Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

An Improved Version of Logistic Bayesian LASSO for Detecting Rare Haplotype-Environment Interactions with Application to Lung Cancer

Yuan Zhang and Swati Biswas

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA.

ABSTRACT: The importance of haplotype association and gene–environment interactions (GxE) in the context of rare variants has been underlined in voluminous literature. Recently, a software based on logistic Bayesian LASSO (LBL) was proposed for detecting GxE, where G is a rare (or common) haplotype variant (rHTV)–it is called LBL-GxE. However, it required relatively long computation time and could handle only one environmental covariate with two levels. Here we propose an improved version of LBL-GxE, which is not only computationally faster but can also handle multiple covariates, each with multiple levels. We also discuss details of the software, including input, output, and some options. We apply LBL-GxE to a lung cancer dataset and find a rare haplotype with protective effect for current smokers. Our results indicate that LBL-GxE, especially with the improvements proposed here, is a useful and computationally viable tool for investigating rare haplotype interactions.

KEYWORDS: logistic Bayesian LASSO, rare variants, rHTV, GxE, GWAS, MCMC, retrospective likelihood

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Zhang and Biswas. An Improved Version of Logistic Bayesian LASSO for Detecting Rare Haplotype-Environment Interactions with Application to Lung Cancer. *Cancer Informatics* 2015:14(S2) 11–16 doi: 10.4137/CIN.S17290.

RECEIVED: November 18, 2014. **RESUBMITTED:** December 23, 2014. **ACCEPTED FOR PUBLICATION:** December 25, 2014.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

TYPE: Review

FUNDING: 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 authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: swati.biswas@utdallas.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Genomewide-association studies (GWAS) have been successful in identifying common variants associated with complex human diseases in the past decade; however, the variants found explain only a small fraction of the overall genetic contribution to disease.[1,2] Inevitably, this shortcoming provoked a heated debate over the issue of missing heritability.[3–5] It is now strongly believed that one underlying cause of missing heritability is rare variants, which were not captured by GWAS but now can be genotyped using next-generation sequencing (NGS) technologies. In this regard, rare haplotype variants (rHTVs) are a rich source of rare variants. In fact, they are available not only from NGS data but also from GWAS data accumulated so far as rHTV can result from combinations of common single nucleotide polymorphisms (SNPs). Thus, there is a great deal of wealth that awaits mining from the GWAS data to explore the common disease rare variant hypothesis. Also, haplotypes are biologically relevant and methods based on them may be more powerful than SNP-based procedures, especially when there is a group of alleles operating in concert.[6,7] Thus, methods and associated software that can handle the challenging problem of rHTV association are very much needed.

In addition to rare variants, another potential cause of "missing heritability" is identified to be gene–environment interaction (GxE), and thus currently is a subject of vigorous research. Failure to properly account for GxE has resulted in several investigations missing the opportunity to detect variants that work only in the presence of certain environmental covariates.[8]

Thus, it is clear that there is a great need for developing methods for detecting GxE, where G is rHTV. Aiming to fill this gap, Biswas et al.[9] designed a software, LBL-GxE, based

on logistic Bayesian LASSO (LBL).[10] Although it performs well, it is relatively computationally intensive and can only handle single environmental covariate with two levels. Therefore, in this article, we propose an improved version of LBL-GxE, which is computationally efficient and can also handle multiple environmental covariates, each with multiple levels. We also describe the software and apply it to a lung cancer dataset.

Method

First, we briefly review the original version of LBL-GxE, in particular, the likelihood used therein. Next we discuss how the likelihood computation is modified in the proposed version to achieve computational savings. Then we describe the software, including its inputs and outputs.

