RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases

Shuo Jiao,[1]* Li Hsu,[1] Stéphane Bézieau,[2] Hermann Brenner,[3] Andrew T. Chan,[4,5] Jenny Chang-Claude,[6] Loic Le Marchand,[7] Mathieu Lemire,[8] Polly A. Newcomb,[1,9] Martha L. Slattery,[10] and Ulrike Peters[1,9]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington; [2]Service de Génétique Médicale, CHU Nantes, Nantes, France; [3]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany; [4]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts; [5]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; [6]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany; [7]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii; [8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada; [9]School of Public Health, University of Washington, Seattle, Washington; [10]Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah

Received 5 March 2013; Revised 4 April 2013; accepted revised manuscript 30 April 2013.
Published online 29 May 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21735

ABSTRACT: **Identification of gene-environment interaction (G × E) is important in understanding the etiology of complex diseases. However, partially due to the lack of power, there have been very few replicated G × E findings compared to the success in marginal association studies. The existing G × E testing methods mainly focus on improving the power for individual markers. In this paper, we took a different strategy and proposed a set-based gene-environment interaction test (SBERIA), which can improve the power by reducing the multiple testing burdens and aggregating signals within a set. The major challenge of the signal aggregation within a set is how to tell signals from noise and how to determine the direction of the signals. SBERIA takes advantage of the established correlation screening for G × E to guide the aggregation of genotypes within a marker set. The correlation screening has been shown to be an efficient way of selecting potential G × E candidate SNPs in case-control studies for complex diseases. Importantly, the correlation screening in case-control combined samples is independent of the interaction test. With this desirable feature, SBERIA maintains the correct type I error level and can be easily implemented in a regular logistic regression setting. We showed that SBERIA had higher power than benchmark methods in various simulation scenarios, both for common and rare variants. We also applied SBERIA to real genome-wide association studies (GWAS) data of 10,729 colorectal cancer cases and 13,328 controls and found evidence of interaction between the set of known colorectal cancer susceptibility loci and smoking.**
Genet Epidemiol 37:452–464, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: **gene-environment interaction; set based; correlation screening; GWAS; rare variants**

Introduction

Both genetic (G) and environmental (E) factors impact common complex diseases, such as cancer, diabetes, or cardiovascular diseases. For most of these diseases, several environmental factors and a rapidly increasing number of genetic factors have been identified [Hindorff et al., 2009]. However, little is understood about the interplay between G and E. Some exceptions include an observed interactions between smoking and the *GSTM1* deletion and a tag SNP in *NAT2* in bladder cancer [García-Closas et al., 2005; Rothman et al., 2010], *ADH7* variants and alcohol consumption in upper aerodigestive cancers [Hashibe et al., 2008], or *GRIN2A* variants and coffee consumption in Parkinson's disease [Hamza et al., 2011].

Although measurement error and data harmonization issues across studies for the environmental factors may have contributed to the limited numbers of confirmed gene-environment interactions (G × E), probably more importantly, the statistical power to detect an interaction is much smaller compared to detecting a main effect. In fact, it has been shown that the detection of an interaction needs at least approximately four times as many subjects as are needed to detect a main genetic effect of comparable effect size [Smith and Day, 1984]. A number of methods have been proposed to enhance the power of detecting G × E, which includes the case-only test [Chatterjee and Carroll, 2005; Piegorsch et al., 1994], the empirical Bayes method [Mukherjee and Chatterjee, 2008], and the Bayesian model averaging method [Li and Conti, 2009]. Two types of screening methods have also been proposed to reduce the multiple testing burden in

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Shuo Jiao, Cancer Prevention Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., M4-B402, Seattle, WA 98109. E-mail: sjiao@fhcrc.org

genome-wide G × E search: the correlation-based screening [Murcray et al., 2009] and the marginal association based screening [Kooperberg and Leblanc, 2008]. Toward this end, several recent methods were developed to combine and take advantage of different screening and testing techniques, such as the hybrid method by Murcray et al. [2011] and cocktail method by Hsu et al. [2012].

The above-mentioned efforts focus on improving the power of detecting G × E for individual markers. On the other hand, the set-based association testing has attracted increasing interest. A set-based method can not only enhance the power by aggregating multiple signals in the same set, but also greatly reduce the number of tests to be performed and thus reduce the multiple testing burden. Most of the existing set-based methods are for detecting genetic main effects, which means testing the association between a set of SNPs and a phenotype. Tzeng et al. [2011] provided a nice summary of those methods, which include burden tests that compute weighted sum of genotypes across markers [Gauderman et al., 2007; Li et al., 2009; Wang and Abbott, 2008; Wang and Elston, 2007], methods that exploit the pairwise genetic similarity among samples [Beckmann et al., 2005; Dempfle et al., 2007; Mukhopadhyay et al., 2010; Schaid et al., 2005; Tzeng et al., 2003, 2009; Wei et al., 2008; Wessel and Schork, 2006], variance component methods [Goeman et al., 2004; Kwee et al., 2008; Neale et al., 2011; Schaid 2010; Tzeng and Zhang, 2007; Wu et al., 2010], a method that combines *P*-values within a gene [Liu et al., 2010], group additive regression [Luan and Li, 2008], Tukey's model [Chatterjee et al., 2006], and an entropy-based method [Zhao et al., 2005]. Set-based methods have drawn more attention in the sequencing studies because of the rarity of the variants, for example, several variations of the burden tests [Han and Pan, 2010; Li and Leal, 2008, 2009; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Price et al., 2010] and variance component tests [Neale et al., 2011; Wu et al., 2011] have been proposed for sequencing data. In contrast, few methods have been proposed for set-based G × E tests. Tzeng et al. [2011] developed a method to test for interaction between a set of markers and an environment variable by extending the set-based genetic similarity method to the G × E setting [Tzeng et al., 2011]. As there is no competing method, they compared the new method with the benchmark minimum *P*-value method and their method showed favorable performance. However, their method was designed for a continuous outcome and cannot be applied to a case-control study for complex diseases.

