A SCREENING-TESTING APPROACH FOR DETECTING GENE-ENVIRONMENT INTERACTIONS USING SEQUENTIAL PENALIZED AND UNPENALIZED MULTIPLE LOGISTIC REGRESSION

H. ROBERT FROST*[1,2,3], ANGELINE S. ANDREW[1,2], MARGARET R. KARAGAS[1,2],
AND JASON H. MOORE[1,2,3]

[1]*Institute for Quantitative Biomedical Sciences,*
Geisel School of Medicine, Dartmouth College,
Lebanon, NH 03756

[2]*Section of Biostatistics and Epidemiology, Department of Community and Family Medicine,*
Geisel School of Medicine, Dartmouth College,
Lebanon, NH 03756

[3]*Department of Genetics,*
Geisel School of Medicine, Dartmouth College,
Hanover, NH 03755

Gene-environment (G×E) interactions are biologically important for a wide range of environmental exposures and clinical outcomes. Because of the large number of potential interactions in genome-wide association data, the standard approach fits one model per G×E interaction with multiple hypothesis correction (MHC) used to control the type I error rate. Although sometimes effective, using one model per candidate G×E interaction test has two important limitations: low power due to MHC and omitted variable bias. To avoid the coefficient estimation bias associated with independent models, researchers have used penalized regression methods to jointly test all main effects and interactions in a single regression model. Although penalized regression supports joint analysis of all interactions, can be used with hierarchical constraints, and offers excellent predictive performance, it cannot assess the statistical significance of G×E interactions or compute meaningful estimates of effect size. To address the challenge of low power, researchers have separately explored screening-testing, or two-stage, methods in which the set of potential G×E interactions is first filtered and then tested for interactions with MHC only applied to the tests actually performed in the second stage. Although two-stage methods are statistically valid and effective at improving power, they still test multiple separate models and so are impacted by MHC and biased coefficient estimation. To remedy the challenges of both poor power and omitted variable bias encountered with traditional G×E interaction detection methods, we propose a novel approach that combines elements of screening-testing and hierarchical penalized regression. Specifically, our proposed method uses, in the first stage, an elastic net-penalized multiple logistic regression model to jointly estimate either the marginal association filter statistic or the gene-environment correlation filter statistic for all candidate genetic markers. In the second stage, a single multiple logistic regression model is used to jointly assess marginal terms and G×E interactions for all genetic markers that pass the first stage filter. A single likelihood-ratio test is used to determine whether any of the interactions are statistically significant. We demonstrate the efficacy of our method relative to alternative G×E detection methods on a bladder cancer data set.

1. Introduction

A significant body of recent research in the statistical genetics and genetic epidemiology communities has focused on the detection of statistical interactions between genetic markers and environmental variables (G×E interactions) using genome-wide association (GWA) data.[1]

Such data sets are comprised by the measurements of thousands to over one million genetic markers, typically single nucleotide polymorphisms (SNPs), along with relevant clinical and environmental variables on a set of human subjects that number in the thousands to hundreds-of-thousands for large GWA studies. Since the number genetic markers, and therefore the number of potential G×E interactions for a single environmental variable, is usually larger than the number of subjects, statistical testing of G×E interactions has typically been accomplished by fitting separate models for each genetic marker and applying multiple hypothesis correction (MHC) to the generated p-values to control the type I error rate. Although a G×E interaction can be defined as a departure from additivity on either a log odds or absolute risk scale, we focus on the former type of interaction in this paper. Statistically, such an interaction is commonly tested using a logistic regression model of the form:

$$logit(P(D=1|G,E)) = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE \qquad (1)$$

where D is a binary outcome variable, E is the environmental variable and G is one of the genetic markers. In this paper, we assume that both D and E are binary, e.g., disease case/control status and exposed/non-exposed indicator, and that G represents a SNP specified using additive coding, i.e., 0, 1 or 2 based on the number of copies of the minor allele. Using this model, the null hypothesis of no G×E interaction on a log odds scale can be specified as $H_0 : \beta_{GE} = 0$ with significance tested via either a Wald test associated with $\hat{\beta}_3$ or a likelihood ratio test. Variations on this basic approach that also use one model per potential G×E interaction include the case-only gene-environment association test, the test of marginal association and the combined test of marginal gene association and G×E interaction.[2]

Although methods that test a separate model for each potential G×E interaction (so-called one-step methods) are easy to understand, simple to implement and can, in many instances, identify biologically plausible interactions, they have two serious drawbacks. First, the power to detect G×E interactions, already much lower than the power to detect main effects at a given sample size,[3] is severely degraded for even a moderate number of genetic markers due to the requirement for MHC to control the type I error rate across the separate models. Second, because each interaction is assessed independently, the estimated interaction coefficients will be biased if associations exist between the genetic markers.

