Published in final edited form as:

Genet Epidemiol. 2014 January ; 38(1): 31–41. doi:10.1002/gepi.21773.

Detecting Rare Haplotype-Environment Interaction with Logistic Bayesian LASSO

Swati Biswas^{1,*,+}, Shuang Xia^{2,+}, and Shili Lin^{2,*}

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

²Department of Statistics, The Ohio State University, Columbus, OH 43210

Abstract

Two important contributors to missing heritability are believed to be rare variants and gene-environment interaction (GXE). Thus, detecting GXE where G is a rare haplotype variant (rHTV) is a pressing problem. Haplotype analysis is usually the natural second step to follow up on a genomic region that is implicated to be associated through single nucleotide variants (SNV) analysis. Further, rHTV can tag associated rare SNV and provide greater power to detect them than popular collapsing methods. Recently we proposed Logistic Bayesian LASSO (LBL) for detecting rHTV association with case-control data. LBL shrinks the unassociated (especially common) haplotypes towards zero so that an associated rHTV can be identified with greater power. Here we incorporate environmental factors and their interactions with haplotypes in LBL. As LBL is based on retrospective likelihood, this extension is not trivial. We model the joint distribution of haplotypes and covariates given the case-control status. We apply the approach (LBL-GXE) to the Michigan, Mayo, AREDS, Pennsylvania Cohort Study on Age-related Macular Degeneration (AMD). LBL-GXE detects interaction of a specific rHTV in CFH gene with smoking. To the best of our knowledge, this is the first time in the AMD literature that an interaction of smoking with a specific (rather than pooled) rHTV has been implicated. We also carry out simulations and find that LBL-GXE has reasonably good powers for detecting interactions with rHTV while keeping the type I error rates well-controlled. Thus, we conclude that LBL-GXE is a useful tool for uncovering missing heritability.

Keywords

Age-related macular degeneration; Complement Factor H gene; GXE; GWAS; LBL; MCMC; Missing Heritability; Rare variants; Regularization; Retrospective Likelihood

Co-corresponding authors 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, Fax: (972) 883-6622, swati.biswas@utdallas.edu Shili Lin, PhD, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, Tel: (614) 292-7404, Fax: (614) 292-2096, shili@stat.osu.edu.

[†]Co-first authors

WEB RESOURCES

The URLs for software package is as follows:

- LBL: <http://www.stat.osu.edu/~statgen/SOFTWARE/LBL> and <http://www.utdallas.edu/~swati.biswas>

The authors have no conflict of interest to declare.

Introduction

Genome-wide association studies (GWAS) in the past decade have led to the identification of many common SNPs that are believed to be associated with complex diseases. However, the associated genes tend to have small effects on diseases (odds ratios about 1.1–1.5) and the variations explain only a small proportion of the disease burden in the population (5–10%) [Maher et al., 2008, Manolio et al., 2009]. These revelations set off a vigorous debate and raised the question of where to find the “missing heritability” [Goldstein, 2009, Hirschhorn, 2009, Kraft, 2009, Manolio et al., 2009]. This led to the contention that the search for the genetic burden of complex diseases should not only focus on common, but also on rare, variants.

This shift in hypothesis started a race to discover rare Single Nucleotide Variants (rSNV). The race has been catalyzed with the rapid availability of the Next Generation Sequencing (NGS) technology, and many statistical methods have been developed for detecting association of rSNV with complex diseases accordingly. On the other hand, there is another type of rare variants, namely rare haplotype variants (rHTV), that are readily available from the amassed GWAS data, as rHTV can result from combinations of common SNPs. Haplotype analysis is usually the natural second step to follow up on a genomic region that has been implicated to harbor significantly associated variants from single marker analysis [Wang et al., 2012], and rHTVs are frequently found in such analyses [Liu et al., 2005, Wang et al., 2012]. Further, rHTV formed by common SNVs are more likely to tag rSNVs not genotyped in GWAS and provide greater power to detect rSNVs than popular collapsing methods [Li et al., 2010, Lin et al., 2012]. Hence, even in studies where haplotype association is not the goal per se, haplotypes provide a biologically sensible way to examine association with rSNVs to gain power. Therefore, there may be a great deal of potential in the GWAS data to explore the “Common Disease Rare Variant” hypothesis, be it for the purpose of association testing with rSNV (via rHTV) or rHTV themselves. As secondary analyses become more popular as an economical way to fully explore and utilize existing data, it is anticipated that the wealth of GWAS data will be given a second look, and it is important that powerful statistical methods for detecting rHTV are ready to meet the increasing demands.

Typically, when rHTVs are present in data, they are either not considered at all [Klein et al., 2005] or they are pooled into a single variant [Spencer et al., 2007]. The former misses the opportunity to investigate the relevance of rHTV whereas the latter can suffer from power loss, especially if the effects of the pooled rHTV are of different directions (i.e. both risk and protective haplotypes exist) [Biswas and Lin, 2012]. In recent years, newer approaches have been proposed in which individual rHTVs and their directional effects are taken into consideration [Guo and Lin, 2009, Li et al., 2010, Zhu et al., 2010, Biswas and Lin, 2012, Lin et al., 2012]. Among them is Logistic Bayesian LASSO (LBL) [Biswas and Lin, 2012, Biswas and Papachristou, 2013, Xia and Lin, 2013], a Bayesian version of penalized regression. This approach is able to downplay the effects of unassociated (especially common) haplotypes to achieve enough noise reduction so that the signals contained in the associated rare haplotypes can be more easily detected. In an application to age-related

macular degeneration (AMD), LBL was able to implicate a specific rare haplotype for the first time in the AMD literature [Biswas and Lin, 2012].

In addition to rare variants, gene-environment interaction (GXE) is believed to be another important contributor to missing heritability [Thomas, 2010]. Many statistical methods have been proposed to detect GXE where the “gene” is either taken to be common SNPs [Chatterjee and Carroll, 2005, Kraft, 2007, Mukherjee and Chatterjee, 2008, Wakefield et al., 2010] or haplotypes [Kwee et al., 2007, Chen et al., 2008, Lobach et al., 2008, Chatterjee et al., 2009, Chen et al., 2009, Hein et al., 2009]. However, many GXEs remain elusive due to the “curse of dimensionality” as the number of variables representing such interactions can grow prohibitively large [Wakefield et al, 2010].