A natural approach to developing a set-based G × E test is directly extending the set-based main effect test by treating the interaction term (usually the product of G and E) as a new genetic variable. For example, the existing burden test computes the (un)weighted sum of the genotypes (minor alleles counts) across SNPs in the set and test whether the sum is associated with the phenotype. A simple extension of burden tests to the G × E setting would be to sum the interaction terms (products) of G and E instead of summing over the G's alone. However, this kind of approach has several disadvantages. First, assumptions that are reasonable for main effects may not be reasonable for G × E, i.e., the power of burden tests for rare variants depends on the assumption that most rare missense variants are deleterious but it is not reasonable to assume all G × E's have the same direction. In addition, this simple extension fails to exploit some unique characteristics of G × E. For instance, one major difficulty in the set-based main effect test is the lack of prior information on which SNPs are null and what directions the effects are. In contrast, this valuable information can be partially obtained for interaction effect from established screening statistics for G × E tests.

To overcome the aforementioned drawbacks, we proposed a novel set-based gene-environment interaction (SBERIA) test for case-control studies. The proposed method uses the correlation between the environmental variable and the SNPs in a set as a guide to aggregate the genotypes. The aggregated genotype is then used to test for interaction in a regular logistic regression model. SBERIA is easy to implement and efficient in computation. It can be applied to both common and rare variants. We demonstrate through simulation that our proposed method is more powerful compared to the benchmark methods under a wide range of scenarios, including both genome-wide association studies (GWAS) and rare variant settings. We also applied SBERIA to real GWAS data and found evidence of interaction between the set of previously identified colorectal cancer (CRC) susceptibility loci and smoking.

Material and Methods

Notations and Models

Suppose there are N subjects and the disease status is denoted by D_i (= 0 or 1) for subject i, $i = 1, \ldots, N$. Assume E_i is the environmental variable, $\mathbf{X_i} = (X_{i1}, \ldots, X_{iq})$ is a vector of q potential confounder covariates, and $\mathbf{G_i} = (G_{i1}, \ldots, G_{ip})$ is a vector of p genetic markers. The interaction model between the set of p markers and the environmental variable is

$$\text{logit}(D_i) = \alpha_0 + \alpha_1 E_i + \mathbf{G_i}\alpha_2 + \mathbf{X_i}\alpha_3 + E_i\mathbf{G_i}\beta, \qquad (1)$$

where logit() is the logit link function; α_0 is the intercept; α_1 is the coefficient for the main effect of E_i; α_2 is the px1 vector of coefficients for $\mathbf{G_i}$;α_3 is the qx1 vector of coefficients for $\mathbf{X_i}$; $E_i\mathbf{G_i} = (E_iG_{i1}, \ldots, E_iG_{ip})$; $\beta = (\beta_1, \ldots, \beta_p)^T$ is the px1 vector of interaction coefficients. The null hypothesis for interaction effects is $H_0 : \beta = 0$.

Two Benchmark Methods

A typical method of testing $H_0 : \beta = 0$ is the likelihood ratio test, which compares the likelihood of models (1) with and without the interaction terms and then tests the hypothesis with a p degree of freedom (DF) chi-square test. We will refer this test as the LR test in the rest of the paper. A problem of the LR test in this case is that the relatively large number of markers or high LD among markers could result in numerical

instability, leading to inflated type I error, which we will show in the simulation.

Another commonly used method is the so-called minimum *P*-value (min-p) method. The min-p method tests interaction for each marker j in the set individually with the following model:

$$\text{logit}(D_i) = \gamma_0 + \gamma_1 E_i + \gamma_2 G_{ij} + \mathbf{X_i}\gamma_3 + \beta_j E_i G_{ij}, \quad (2)$$

and the hypothesis to be tested is $H_0 : \beta_j = 0$, for $j = 1, \ldots, p$. From the p interaction *P*-values for the p SNPs, the min-p method selects the smallest *P*-value and corrects it for multiple comparisons using permutation or by estimating the effective number of DF [Gao et al., 2008; Moskvina and Schmidt, 2008]. In our simulation, we will use 10,000 permutations to determine the corrected *P*-value for the min-p method. As we can see, the min-p method avoids the problem of potential large number of predictors in the LR test by modeling each marker individually instead of jointly. However, the min-p method is not efficient in situations where causal SNPs are in LD with multiple SNPs or when multiple independent signals exist in the set, as it only considers the minimum *P*-value.

The SBERIA Method

The main motivation for performing a set-based analysis is that aggregating signals of markers can potentially boost the power. However, as described in the Introduction, one difficulty in the signal aggregation is how to tell signals from noise and how to determine the direction of the signals. In the set-based main effect tests, there have been several attempts trying to solve this issue. Han and Pan [2010] used the signs of the marginal effect to determine the direction of the main effect [Han and Pan, 2010]. Lin and Tang [2011] used the corresponding regression coefficient plus a constant as the weight for each marker [Lin and Tang, 2011]. Cai et al. [2012] proposed to weight each marker based on the *z*-score of its effect [Cai et al., 2012]. One common characteristic of these methods is that the statistics used to weight the markers are not independent of the main effect test. Hence, permutation is needed to estimate the null distribution and maintain the correct type I error, which is computationally intensive. Fortunately for G × E, there are screening statistics that are informative for weighting the markers but still independent with the interaction test. Therefore, it would be appealing to take advantage of this desirable feature of the G × E test.

Correlation screening has been established as an efficient screening tool for the G × E test [Murcray et al., 2009]. Let us consider the following simple example to see the rationale of the correlation screening. Suppose there is a rare disease D, an environmental variable E (= 0 or 1), and a genetic variable G (= 0 or 1). G and E are assumed to be independent in controls (and because of the rarity of the disease, also approximately independent in general population). Assume there is a positive interaction between E and G such that the disease risk would only increase when both E = 1 and G = 1. Then we expect to see more E = 1 and G = 1 combinations in the cases, which means G and E will be positively correlated in the cases. As G and E are independent in controls, they will be also positively correlated in the combined case-control samples. On the other hand, if E and G impact D independently without interaction, it can be shown (supplementary material) that E and G are approximately when the disease is rare. From this simple example, we can see that the correlation between G and E combined case-control samples can be useful as a screening statistic for interaction between G and E. In addition, the direction of the correlation can inform the direction of the interaction. More importantly, as the correlation screening is conducted on the case-control combined samples and it does not use the phenotype information, it has been shown both by Murcray et al. [Murcray et al., 2009] and Dai et al. [Dai et al., 2012] that the correlation screening in combined case-control samples is asymptotically independent of the G × E test, no matter whether G and E are independent or not. This motivates us to propose the following method.