To address the poor statistical power of standard one-step methods, researchers have recently explored two-stage, or screening-testing, methods.[4–9] Screening-testing methods first filter the set of candidate genetic markers and, for the markers that pass the first stage filter, test G×E interactions. As long as the statistic used to filter the genetic markers in the first stage is statistically independent of the second stage test statistic under the null hypothesis, type I error rates will be correctly controlled with MHC applied to just the smaller number of hypotheses that pass the first stage filter.[7,10] Two popular independent filters for G×E interaction detection are the marginal association filter[4] and the gene-environment correlation filter.[5] The marginal association filter measures the statistical association between the outcome variable and the genetic marker using a logistic regression model of the form:

$$logit(P(D=1|G)) = \beta_0 + \beta_G G \qquad (2)$$

where the filter statistic is the p-value associated with the $\hat{\beta}_1$ coefficient estimate. The gene-environment correlation filter measures the statistical association between the environmental variable and each genetic marker using a logistic regression model of the form:

$$logit(P(E = 1|G)) = \beta_0 + \beta_G G \tag{3}$$

This is the same model used with the case-only gene-environment association test for G×E interaction, but fit, for the correlation filter, using pooled cases and controls instead of just cases. Similar to the marginal association filter, the correlation filter uses the p-value associated with the $\hat{\beta_G}$ coefficient estimate as a filter statistic. To be effective at improving power, a filter statistic must not only be independent of the second stage test statistic under the null hypothesis, but must also be associated with the test statistic under the alternative hypothesis of G×E interaction. While the first requirement has been proven for both the marginal association and correlation filter statistics in the context of G×E interaction detection using logistic regression models of the form in equation 1,[7] there is no guarantee that the second requirement will hold for the data set under analysis. For some data sets, the marginal association filter will be optimal, for others the correlation filter will perform best and success has been reported using an ensemble of both filter types, e.g., the cocktail method.[8] Although current screening-testing methods can be effective at improving G×E detection power, the fact that these methods use separate models for each candidate genetic marker during both the filter and testing stages means that both MHC correction and omitted variable bias remain issues. For data sets with small sample-to-marker or signal-to-noise ratios, even the reduced MHC penalty after filtering is sufficient to negate G×E detection power.

Another recently developed approach to G×E interaction detection involves the use of penalized regression to jointly estimate all possible G×E interactions as well as main effects in a single model. Such penalization approaches typically enforce a hierarchical constraint that will only consider interaction terms for significant main effects. Many variations of the joint penalized model approach exist, including the hierarchical LASSO by Bien et al.,[11] the penalized hierarchical approach of Liu et al.,[12] the progressive hierarchical penalization approach of Zhu et al.,[13] the multi-stage LASSO method of Wu et al.[14] and approaches that fit a single LASSO-penalized model with all possible marginal and interaction terms (termed all-pairs LASSO (APL) by Bien et al.)[11] The approach of Wu et al.[14] is especially relevant since it employs a LASSO penalized multiple logistic regression model in the first stage to filter genetic markers based on marginal association and then tests for interactions in a second stage model. The fact that Wu et al. use LASSO penalization in the second stage model, however, means that their approach cannot generate valid measures of interaction statistical significance and is therefore not a valid screening-testing method. Also, Wu et al. focus on gene-gene as opposed to gene-environment interactions. Methods that fit a single penalized model have the significant benefit of jointly estimating all potential G×E interactions along with marginal gene and environmental effects and can therefore be very effective for prediction; however, the shrunken coefficients may be severely biased with unclear statistical significance. Although some authors, e.g. Wu et al.,[14] advocate refitting a non-penalized model for just the interaction terms with non-zero coefficients in the penalized model to generate more meaningful coefficients and measures of statistical significance, this approach fails to account for the prior penalized selection process and thus cannot correctly compute statistical significance or interaction effect size.

To address the limitations of inadequate power and biased coefficient estimation associated with one-step approaches, we have developed a novel G×E interaction detection method that combines aspects of screening-testing with hierarchical penalized regression. In the first stage, our approach uses a single elastic net-penalized multiple logistic regression model to jointly estimate either the marginal association filter statistic or the gene-environment correlation filter statistic for all candidate genetic markers. In the second stage, a single multiple logistic regression model is used to jointly assess marginal effects and G×E interactions for all genetic markers that pass the first stage filter. An important feature of our approach is that a single omnibus test can be used to detect the presence of statistically significant G×E interactions. As we demonstrate using a bladder cancer genotype data set with smoking status as the environmental variable, our method provides the statistical benefits of joint estimation along with significantly improved G×E detection power relative to competing approaches.