Although there have been ample activities lately on detecting rare variants, relatively less attention has been paid to interactions of rSNV or rHTV with environmental factors. This is not surprising given that research on rare variants is a new area, and although many solutions for detecting rare variants have been proposed, there are still on-going debates as to which methods are more promising. The difficulty is magnified by bringing in environmental factors into this problem, as one would not only need to deal with infrequent events but also has to address high dimensionality. Nevertheless, as it has been discussed amply in GXE research for common variants, failure to incorporate such interacting effects can lead to suboptimal results [Pan, 2010].

This paper is aimed to fill this gap by incorporating environmental factors and their interacting effects with haplotypes into the LBL framework for case control data. LBL is based on retrospective likelihood, that is, it models the probability of haplotypes conditional on the case control status. This likelihood is more appropriate here than the commonly used prospective likelihood as haplotypes are not observed directly, and for a sample of unrelated individuals, haplotypes may not be deduced unambiguously from the observed genotype data. In such situations, inference using prospective and retrospective likelihoods are not equivalent, and the use of a prospective likelihood for a case control design may be less efficient [Carroll et al., 1995].

However, a retrospective likelihood is more complicated than its simpler prospective counterpart. In particular, it involves modeling the joint distribution of haplotypes and covariates. To facilitate this modeling, we assume independence of gene (haplotypes) and environmental covariates in the control population, which is a weaker assumption than the assumptions of gene-environment independence in the target population and the disease under study being rare, frequently adopted in literature [Chatterjee and Carroll, 2005, Kwee et al, 2007, Mukherjee and Chatterjee, 2008]. We apply this extension, LBL-GXE, to GWAS data from Michigan, Mayo, AREDS, Pennsylvania (MMAP) Cohort Study on AMD and find a significant interaction between a rare haplotype and smoking. In the AMD literature, although there have been hints of rare gene-environment interactions, due to lack of appropriate statistical methods, interactions of smoking with specific rHTVs were not studied [Spencer et al., 2007]. To study LBL-GXE thoroughly, we also carry out extensive simulations to investigate its type I error rate and power for detecting interactions under a variety of factors, including the marginal effect size of the interacting rare haplotypes, the

marginal effect of the interacting environmental factor, and the distribution of the environmental factor.

Methods

Here we describe the LBL-GXE methodology, including the likelihood, priors, estimation of posterior distributions using Markov chain Monte Carlo (MCMC) methods, and association test procedure. This closely follows the methodology developed for the original LBL [Biswas and Lin, 2012]. In particular, the retrospective likelihood used therein is adopted here but with the addition of the environmental covariates and interactions. For this extension, here we have to additionally model the distribution of covariates in the case and control populations separately. The regularization of regression coefficients is achieved through Bayesian LASSO as in the original LBL.

Retrospective Likelihood

Consider a case-control sample of size n , of which first n_1 are cases and the rest $n - n_1 = n_2$ are controls. Let $Y_i = 1/0$ denote the case/control status of the i th individual, $i = 1, \dots, n$, and $\mathbf{Y} = (Y_1, \dots, Y_n)$. Suppose L SNPs within a haplotype block are of interest. Denote the genotypes of the i th individual at these SNPs by G_i , and $\mathbf{G} = (G_1, \dots, G_n)$. Further, let Z_i denote the missing (phased) haplotype pair of the i th individual that is compatible with the observed genotype data, and $\mathbf{Z} = (Z_1, \dots, Z_n)$. Note that Z_i 's are unobservable for many individuals as phase information is usually not deductible from the genotype data, G_i . Suppose we have a vector of environmental covariates $\mathbf{E}$ and covariate value(s) of the i th individual is $\mathbf{E}_i$. Then the retrospective likelihood using the complete data is written as:

$$L_C(\Psi) = \prod_{i=1}^{n_1} P(Z_i, \mathbf{E}_i, Y_i=1, \Psi) P(\mathbf{E}_i | Y_i=1, \Psi) \prod_{i=n_1+1}^n P(Z_i, \mathbf{E}_i, Y_i=0, \Psi) P(\mathbf{E}_i | Y_i=0, \Psi), \quad (1)$$

where $\Psi = (\beta, \gamma)$ consists of regression coefficients (β) and parameters associated with haplotype frequencies (γ), which we will explicitly define in the rest of this section.

In the following, we suppress the subscript i and the parameter vector Ψ where ambiguity does not occur. Assuming independence of $\mathbf{Z}$ and $\mathbf{E}$ in the control population, we let $a_Z = P(\mathbf{Z} | Y=0)$ denote the frequency of haplotype pair $\mathbf{Z}$. On the other hand, $\mathbf{Z}$ and $\mathbf{E}$ may be correlated in the case population, therefore the haplotype pair frequency depends on $\mathbf{E}$, and thus is denoted as $b_{Z|\mathbf{E}} = P(\mathbf{Z} | \mathbf{E}, Y=1)$. We can express $b_{Z|\mathbf{E}}$ in terms of a_Z and the odds of disease for a given $\mathbf{Z}$ and $\mathbf{E}$, $\theta_{Z,\mathbf{E}} (= P(Y=1 | \mathbf{Z}, \mathbf{E}) / P(Y=0 | \mathbf{Z}, \mathbf{E}))$:

$$b_{Z|\mathbf{E}} = P(\mathbf{Z} | \mathbf{E}, Y=1) = \frac{P(Y=1 | \mathbf{Z}, \mathbf{E}) P(\mathbf{Z}, \mathbf{E})}{\sum_H P(Y=1 | H, \mathbf{E}) P(H, \mathbf{E})} = \frac{\theta_{Z,\mathbf{E}} a_Z}{\sum_H \theta_{H,\mathbf{E}} a_H}, \quad (2)$$

where H is the set of all possible haplotype pairs. Suppose there are m haplotypes and $\mathbf{f} = (f_1, f_2, \dots, f_m)$ denote their frequencies in the control population. We model a_Z for a haplotype pair $\mathbf{Z} = z_k/z_{k'}$ as follows:

$$a_Z(\gamma) = P(Z = z_k / z_{k'} | Y = 0, \gamma) = \delta_{kk'} df_k + (2 - \delta_{kk'}) (1 - d) f_k f_{k'}, \quad (3)$$

where $\delta_{kk'} = 1(0)$ if $z_k = z_{k'} (z_k \neq z_{k'})$, $\gamma = \{f, d\}$, and $d \in (-1, 1)$ is the within-population inbreeding coefficient that can be used to capture excess/reduction of homozygosity [Weir, 1996]. By modeling the frequency in this way, we do not need to make the assumption of Hardy-Weinberg Equilibrium (HWE). We treat f and d as unknown parameters and do not assign any values to them. Their posterior distributions will be estimated along with those of β in an MCMC algorithm to be described later in this section.