We first compute the correlation between E_i and G_{ij} ($j = 1, \ldots, p$) in (1) by either fitting a logistic regression (when E_i is binary) or a linear regression (when E_i is continuous) with E_i as the response and G_{ij} as the predictor. Then for each SNP j ($j = 1, \ldots, p$), we get a *Z*-score Z_j for the correlation between E_i and G_{ij}. Then we fit the following logistic regression:

$$\text{logit}(D_i) = \alpha_0 + \alpha_1 E_i + \mathbf{G_i}\alpha_2 + \mathbf{X_i}\alpha_3 + \rho E_i \mathbf{G_i}\hat{\mathbf{w}}, \quad (3)$$

where $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_p)^T$ is the weight vector and $\hat{w}_j = I(|Z_j| > \theta_N)\text{sign}(Z_j) + \varepsilon$. $I(x)$ is an indicator function, which equals 0 when x is false and 1 when x is true. $\text{sign}(x) = 1$ when $x > 0$, -1 when $x < 0$, and 0 when $x = 0$. $\theta_N = o(N^{1/2})$ and ε are prespecified positive constants. The hypothesis of interest is $H_0 : \rho = 0$.

As we can see, $E_i \mathbf{G_i}\hat{\mathbf{w}}$ is the weighted sum of the interaction terms and the weight, which can be 1, –1, or 0 (if we ignore ε), is determined by correlation *Z*-score Z_j. $|Z_j|$ measures the strength of the correlation signal so $I(|Z_j| > \theta_N)$ only selects markers showing correlation signals that are greater than a threshold. $\theta_N = o(N^{1/2})$ because we expected $I(|Z_j| > \theta_N)$ to converge to 0 as $N \to \infty$ when there is no correlation between G and E in the combined sample and converge to 1 when there is correlation. For the selected marker (markers with $I(|Z_j| > \theta_N) = 1$), the direction of the interaction term is determined by the direction of the correlation ($\text{sign}(Z_j)$). This is inspired by the observation that the directions of interaction and correlation tend to agree in the simple example above. The addition of a constant ε ensures that a weight will be assigned if no marker is selected.

θ_N and ε need to be specified for SBERIA. In practice, we found through simulation that the power of SBERIA did not change substantially as θ_N changes for a given N between 2,000 and 20,000 (results not shown). Hence in this paper, we set θ_N to a constant such that $\text{Prob}(|Z_j| > \theta_N) \approx 0.1$ under the null. ε is set to a very small value (0.0001) so that it does not affect the weight if $I(|Z_j| > \theta_N)\text{sign}(Z_j)$ is not 0.

In summary, SBERIA first selects markers of which the correlation signal strength is greater than a threshold. For

the selected markers, we compute a weighted sum of their interaction terms, where the weight = 1 if the corresponding correlation is positive and –1 otherwise. As the correlation statistic is independent of the interaction test, regular logistic regression can be used to test the hypothesis without requiring permutation. The validity of our method is proved in the supplementary material. We also conduct extensive simulation to evaluate the type I error rate and power of SBERIA.

Simulation

To evaluate the performance of SBERIA, we conducted extensive simulation under various settings.

Set-Based G × E in GWAS settings

A Gene-Based Marker Set

We mimicked the real GWAS data by generating a set of markers based on the realistic LD structure within the *SMAD7* gene. *SMAD7*, short for *SMAD* family member 7, is a gene located at 18q21.1. It is known to interact with the TGF-beta receptor and several SNPs in this region have been found to associate with CRC risk [Broderick et al., 2007; Peters et al., 2011; Tenesa et al., 2008; Tomlinson et al., 2008]. *SMAD* spans from 44,700k bp to 44,731k bp and has 48 SNPs from Hapmap II release 24 [The International HapMap Project 2003], which is close to the median number (= 43) of SNPs per gene [Huang et al., 2011]. Out of the 48 SNPs, 21 were genotyped in Illumina Human1M. We extracted the haplotypes of the 21 SNPs from the phased Hapmap data and randomly paired haplotypes such that the simulated marker set maintains the same LD structure as the 21 SNPs in the Hapmap. The LD structure of the 21 SNPs is shown in supplementary Figure S1. We chose two SNPs rs4939827 and rs7351039 from the 21 SNPs and make them the hidden causal SNPs in the simulation. The two SNPs were chosen such that one is common (rs4939827, MAF = 0.49) and one is less common (rs7351039, MAF = 0.08). The two chosen SNPs are not in LD with each other and both SNPs were tagged by some other SNPs. The other 19 SNPs were considered as the marker set in the simulation.

The disease status was generated based on the following model:

$$\text{logit}(D_i) = \alpha_0 + \gamma E_i + \alpha_1 G_{i1} + \alpha_2 G_{i2} + \beta_1 E_i G_{i1} + \beta_2 E_i G_{i2}, \quad (4)$$

where $\alpha_0 = \exp(-5)$, representing a relatively rare disease. G_{i1} and G_{i2} are the simulated genotypes (= 0, 1, or 2) for rs4939827 and rs7351039, respectively. E_i is the environmental variable. We tried two ways of generating E_i: (1) E_i is continuous: $E_i \sim N(0, 1)$; (2) E_i is binary: $E_i \sim$ Bernoulli(p = 0.3).

Type I error. To evaluate the type I error rate, we set $\beta_1 = \beta_2 = 0$ in (4). We let $\alpha_1 = \alpha_2 = 0$ or log(1.5). As described above, we used the four different ways to generate E_i. For each simulation scenario, we randomly generated 1,000 cases and 1,000 controls. Then we performed the set-based G × E tests using the LR test, the min-p method and SBERIA. The procedure was repeated 2,000 times to estimate the type I error rate with significance level 0.05.