2. Methods

2.1. *Proposed screening-testing method for G×E interaction detection*

2.1.1. *Screening stage*

Our approach filters the set of measured genetic markers using a penalized multiple logistic regression model that jointly computes a filter statistic, either the marginal association filter[4] or the gene-environment correlation filter,[5] for all measured genetic markers. For the marginal association filter, a penalized multiple logistic regression model of the following form is used:

$$logit(P(D=1|G)) = \beta_0 + \beta_{G_1}G_1 + ... + \beta_{G_p}G_p \tag{4}$$

This model is fit using an elastic net[15] penalty via the *glmnet* R package implementation.[16] This procedure computes coefficient estimates to maximize an objective function with both L1, i.e., LASSO, and L2, i.e., ridge, penalties:

$$-\frac{log(L(\beta_1, ..., \beta_p|G)}{n} + \lambda(\frac{1-\alpha}{2}\sum_{i=1}^{p}\beta_{G_i}^2 + \alpha\sum_{i=1}^{p}|\beta_{G_i}|) \tag{5}$$

where the α coefficient is the elastic net mixing parameter ($\alpha = 1$ corresponds to just LASSO penalization and $\alpha = 0$ corresponds to just ridge penalization). The elastic net penalty parameter λ can be selected according to cross-validation or to achieve a specific number of non-zero coefficients. For the gene-environment correlation filter, a penalized multiple logistic regression model of the following form is used:

$$logit(P(E=1|G)) = \beta_0 + \beta_{G_1}G_1 + ... + \beta_{G_p}G_p \tag{6}$$

Modeling fitting in this case follows the same approach used for the model in equation 4.

2.1.2. *Testing stage*

To test for G×E interactions, a single multiple logistic regression model is fit using marginal and interaction terms for all genetic markers selected during the screening stage:

$$logit(P(D=1|G,E)) = \beta_0 + \beta_E E + \beta_{G_1} G_1 + ... + \beta_{G_p} G_p + \beta_{G_1 E} G_1 E + ... + \beta_{G_p E} G_p E \quad (7)$$

If desired, covariates can also be included in this model. To determine if any of the G×E interaction coefficients are statistically significant, the null hypothesis $H_0 : \beta_{G_1 E} = ... = \beta_{G_p E} = 0$ is tested using a LR test between a version of the model in equation 7 without the interaction terms and the model with interaction terms. If p-value from this LR test is significant, this indicates that at least one of the G×E interaction coefficients is significantly non-zero. The interactions can then be prioritized for further investigation based on the estimated interaction coefficient size and the associated Wald test p-values, perhaps after MHC. If the p-value from this LR test is not significant, no further investigation is performed.

If only one of the two supported filter statistics is used during the first stage, then the model in equation 7 is fit just a single time and only one LR test is performed to detect potential G×E interactions. If both filters are applied, the model in equation 7 is fit separately using the output from each filter, LR tests are performed on both models and the generated p-values are adjusted via the Bonferroni method, i.e., 2*p-value. If neither model has a significant LR p-value after MHC, no further investigation is performed, otherwise, the model with the most significant LR test result is used.

Although an ensemble approach, similar to the cocktail method,[8] could be adopted that combines the results from the marginal association and gene-environment correlation filters to build a single stage two model, such an approach would eliminate a key benefit of filtering based on a single penalized regression model, namely the reduction of multi-collinearities. Because the marginal association and gene-environment correlation models would be estimated separately, each model could identify genetic markers highly correlated with the predictors output by the other model, resulting in estimation instability for the second stage multiple logistic regression model.

2.1.3. *Interpretation*

Because the coefficients in a multiple logistic regression model with interaction terms represent conditional effects on the log-odds of the outcome variable, they do not have a straight-forward interpretation. This complex, conditional interpretation can be seen as a disadvantage of the proposed approach relative to separate logistic regression models for each interaction. Specifically, the effect size of each interaction term must be evaluated by considering the estimated coefficients for both the G_i predictor and the G_iE predictor. To be precise, the change in the log odds of the outcome per change in the number of minor allele copies (assuming additive coding) when there is no environmental exposure and all other predictors are held constant is represented by the estimated coefficient for the G_i predictor and the change in the log odds of the outcome per change in the number of minor allele copies when there is environmental exposure is represented by the sum of the estimated coefficients for the G_i and G_iE predictors. In more simplified terms, the estimated coefficient for the G_iE predictor reflects increased risk of disease for environmentally exposed individuals who carry the risk allele compared to unexposed risk allele carriers.