We model the odds of disease using a logistic regression model: $\theta_{ZE} = \exp(\mathbf{X}_{ZE}\beta)$, where the (row) design vector $\mathbf{X}_{ZE} = (1, \mathbf{X}_Z, \mathbf{X}_E, \mathbf{X}_Z\mathbf{X}_E)$ includes haplotype values, environmental covariates, and their interactions, and β consists of the corresponding regression coefficients including intercept β_0 . Specifically, for an omnibus test, $\mathbf{X}_Z = (x_1, x_2, \dots, x_{m-1})$, 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. To be concrete, we consider a categorical covariate; derivation for a continuous covariate can be similarly carried out. Let $\mathbf{X}_E$ consists of usual dummy variables and $\mathbf{X}_Z\mathbf{X}_E$ is obtained by (scalar) multiplication of $\mathbf{X}_Z$ and $\mathbf{X}_E$. Suppose there are c levels of covariate(s). Then, excluding the baseline haplotype and covariate level, there are a total of $1 + (m - 1) + (c - 1) + (m - 1)(c - 1) = mc$ regression coefficients.

It remains to model $P(E | Y)$ to fully specify (1). For this, we will use the Bayes rule. Consider

$$P(Y=1|E) = \sum_H \theta_{H,E} P(Y=0|E, H) P(H|E) = \sum_H \theta_{H,E} a_H P(Y=0|E).$$

Thus,

$$P(Y=0|E) = 1 - P(Y=1|E) = 1 - \sum_H \theta_{H,E} a_H P(Y=0|E).$$

By combining the terms involving $P(Y=0|E)$, we have

$$P(Y=0|E) = 1 / \left(1 + \sum_H \theta_{H,E} a_H \right)$$

and

$$P(Y=1|E) = \sum_H \theta_{H,E} a_H / \left(1 + \sum_H \theta_{H,E} a_H \right).$$

Now assuming fixed marginal distributions for covariate and disease status, i.e., fixed values of $P(E)$ and $P(Y=1)$, we have

$$P(\mathbf{E}|Y=1) \propto P(Y=1|\mathbf{E})P(\mathbf{E}) = \frac{\sum_H \theta_{H,E} a_H P(\mathbf{E})}{1 + \sum_H \theta_{H,E} a_H}, \quad (4)$$

and

$$P(\mathbf{E}|Y=0) \propto (Y=0|\mathbf{E})P(\mathbf{E}) = \frac{P(\mathbf{E})}{1 + \sum_H \theta_{H,E} a_H}. \quad (5)$$

Thus we can rewrite the complete data likelihood in (1) using (2), (4) and (5) as:

$$L_C(\Psi) \propto \prod_{i=1}^n \frac{a_{Z_i}}{(1 + \sum_H \theta_{H,E_i} a_H)} \prod_{i=1}^{n_1} \theta_{Z_i, E_i},$$

where $\Psi = (\beta, \gamma = \{f, d\})$. We note that the above likelihood is similar to the retrospective likelihood in Kwee et al. [2007], however, their likelihood was derived under different assumptions: G-E independence in the target population, rare disease, and HWE.

Priors

The priors and the rest of the inference follow closely from Biswas and Lin [2012]. Here we describe them briefly for the sake of completeness. We use Bayesian LASSO to regularize the β coefficients by assigning each of them a double-exponential prior distribution with hyper-parameter λ :

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

The mean of this distribution is zero. So, apriori all regression coefficients are zero on average, which serves as penalty and aids in their shrinkage. The variance of this distribution is $2/\lambda^2$ and so λ controls the degree of penalty. We let λ follow a Gamma(a, b) distribution:

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-\lambda b), 0 < \lambda < \infty.$$

We set $a = b = 20$ following Biswas and Lin [2012] as this setting leads to a realistic range for odds ratios. For f , we use Dirichlet(1, 1, ..., 1) prior consisting of m parameters for m haplotype frequencies. Finally, for d , we note that its value is dependent on f because of the constraint that $d > -f_k/(1 - f_k)$, for all k , imposed by the condition that $a_Z(\gamma) \geq 0$. As $|d| \leq 1$, we set the conditional prior distribution of d given f to be uniform on the range $(\max_k \{-f_k/(1 - f_k)\}, 1)$.

Posterior Distributions and Association Testing

The joint posterior distribution of all parameters is

$$\pi(\beta, \lambda, \mathbf{f}, d, \mathbf{Z} | \mathbf{Y}, \mathbf{G}, \mathbf{E}) \propto L_C(\Psi) \pi(\beta | \lambda) \pi(\lambda) \pi(\mathbf{f}) \pi(d | \mathbf{f}).$$

We use MCMC methods to estimate the posterior distributions of all parameters. Specifically, given the parameter estimates at the t th iteration (denoted by superscript (t)), we update all parameters at the $(t + 1)$ th MCMC iteration as follows:

Update β_j : Use Metropolis-Hastings algorithm with proposal distribution of double exponential with mean equal to the current value, $\beta_j^{(t)}$, and standard deviation $\sqrt{|\beta_j^{(t)}|}$.

Update λ : The conditional posterior distribution of λ is Gamma

$(a + mc, \sum_{j=0}^{mc-1} |\beta_j| + b)$ distribution so we use Gibbs sampler to sample λ .

Update $\mathbf{Z}_i$: Use Gibbs sampler to sample from the conditional (discrete) probability distribution of all possible haplotype pairs for the i th individual.

Update $\mathbf{f}$: Use Metropolis-Hastings algorithm with Dirichlet proposal distribution,

$Dir(a_1, a_2, \dots, a_m)$, where $a_k / \sum_{k=1}^m a_k = f_k^{(t)}$, $k = 1, \dots, m$.

Update d : Use Metropolis-Hastings algorithm with uniform proposal distribution on the range given by $d^{(t)} \pm 0.05$ subject to the constraint that $\max_k \{-f_k / (1 - f_k)\} < d^{(t+1)} < 1$.

After obtaining the posterior distributions of β_j , we test the following hypotheses:

$$H_0: |\beta_j| \leq \varepsilon = 0.1 \text{ versus } H_a: |\beta_j| > 0.1, j = 0, \dots, (mc - 1).$$

Here H_0 corresponds to no association of effect j with the disease under study. The choice of ε is justified as follows: since $\beta = 0.1$ leads to an odds ratio of 1.1, we believe that such a small empirical odds ratio can be treated as no association. We carry out this test using Bayes Factor (BF), which is the ratio of posterior odds to prior odds of H_a to H_0 . To calculate prior odds, we use the marginal prior distribution of β_j obtained by integrating out λ . Posterior odds is calculated from the estimated posterior distribution of β_j . Following Biswas and Lin [2012], if BF exceeds 2, we reject H_0 and conclude that the effect j is associated with the disease.