Power. To evaluate the power, we set β_1 = log(1.05), log(1.10), log(1.15), log(1.20), log(1.25), or log(1.3) when E_i is continuous and β_1 = log(1.1), log(1.2), log(1.3), log(1.4), log(1.5), or log(1.6) when E_i is binary. The values of β_1 were chosen such that the power was in a reasonable range. For each value of β_1, β_2 can take three values β_1, $-\beta_1$, or 0, which represents situations where two signals are in the same direction, in the different direction, or when there is only one signal, respectively. The main effects α_1 and α_2 were set to 0. We also tried other values for the main effects and the results were quantitatively similar. Same as above, we randomly generated 1,000 cases and 1,000 controls. We evaluated SBERIA, the min-p method and the LR test for the power performance. Each parameter setting for the simulation was repeated 2,000 times and we used significance level 0.05.

A Set of Independent Markers

In the simulation above, the 21 SNPs in the set were not independent with each other as they were generated based on the LD structure in the *SMAD7* gene. In addition to grouping SNPs by genes, there are other ways of forming a marker set in practice. For example, it is common practice to pull together previously identified susceptibility loci for a given trait and study them as a set. To mimic this situation, we generated 20 independent SNPs. For each SNP, its MAF is generated from uniform distribution $U(0.1, 0.5)$ under Hardy-Weinberg equilibrium. We randomly chose two SNPs as potential causal SNP for G × E. The disease status was generated based on the following model:

$$\text{logit}(D_i) = \alpha_0 + \gamma E_i + \sum_{j=1}^{20} \alpha_j G_{ij} + \beta_1 E_i G_{i1} + \beta_2 E_i G_{i2}, \quad (5)$$

where G_{i1} and G_{i2} are the genotypes for two chosen causal G × E SNPs. The main effects α_j's were generated from $U(\log(1.05), \log(1.5))$.

A wide adopted way of summarizing information from previously identified susceptibility loci is to calculate the genetic risk score (GRS), which is the sum of risk alleles from all SNPs. Hence in this simulation scenario, we also tried to perform the set-based G × E test by computing GRS and test the interaction between GRS and E using a regular logistic regression. The same parameters and procedures as the first simulation scenario were used to evaluate type I error and power for SBERIA, the min-p method, LR test, and the GRS method.

Correlated G and E

G and E were assumed to be independent in the simulation so far, which is a reasonable assumption in real applications [Cornelis et al., 2012]. However, in rare situations, G and E can be correlated in the general population.

As shown in Murcray et al. [2011] and Hsu et al. [2012], the correlation screening is not efficient when G and E are negatively correlated. Hence, they proposed to use some combinations of correlation screening and marginal screening (which uses the marginal association test of each SNP as a screening for interaction test). In the current simulation scenario, we also tried a simple modification to SBERIA that combines correlation and marginal screenings in way similar to Gauderman et al. [2012]. Specifically, instead of using $\hat{w}_j = I(|Z_j| > \theta_N)\text{sign}(Z_j) + \varepsilon$ in (3), we define

$$\hat{w}_j = I(S_j > \tau_N)\text{sign}(C_j) + \varepsilon, \quad (6)$$

where $S_j = Z_j^2 + M_j^2$ and M_j is the Wald statistic of the marginal association for marker j ($j = 1, \ldots, p$); $C_j = Z_j$ if $Z_j^2 > M_j^2$ else $C_j = M_j$. τ_N is also defined such that $\text{Prob}(S_j > \tau_N) = 0.1$ under the null.

The same settings were used as the first simulation scenario when $\beta_1 = \beta_2$, except that E_i was generated to be correlated with G. We considered two scenarios:

1. E is correlated with the two causal SNPs. In this setting, E_i is either positively correlated with G_{i1} and G_{i2}: $\text{logit}(E_i) = \text{logit}(0.3) + b_1 G_{i1} + b_2 G_{i2}$, where $b_1 = b_2 = \log(1.2)$ or E_i is negatively correlated with G_{i1} and positively correlated with G_{i2} ($b_1 = -b_2 = -\log(1.2)$).
2. E is correlated with two random selected null SNPs. Similar as above, E_i can also be positively or negatively correlated with the two null SNPs.

Same procedures as before were used to evaluate the type I error and power of SBERIA, the min-p method, LR test, and the modification to SBERIA.

Set-Based G × E in Rare Variant Setting

We also conducted simulations to evaluate the performance of SBERIA if the variants in the marker set are less common, as in sequencing data. In the simulation experiment, we followed the simulation setup proposed in Lin and Tang [2011] to generate the genotypes for rare variants [Lin and Tang, 2011]. Specifically, we generated 10 variants G_{ij} ($j = 1, \ldots, 10$) with MAF = $0.005 \times j$ under Hardy-Weinberg equilibrium. As it is less likely for rare variants to correlate with the environmental variable, we generated E_i either as a continuous variable from $N(0,1)$ or a binary environmental variable E_i from Bernoulli(0.3).

The disease status was generated from the following model:

$$\text{logit}(D_i) = \alpha_0 + \gamma E_i + \sum_{j=1}^{10} \alpha_j G_{ij} + \sum_{j=1}^{10} \beta_j E_i G_{ij}, \quad (7)$$

where α_0 is set to $\exp(-5)$ and γ is set to be $\log(1.2)$ as in the GWAS simulation. As there is no competing set-based G × E method in the rare variant setting, in addition to min-p and LR test, we decided to compare SBERIA with the simple extension of the burden test as described in the Introduction. We will denote this method as burden G × E. Specifically, burden G × E creates a new variable $G_i = \sum_{j=1}^{10} G_{ij}$, which is the total number of minor alleles across the 10 rare variants. Then it tests the interaction by fitting the following model and tests $H_0 : \lambda = 0$:

$$\text{logit}(D_i) = \alpha_0 + \gamma E_i + \sum_{j=1}^{10} \alpha_j G_{ij} + \lambda E_i G_i \quad (8)$$

Type I error. To evaluate the type I error rate, the coefficients β_j's ($j = 1, \ldots, 10$) were set to 0. We randomly generated α_j's from a uniform distribution $U(\log(1.2), \log(3))$. As before, we randomly sampled 1,000 cases and 1,000 controls. The procedures were replicated 2,000 times to estimate the type I error rate for SBERIA, min-p, LR test, and burden G × E with significance level 0.05.