2.2. *Bladder cancer data*

We analyzed genetic variation in hypothesized cancer susceptibility genes and cigarette smoking in a population-based case-control study of bladder cancer. Detailed methods have been described previously in Karagas et al.[17] and Andrew et al.[18] Briefly, the cases were New Hampshire residents of ages 25 to 74 years, diagnosed with bladder cancer from July 1, 1994 to June 30, 2001 identified via the New Hampshire State Cancer Registry. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation, while controls aged 65 and older were chosen from data provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. The overwhelming majority ($\sim 98\%$) of the subjects were of Caucasian origin. Given the large proportion of Caucasian subjects, population structure should not be an issue for this data set, as confirmed in Andrew et al.[19] We interviewed a total of 857 patients with bladder cancer, which was 85% of the cases confirmed to be eligible for the study, and 1191 controls without cancer. Informed consent was obtained from each participant and all procedures and study materials were approved by the Committee for the Protection of Human Subjects at Dartmouth College. DNA was isolated from peripheral circulating blood lymphocyte or buccal specimens using Qiagen genomic DNA extraction kits (QIAGEN Inc., Valencia, CA). Genotyping was performed on all DNA samples of sufficient concentration using the GoldenGate Assay system (Illumina, Inc., San Diego, CA). Out of the submitted samples, 99.5% were successfully genotyped, and samples repeated on multiple plates yielded the same call for 99.9% of the SNPs.[20] Excluding subjects who did not have genotype calls for more than 50% of the SNPs, and one additional case due to missing data on smoking status, resulted in 610 cases and 865 controls included in our analysis. After removing SNPs with missing genotype values for more than 10% of the 1475 samples, we analyzed genotype data for a total of 1488 SNPs. Remaining missing genotype values in this dataset were imputed using a simple frequency-based approach in which the missing value was set to the most common genotype in the study population. Genotyped SNPs were mostly those included on the Illumina Cancer Panel, representing ≈ 400 hypothesized cancer-related genes. SNPs were selected within coding, intronic and flanking regions hypothesized to be potentially functional for the genes of interest, including a median of three SNPs per gene.

2.3. *G×E interaction detection for bladder cancer data*

To support comparative evaluation of our proposed method using the bladder cancer data set described above, potential G×E interactions between smoking status (recoded as never (0) or ever (1)) and SNPs relative to bladder cancer case/control status were computed using the proposed G×E interaction detection method and standard one-step and two-stage approaches. Analysis details for each method are outlined in Sections 2.3.1-2.3.3 below. For all methods, age and gender were included as covariates in the logistic regression models used to test for G×E interactions, i.e., the models specified by equations 1 and 7.

2.3.1. *One-step G×E interaction test for bladder cancer data*

For each of the 1488 analyzed SNPs, a SNP-smoking interaction was tested using a separate logistic regression model of the form specified in equation 1 above with MHC performed using the false discovery rate (FDR) method of Benjamini and Hochberg.[21]

2.3.2. *Standard two-stage G×E interaction test for bladder cancer data*

The 1488 analyzed SNPs were first filtered using either the marginal association filter statistic, as computed via the logistic regression model in equation 2, or the gene-environment correlation filter statistic, as computed via the logistic regression model in equation 3. So that the results from the two-stage method would be comparable to those generated by our proposed G×E detection method, the number of SNPs retained after the first stage filtering was fixed at the same number used for the proposed method (103 SNPs, see Section 2.3.3 below). For the SNPs that passed the first stage filter, the presence of a SNP-smoking interaction was tested using the same logistic regression model and MHC method employed for the one-step G×E test.

2.3.3. *Proposed G×E interaction detection method for bladder cancer data*

The screening-testing approach outlined in Section 2.1 above was executed on the bladder cancer data using both the marginal association filter and the gene-environment correlation filter. For each filter, the screening stage penalized logistic regression model was fit with the elastic net mixing parameter α set to .999 to provide estimation stability via a small L2 penalty[16] and the λ penalty parameter was set to achieve a ratio of observations-to-predictors in the unpenalized stage 2 model of 7 (the middle of the 5-to-9 recommended by Vittinghoff and McCulloch for multiple logistic regression[22]). For the 1475 analyzed subjects, the observations-to-predictors ratio of 7 allowed approximately 103 SNPs to be kept after screening with corresponding λ values of 0.0112 and 0.0118 for the gene-environment correlation and marginal association filters, respectively.

