman N, Rebbeck TR, Hayes RB, Chow WH, Zheng W, Hein DW, Garcia-Closas M. Impact of misclassification in genotype-exposure interaction studies: Example of N-

Acetyltransferase 2 (NAT2), smoking, and bladder cancer. Cancer Epidemiol Biomarkers Prev. 2004; 13:1543–1546. [PubMed: 15342459]

30. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]
31. Burkett K, Graham J, McNeney B. Software for likelihood inference of trait associations with SNP haplotypes and other attributes. J Stat Softw. 2006; 16:1–19.
32. Scott, AJ., Wild, CJ. Fitting logistic regression models in case-control studies with complex sampling. In: Chambers, RL., Skinner, CJ., editors. Analysis of survey data. Wiley, Chichester; England: 2003. p. 109-120.
33. Xia S, Lin S. Detecting longitudinal effects of haplotypes and smoking on hypertension using B-Splines and Bayesian LASSO. BMC Proceedings. 2014; 8:S85. 2014. [PubMed: 25519413]
34. 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]

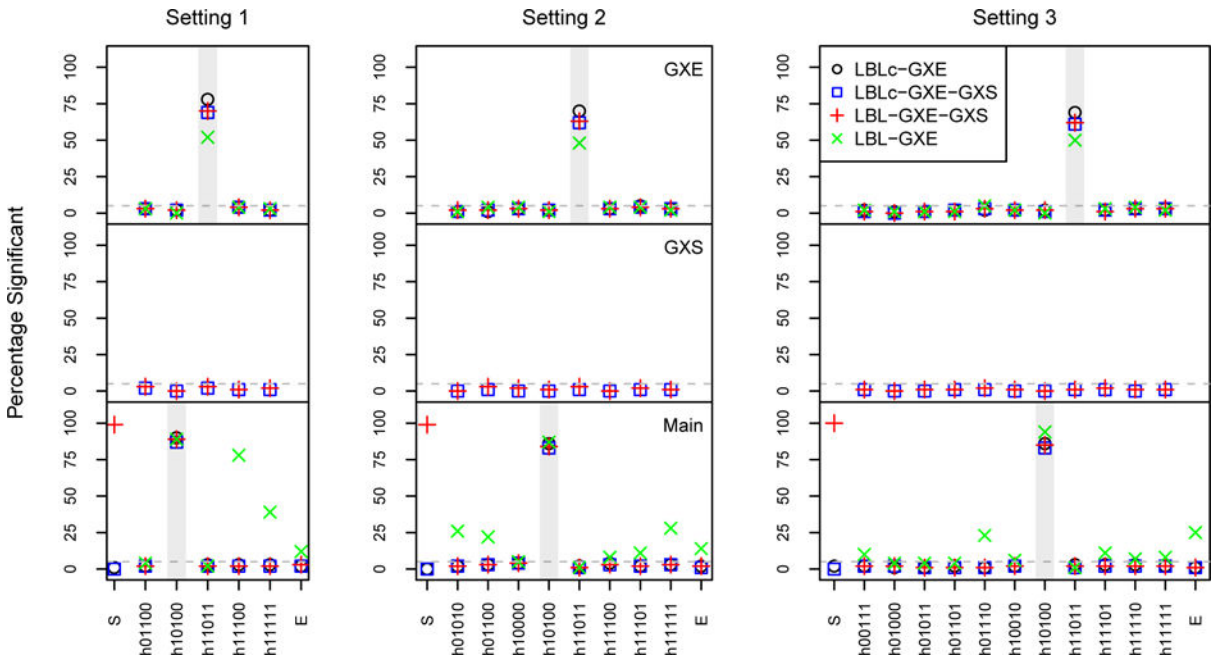


**Figure 1.**

Powers (in gray shadow) and type I error rates of LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS, and LBL-GXE for Scenario 1 (OR.R1 = 3, OR.R2XE = 3, and all other ORs = 1). Each plot has three panels for main effects (bottom row), interactions of the corresponding haplotypes with S (middle row), and interactions of the corresponding haplotypes with E (top row). 5% is marked by a gray horizontal dashed line. The haplotype frequencies are listed in Table 1.