The original version of LBL-GxE. Suppose we have a case–control sample of size n consisting of n_1 cases and n_2 controls. Let $Y_i = 1/0$ denote the case/control status of the ith individual, $i = 1, \ldots, n$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Let G_i denote the genotype of the ith individual and $\boldsymbol{G} = (G_1, \ldots, G_n)$. As haplotype pair of a person may not be deduced unambiguously from genotypes, we further let Z_i denote the missing (phased) haplotype pair of the ith individual and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$. Next we denote the vector of environmental covariates of individual i by $\boldsymbol{E}_i$. LBL-GxE is based on a retrospective likelihood as follows:

$$\begin{aligned} L_c(\Psi) &= \prod_{i=1}^{n_1} P(Z_i, \boldsymbol{E}_i \mid Y_i = 1, \Psi) \prod_{i=n_1+1}^{n} P(Z_i, \boldsymbol{E}_i \mid Y_i = 0, \Psi) \\ &= \prod_{i=1}^{n_1} P(Z_i \mid \boldsymbol{E}_i, Y_i = 1, \Psi) P(\boldsymbol{E}_i, Y_i = 1, \Psi) \\ &\quad \prod_{i=n_1+1}^{n} P(Z_i \mid \boldsymbol{E}_i, Y_i = 0, \Psi) P(\boldsymbol{E}_i \mid Y_i = 0, \Psi), \end{aligned}$$

where Ψ is a vector consisting of regression coefficients and parameters associated with haplotype frequencies. c in the subscript of the likelihood L refers to "complete" as this is a likelihood of the complete data, which includes the missing data $\boldsymbol{Z}$, making $\boldsymbol{G}$ redundant.

The individual terms in the above likelihood can be expressed in terms of frequencies of each haplotype pair in the control population, $P(Z|\boldsymbol{E}, Y = 0, \Psi)$, and the odds of disease for given haplotype pair and covariates, $P(Y = 1|Z, \boldsymbol{E}, \Psi)/P(Y = 0|Z, \boldsymbol{E}, \Psi)$. Further, the former can be written in terms of frequencies of haplotypes $\boldsymbol{f} = \{f_1, \ldots, f_m\}$ (assuming m possible haplotypes in the population) and within-population inbreeding coefficient d, which captures excess/reduction of homozygosity. By modeling haplotype pair frequencies using d, there is no need to make the assumption of Hardy–Weinberg equilibrium. The odds of disease is modeled using logistic regression whose regression coefficients ($\boldsymbol{\beta}$) capture the effects of haplotypes, covariate levels, and their interactions.

Next, prior distributions are assigned to the parameters ($\boldsymbol{\beta}$, $\boldsymbol{f}$, d). Each β coefficient is penalized through Bayesian LASSO by assigning a double exponential prior centered at 0 with hyper-parameter λ. We then let λ follow Gamma($a = 20$, $b = 20$) distribution. For $\boldsymbol{f}$, we use Dirichlet (1, …, 1) prior consisting of a total of m 1's for the m haplotypes. For d, the prior is Uniform $(\max_k\{-f_k/(1-f_k)\}, 1)$.

The posterior distributions are estimated using Markov chain Monte Carlo (MCMC) methods. The main goal of the method is to carry out association test rather than effect estimation (although posterior means and credible sets of all parameters are provided by the method). The inference for association is carried out using Bayes factor (BF). If BF for a certain effect exceeds 2, that effect is considered as significant. More details about the likelihood, priors, posterior estimation, and inference can be found in Biswas and Lin[9] and Biswas et al.[10]

From a computational point of view, we note that as Z_i's are usually unobservable, they are updated using Gibbs sampler in every MCMC iteration for all persons whose haplotypes cannot be deduced unambiguously from their genotypes, and this is computationally demanding. Further, even though, in principle, the method can handle multiple covariates, each with multiple levels, the software could only handle one binary covariate, thus limiting its practical utility.