3. Results

3.1. *One-step G×E interaction test results*

Table 1 shows the ten most significant smoking-SNP interactions computed via the one-step method detailed in Section 2.3.1. The $\hat{\beta}_{GE}$ values in the table represent the estimated interaction term coefficients from the logistic regression model specified in equation 1 with age and gender as covariates. The p-values were generated via a LR test comparing a model without the G×E term to the model with the G×E term and the false discovery rate (FDR) values were generated for all p-values using the method of Benjamini and Hochberg.[21] Although some of the interaction LR p-values appear significant, after MHC to control the FDR, all findings appear consistent with H_0. In addition to the poor power after MHC, half of the top ten interactions returned by the one-step method involve highly correlated SNPs from the same gene, MASP1; a direct result of testing each interaction in a separate regression model.

Table 1. Ten most significant smoking-SNP interactions computed using the standard one-step method.

dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR
rs12635264	MASP1	-0.604	0.00142	0.848
rs13089330	MASP1	-0.596	0.00165	0.848
rs13094773	MASP1	-0.556	0.00332	0.848
rs2972418	GHR	-0.519	0.00383	0.848
rs9282553	ABCA6	-1.07	0.00419	0.848
rs3864099	MASP1	-0.541	0.00437	0.848
rs4376034	MASP1	-0.536	0.00447	0.848
rs3213216	IGF2	0.585	0.00457	0.848
rs3217773	CCNA2	0.6	0.0056	0.848
rs2229765	IGF1R	0.497	0.00589	0.848

3.2. *Standard two-stage G×E interaction test results*

Table 2. Ten most significant smoking-SNP interactions computed via the standard two-stage G×E detection method using either a marginal association filter or a gene-environment correlation filter.

Marginal association filter					G-E correlation filter				
dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR	dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR
rs2233679	PIN1	-0.381	0.0463	0.87	rs6347	SLC6A3	0.512	0.0219	0.983
rs2266690	CCNH	-0.376	0.0585	0.87	rs4696480	TLR2	-0.418	0.0255	0.983
rs5923	LCAT	-0.774	0.0711	0.87	rs1126667	ALOX12	-0.356	0.0492	0.983
rs3755557	GSK3B	-0.375	0.0776	0.87	rs1127717	FTHFD	0.442	0.0644	0.983
rs1799802	ERCC4	-0.652	0.101	0.87	rs2855262	SOD3	-0.354	0.0656	0.983
rs3937387	BZRP	0.27	0.144	0.87	rs998074	IGF2R	0.309	0.0947	0.983
rs1051740	EPHX1	0.298	0.153	0.87	rs7921327	AKR1C3	0.327	0.0983	0.983
rs5742926	PMS1	-0.362	0.154	0.87	rs676387	HSD17B1	-0.34	0.1	0.983
rs12801239	KIRREL3	0.272	0.167	0.87	rs998075	IGF2R	0.298	0.107	0.983
rs3448	GPX1	-0.277	0.187	0.87	rs4817027	BIC	-0.341	0.111	0.983

Table 2 displays the ten most significant smoking-SNP interactions computed via the standard two-stage method using either a marginal association filter or a gene-environment correlation filter, as detailed in Section 2.3.2. In this case, the top ten interactions returned by the two filters are completely disjoint. Although the first stage filter reduced the number of smoking-SNP interactions tested via logistic regression models from the 1488 examined by the one-step method to only 103, none of the results appear significant after MHC. In fact, the top uncorrected p-values after filtering, although independently significant, are substantially higher than the top uncorrected p-values found when all SNPs are tested via the one-step approach, indicating that there is only a weak correlation between the two filter statistics and G×E interaction (as tested using logistic regression models of the form in equation 1) for this data set under the alternate hypothesis of G×E interaction.

3.3. *Proposed G×E interaction detection method results*

Table 3 shows the significant smoking-SNP interactions computed using the proposed method, as detailed in Section 2.3.3. Specifically, the table includes interactions whose Wald test p-

Table 3. Smoking-SNP interactions computed via the proposed G×E detection method with Wald test p-values for the estimated interaction coefficients in the test stage multiple logistic regression model below 0.05. LR test p-values are the Bonferroni-corrected p-values from a likelihood ratio test comparing a test stage model without interaction terms to a model with interaction terms.