To analyze a sample with LBL-GXE, first we pre-process the sample using Hapassoc software [Burkett et al., 2006]. In particular, a list of haplotypes that are compatible with each person's genotype is returned from this pre-processing and is used by LBL. Then we analyze the sample using LBL-GXE with the model consisting of all main effects of haplotypes and covariates, and all interaction effects of haplotypes with covariates. The starting values of the parameters in the MCMC algorithm are set to be $\beta = 0.01$, $\lambda = 1$, and $d = 0$. The starting values of $\mathbf{f}$ are set to be the frequency estimates returned by Hapassoc, which are the maximum likelihood estimators. The algorithm is not sensitive to starting values as long as convergence of the chain is ensured for which we use trace plots and R^2 diagnostic statistic [Gelman et al., 2003].

Application to MMAP AMD data

Previous literature on AMD indicates presence of rare haplotypes and their potential interaction with smoking [Li et al., 2006, Spencer et al., 2007]. However, due to lack of statistical methods that are powerful enough to detect these effects, until recently, rare haplotypes have been only studied by pooling them. In Biswas and Lin [2012], for the first time, we studied rare haplotypes individually and found evidence for association with specific rare haplotypes in Complement Factor H (CFH) gene. The data used therein was the National Eye Institute's Age-Related Eye Disease Study (AREDS) Research Group [Age-Related Eye Disease Study Research Group, 1999]. As that dataset is too small (459 subjects) for detecting interaction effects, especially with rare haplotypes, we obtained MMAP dataset from the database of Genotypes and Phenotypes (dbGaP). This dataset consists of data from four sites — University of Michigan, Mayo Clinic, AREDS, and University of Pennsylvania. There is an overlap of some subjects between the AREDS site of MMAP with the original AREDS data used in Biswas and Lin [2012], however, the two datasets differ in terms of the available SNPs, as explained below.

Spencer et al. [2007] analyzed an 8-SNP haplotype block in the CFH gene on chromosome 1 consisting of the following SNPs (in order): rs3753394, rs529825, rs800292, rs3766404, rs1061170, rs203674, rs3753396, and rs1065489. Their sample consisted of 548 cases and 248 controls from Vanderbilt University Medical Center and Duke University Medical Center, and they fitted a standard logistic regression model to the data. They pooled the rare haplotypes with frequencies less than 5% into a single super variant. In their full regression model with interactions, they found the main effect of the pooled rare haplotypes as well as the interaction of it with smoking to be significant. Klein et al. [2005] conducted a GWAS with 96 cases and 50 controls from the AREDS data and implicated two significant SNPs in the CFH gene, namely, rs380390 and rs1329428. They then followed up with a haplotype analysis with six SNPs in the region (including the two significant SNPs) using inferred haplotypes for each individual. They ignored the haplotypes with frequencies less than 1% and found that the haplotype containing the risk allele at rs380390 conferred highest risk. In Biswas and Lin [2012], we used the AREDS data and combined five SNPs from the Spencer et al. study that were available in the AREDS data with the two significant SNPs from the Klein et al. study to form a 7-SNP haplotype block: rs3753394, rs800292, rs203674, rs3753396, rs380390, rs1329428, and rs1065489. This resulted in 311 cases and 148 controls with observed genotypes for these SNPs. Using LBL, we found four significant haplotypes including one rare (frequency = 0.01) and one borderline rare (frequency = 0.046).

Following the above general background, and in particular, Spencer et al.'s result on significant interaction of the pooled variant of rare haplotypes with smoking, we focus here on the same region of CFH gene. Of the SNPs mentioned above, five SNPs were available in the MMAP data: rs3753394, rs800292, rs3766404, rs1329428 and rs1065489. To obtain a relatively more stable model with limited sample size, we decided to drop rs3753394 as it is the least significant in the Spencer et al. study as well as in our data. Therefore, we focus on analyzing a 4-SNP region consisting of rs800292, rs3766404, rs1329428, and rs1065489. There are a total of 1543 subjects with 938 cases and 605 controls.

The four SNPs considered here are compatible with 12 possible haplotypes. We used “Ever Smoked” with two levels (Yes/No) as a covariate so there are a total of 12 main effects and 11 interaction effects apart from the intercept term in the model. We chose the baseline haplotype to be GAGA as it does not contain risk alleles reported by Spencer et al. [2007] and Klein et al. [2005] at the respective SNPs. Also, it has similar estimated frequencies for cases and controls. Further, this haplotype corresponds to the same baseline used by Biswas and Lin [2012] and Spencer et al. [2007] although we note that the labeling of alleles in datasets used therein are different from those in the MMAP dataset. The results of applying LBL are reported in Table 1. The strongest evidence for association is for the risk haplotype GAGC and the protective haplotype GAAC, which is rare among the cases (frequency < 0.05). There are four additional protective haplotypes: AAAC, AAGA, AGAC, and GGAC, with the middle two also being rare. The evidence for the risk haplotype and the two protective haplotypes AAAC and GGAC is consistent with the results of Spencer et al. [2007] wherein the authors found evidence for two protective and one risk haplotypes. They also found an interaction effect of pooled rare haplotypes and smoking with the estimated beta coefficient being negative (protective effect). In our analysis, we found the interaction $GAAA \times \text{smoking}$ to be significant where the haplotype is rare and the interaction confers a risk effect. As rare haplotypes were pooled in Spencer et al. [2007], it is possible that their pooled term consisted of effects of protective as well as risk types. Comparing our results with Biswas and Lin [2012], the results for the main effects of GAAC, GGAC, and GAGC are consistent with it.

We also compare the above fitted model with the results of a model containing main effects only, i.e., no interactions, fitted using LBL. In the latter model, the haplotypes that were found to be significant above (GAGC, GAAC, AAAC, AAGA, AGAC, and GGAC) are still significant and in the same direction. Additionally, smoking is also significant, conferring risk effect. However, GAAA is not significant in this model as well as in the model with interaction terms. In other words, GAAA confers risk on AMD only in the presence of smoking. Thus, we can see that without fitting the full model with interactions, we would have missed the effect of the rare haplotype GAAA on AMD entirely.