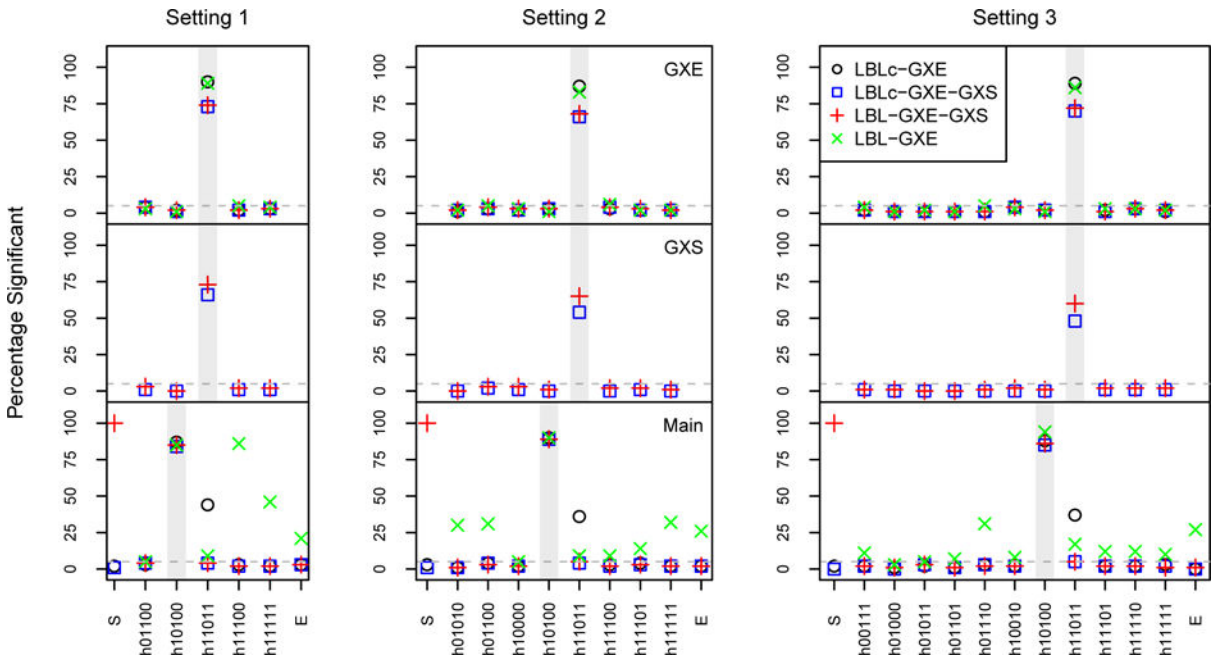


**Figure 2.**

Powers (in gray shadow) and type I error rates of LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS, and LBL-GXE for Scenario 2 (OR.R1 = 3, OR.R2XS = 3, OR.R2XE = 3, and all other ORs = 1). Each plot has three panels for main effects (bottom row), interactions of the corresponding haplotypes with S (middle row), and interactions of the corresponding haplotypes with E (top row). 5% is marked by a gray horizontal dashed line. The haplotype frequencies are listed in Table 1.