The improved version of LBL-GxE. To reduce the computation time, we get around the need for updating Z_i in every iteration. Following Kwee et al.[11], we consider the observed (rather than complete) data likelihood and sum over all haplotype pairs that are compatible with the observed genotypes. For ith individual, suppose there are $S(G_i)$ haplotype pairs that are compatible with the observed genotype G_i; ie, let $S(G_i)$ denote the set of all possible haplotype pairs for G_i. Further, let Z_{ir} denote the rth component of $S(G_i)$ or rth compatible haplotype for individual i. Then the retrospective observed data likelihood is written as

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

As all compatible Z_{ir}s are summed over in the above likelihood, there is no need to update Z_i any more in every iteration, which leads to savings in computation time. The rest of the modeling remains the same as in the original version. The equivalence of the two versions is justified as follows. The original version of LBL-GxE accounts for all possible haplotype pairs of each person by summing over them in the MCMC algorithm (through update of missing Z_i at every iteration), while the improved version achieves the same goal

by summing over them in the likelihood itself (through sum over $S(G_i)$).

In addition, in the improved version of the software, we enable handling of multiple covariates, each with as many levels as needed.

Software description. Both original and updated versions of LBL-GxE (Versions 1.0 and 1.1) can be found at http://www.utdallas.edu/~swati.biswas/. LBL-GxE is an R package, which calls a C program internally for carrying out the MCMC algorithm. Here we discuss how to run LBL-GxE. The first required input is "dat", which is a data frame consisting of non-SNP and SNP data. The non-SNP data has affection status in the first column and environmental covariates, if any, in other columns. The SNP data can be in the allelic format if the "allelic" argument (optional) is set to be TRUE (default) or in the genotypic format if this argument is FALSE. Any missing allelic data can be coded as NA or "", and missing genotypic data should be coded as "" if both alleles are missing or with one allele name if only one allele is available, eg, "A" if the other allele is missing. The second required input is "numSNPs", the number of SNPs in the haplotype region under study.

The software package comes with an example data frame LBL.ex, which has affection status *affected* in the first column, a two-level covariate *smoke* in the second column, and five allelic formatted SNPs in the rest of the columns. The first five rows of LBL.ex are as follows:

AFFECTED	SMOKE	M1.1	M1.2	M2.1	M2.2	M3.1	M3.2	M4.1	M4.2	M5.1	M5.2
1	1	1	1	0	0	0	0	1	1	1	1
1	0	1	1	1	0	1	0	1	1	1	1
1	1	1	1	0	0	0	0	1	1	1	1
1	0	1	1	1	1	1	1	0	0	0	0
1	0	1	0	0	1	0	1	1	0	1	0

Next, to analyze the input data (eg, LBL.ex) using LBL-GxE, the R command is

>LBL(LBL.ex, numSNPs = 5)

There are some other optional arguments in the function LBL. Following is a brief description of some of these arguments:

- maxMissingGenos: maximum number of single-locus genotypes with missing data to be allowed for each subject. Default is 1.
- haplo.baseline: the haplotype to be used for baseline coding. Default is the most frequent haplotype.
- cov.baseline: the baseline to be used for each covariate. Default is a vector of zeros (ie, the covariate category labeled as 0) whose length is the number of covariates.
- interaction: an indicator of whether or not to model GxEs. Default is TRUE.
- seed: the seed to be used for the MCMC sampling scheme. Default is NULL, ie, system generates a seed automatically, which may change each time the code is run.
- burn.in: the burn-in period of the MCMC sampling scheme. Default is 20,000.
- num.it: the total number of MCMC iterations, including the burn-in iterations. Default is 50,000.

LBL-GxE internally calls pre.hapassoc function from the R package Hapassoc[12] to pre-process the input dataset and thereby borrows directly some of the above-mentioned arguments and data formats from pre.hapassoc, such as dat, numSNPs, maxMissingGenos, haplo.baseline, and allelic. Users do not need to run pre.hapassoc separately as a call to it is built into the software.