Using LBL, we are able to show that the odds of AMD may increase more than 3 times if a person ever smoked and has the rare haplotype GAAA. To the best of our knowledge, this is the first time that an interaction with a specific rare haplotype in the CFH gene has been implicated in the AMD literature. In Figure 1, we show a cladogram depicting the relationships between all significant effects found here. Starting with the baseline haplotype GAGA, it shows how changes in one SNP position leads to an associated haplotype at each step. The protective haplotype GAAC can be viewed as evolving from the GAGC risk haplotype by changing the third SNP from a G (risk allele detected in Klein et al.) to an A, indicating that the A nucleotide in the third SNP may play a significant protective role. Further, note that this nucleotide, together with the C nucleotide in the fourth SNP, are preserved through the other three protective haplotypes “descending” from it. Also, no one single SNP is sufficient for differentiating between risk and protective effects. This demonstrates, as is well-known, that from a biological point of view, a haplotype is more than just a combination of SNPs. Further, there may exist cis-acting effects of two mutations

in the rare protective haplotype GAAC, exemplifying the notion of increased power of a haplotype approach in detecting association in such a scenario.

Simulation Study

We carry out extensive simulations to study the properties of LBL-GXE for detecting rare haplotype $\times$ environment interaction as well as the main effects. In particular, we simulate data under three settings as listed in Table 2. Each setting has two rare causal haplotypes of frequencies 0.005 and 0.01 that we refer to as R1 and R2, respectively. There is a binary environmental covariate, E, and its interaction with R2 is denoted by R2XE. For Setting 1, we also investigate some scenarios with a common causal haplotype, C (also shown in Table 2), and its interaction with E, CXE. To investigate how the power for detecting main and interaction effects vary with the true underlying effects, i.e., odds ratios (OR), of R1, R2, C, E, R2XE, and CXE, and prevalence of E, we use the following values of these parameters and investigate many (although not all) of their combinations:

- OR of R1 (OR.R1): 1, 3
- OR of R2 (OR.R2): 1, 2, 2.5, 3, 4
- OR of C (OR.C): 1, 1.3, 1.5, 1.8
- OR of E (OR.E): 1, 1.5, 1.8, 2
- OR of R2XE (OR.R2XE): 1, 2, 3, 4, 5
- OR of CXE (OR.CXE): 1, 1.8
- prevalence of E (p_E): 0.1, 0.25, 0.5

We consider a combination of the above parameter values as a separate simulation scenario. For Setting 1, in scenarios in which C is causal (OR.C > 1), R1 is set to be non-causal so that there are at most two causal main haplotype effects, R2 and C. In the other scenarios considered, R1 and R2 are set to be the risk haplotypes. In all scenarios, OR for all main and interaction variables not listed above is set to be 1 (i.e. no effect). Note that OR of 1 for C, R2, and R2XE in several of the scenarios are used to investigate the type I error.

We generate samples of sizes 1000, 1500, or 2000 with equal number of cases and controls for each scenario under each setting. For each individual, the unphased genotypes at the five SNPs and case/control status are generated in the following manner. First we simulate the phased haplotype pair, say Z^* , using the frequencies given in Table 2 and assuming $d = 0$ (HWE). Then we generate a covariate value, say E^* , using one of the p_E parameters listed above. Next, we find the probability (p) that the individual is case or control using the following logistic regression model:

$$\log \frac{p}{1-p} = \mathbf{X}_{Z^*, E^*} \boldsymbol{\beta},$$

where $\mathbf{X}_{Z^*, E^*}$ is the design vector consisting of intercept, generated haplotypes, covariate, and their interactions, as described in the Methods section. The most frequent haplotype is

set as the baseline. For setting the value of the intercept, we use baseline prevalence of 0.1, i.e., $\beta_0 = \log(0.1/0.9)$. For other β coefficients, we use the corresponding ORs mentioned above. After generating a value for p , the individual is assigned as case with probability p . After the assignment, the phase information is removed and only genotypes are retained. This process is repeated to generate the required numbers of cases and controls.

For each simulation scenario, 500 samples are generated and analyzed by LBL-GXE. Then, we calculate the proportion of times (out of 500) each effect is inferred to be associated, that is, its BF exceeds 2. This is power or type I error depending on whether the corresponding true OR is not equal, or equal, to 1.

We first investigate scenarios in which there are at most two rare risk haplotypes, R1 and R2. Figure 2 shows the power for detecting R1, R2, E, and R2XE as functions of OR.R2 for Setting 2 (9 haplotypes in the simulation model), $n=2000$, OR.E = 1.8, OR.R2XE = 4, and $p_E = \{0.1, 0.25, 0.5\}$. As expected, we see that the power for R1 and E are not affected by OR.R2 values and the power for detecting R2 increases with increasing OR.R2. However, the power for R2XE decreases with increasing OR.R2 values, an indication that the effects of R2 (main) and R2XE (interaction) are confounded. Another interesting observation is that the power for E and R2XE decrease with decreasing p_E while the power for R1 is higher for smaller p_E . Although the latter seems counter intuitive, it is actually quite explainable: As E becomes more prevalent, there are more individuals affected due to the environmental factor or R2XE, therefore, the number of cases in the sample due to the effect of R1 is smaller, reducing the power to detect R1. On the other hand, the power for R2 is not much affected by p_E value, which can be explained by the balancing act between increased number of cases due to R2XE but decreased number of cases due to R2. From the upper right sub-figure (plotting the properties of R2), we can see that the type I error is well under control to be around 0.05 when OR.R2 = 1 (also see the top segment of Table 3). Similar patterns are seen for other sample sizes and the other two settings (shown in Supplementary Materials).