Marginal association filter					G-E correlation filter				
Corrected LR p-value: 0.174					Corrected LR p-value: 0.022				
dbSNP ID	Gene name	$\hat{\beta}_{GE}$	Wald p-val	FDR	dbSNP ID	Gene name	$\hat{\beta}_{GE}$	Wald p-val	FDR
rs2233679	PIN1	-1.69	0.00102	0.108	rs3213223	IGF2	-1.21	0.00103	0.0585
rs26279	MSH3	1.49	0.0021	0.111	rs2266690	CCNH	-1.06	0.00118	0.0585
rs1584415	*na*	-0.978	0.00703	0.18	rs861539	XRCC3	0.906	0.00169	0.0585
rs9642880	CASC11	-1.08	0.0095	0.18	rs1381841	GSK3B	-1.26	0.00238	0.062
rs698090	MASP1	0.969	0.0111	0.18	rs4696480	TLR2	-0.596	0.018	0.294
rs8173	STK6	-1.12	0.0112	0.18	rs2877796	RGS6	-0.598	0.0194	0.294
rs4986765	BRIP1	-1.01	0.0119	0.18	rs113515	TSPO	0.592	0.0198	0.294
rs2676530	HSD17B1	-0.883	0.0176	0.233	rs7921327	AKR1C3	0.616	0.0234	0.304
rs760589	OPRD1	-0.872	0.0224	0.253	rs4619	IGFBP1	-0.553	0.03	0.337
rs4988340	BRIP1	-0.846	0.0278	0.253	rs869975	GPX3	-1.28	0.0324	0.337
rs3740066	ABCC2	0.744	0.0284	0.253	rs1059519	GDF15	0.636	0.0378	0.343
rs13167280	TERT	-1.12	0.0287	0.253	rs2233679	PIN1	-0.543	0.0403	0.343
rs8037	KRT23	0.763	0.0351	0.286	rs1126667	ALOX12	-0.498	0.0458	0.343
rs3847862	CELA1	-0.699	0.0465	0.324	rs6347	SLC6A3	0.625	0.0465	0.343
rs1650697	MSH3	-0.944	0.0499	0.324	rs872072	TEP1	0.5	0.0495	0.343

values for the estimated interaction coefficients in the test stage model specified in equation 7 are below 0.05. The corresponding FDR value was computed for the family of Wald tests on all interaction coefficients. The second stage model was fit for the genetic markers generated using both the marginal association filter (equation 4) and the gene-environment correlation filter (equation 6) screening stage models. Similar to the results from the standard two-stage method, the marginal association and gene-environment correlation filters generate largely independent sets of smoking-SNP interactions. In both cases, the presence of smoking-SNP interactions in the test stage model was assessed using a LR test comparing the likelihood of the model without interaction terms to the likelihood of the model with interaction terms. Because LR tests were performed for both test stage models, a Bonferroni correction was applied to each LR p-value. After MHC, only the LR test for the model fit using the 103 SNPs output by the gene-environment correlation filter was significant (adjusted p-value=0.022). To measure the quality of the SNPs in this significant second stage model, a Hardy Weinberg test of equilibrium was performed among the controls, resulting in an average test p-value of 0.48.

Further investigation of the most significant smoking-SNP interactions from the test stage model for the gene-environment correlation filter revealed several SNPs with prior evidence of association with bladder cancer and/or smoking in independent populations, most notably, a confirmed interaction between smoking and cyclin H (CCNH).[23] SNPs in Toll-like receptor 2 (TLR2) increased overall bladder cancer risk, however, the smoking interaction was not statistically significant in this smaller study.[24] Variation in the regulator of G-protein signaling 6 (RGS6) reduced bladder cancer risk, with suggestion of an interaction with smoking.[25] The AKR1C3 association is consistent across several studies[26,27] with a potential relationship with smoking.[28] Likewise, our TEP1 bladder cancer association,[27] was independently confirmed.[29]

An interaction with smoking may explain some of the heterogeneity observed among prior studies of the X-ray repair complementing defective in chinese hamster 3 (XRCC3) SNP.[30] SLC6A3 variations lead to stress-induced cigarette craving.[31] While data on SNP associations are lacking, growth differentiation factor 15 (GDF15) is being promoted as a biomarker of urothelial cell cancer.[32] Insulin growth factor 2 (IGF2) is over-expressed in bladder tumors.[33]