The outputs from LBL-GxE are the following:

- BF: Bayes factors for all regression coefficients.
- OR: estimated odds ratios (this is $\exp(\hat{\beta})$).
- CI.OR: 95% credible sets for the ORs.
- freq: posterior means of the haplotype frequencies.
- CI.freq: 95% credible sets for the haplotype frequencies.
- CI.lambda: 95% credible set for λ, which is a hyperparameter of the prior distribution of regression coefficients.
- CI.D: 95% credible set for d, which is within-population inbreeding coefficient and is used to model Hardy–Weinberg disequilibrium, if present.

Usually BF and OR are of most interest, while CI.lambda and CI.D are not of direct interest.

Results

Analysis of lung cancer data. We use the updated version of LBL-GxE to analyze lung cancer data downloaded from the National Institute of Health's database of Genotypes and Phenotypes (dbGaP). These data were collected in the Environment and Genetics in Lung cancer Etiology (EAGLE) study and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. There are a total 2728 cases and 2821 controls. As smoking is an established risk factor for lung cancer, we use it as an environmental covariate in our model. It has three levels: never smoker, former smoker, and current smoker.

Rotunno et al.[13] conducted a haplotype analysis with eight SNPs (rs2854455, rs3766934, rs2292566, rs2260863, rs2234922, rs34143170, rs2292568, rs1051741) in gene EPHX1 on Chromosome 1. They found one protective and one risk haplotypes to be significantly associated with lung cancer. They did not study interaction effects of haplotypes with smoking. Some other authors have investigated the SNP rs2234922 but have found conflicting evidence against the direction of the effect.[14–20] Few papers have also explored interaction of rs2234922 with smoking but found it to be non-significant.[20,21] One paper found interaction of this SNP with smoking to increase risk with smoking

modeled as non-smoker, light smoker, and heavy smoker.[22] These results, combined together, seem to indicate that there may be multiple causal SNPs in cis formation, which may be possibly interacting with smoking to affect the risk of lung cancer.

Of the eight SNPs used in the Rotunno et al study, only two are present in dbGap lung cancer data: rs2234922 and rs1051741, and they are compatible with four possible haplotypes. We first combined the second and third levels of smoking-status to form a combined "smoker" level. We used the most frequent haplotype, AC, to be the baseline. LBL-GxE reveals no significant main or interaction effect other than a highly significant main effect of smoking.

Next, we used the original three-level smoking-status variable and redid the analysis. We used the same baseline haplotype (AC). The results are reported in Table 1. As expected, smoking is a highly significant risk factor for both current and former smokers. However, for current smokers, the interaction effect with AT, a rare haplotype, is also found to be significant (BF >2). The carriers of AT haplotype have about four times lower risk of lung cancer than non-carriers among current smokers.

Finally, we also fit a model with two covariates–smoking (three levels) and sex (two levels). The results are similar to as reported in Table 1. The main effect of sex or its interaction effects with haplotypes are not significant.

Performance comparison. Here we illustrate the equivalence of the two versions in terms of accuracy of estimates and inference. For this, we consider three settings with 6, 9, and 12 haplotypes as used in Biswas et al.[9] and as shown in Table 2. We generate 100 samples under each setting, and analyze each sample by both versions. Then we calculate difference in regression estimates ($\hat{\beta}$) from the two versions, and calculate mean and standard deviation of the differences over 100 replicates for each setting. As hypothesis test is the main goal of this method, we also calculate the percentage of replicates in which each regression coefficient is found to be significant (BF >2) by each version. Then we report the difference in this percentage for the two versions for each setting. This essentially estimates the difference in power (if there is true association) or type I error rates (if there is no association) for each effect. The results are given in Table 2. Note that because of random variability associated with MCMC updates, different runs of the same sample by even the same version will give slightly different results. It is clear that the improved version has practically the same accuracy, power, and type I error rate as the original version. Also, we see that this result holds irrespective of the number of haplotypes.