Power. To evaluate the power, we first randomly selected m (= 8, 5, or 2) markers as the causal variants from the 10 variants. Then we randomly generated the effect size of the selected variants from $U(\log(1.2)c, \log(3)c)$. As the sample size of our simulation is only 2,000, c was chosen to be 1.5 such that the power was in a reasonable range. In practice, the sample size should be much larger to study rare variant. As we can see, in this way all effects are positive, which may not be realistic in G × E setting. Hence, we randomly set the direction of the interaction effect for a subset of causal SNPs to negative (proportion = 0.2, 0.4, or 0.5). One thousand cases and 1,000 controls were generated and the power was estimated from 2,000 replications with significance level 0.05. We only presented the results from the binary E_i's because the results from continuous E_i's were similar.

A Real Data Application

To evaluate the performance of SBERIA in real application, we applied SBERIA to the GWAS data of Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). Specifically, GECCO included the following nested case-control studies in prospective US cohorts Health Professionals Follow-up Study (HPFS); Multiethnic Cohort Study (MEC); Nurses' Health Study (NHS); Physician's Health Study (PHS); Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO); VITamins And Lifestyle (VITAL); Woman's Health Initiative (WHI); and the following case-control studies from the US, Canada, and Europe [Colorectal Cancer Studies 2&3 (Colo2&3); Darmkrebs: Chancen der Verhuetung durch Screening (DACHS); Diet, Activity and Lifestyle Survey (DALS); Ontario Familial Colorectal Cancer Registry (OFCCR); and Postmenopausal Hormone Study-Colon Cancer Family Registry (PMH-CCFR). Numbers of cases and controls, age, and sex distributions are listed in supplementary Table S1. Study-specific descriptions, including eligibility and matching criteria, is available in Peters et al. [2013]. CRC cases were defined as colorectal adenocarcinoma and confirmed by medical records, pathology reports, or death certificates. Colorectal adenoma cases were confirmed by medical records, histopathology, or pathologic reports. Controls for adenoma cases had a negative colonoscopy (except for NHS and HPFS controls matched to cases with distal adenoma, which either had a negative sigmoidoscopy

Table 1. Previously identified CRC susceptibility loci

SNP	Ref[a]	Chromosome	Count allele	CAF[b]	OR[c] (95% CI[d])
rs6691170	Houlston et al. [2010]	1q41	G	0.63	0.94 (0.92–0.97)
rs6687758	Houlston et al. [2010]	1q41	A	0.80	0.92 (0.89–0.94)
rs10936599	Houlston et al. [2010]	3q26.2	C	0.77	1.08 (1.04–1.10)
rs1321311	Dunlop et al. [2012]	6p21	A	0.25	1.10 (1.07–1.13)
rs16892766	Tomlinson et al. [2008]	8q23.3	A	0.92	0.80 (0.76–0.84)
rs10505477	Zanke et al. [2007]	8q24	A	0.50	1.17 (1.12–1.23)
rs6983267	Tomlinson et al. [2007], Zanke et al. [2007], Haiman et al. [2007], Hutter et al. [2010]	8q24	G	0.50	1.21 (1.18–1.24)
rs7014346	Tenesa et al. [2008]	8q24	A	0.36	1.19 (1.15–1.23)
rs719725	Zanke et al. [2007], Kocarnik et al. [2010]	9p24	A	0.62	1.07 (1.03–1.12)
rs10795668	Tomlinson et al. [2008]	10p14	A	0.31	0.89 (0.86–0.91)
rs3824999	Dunlop et al. [2012]	11q13.4	G	0.51	1.08 (1.05–1.10)
rs3802842	Tenesa et al. [2008]	11q23	A	0.71	0.90 (0.87–0.93)
rs7136702	Houlston et al. [2010]	12q13.13	C	0.68	0.94 (0.93–0.96)
rs11169552	Houlston et al. [2010]	12q13.13	C	0.73	1.09 (1.05–1.11)
rs4444235	Tomlinson et al. [2011], Houlston et al. [2008]	14q22.2	C	0.46	1.09 (1.06–1.12)
rs1957636	Tomlinson et al. [2011]	14q22.2	C	0.59	0.92 (0.90–0.95)
rs16969681	Tomlinson et al. [2011]	15q13	C	0.91	0.84 (0.80–0.90)
rs4779584	Tomlinson et al. [2011], Jaeger et al. [2008]	15q13	C	0.82	0.87 (0.84–0.91)
rs11632715	Tomlinson et al. [2011]	15q13	A	0.48	1.12 (1.08–1.16)
rs9929218	Houlston et al. [2008]	16q22.1	A	0.30	0.91 (0.89–0.94)
rs4939827	Tenesa et al. [2008], Broderick et al. [2007]	18q21	C	0.48	0.83 (0.81–0.86)
rs10411210	Houlston et al. [2008]	19q13.1	C	0.90	1.15 (1.10–1.20)
rs961253	Tomlinson et al. [2011], Houlston et al. [2008]	20p12.3	A	0.36	1.12 (1.09–1.15)
rs4813802	Tomlinson et al. [2011], Peters et al. [2011]	20p12.3	G	0.34	1.09 (1.06–1.12)
rs4925386	Houlston et al. [2010], Peters et al. [2011]	20q13.33	C	0.69	1.08 (1.05–1.10)

[a] Ref, references for identifying allele, and for ORs presented.
[b] CAF, count allele frequency in European decent populations.
[c] OR, odds ratio.
[d] CI, confidence interval.
Only the first reference's OR of the SNPs with more than one reference is shown in the table. The same situation applies for the Studies in Previous Publications column.

or colonoscopy exam). All participants gave written informed consent and studies were approved by their respective Institutional Review Boards. Genotyping were done on various platforms and imputed to Hapmap II. Please see a detailed description of genotyping, quality control, and imputation in GECCO in the supplementary material.