4. Discussion

In this paper, we have detailed a novel approach for G×E interaction detection that combines elements of screening-testing methods with hierarchical penalized regression. Similar to existing screening-testing techniques, our approach first filters all measured genetic markers according to a filter statistic that is independent from the G×E test statistic under H_0, and, for the markers that pass the filter, performs a G×E interaction test. The key difference between our approach and existing two-stage methods lies in the structure of the screening and test stage models and the associated statistical G×E interaction tests. Whereas standard two-stage methods fit a separate logistic regression model for each potential G×E interaction in both the screening and testing stages, our method jointly evaluates all markers in a single multiple logistic regression model during both the screening and test stages. Because the number of measured markers is typically much larger than the number of subjects, the screening stage model must be fit using penalization and our approach employs an elastic net penalty that combines L1 and L2 penalty terms.[15] The use of penalized multiple logistic regression enables either the marginal association filter statistic[4] or the gene-environmental correlation filter statistic[5] to be jointly computed for all markers, and, because LASSO-penalization tends to retain only one predictor from a set of correlated predictors,[34] the set of terms with non-zero coefficients will contain few significant collinearities. Generating a fairly small set of high-quality candidate markers in the screening stage that is free from collinearities is critical when attempting to fit a single unpenalized multiple logistic regression for these markers in the test stage. Assessing G×E interactions in the test stage using a single multiple logistic regression model, as opposed to separate models for each interaction, has two major benefits. First, estimating coefficients jointly decreases the bias associated with omitted predictors in regression. Second, and most importantly, fitting a single model for all markers that pass the screening stage enables the use of a single omnibus test to assess whether any statistically significant G×E interactions exist. Use of just one statistical test completely eliminates the penalty of MHC on power for basic G×E interaction detection. If the number of markers kept after screening is relatively small and the filter statistic correctly retains those markers with high likelihood of being in a G×E interaction, it is quite reasonable to limit inference to a single omnibus test. Wald test p-values and effect size estimates are then used to prioritize the interactions for further investigation and experimental validation. In situations with sample size constraints or poor data quality, a single omnibus test on a filtered set of markers may in fact be the only adequately powered test of G×E interactions.

The benefits of our proposed method relative to standard approaches are clearly demonstrated by the analysis of the bladder cancer data set for smoking-SNP interactions. Neither the one-step nor the standard two-stage methods were able to find any statistically significant smoking-SNP interactions after MHC. The inability of the one-step and two-stage methods to identify significant interactions mirrors the results from other investigations into smoking-SNP interactions relative to bladder cancer, such as the recent study by Figueroa et al.[35] that failed to find significant additive or multiplicative interactions after MHC using a one-step analysis. Our proposed method, on the other hand, successfully found evidence of statistically significant interactions when using the gene-environment correlation filter, as evidenced by the corrected LR test p-value of 0.022 and multiple interactions coefficients with Wald test FDR values below 0.1. The significant interactions identified in this model have not been previously discovered via statistical G×E interaction tests using this data set. A subsequent investigation of this significant test stage model found biological support in the research literature for many of the most significant smoking-SNP interactions.

Although our approach has important methodological and statistical benefits relative to existing G×E interaction detection methods, there are some key limitations to note. First, interpretation of interaction coefficients may be more difficult using a joint model than when using separate models per interaction. Second, the use of an omnibus test just indicates that at least one of the G×E interactions is significant, it does not specify which interaction; unless MHC is applied to the Wald test p-values, these can be only be used for qualitative prioritization and not as strict measures of statistical significance. Finally, the evaluation detailed in this paper was for a data set with a small number of markers; it will be important to assess how well the method scales to genomic data sets measuring upwards of one million markers. For such large data sets, the computational complexity of the elastic net implementation may be a key constraint. In future work, it will important to test our approach on a diverse collection of GWAS data sets for a range of different environmental exposures and outcome variables.

Acknowledgement

Funding: National Institutes of Health R01 grants LM010098, LM011360, LM009012, EY022300, GM103506 and GM103534. Conflict of Interest: None declared.

References

1. D. J. Hunter, *Nat Rev Genet* **6**, 287 (Apr 2005).
2. A. Ziegler and I. R. König, *A statistical approach to genetic epidemiology*, 2nd edn. (Wiley-VCH, Weinheim, 2010).
3. D. Wahlsten, *Behavioral and Brain Sciences* **13**, 109 (Mar 1990).
4. C. Kooperberg and M. Leblanc, *Genet Epidemiol* **32**, 255 (Apr 2008).
5. C. E. Murcray, J. P. Lewinger and W. J. Gauderman, *Am J Epidemiol* **169**, 219 (Jan 2009).
6. C. E. Murcray, J. P. Lewinger, D. V. Conti, D. C. Thomas and W. J. Gauderman, *Genet Epidemiol* **35**, 201 (Apr 2011).
7. J. Y. Dai, C. Kooperberg, M. Leblanc and R. L. Prentice, *Biometrika* **99**, 929 (Dec 2012).
8. L. Hsu, S. Jiao, J. Y. Dai, C. Hutter, U. Peters and C. Kooperberg, *Genet Epidemiol* **36**, 183 (Apr 2012).
9. J. Millstein, *Front Genet* **4**, p. 306 (2013).
10. R. Bourgon, R. Gentleman and W. Huber, *Proc Natl Acad Sci U S A* **107**, 9546 (May 2010).
11. J. Bien, J. Taylor and R. Tibshirani, *The Annals of Statistics* **41**, 1111 (2013).