Computation time comparison. We carried out all computations reported in this article on a 3.60 GHz Xeon processor under Linux operating system with 15.55 GB RAM. In the above lung cancer analysis with the two-level smoking-status variable, the updated version takes 218 seconds, while the original version takes as long as 758 seconds. The computation time is actually directly related to the number of haplotypes. As the lung cancer data have only four haplotypes, we also compared computation time of the original and the improved versions on simulated datasets with a larger number of haplotypes.

Specifically, we consider the same three settings with 6, 9, and 12 haplotypes used above and earlier in Biswas et al.[9] Three datasets are simulated accordingly, each having a simulated two-level covariate (as the original version cannot handle more than two levels). The fitted models include main effects of haplotypes, covariates, and their interactions.

A comparison of computation times is shown in Table 3. It is clear that the improved version provides substantial savings of time, and the savings increase as the number of haplotypes increases.

Table 1. Analysis of lung cancer data. The haplotype frequency estimates are obtained using Hapassoc.

TYPE	OVERALL FREQ	CASE FREQ	CONTROL FREQ	OR	BF
AT	0.0024	0.0026	0.0022	1.17	0.48
GC	0.0943	0.0938	0.0948	0.91	0.12
GT	0.0960	0.0899	0.1019	0.95	0.10
Former smoker	0.4300	0.4600	0.4000	3.53	>100*
Current smoker	0.4000	0.4600	0.3500	4.18	>100*
AT X former smoker	–	–	–	1.41	0.61
GC X former smoker	–	–	–	1.14	0.17
GT X former smoker	–	–	–	0.88	0.18
AT X current smoker	–	–	–	0.25	3.28*
GC X current smoker	–	–	–	1.02	0.11
GT X current smoker	–	–	–	0.93	0.13

Note: *BF >2.
Abbreviation: Freq, frequency.

Discussion

Rare variants and GxE have been heralded as keys to solving the pressing problem of the so-called "missing heritability". Although some methods have been proposed for dealing with these two problems when G is rHTV, computation time is a usual limitation of haplotype analysis, especially when interaction terms are fitted. So, the work in this paper of reducing the computing time is important from a practical point of view. When applied to the lung cancer data with the two-level smoking-status variable, the improved version of LBL-GxE saves 71% of the time used by the original version. The savings are even more as the number of haplotypes increases. Especially, when applied to a simulated dataset with 2000 samples and 12 haplotypes, the updated version saves as much as 85% of the time cost by the original version.

Table 2. Comparison of performance between the original (Version 1.0) and the improved (Version 1.1) versions of LBL-GxE. Mean and SD are mean and standard deviation (over 100 replicates) of the difference in regression estimates ($\hat{\beta}$) from the two versions. %(BF >2) is the difference in powers (for effects with OR >1) or type I error rates (for effects with OR = 1); it is the difference in the percentages of replicates in which each regression coefficient is found to be significant (BF >2) for the two versions.