A number of loci have been identified to associate with CRC risk [Broderick et al., 2007; Dunlop et al., 2012; Houlston et al., 2008, 2010; Jaeger et al., 2008; Peters et al., 2011; Tenesa et al., 2008; Tomlinson et al., 2007, 2008, 2011; Zanke et al., 2007]. These CRC susceptibility loci are useful for genetic risk profiling and allow the stratification of population subgroups at different genetic risks [Lubbe et al., 2012]. To get a more comprehensive understanding of CRC risk, it is also of interest to explore possible interactions between the genetic risk factors and environmental variables. In this paper, we included the genotypes of 25 known CRC loci (Table 1) in GECCO and treated them as a marker set. We then tested for interaction between this marker set and smoking status (ever/never). Smoking status is a dichotomous variable harmonized across all studies. Please see the supplementary material for details of the harmonization procedure.

Specifically, we created a pooled dataset of 10,729 cases and 13,328 controls for the 25 known CRC loci by combining the studies in GECCO. Each directly genotyped SNP was coded as 0, 1, or 2 copies of the variant allele. For imputed SNPs, we used the expected number of copies of the variant allele (the "dosage"). Both genotyped and imputed SNPs are treated as continuous variable (i.e., log-additive effects). We then applied SBERIA in (3) to the pooled dataset. The covariates $\mathbf{X}$ we adjusted for include age, sex, the first three principle components, study indicators, and the interaction between principle components and study indicator. As a comparison, we also tried two possible benchmark methods: the min-p method, which computes the interaction *P*-value for each of the 25 SNPs separately and selects the minimum *P*-value while correcting for multiple comparisons using the Bonferroni method; the second alternative method is to compute GRS and test the interaction between GRS and smoking status using a regular logistic regression.

Results

Set-Based G × E in GWAS Settings

A Gene-Based Marker Set

The estimated type I error for SBERIA, min-p, and the LR test are summarized in Table 2. It can be seen that both SBERIA and the min-p method always maintain the correct type I error (0.05). However, the LR test generally gives inflated type I error, which is a result of its numerical instability due to the relatively large number of variables. Figure 1 shows the power comparison results for this simulation scenario. It can be seen that when $\beta_1 = \beta_2$, SBERIA has better power than both min-p and the LR test. The average power gain of SBERIA over min-p is 13.9% with a range of –6% to 24.3% (excluding data points where the power of min-p is less than 0.1 to prevent numeric instability). Also as expected, SBERIA is still more powerful than the min-p method (average percent of

Table 2. Type I error rate (95% CI) for SBERIA, min-p, and LR test in simulation scenario 1 (a gene-based marker set) in GWAS settings

	SBERIA	Min-p	LR test
E_i is continuous and independent of G_{i1} and G_{i2}			
$\alpha_1 = \alpha_2 = 0$	0.044 (0.035 0.052)	0.045 (0.036 0.054)	0.061 (0.051 0.071)
$\alpha_1 = \alpha_2 = \log(1.5)$	0.045 (0.036 0.054)	0.042 (0.033 0.05)	0.062 (0.052 0.073)
E_i is binary and independent of G_{i1} and G_{i2}			
$\alpha_1 = \alpha_2 = 0$	0.049 (0.04 0.058)	0.046 (0.037 0.055)	0.070 (0.059 0.081)
$\alpha_1 = \alpha_2 = \log(1.5)$	0.042 (0.033 0.051)	0.046 (0.036 0.055)	0.063 (0.052 0.074)

power gain is 12.5% with a range from 5.2% to 19.5%) when the two causal SNPs have interaction effects in opposite directions ($\beta_1 = -\beta_2$), which demonstrates that the correlation screening is able to predict the direction of interaction effect fairly well. With inflated type I error, the LR test still only gives power that was close to or less than SBERIA. For the scenario where there was only one causal SNP ($\beta_2 = 0$), SBERIA still performs better than the other two methods (average percent of power gain over min-p is 11.5% with a range from 0.4% to 18.2%). This could be attributed to the fact that SBERIA aggregates information from several LD SNPs of the causal variant and thereby increases the power. The advantage of SBERIA is more apparent if one considers the fact that the min-p method requires often time-consuming permutation to get the corrected *P*-value (otherwise the simple Bonferroni correction using the number of markers in the set would be too conservative).

A Set of Independent Markers

The type I error for this simulation scenario was summarized in Table 3. All except the LR test maintain the correct type I error. From Figure 2, it can be seen that SBERIA almost always gives the best power. When $\beta_1 = \beta_2$, the average percent of power gain of SBERIA over min-p is 24.8% with a range from 8.2% to 49.0%; when $\beta_1 = -\beta_2$, the average percent of power gain of SBERIA over min-p is 15.9% with a range from 3.9% to 27.6%; when $\beta_2 = 0$, the average percent of power gain of SBERIA over min-p is 10.7% with a range from –8.6% to 25.2%. The GRS method always gives the lowest power.

Correlated G and E

From Table 4, it can be seen that only LR test gives inflated type I error. As SBERIA uses the correlation between G and E in case-control combined samples as the screening tool, it is expected that the power of SBERIA would be impacted if gene-environment correlation exists in the general population. The top left plot of Figure 3 show that the power of SBERIA is further boosted if the gene-environment correlations in the general population are positive for both causal variants. As shown in the top right plot of Figure 3, the power of SBERIA drops if the correlation in the general population is in a different direction compared with the interaction, which is in line with expectation. As expected, the simple modification of SBERIA shows a desirable performance in this case. Compared with the unmodified version, it has almost the same magnitude of power gain when the correlations are positive and have little power loss when the correlation is in a different direction compared to the interaction. If there are correlation between null SNPs and E, the two plots on the bottom of Figure 3 shows that the power advantage of SBERIA and SBERIA-M is reduced, which is expected because the correlation between null SNPs and E would make the null SNPs more likely to be selected and therefore dilute the interaction signal. It is worth noting, however, that gene-environment correlation in population is relatively rare in real applications [Cornelis et al., 2012].