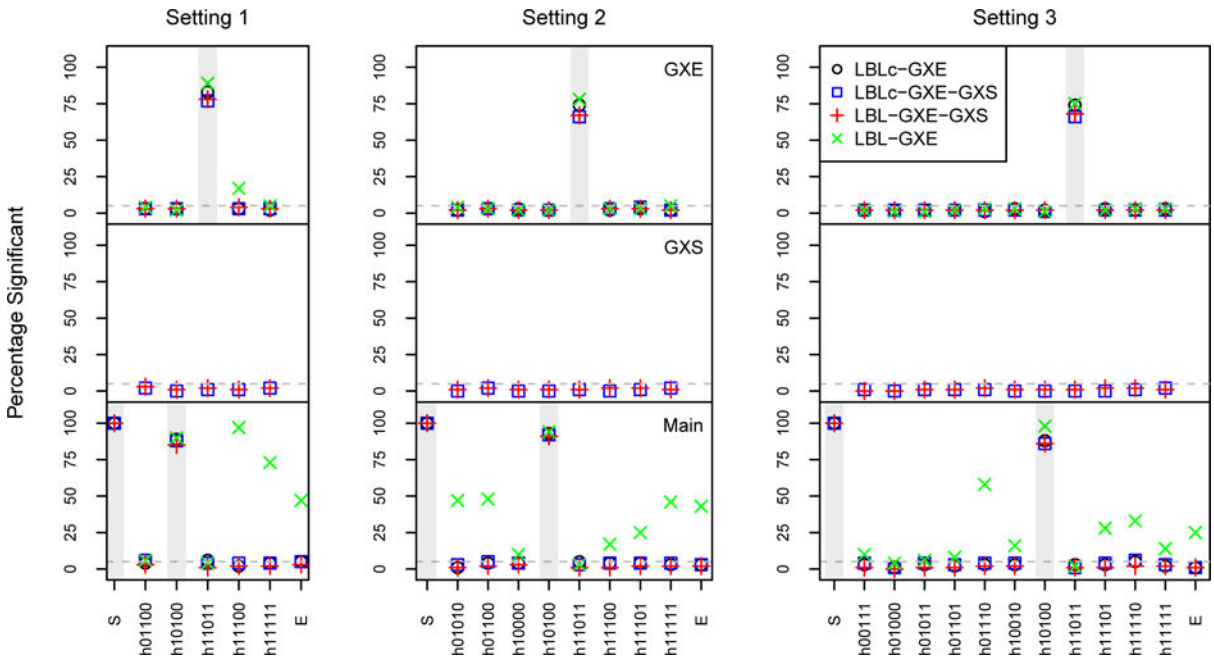


**Figure 3.**

Powers (in gray shadow) and type I error rates of LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS, and LBL-GXE for Scenario 3 (OR.R1 = 3, OR.S = 3, OR.R2XE = 3, and all other ORs = 1). Each plot has three panels for main effects (bottom row), interactions of the corresponding haplotypes with S (middle row), and interactions of the corresponding haplotypes with E (top row). 5% is marked by a gray horizontal dashed line. The haplotype frequencies are listed in Table 1.