SETTING 1					SETTING 2					SETTING 3				
EFFECT	OR	MEAN	SD	%(BF >2)	EFFECT	OR	MEAN	SD	%(BF >2)	EFFECT	OR	MEAN	SD	%(BF >2)
h1	1	0.00	0.05	0.01	h1	1	0.00	0.05	0.00	h1	1	0.00	0.05	0.00
h2	3	0.01	0.25	0.00	h2	1	0.00	0.04	0.00	h2	1	0.00	0.06	0.01
h3	1	0.00	0.14	0.01	h3	3	0.00	0.05	0.00	h3	1	0.00	0.06	−0.01
h4	1	0.00	0.03	0.00	h4	1	0.01	0.14	−0.01	h4	1	0.00	0.03	0.00
h5	1	0.00	0.05	0.00	h5	1	0.00	0.09	0.01	h5	1	0.00	0.03	0.00
E	1	0.00	0.06	0.00	h6	1	0.00	0.05	0.00	h6	1	0.00	0.05	0.00
h1xE	1	0.00	0.06	0.00	h7	1	0.00	0.06	0.00	h7	3	0.00	0.12	−0.02
h2xE	1	0.00	0.10	0.01	h8	1	0.00	0.05	0.01	h8	1	0.00	0.17	0.00
h3xE	3	0.00	0.10	0.03	E	1	0.00	0.04	0.00	h9	1	0.00	0.05	0.00
h4xE	1	0.00	0.05	0.01	h1xE	1	0.00	0.05	0.00	h10	1	0.00	0.04	0.00
h5xE	1	0.00	0.03	0.00	h2xE	1	0.00	0.04	0.00	h11	1	0.00	0.03	0.00
–	–	–	–	–	h3xE	1	0.00	0.09	0.01	E	1	0.00	0.05	−0.01
–	–	–	–	–	h4xE	1	0.00	0.18	0.00	h1xE	1	0.00	0.05	0.00
–	–	–	–	–	h5xE	3	0.00	0.19	0.00	h2xE	1	0.00	0.04	0.00
–	–	–	–	–	h6xE	1	0.00	0.04	0.00	h3xE	1	0.00	0.03	0.00
–	–	–	–	–	h7xE	1	0.00	0.04	0.00	h4xE	1	0.00	0.04	−0.01
–	–	–	–	–	h8xE	1	0.00	0.05	−0.01	h5xE	1	0.00	0.08	0.00
–	–	–	–	–	–	–	–	–	–	h6xE	1	0.00	0.03	0.00
–	–	–	–	–	–	–	–	–	–	h7xE	1	0.00	0.11	0.00
–	–	–	–	–	–	–	–	–	–	h8xE	3	0.00	0.10	0.02
–	–	–	–	–	–	–	–	–	–	h9xE	1	0.00	0.03	0.00
–	–	–	–	–	–	–	–	–	–	h10xE	1	0.00	0.03	0.00
–	–	–	–	–	–	–	–	–	–	h11xE	1	0.00	0.07	0.00

In addition, the new version can handle multiple environmental covariates, each with multiple levels. Our real data analyses illustrate the importance of this extension aptly as a significant interaction effect of smoking with an rHTV was only detected when smoking was modeled as a three-level covariate. This type of modeling is consistent with the literature as smoking is typically modeled using three levels (either the same way as ours or as non-smoker, light smoker, and heavy smoker).[20–22] These analyses are also an important contribution to the lung cancer literature as there have been conflicting results on the studied region. Our results point to potential involvement of a rare rHTV interacting with smoking. Specifically, the protective effect we found adds to the growing evidence to support why some smokers have lower risk of lung cancer compared to other smokers.

Table 3. Comparison of computation time (in seconds) between the original (Version 1.0) and improved (Version 1.1) versions of LBL-GxE.

DATA	SAMPLE SIZE	# HAPLOTYPES	VERSION 1.0*	VERSION 1.1
Lung cancer data: two-level smoking	5549	4	758	218
Lung cancer data: three-level smoking	5549	4	–	312
Lung cancer data: three-level smoking and sex	5549	4	–	387
Simulated data 1: two-level covariate	2000	6	341	127
Simulated data 2: two-level covariate	2000	9	906	200
Simulated data 3: two-level covariate	2000	12	2123	308

Note: *Version 1.0 can only handle one covariate with two levels.

It should be noted that a key assumption of LBL-GxE is gene–environment independence. We checked this assumption using a test of independence between the haplotypes considered here and smoking (for both two and three levels), and it appears to be satisfied. However, when the assumption is violated, this method may not be valid. Our ongoing work involves extending the approach to deal with scenarios when this gene–environment independence assumption may not hold.

Acknowledgments

The lung cancer dataset used for the analyses described in this manuscript was obtained from the EAGLE and PLCO Study of lung cancer found at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000093.v2.p2 through dbGaP accession number phs000093.v2.p2. The authors would like to thank EAGLE and PLCO participants and researchers for their valuable contribution to this research. Also, the authors are thankful to the anonymous reviewers for their constructive comments.