To investigate more in-depth the behavior of LBL for detecting interaction, in Figure 3, we plot for Setting 1 (6 haplotypes) the power for R2XE against varying values of OR.R2 (1, 2, 3, 4 but fixing OR.E = 1 and OR.R2XE = 4), OR.R2XE (1, 2, 3, 4, 5 but fixing OR.R2 = 1 and OR.E = 1 or 1.5), and OR.E (1, 1.5, 1.8, 2 but fixing OR.R2 = 1 and OR.R2XE = 4) with $p_E = 0.5$, and sample sizes of 1000, 1500, or 2000. We see that the power generally decreases as OR.R2 or OR.E increases. Some exceptions occur when comparing OR.R2 = 1 vs. 2 ($n = 1000$, or 2000) and OR.E = 1.8 vs. 2 ($n = 1500$). These findings are consistent with what we observed in the lower-left plot of Figure 2 except that the non-null effect of the haplotype on the power to detect interaction effect is much more pronounced when there is a smaller number of haplotypes (six vs. nine) in the setting. Also, from the two plots on the right panel, we see that the power is higher for OR.E = 1 than 1.5, again consistent with the observation on the effect of R2 on detecting interaction effect. The right two plots also indicate that the type I error rate for R2XE (when OR.R2XE = 1) is acceptable at around 0.05 (also see the middle segment of Table 3). In summary, from these figures and results in the Supplementary Materials, we may conclude that LBL has reasonably good power for detecting R2XE of OR 3 or more with sample sizes of 1500 or more. However, the power depends on the prevalence of the covariate and the true main effects of the interacting

haplotype and covariate in addition to the effect of the interaction itself. We note that these observations on the power for detecting GXE are similar to those reported in the literature earlier, e.g., Gauderman [2002].

We now turn our attention to investigating scenarios in which there are both common and rare risk haplotypes (C and R2). Similar to Figure 2, in Figure 4, we plot for Setting 1, the power for detecting C, R2, E, and CXE as functions of OR.C with $n=1500$, OR.R2 = 2.5, OR.E = 1.8, OR.CXE = 1.8, OR.R2XE = 1, and $p_E = \{0.1, 0.25, 0.5\}$. The power for detecting C is very high, increases with increasing OR.C, as expected, and does not depend on p_E value. The power for detecting CXE is also high, however, consistent with Figure 2, it decreases as OR.C increases. CXE's power also depends to some extent on the p_E value. Also, power for detecting R2 and E are not much affected by OR.C value. The type I error rate for C (when OR.C = 1) in the top-left plot are also acceptable (also see the bottom segment of Table 3). More plots from Setting 1 involving effects of C, R2, and CXE are in the Supplementary Materials.

Discussion

Rare variants and GXE are widely believed to be important contributors to the so-called “missing heritability”. Each is a difficult problem on its own making their combination highly challenging. Thus, even though some amount of research has been carried out in each of these two problems separately, only a handful of works have tried to tackle the two problems together, especially when G is rHTV. Although haplotype analysis is important in its own right due to biological significance of haplotypes, rHTVs have lately gained more attention as it is hypothesized that rHTVs formed by common SNVs are more likely to tag rSNVs not genotyped in GWAS [Li et al., 2010, Lin et al., 2012]. Here we extended LBL that has been proposed earlier for detecting rHTV association to incorporate GXE. We applied LBL-GXE to the GWAS data on AMD and found an interaction effect of an rHTV with smoking. Specifically, those who ever smoked and have this rHTV are estimated to be 3.76 times more likely to develop AMD than those who have never smoked or do not have this rHTV. Smoking is a known risk factor for AMD and it is indicated in the literature that it can interact with certain genetic variants to modify one's risk of AMD [Schmidt et al., 2006, Seddon et al., 2006, Spencer et al., 2007, Seitsonen et al., 2008]. For example, the overall effect of a variant in the LOC387715 gene on AMD is driven primarily by a strong association in smokers [Schmidt et al., 2006]. Also, Seitsonen et al. [2008] found that smoking affects the risk of AMD through interaction with the C3 genotype. Along the same line, here we found the interaction of the GAAA haplotype (from 4 SNPs in the CFH gene) and smoking to be significant even though the main effects of each of them are not significant. The effect of GAAA, an rHTV, has not been studied before. The closest result to this finding is in Spencer et al. [2007], and the results are somewhat consistent, although their interaction term had an opposite (protective) effect compared to ours (risk), most likely due to their genetic variant being pooled rHTVs. As the directions of effects of individual rHTVs pooled together by Spencer et al. is not known, comparison of effects in the two studies seems difficult. Thus, our finding needs to be replicated, preferably with a larger sample size, using a statistical method that is powerful enough for handling rHTVs.

Our simulation study showed that LBL-GXE has reasonably good power for detecting interactions with rHTV. We found that it is easier to detect an interaction effect if one or both of these main effects are small, i.e., when the haplotype and/or covariate affect the risk mostly through interaction. This is not surprising since the affecteds in samples simulated under those scenarios were due to the existence of both risk factors. This was the case in our AMD data analysis, and it is likely that the weak main effects of GAAA and smoking made it easier to detect their interaction with a sample size of about 1500 only. As supported by our simulations, LBL-GXE has typically high power to detect interaction effects of OR 3 – 4 with such sample sizes when the corresponding main effects are small. Simulations also showed, as expected, that it is easier to detect interaction if the covariate is more prevalent. Again this is illustrated in our data analysis as the covariate smoking was about 50% prevalent in the data.

With regard to setting values of various parameters such as a , b , ε , and tuning parameters in the proposal distributions of f and d , we used the same values as in Biswas and Lin [2012]. We see in our simulations that the results are well-calibrated in the sense that the type I errors are well-controlled and the convergence is good (checked by monitoring the trace plots and the R^2 diagnostic statistic [Gelman et al., 2003]) with these parameters even after incorporating the interaction terms in LBL. The robustness of these parameter values have been explored in Biswas and Lin [2012] by varying the values of these parameters and finding that the results are generally insensitive to the specified parameter values. LBL retains good performance with these parameter specifications in the simulated dataset of the Genetic Analysis Workshop 18 also [Biswas and Papachristou, 2013]. For updating the beta parameters, we also explored a proposal distribution of multivariate normal with a constant variance for all beta parameters in all iterations. This allows all beta parameters to be updated together, however, we found that the convergence of the chain was slower with that approach. Nonetheless, after ensuring convergence with longer runs, the results from both approaches were similar, including powers, type I errors, and results of AMD analysis. This result further adds to the detailed sensitivity analysis on LBL conducted in Biswas and Lin [2012].

The computational intensity of LBL has increased with the inclusion of covariates and interactions. The computation time depends on the number of haplotypes, m (rather than the number of SNPs) and the number of covariate levels, c . As the model in LBL-GXE includes all possible interaction terms of each haplotype with each covariate level, the total computation time depends on mc . For our AMD data analysis with 12 haplotypes and 2 levels of covariate, LBL-GXE took 25.85 minutes to run on a 3.60 GHz Xeon processor under Linux operating system with 15.55 GB RAM.