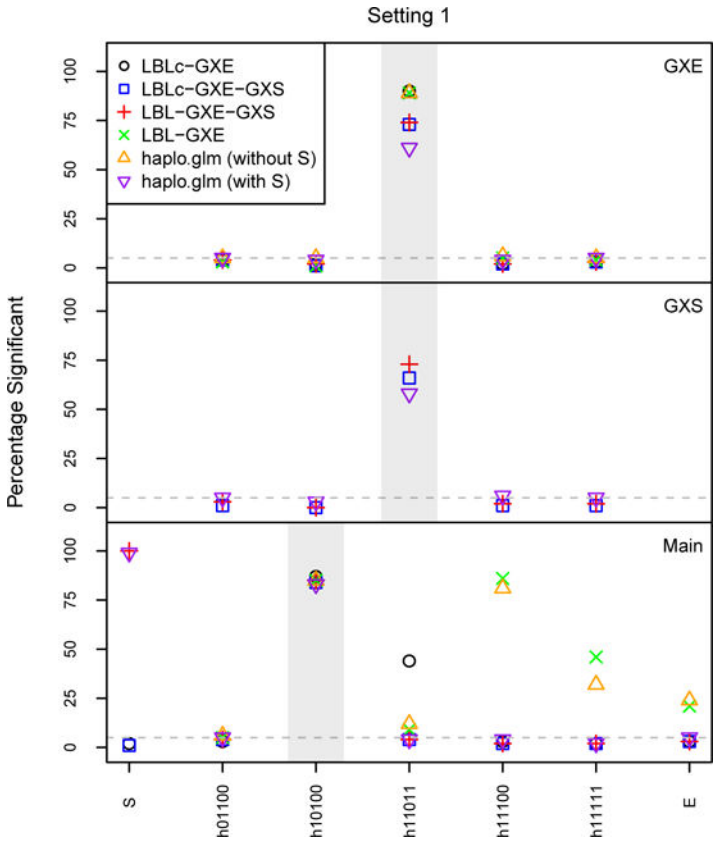


**Figure 4.**

Powers (in gray shadow) and type I error rates of LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS,, LBL-GXE, and haplo.glm (with and without S) for Scenario 2 (OR.R1 = 3, OR.R2XS = 3, OR.R2XE = 3, and all other ORs = 1). Each plot has three panels for main effects (bottom row), interactions of the corresponding haplotypes with S (middle row), and interactions of the corresponding haplotypes with E (top row). 5% is marked by a gray horizontal dashed line. The haplotype frequencies are listed in Table 1.