Set-Based G × E in Rare Variant Setting

From Table 5, it can be seen that both SBERIA and the burden G × E test maintain the correct type I error. However, the min-p method seems to be conservative, which could be due to the rarity of the SNPs. On the other hand, the LR test is highly inflated. Figure 4 shows the power comparison between various methods. LR test always has the best power, however, given its highly inflated type I error, it is not applicable in practice. It can be seen that SBERIA is always more powerful than the min-p and burden G × E method in the simulation. The advantage of SBERIA is most obvious when around half of the causal loci have negative interaction with E and the others have positive interaction. Again, this shows that correlation screening did a good job informing us about the direction of interaction effects.

A Real Data Application

The results for testing for interaction between the known CRC marker set and smoking status using GECCO GWAS data are summarized in Table 6. It can be seen that SBERIA reaches the significance level 0.05 and the GRS method also gives a *P*-value close to the significance level. Hence, there is evidence that the genetic risk of CRC is interacting with the smoking status. On the other hand, SBERIA gives a more significant *P*-value compared to the min-p and the GRS method, which demonstrates the potential advantage of SBERIA. In addition, when exploring which SNPs contribute to the interaction signal in the marker set, we found that rs10936599 shows the strongest evidence—it was selected by the correlation screening of SBERIA and it also has the smallest interaction *P*-value in min-p.

Discussion

In this paper, we proposed a novel method to test for interaction between a set of markers and an environmental variable in case-control studies. SBERIA takes advantage of the unique features of G × E test by using the correlation screening to inform the aggregation of interaction effects

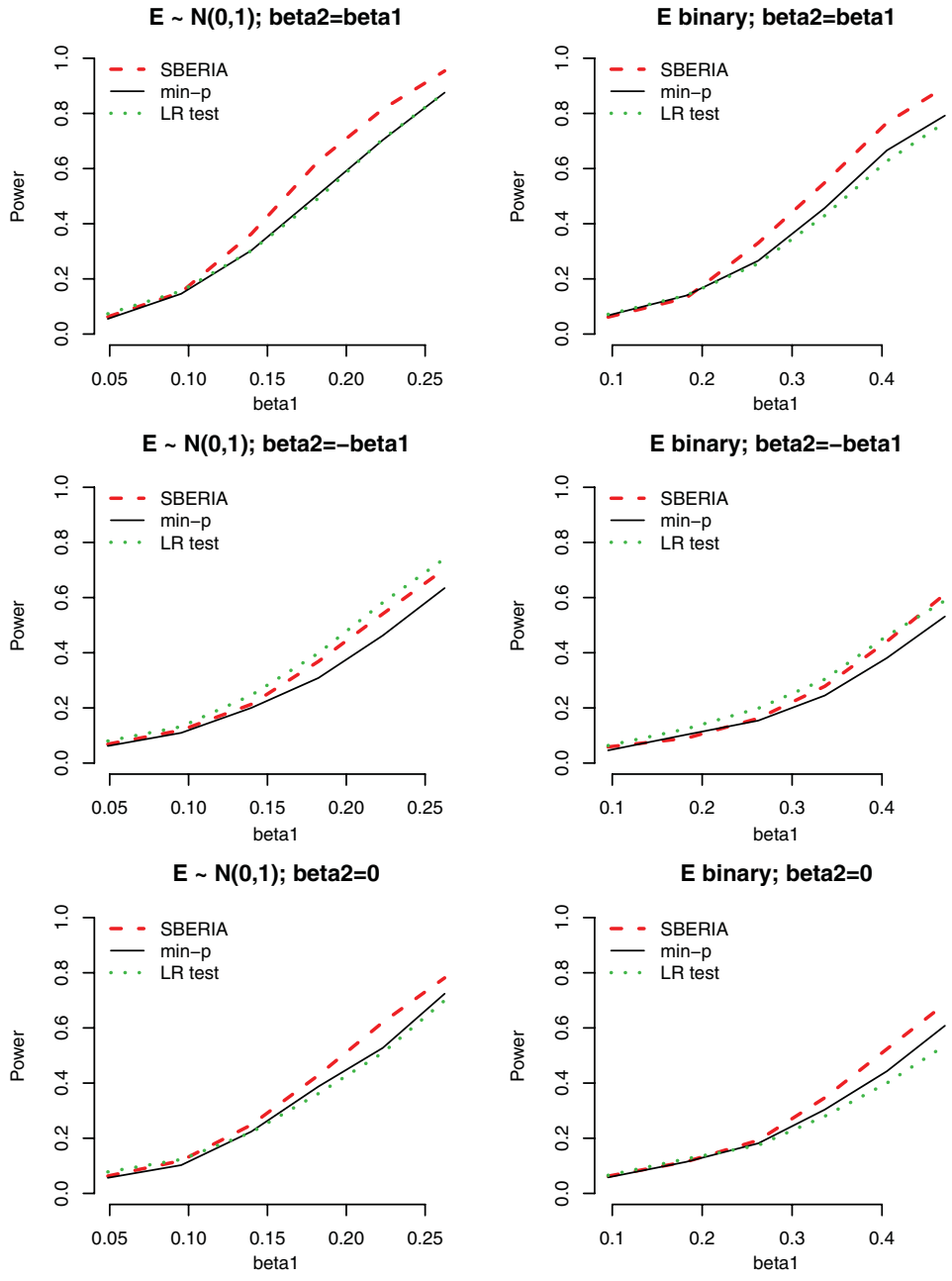

Figure 1. Power comparison between SBERIA, the min-p method, and the LR test in simulation scenario 1 (a gene-based marker set) of GWAS settings. The three plots on the left are results when E_i was generated as continuous variable and the plots on the right are for binary E_i's. The top plots are for simulation scenarios where $\beta_1 = \beta_2$; the plots in the middle are for scenarios where $\beta_1 = -\beta_2$; the bottom plots are for scenarios where $\beta_2 = 0$.