In summary, we have proposed a novel and promising method for detecting interactions with rHTVs. This approach can be applied not only to data currently being generated through next-generation sequencing technology but also to GWAS data. This makes LBL one of only few tools available to explore rare variants from the hugely amassed GWAS data, which were initially thought to be of little use for the purpose of testing “Common Disease Rare Variant” hypothesis. Given the amount of time, effort, and money that went into generating the GWAS data, it behooves us to exploit its full richness in addition to the

newly generated sequencing data. Our application on AMD here and previously has amply illustrated that this is, in fact, possible with powerful enough new statistical approaches such as LBL. To enrich this toolkit further, we are working on extending LBL to handle other data types, for example, family data. In this process, it is also important to compare the performance of LBL with other recently proposed methods for rHTVs (with or without interactions), as their software become more readily available, to investigate relative merits of each approach, and this is also one of our future plans.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the grant R03CA171011-01 from NCI, the grant DMS-1208968 from NSF and by allocations of computing times from the Ohio Supercomputer Center and the Texas Advanced Computing Center at the University of Texas at Austin. The MMAP dataset used for the analyses described in this manuscript was obtained from the NEI Study of Age-Related Macular Degeneration (NEI-AMD) Database found at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000182.v2.p1 through dbGaP accession number phs000182.v2.p1. Funding support for NEI-AMD was provided by the National Eye Institute. The authors would like to thank NEI-AMD participants and the NEI-AMD Research Group for their valuable contribution to this research. The authors are also thankful to the two anonymous reviewers for their helpful comments and suggestions, which led to improvement of the manuscript.

References

- Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study. AREDS): Design Implications. *Control Clin Trials*. 1999; 20:573–600. [PubMed: 10588299]
- 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, Papachristou C. Evaluation of Logistic Bayesian LASSO for Identifying Association with Rare Haplotypes. *BMC Proceedings*. 2013 (in press).
- Burkett K, Graham J, McNeney B. hapassoc: Software for likelihood inference of trait associations with SNP haplotypes and other attributes. *J Stat Softw*. 2006; 16:1–19.
- Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *J Am Statist Assoc*. 1995; 90:157–169.
- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.
- Chatterjee N, Chen YH, Luo S, Carroll RJ. Analysis of Case-Control Association Studies: SNPs, Imputation and Haplotypes. *Stat Sci*. 2009; 24:489–502. [PubMed: 20543902]
- Chen J, Chatterjee N. Haplotype-based association in cohort and nested case-control studies. *Biometrics*. 2006; 62:28–35. [PubMed: 16542226]
- 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]
- Chen YH, Chatterjee N, Carroll RJ. Shrinkage Estimators for Robust and Efficient Inference in Haplotype-Based Case-Control Studies. *J Am Stat Assoc*. 2009; 104:220–233. [PubMed: 19430598]
- Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*. 2002; 21:35–50. [PubMed: 11782049]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. 2nd edition. Boca Raton: Chapman and Hall/CRC; 2003.

- Goldstein DB. Common genetic variation and human traits. *New Eng J Med*. 2009; 360:1696–1698. [PubMed: 19369660]
- Guo W, Lin S. Generalized Linear Modeling with Regularization for Detecting Common Disease Rare Haplotype Association. *Genet Epidemiol*. 2009; 33:308–316. [PubMed: 19025789]
- Hein R, Beckmann L, Chang-Claude J. Comparison of Different Haplotype-Based Association Methods for Gene-Environment. GxE. Interactions in Case-Control Studies when Haplotype-Phase Is Ambiguous. *Hum Hered*. 2009; 68:252–267. [PubMed: 19622892]
- Hirschhorn JN. Genomewide association studies - illuminating biologic pathways. *New Eng J Med*. 2009; 360:1699–1701. [PubMed: 19369661]
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005; 308:385–389. [PubMed: 15761122]
- Kraft P, Yen Y-C, Stram DO, Morrison CJ, Gauderman WJ. Exploiting Gene-Environment Interaction to Detect Genetic Associations. *Hum Hered*. 2007; 63:111–119. [PubMed: 17283440]
- Kraft P, Hunter DJ. Genetic risk prediction - are we there yet? *New Eng J Med*. 2009; 360:1701–1703. [PubMed: 19369656]
- 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]
- Li M, Atmaca-Sonmez P, Othman M, Branham KE, Khanna R, Wade MS, Li Y, Liang L, Zarepari S, Swaroop A, Abecasis GR. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nature Genetics*. 2006; 38:10491054.
- Li Y, Byrnes AE, Li M. To Identify Associations with Rare Variants, Just WHaIT: Weighted Haplotype and Imputation-Based Tests. *Am J Hum Genet*. 2010; 87:728–735. [PubMed: 21055717]
- Lin W-Y, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. Haplotype-based methods for detecting uncommon causal variants with common SNPs. *Genetic Epidemiology*. 2012; 36:572–582. [PubMed: 22706849]
- Liu P-Y, Zhang Y-Y, Lu Y, Long J-R, Shen H, Zhao L-J, Xu F-H, Xiao P, Xiong D-H, Liu Y-J, Recker RR, Deng H-W. A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. *J Med Genet*. 2005; 42:221–227. [PubMed: 15744035]
- Lobach I, Raymond J, Carroll RJ, Spinka C, Gail MH, Chatterjee N. Haplotype-Based Regression Analysis and Inference of Case-Control Studies with Unphased Genotypes and Measurement Errors in Environmental Exposures. *Biometrics*. 2008; 64:673–684. [PubMed: 18047538]
- Maher B. Personal genomes: the case of the missing heritability. *Nature*. 2008; 456:18–21. [PubMed: 18987709]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
- 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]
- Pan W. Statistical Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. *Hum Hered*. 2010; 69:131–142. [PubMed: 19996610]
- Schmidt S, Hauser MA, Scott WK, Postel EA, Agarwal A, Gallins P, Wong F, Chen YS, Spencer K, Schnetz-Boutaud N, Haines JL, Pericak-Vance MA. Cigarette smoking strongly modifies the association of LOC387715 and age-related macular degeneration. *Am J Hum Genet*. 2006; 78:852–864. [PubMed: 16642439]