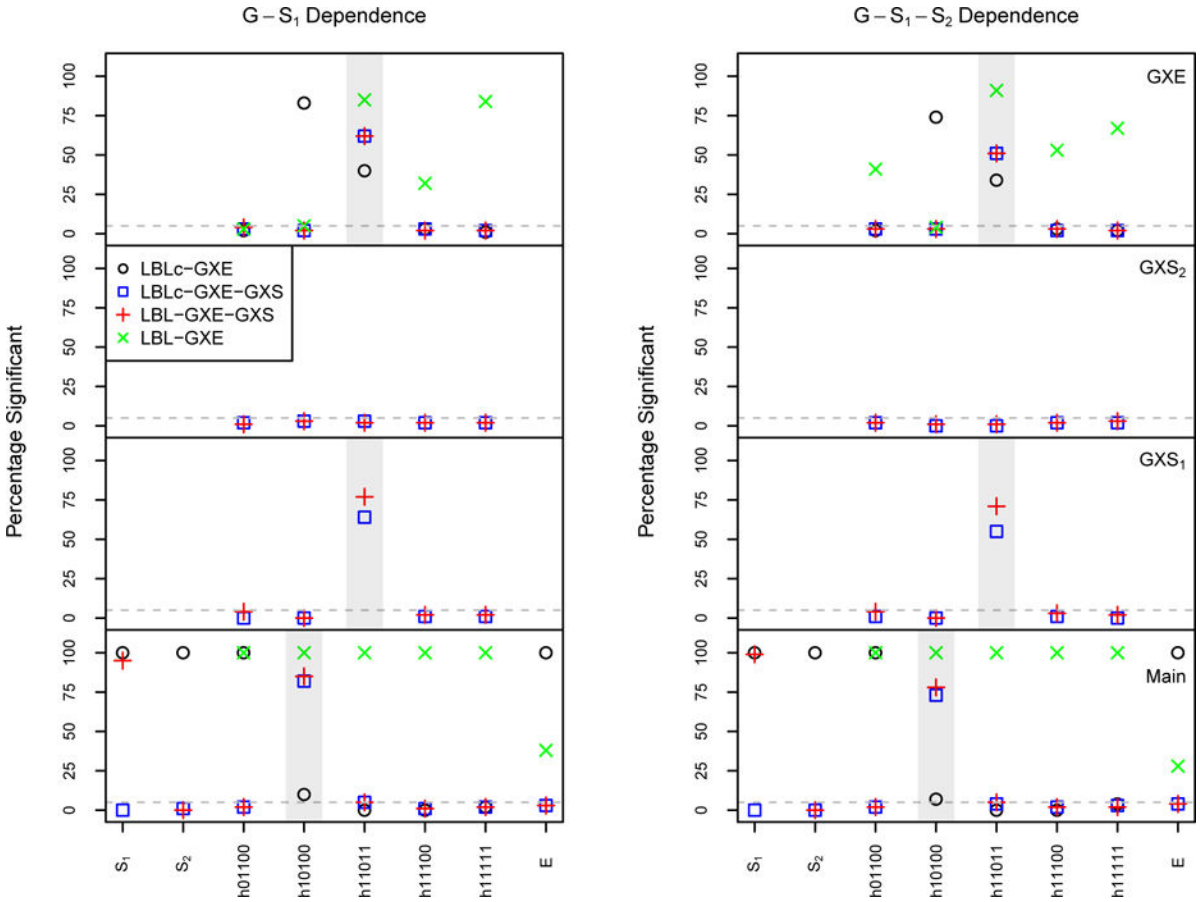


**Figure 5.**

Powers (in gray shadow) and type I error rates of LBLc-GXE, LBLc-GXE-GXS, LBL-GXE-GXS, and LBL-GXE when there are two stratifying variables $S_1$ and $S_2$, where $p_{S_1}=0.15$, $p_{S_2}=0.5$, $p_{E|S_2=0}=0.15$, $p_{E|S_2=1}=0.2$, OR.R1 = 3, OR.R2XS$_1$ = 5, OR.R2XE = 4, and all other ORs = 1. Each plot has four panels for main effects (bottom row), interactions of the corresponding haplotypes with S$_1$ (second from bottom row), interactions of the corresponding haplotypes with S$_2$ (third from bottom row), and interactions of the corresponding haplotypes with E (top row). 5% is marked by a gray horizontal dashed line. The haplotype frequencies are listed in Table 2.

**Table 1**

Simulation Setup for One Stratifying Variable: OR under association scenarios 1 – 6 and frequencies of haplotypes and environmental covariate in each stratum.

| | | Association Scenarios (OR) | | | | | | Freq | |
|---|---|---|---|---|---|---|---|---|---|
| **Setting** | **Hap** | **1** | **2** | **3** | **4** | **5** | **6** | **S = 0** | **S = 1** |
| 1 | 01100 | – | – | – | – | – | – | 0.35 | 0.25 |
| | 10100 (R1) | 3 | 3 | 3 | 3 | 3 | – | 0.01 | 0.005 |
| | 11011 (R2) | 3 (E) | 3 (S), 3 (E) | 3 (E) | 3 (S) | 3 (S), 3 (E) | – | 0.01 | 0.02 |
| | 11100 | – | – | – | – | – | – | 0.03 | 0.28 |
| | 11111 | – | – | – | – | – | – | 0.05 | 0.17 |
| | 10011 | – | – | – | – | – | – | 0.55 | 0.275 |
| | E | – | – | – | – | 1.5 | – | 0.3 | 0.7 |
| | S | – | – | 3 | – | – | – | 0.7* | 0.3** |
| 2 | 01010 | – | – | – | – | – | – | 0.02 | 0.1 |
| | 01100 | – | – | – | – | – | – | 0.18 | 0.32 |
| | 10000 | – | – | – | – | – | – | 0.13 | 0.03 |
| | 10100 (R1) | 3 | 3 | 3 | 3 | 3 | – | 0.01 | 0.005 |
| | 11011 (R2) | 3 (E) | 3 (S), 3 (E) | 3 (E) | 3 (S) | 3 (S), 3 (E) | – | 0.01 | 0.02 |
| | 11100 | – | – | – | – | – | – | 0.15 | 0.03 |
| | 11101 | – | – | – | – | – | – | 0.06 | 0.11 |
| | 11111 | – | – | – | – | – | – | 0.05 | 0.15 |
| | 10011 | – | – | – | – | – | – | 0.39 | 0.235 |
| | E | – | – | – | – | 1.5 | – | 0.3 | 0.7 |
| | S | – | – | 3 | – | – | – | 0.7* | 0.3** |
| 3 | 00111 | – | – | – | – | – | – | 0.03 | 0.11 |
| | 01000 | – | – | – | – | – | – | 0.01 | 0.03 |
| | 01011 | – | – | – | – | – | – | 0.03 | 0.07 |
| | 01101 | – | – | – | – | – | – | 0.03 | 0.09 |
| | 01110 | – | – | – | – | – | – | 0.22 | 0.06 |
| | 10010 | – | – | – | – | – | – | 0.11 | 0.05 |