Author Contributions

Conceived and designed the experiments: SB. Analyzed the data: YZ. Wrote the first draft of the manuscript: YZ. Contributed to the writing of the manuscript: SB. Agree with manuscript results and conclusions: YZ, SB. Jointly developed the structure and arguments for the paper: SB. Made critical revisions and approved final version: SB. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Maher B. Personal genomes: the case of the missing heritability. *Nature*. 2008;456:18–21.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–53.
3. Goldstein DB. Common genetic variation and human traits. *New Engl J Med*. 2009;360:1696–8.
4. Hirschhorn JN. Genomewide association studies – illuminating biologic pathways. *New Engl J Med*. 2009;360:1699–701.
5. Kraft P, Hunter DJ. Genetic risk prediction – are we there yet? *New Engl J Med*. 2009;360:1701–3.
6. Clark AG. The role of haplotypes in candidate gene studies. *Genet Epidemiol*. 2004;27:321–33.
7. Schaid DJ. Genetic epidemiology and haplotypes. *Genet Epidemiol*. 2004;27:317–20.
8. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010;11:259–72.
9. Biswas S, Xia S, Lin S. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. *Genet Epidemiol*. 2014;38:31–41.
10. Biswas S, Lin S. Logistic Bayesian LASSO for identifying association with rare haplotypes and appli cation to age-related macular degeneration. *Biometrics*. 2012;68:587–97.
11. 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.
12. 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.
13. Rotunno M, Yu K, Lubin JH, et al. Phase I metabolic genes and risk of lung cancer: multiple polymorphisms and mRNA expression. *PLoS One*. 2009;4:e5652.
14. Graziano C, Comin CE, Crisci C, et al. Functional polymorphisms of the microsomal epoxide hydrolase gene: a reappraisal on a early-onset lung cancer patients series. *Lung Cancer*. 2009;63:187–93.
15. Gsur A, Zidek T, Schnattinger K, et al. Association of microsomal epoxide hydrolase polymorphisms and lung cancer risk. *Br J Cancer*. 2003;89:702–6.
16. Persson I, Johansson I, Lou YC, et al. Genetic polymorphism of xenobiotic metabolizing enzymes among Chinese lung cancer patients. *Int J Cancer*. 1999;81:325–9.
17. To-Figueras J, Gene M, Gomez-Catalan J, Pique E, Borrego N, Corbella J. Lung cancer susceptibility in relation to combined polymorphisms of microsomal epoxide hydrolase and glutathione S-transferase *PI*. *Cancer Lett*. 2001;173:155–62.
18. Voho A, Metsola K, Anttila S, et al. EPHX1 gene polymorphisms and individual susceptibility to lung cancer. *Cancer Lett*. 2006;237:102–8.
19. Wu X, Gwyn K, Amos CI, Makan N, Hong WK, Spitz MR. The association of microsomal epoxide hydrolase polymorphisms and lung cancer risk in African-Americans and Mexican-Americans. *Carcinogenesis*. 2001;22:923–8.
20. Zhao H, Spitz MR, Gwyn KM, Wu X. Microsomal epoxide hydrolase polymorphisms and lung cancer risk in non-Hispanic whites. *Mol Carcinog*. 2002;33:99–104.
21. Lee WJ, Brennan P, Boffetta P, et al. Microsomal epoxide hydrolase polymorphisms and lung cancer risk: a quantitative review. *Biomarkers*. 2002;7:230–41.
22. Timofeeva M, Kropp S, Sauter W, et al; LUCY-Consortium. Genetic polymorphisms of MPO, GSTT1, GSTM1, GSTP1, EPHX1 and NQO1 as risk factors of early-onset lung cancer. *Int J Cancer*. 2010;127:1547–61.