12. J. Liu, J. Huang, Y. Zhang, Q. Lan, N. Rothman, T. Zheng and S. Ma, *Genomics* **102**, 189 (Oct 2013).
13. R. Zhu, H. Zhao and S. Ma, *Genet Epidemiol* **38**, 353 (May 2014).
14. T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel and K. Lange, *Bioinformatics* **25**, 714 (Mar 2009).
15. H. Zou and T. Hastie, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 301 (2005).
16. J. H. Friedman, T. Hastie and R. Tibshirani, *Journal of Statistical Software* **33**, 1 (Feb 2010).
17. M. R. Karagas, T. D. Tosteson, J. Blum, J. S. Morris, J. A. Baron and B. Klaue, *Environ Health Perspect* **106 Suppl 4**, 1047 (Aug 1998).
18. A. S. Andrew, J. Gui, T. Hu, A. Wyszynski, C. J. Marsit, K. T. Kelsey, A. R. Schned, S. A. Tanyos, E. M. Pendleton, R. M. Ekstrom, Z. Li, M. S. Zens, M. Borsuk, J. H. Moore and M. R. Karagas, *BJU International* , n/a (2014).
19. A. S. Andrew, T. Hu, J. Gu, J. Gui, Y. Ye, C. J. Marsit, K. T. Kelsey, A. R. Schned, S. A. Tanyos, E. M. Pendleton, R. A. Mason, E. V. Morlock, M. S. Zens, Z. Li, J. H. Moore, X. Wu and M. R. Karagas, *PLoS One* **7**, p. e51301 (2012).
20. A. S. Andrew, H. H. Nelson, K. T. Kelsey, J. H. Moore, A. C. Meng, D. P. Casella, T. D. Tosteson, A. R. Schned and M. R. Karagas, *Carcinogenesis* **27**, 1030 (May 2006).
21. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* , 289 (1995).
22. E. Vittinghoff and C. E. McCulloch, *Am J Epidemiol* **165**, 710 (Mar 2007).
23. M. Chen, A. M. Kamat, M. Huang, H. B. Grossman, C. P. Dinney, S. P. Lerner, X. Wu and J. Gu, *Carcinogenesis* **28**, 2160 (Oct 2007).
24. V. Singh, N. Srivastava, R. Kapoor and R. D. Mittal, *Arch Med Res* **44**, 54 (Jan 2013).
25. D. M. Berman, Y. Wang, Z. Liu, Q. Dong, L.-A. Burke, L. A. Liotta, R. Fisher and X. Wu, *Cancer Res* **64**, 6820 (Sep 2004).
26. J. D. Figueroa, N. Malats, M. García-Closas, F. X. Real, D. Silverman, M. Kogevinas, S. Chanock, R. Welch, M. Dosemeci, Q. Lan, A. Tardón, C. Serra, A. Carrato, R. García-Closas, G. Castaño-Vinyals and N. Rothman, *Carcinogenesis* **29**, 1955 (Oct 2008).
27. A. S. Andrew, J. Gui, A. C. Sanderson, R. A. Mason, E. V. Morlock, A. R. Schned, K. T. Kelsey, C. J. Marsit, J. H. Moore and M. R. Karagas, *Hum Genet* **125**, 527 (Jun 2009).
28. T. Hu, Q. Pan, A. S. Andrew, J. M. Langer, M. D. Cole, C. R. Tomlinson, M. R. Karagas and J. H. Moore, *BioData Min* **7**, p. 5 (2014).
29. J. Chang, C. P. Dinney, M. Huang, X. Wu and J. Gu, *PLoS One* **7**, p. e30665 (2012).
30. Q. Ma, Y. Zhao, S. Wang, X. Zhang, J. Zhang, M. Du, L. Li and Y. Zhang, *Tumour Biol* **35**, 1473 (Feb 2014).
31. J. Erblich, C. Lerman, D. W. Self, G. A. Diaz and D. H. Bovbjerg, *Pharmacogenomics J* **4**, 102 (2004).
32. V. L. Costa, R. Henrique, S. A. Danielsen, S. Duarte-Pereira, M. Eknaes, R. I. Skotheim, A. Rodrigues, J. S. Magalhães, J. Oliveira, R. A. Lothe, M. R. Teixeira, C. Jerónimo and G. E. Lind, *Clin Cancer Res* **16**, 5842 (Dec 2010).
33. G. Pignot, A. Vieillefond, S. Vacher, M. Zerbib, B. Debre, R. Lidereau, D. Amsellem-Ouazana and I. Bieche, *Br J Cancer* **106**, 1177 (Mar 2012).
34. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 273 (2011).
35. J. D. Figueroa *et al.*, *Carcinogenesis* **35**, 1737 (Aug 2014).