| | | Association Scenarios (OR) | | | | | | Freq | |
|---|---|---|---|---|---|---|---|---|---|
| **Setting** | **Hap** | **1** | **2** | **3** | **4** | **5** | **6** | **S = 0** | **S = 1** |
| | 10100 (R1) | 3 | 3 | 3 | 3 | 3 | – | 0.01 | 0.005 |
| | 11011 (R2) | 3 (E) | 3 (S), 3 (E) | 3 (E) | 3 (S) | 3 (S), 3 (E) | – | 0.01 | 0.02 |
| | 11101 | – | – | – | – | – | – | 0.13 | 0.05 |
| | 11110 | – | – | – | – | – | – | 0.18 | 0.08 |
| | 11111 | – | – | – | – | – | – | 0.05 | 0.15 |
| | 10001 | – | – | – | – | – | – | 0.19 | 0.285 |
| | E | – | – | – | – | 1.5 | – | 0.3 | 0.7 |
| | S | – | – | 3 | – | – | – | 0.7* | 0.3** |

An OR followed by (S) is an interaction effect between that haplotype and stratifying variable, an OR followed by (E) is an interaction effect between that haplotype and covariate, otherwise it denotes the main effect. An OR of "–" denotes null effect (OR = 1). Haplotype frequencies are different for S = 0 and S = 1 groups. Freq: frequency.

*$P(S = 0)$,

**$P(S = 1)$.

**Table 2**

Simulation Setup for Two Stratifying Variables: OR and haplotype frequencies under two types of G-S dependence.

| | | Frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | G-$S_1$ Dependence | | | | G-$S_1$-$S_2$ Dependence | | | |
| **Hap** | **OR** | **Stra 1** | **Stra 2** | **Stra 3** | **Stra 4** | **Stra 1** | **Stra 2** | **Stra 3** | **Stra 4** |
| 01100 | – | 0.35 | 0.35 | 0.25 | 0.25 | 0.27 | 0.24 | 0.32 | 0.27 |
| 10100 (R1) | 3 | 0.01 | 0.01 | 0.005 | 0.005 | 0.01 | 0.008 | 0.005 | 0.004 |
| 11011 (R2) | 5 ($S_1$), 4 (E) | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.007 | 0.02 | 0.013 |
| 11100 | – | 0.03 | 0.03 | 0.28 | 0.28 | 0.09 | 0.125 | 0.22 | 0.29 |
| 11111 | – | 0.05 | 0.05 | 0.17 | 0.17 | 0.15 | 0.11 | 0.07 | 0.049 |
| 10011 | – | 0.55 | 0.55 | 0.275 | 0.275 | 0.47 | 0.51 | 0.365 | 0.375 |

An OR followed by ($S_1$) is an interaction effect between that haplotype and stratifying variable, an OR followed by (E) is an interaction effect between that haplotype and covariate, otherwise it denotes the main effect. An OR of "–" denotes null effect (OR = 1). Under G-$S_1$ dependence, haplotype frequencies are the same in strata 1 and 2, and also the same in 3 and 4. Under G-$S_1$-$S_2$ dependence, haplotype frequencies are different in different strata. Stra: Stratum.