- Seddon JM, George S, Rosner B, Klein ML. CFH gene variant, Y402H, and smoking, body mass index, environmental associations with advanced age-related macular degeneration. *Hum Hered*. 2006; 61:157–165. [PubMed: 16816528]
- Seitsonen SP, Onkamo P, Peng G, Xiong M, Tommila PV, Ranta PH, Holopainen JM, Moilanen JA, Palosaari T, Kaarniranta K, Meri S, Immonen IR, Jarvela IE. Multi-factor effects and evidence of potential interaction between complement factor H Y402H and LOC387715 A69S in age-related macular degeneration. *PLoS One*. 2008; 3:e3833. [PubMed: 19048105]
- Spencer KL, Hauser MA, Olson LM, Schnetz-Boutaud N, Scott WK, Schmidt S, Gallins P, Agarwal A, Postel EA, Pericak-Vance MA, Haines JL. Haplotypes spanning the complement factor H gene are protective against age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2007; 48:4277–4283.
- Thomas D. Gene-Environment-Wide Association Studies: Emerging Approaches. *Nat Rev Genet*. 2010; 11:259–272. [PubMed: 20212493]
- Wakefield J, De Vocht F, Hung RJ. Bayesian Mixture Modeling of Gene-Environment and Gene-Gene Interactions. *Genet Epidemiol*. 2010; 34:16–25. [PubMed: 19492346]
- Wang L, Liu R, Li D, Lin S, Fang X, Ribick M, Backer G, Kain M, Rammohan K, Zheng P, Liu Y. A hypermorphic SP1-binding CD24 Variant Associates with Risk and Progression of Multiple Sclerosis. *Am J Transl Res*. 2012; 4:347–356. [PubMed: 22937211]
- Weir, BS. Genetic data analysis II. Sunderland, MA: Sinauer Associates; 1996.
- Xia S, Lin S. Detecting longitudinal effects of haplotypes and smoking on hypertension using B-Splines and Logistic Bayesian Lasso. *BMC Proceedings*. 2013 (in press).
- Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol*. 2010; 34:171–187. [PubMed: 19847924]

Figure 1.

Cladogram showing changes in various SNP positions (shaded and indicated next to arrow) of baseline GAGA lead to significant haplotypes. The haplotypes in red font are of risk type while the ones in green are of protective type. Freq: Overall Frequency.

Figure 2.

Power for detecting the effects of R1, R2, and E, and the interaction effect of R2XE as functions of OR.R2 for Setting 2 in Table 2 with $n = 2000$, OR.R1 = 3, OR.E = 1.8, OR.R2XE = 4, and all other ORs = 1. Note that in the top right figure, when OR.R2 = 1 (null scenario), the power is actually type I error (the actual numbers are listed in the top segment of Table 3).

Figure 3.

Power for detecting the interaction effect of R2xE as function of OR.R2 (by fixing OR.E = 1 and OR.R2XE = 4), OR.E (by fixing OR.R2 = 1 and OR.R2XE = 4), and OR.R2XE (by fixing OR.R2 = 1 and OR.E = 1 or 1.5) for Setting 1 in Table 2. $p_E = 0.5$, OR.R1 = 3, and all other ORs = 1. Note that in the right two figures, when OR.R2XE = 1 (null scenario), the power is actually type I error (the actual numbers are listed in the middle segment of Table 3).

Figure 4.

Power for detecting the effects of C, R2, and E, and the interaction effect of CXE as functions of OR.C for Setting 1 in Table 2 with $n = 1500$, $OR.R2 = 2.5$, $OR.E = 1.8$, $OR.CXE = 1.8$, and all other ORs = 1. Note that in the top left figure, when $OR.C = 1$ (null scenario), the power is actually type I error (the actual numbers are listed in the bottom segment of Table 3).

Table 1

Analysis of AMD data. The frequency estimates are obtained using Hapassoc [Burkett et al., 2006].

Type	Effect	Overall Freq	Case Freq	Control Freq	OR	BF
Hap	AAAA	0.001	0.000	0.001	0.91	0.70
	AAAC	0.142	0.106	0.197	0.60	26.65*
	AAGA	0.007	0.003	0.013	0.35	2.57*
	AAGC	0.006	0.005	0.007	0.87	0.42
	AGAC	0.008	0.003	0.014	0.32	4.82*
	AGGA	0.0003	0.000	0.001	0.54	0.83
	AGGC	0.0002	0.0004	0.000	0.98	0.63
	GAAA	0.003	0.006	0.000	1.73	0.80
	GAAC	0.044	0.027	0.070	0.44	43.24*
	GAGA	0.164	0.160	0.170	NA	NA
Env	GAGC	0.529	0.616	0.395	1.77	> 100*
	GGAC	0.096	0.073	0.133	0.59	15.93*
	smoke	0.515	0.462	0.548	1.40	0.84
Hap × Env	AAAA × smoke				1.03	0.72
	AAAC × smoke				1.02	0.12
	AAGA × smoke				0.76	0.61
	AAGC × smoke				0.76	0.51
	AGAC × smoke				0.69	0.62
	AGGA × smoke				0.73	0.74
	AGGC × smoke				1.36	0.74
	GAAA × smoke				3.76	3.29*
	GAAC × smoke				1.05	0.23
	GAGC × smoke				0.98	0.07
	GGAC × smoke				1.12	0.18

Hap: Haplotype; Env: Environmental covariate; Freq: Frequency;

* : BF > 2

Table 2

Three simulation settings with varying number of haplotypes. Odds ratios for these haplotypes, a binary covariate, and their interactions are provided in the text.

Setting 1			Setting 2			Setting 3		
Hap	Freq	Hap	Freq	Hap	Freq	Hap	Freq	Freq
01100	0.300	01010	0.060	00111	0.070			
10100 (R1)	0.005	01100	0.250	01000	0.020			
11011 (R2)	0.010	10000	0.080	01011	0.050			
11100	0.155	10100 (R1)	0.005	01101	0.060			
11111 (C)	0.110	11011 (R2)	0.010	01110	0.140			
10011	0.420	11100	0.090	10010	0.080			
		11101	0.085	10100 (R1)	0.005			
		11111	0.100	11011 (R2)	0.010			
		10011	0.320	11101	0.090			
				11110	0.130			
				11111	0.100			
				10001	0.245			

Table 3

Type I error rates corresponding to various null effects in the simulation study. Setting corresponds to the settings described in Table 2. These type I error rates are also plotted in Figures 2, 3, and 4.

Setting	Null Effect	P_E	Sample Size	Type I Error
2 (Figure 2)	R2	0.1	2000	0.016
		0.25	2000	0.022
		0.5	2000	0.046
1 (Figure 3)	R2XE ^a	0.5	1000	0.026
			1500	0.022
			2000	0.016
	R2XE ^b	0.5	1000	0.038
			1500	0.024
			2000	0.012
1 (Figure 4)	C	0.1	1500	0.004
		0.25	1500	0.012
		0.5	1500	0.022

^aThere is no effect of the environmental covariate (OR.E=1).

^bThere is an effect of the environmental covariate (OR.E=1.5).