**Table 3**

Characteristics distributions of the KCS data according to several variables. The percentages are based on unweighted counts.

| | | Cases (*n* = 909) | | Controls (*n* = 936) | |
|---|---|---|---|---|---|
| | | **White (*n* = 652)** | **Black (*n* = 257)** | **White (*n* = 559)** | **Black (*n* = 377)** |
| Age | <45 | 78 (12.0%) | 24 (9.3%) | 65 (11.6%) | 66 (17.5%) |
| | 45–51 | 142 (21.8%) | 74 (28.8%) | 117 (20.9%) | 93 (24.7%) |
| | 55–62 | 208 (31.9%) | 90 (35.0%) | 167 (29.9%) | 101 (26.8%) |
| | 65–74 | 158 (24.2%) | 55 (21.4%) | 152 (27.2%) | 94 (24.9%) |
| | ≥75 | 66 (10.1%) | 14 (5.4%) | 58 (10.4%) | 23 (6.1%) |
| Sex | Female | 277 (42.5%) | 91 (35.4%) | 201 (36.0%) | 191 (50.7%) |
| | Male | 375 (57.5%) | 166 (64.6%) | 358 (64.0%) | 186 (49.3%) |
| Site | Detroit | 571 (87.6%) | 191 (74.3%) | 489 (87.5%) | 309 (82.0%) |
| | Chicago | 81 (12.4%) | 66 (25.7%) | 70 (12.5%) | 68 (18.0%) |
| Smoking | Never | 247 (37.9%) | 84 (32.7%) | 232 (41.5%) | 134 (35.5%) |
| | Former | 225 (34.5%) | 71 (27.6%) | 216 (38.6%) | 120 (31.8%) |
| | Current | 180 (27.6%) | 101 (39.7%) | 121 (19.8%) | 123 (32.6%) |

*J Hum Genet*. Author manuscript; available in PMC 2017 October 20.

**Table 4**

Haplotype frequencies in the KCS data as reported by hapassoc. The five SNPs in this haplotype block are rs1041983, rs1801280, rs1799929, rs1799930, and rs1208.

| | White | | | Black | | |
|---|---|---|---|---|---|---|
| | **Overall Freq** | **Case Freq** | **Control Freq** | **Overall Freq** | **Case Freq** | **Control Freq** |
| CCCGA | – | – | 0.0004 | – | – | – |
| CCCGG | 0.0140 | 0.0177 | 0.0095 | 0.0535 | 0.0486 | 0.0567 |
| CCTAA | 0.0004 | 0.0008 | – | – | – | – |
| CCTGA | 0.0220 | 0.0208 | 0.0230 | 0.0113 | 0.0130 | 0.0101 |
| CCTGG | 0.3987 | 0.3948 | 0.4037 | 0.2387 | 0.2496 | 0.2313 |
| CTCAA | – | – | – | 0.0020 | 0.0024 | 0.0017 |
| CTCGA | 0.2266 | 0.2284 | 0.2244 | 0.1407 | 0.1391 | 0.1418 |
| CTCGG | 0.0046 | 0.0055 | 0.0037 | 0.0908 | 0.0998 | 0.0847 |
| TTCAA | 0.3076 | 0.3014 | 0.3148 | 0.2670 | 0.2685 | 0.2661 |
| TTCGA | 0.0260 | 0.0307 | 0.0206 | 0.1891 | 0.1723 | 0.2004 |
| TTCGG | – | – | – | 0.0069 | 0.0066 | 0.0071 |

Freq: frequency.

“-” indicates the specific haplotype was not found.

## Table 5

Results of analysis of the KCS data*. The five SNPs in this haplotype block are rs1041983, rs1801280, rs1799929, rs1799930, and rs1208.

| | LBLc-GXE | | LBLc-GXE-GXS | |
|---|---|---|---|---|
| | **OR (95% CS)** | **BF** | **OR (95% CS)** | **BF** |
| CCTAA | 1.55 (0.44,9.44) | 0.59 | 1.33 (0.51,5.37) | 0.40 |
| CCTGA | 1.07 (0.73,1.59) | 0.16 | 0.90 (0.53,1.41) | 0.19 |
| CCTGG | 0.99 (0.86,1.15) | 0.02 | 0.90 (0.72,1.11) | 0.12 |
| CTCAA | 1.07 (0.32,3.84) | 0.44 | 1.02 (0.37,2.82) | 0.33 |
| CTCGA | 1.12 (0.96,1.34) | 0.14 | 1.22 (0.96,1.59) | 0.35 |
| CTCGG | **1.81 (1.23,2.67)**[a] | **17.06**[b] | 1.45 (0.84,2.78) | 0.56 |
| TTCGA | 1.10 (0.86,1.42) | 0.12 | 1.10 (0.75,1.68) | 0.17 |
| TTCGG | 0.85 (0.32,1.91) | 0.36 | 0.87 (0.32,1.92) | 0.31 |
| CCCGG | 1.11 (0.78,1.61) | 0.17 | 1.31 (0.81,2.36) | 0.37 |
| former smoking | 1.04 (0.82,1.34) | 0.08 | 1.04 (0.83,1.32) | 0.07 |
| current smoking | **1.45 (1.10,1.92)**[a] | **3.72**[b] | **1.43 (1.09,1.88)**[a] | **3.28**[b] |
| CCCGG X former smoking | 0.94 (0.59,1.45) | 0.18 | 0.96 (0.62,1.46) | 0.16 |
| CCTAA X former smoking | 0.90 (0.17,3.64) | 0.49 | 0.94 (0.29,2.76) | 0.35 |
| CCTGA X former smoking | 1.18 (0.76,1.92) | 0.26 | 1.11 (0.73,1.75) | 0.18 |
| CCTGG X former smoking | 1.03 (0.87,1.22) | 0.04 | 1.02 (0.86,1.21) | 0.03 |
| CTCAA X former smoking | 0.84 (0.16,3.04) | 0.49 | 0.90 (0.27,2.43) | 0.35 |
| CTCGA X former smoking | 0.81 (0.66,1.00) | 0.54 | 0.83 (0.67,1.02) | 0.36 |
| CTCGG X former smoking | **0.61 (0.37,0.99)**[a] | 1.77 | 0.65 (0.39,1.02) | 1.20 |
| TTCGA X former smoking | 1.12 (0.83,1.52) | 0.15 | 1.07 (0.81,1.45) | 0.11 |
| TTCGG X former smoking | 0.92 (0.29,2.53) | 0.40 | 0.96 (0.37,2.26) | 0.30 |
| CCCGG X current smoking | 1.13 (0.74,1.76) | 0.21 | 1.17 (0.79,1.84) | 0.22 |
| CCTAA X current smoking | 2.50 (0.55,28.67) | 0.93 | 1.70 (0.58,10.87) | 0.54 |
| CCTGA X current smoking | 0.94 (0.55,1.53) | 0.21 | 0.91 (0.55,1.43) | 0.19 |
| CCTGG X current smoking | 1.15 (0.96,1.39) | 0.21 | 1.15 (0.96,1.38) | 0.19 |
| CTCAA X current smoking | 1.53 (0.45,8.55) | 0.58 | 1.30 (0.50,4.84) | 0.39 |
| CTCGA X current smoking | 0.95 (0.76,1.17) | 0.07 | 0.96 (0.78,1.18) | 0.06 |
| CTCGG X current smoking | **0.33 (0.18,0.59)**[a] | **>100**[b] | **0.37 (0.20,0.64)**[a] | **>100**[b] |
| TTCGA X current smoking | 0.90 (0.65,1.22) | 0.15 | 0.92 (0.68,1.23) | 0.12 |
| TTCGG X current smoking | 0.87 (0.27,2.26) | 0.40 | 0.90 (0.34,2.01) | 0.31 |
| CCTGG X male | | | **1.21 (1.03,1.43)**[a] | 0.70 |

*Adjusted for stratifying variables (age, sex, race, and site).

[a]95% CS for OR excludes 1;

[b]Bayes Factor (BF) > 2.

Interaction effects of haplotypes with stratifying variables shown only for significant effects.