Table 3. Type I error rate (95% CI) for SBERIA, min-p, LR test, and GRS method in simulation scenario 2 (a set of independent markers) in GWAS settings

SBERIA	Min-p	LR test	GRS test
E_i is continuous and independent of G_{i1} and G_{i2}			
0.044 (0.035 0.054)	0.042 (0.034 0.051)	0.059 (0.049 0.069)	0.044 (0.035 0.053)
E_i is binary and independent of G_{i1} and G_{i2}			
0.050 (0.040 0.059)	0.054 (0.044 0.063)	0.060 (0.049 0.070)	0.050 (0.040 0.059)

within the marker set. Because the correlation screening in combined case-control samples is independent of the interaction test, SBERIA maintains the correct type I error without requiring permutation. SBERIA uses the regular logistic regression model so it is computationally efficient and easy to be implemented. We showed that SBERIA has appealing power compared with the benchmark methods in both GWAS and rare variant settings.

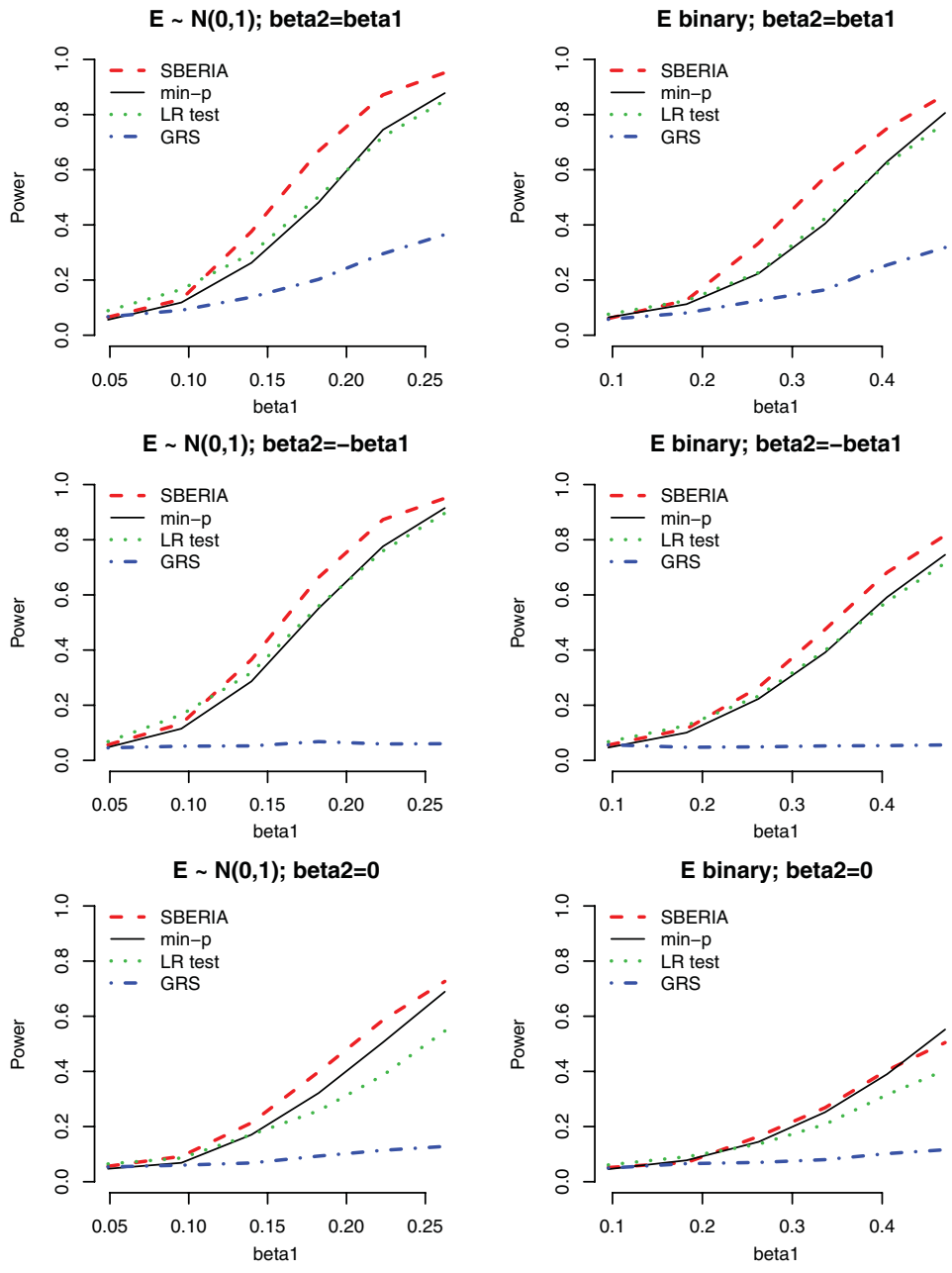

Figure 2. Power comparison between SBERIA, the min-p method, the LR test, and the GRS test in simulation scenario 2 (a set of independent markers) of GWAS settings. The three plots on the left are results when E_i was generated as continuous variable and the plots on the right are for binary E_i's. The top plots are for simulation scenarios where $\beta_1 = \beta_2$; the plots in the middle are for scenarios where $\beta_1 = -\beta_2$; the bottom plots are for scenarios where $\beta_2 = 0$.

Table 4. Type I error rate (95% CI) for SBERIA, min-p, LR test, and the modification of SBERIA in simulation scenario 3 (correlated G and E) in GWAS settings

	SBERIA	Min-p	LR test	SBERIA-modified
E_i is positively correlated with G_{i1} and G_{i2}				
$\alpha_1 = \alpha_2 = 0$	0.052 (0.042 0.062)	0.042 (0.033 0.05)	0.061 (0.051 0.071)	0.054 (0.044 0.064)
$\alpha_1 = \alpha_2 = \log(1.5)$	0.050 (0.040 0.059)	0.050 (0.040 0.059)	0.063 (0.052 0.074)	0.048 (0.038 0.057)
E_i is negatively correlated with G_{i1} and positively correlated with G_{i2}				
$\alpha_1 = \alpha_2 = 0$	0.058 (0.048 0.069)	0.046 (0.036 0.055)	0.062 (0.052 0.073)	0.057 (0.047 0.067)
$\alpha_1 = \alpha_2 = \log(1.5)$	0.044 (0.035 0.054)	0.046 (0.037 0.055)	0.065 (0.054 0.076)	0.052 (0.042 0.062)

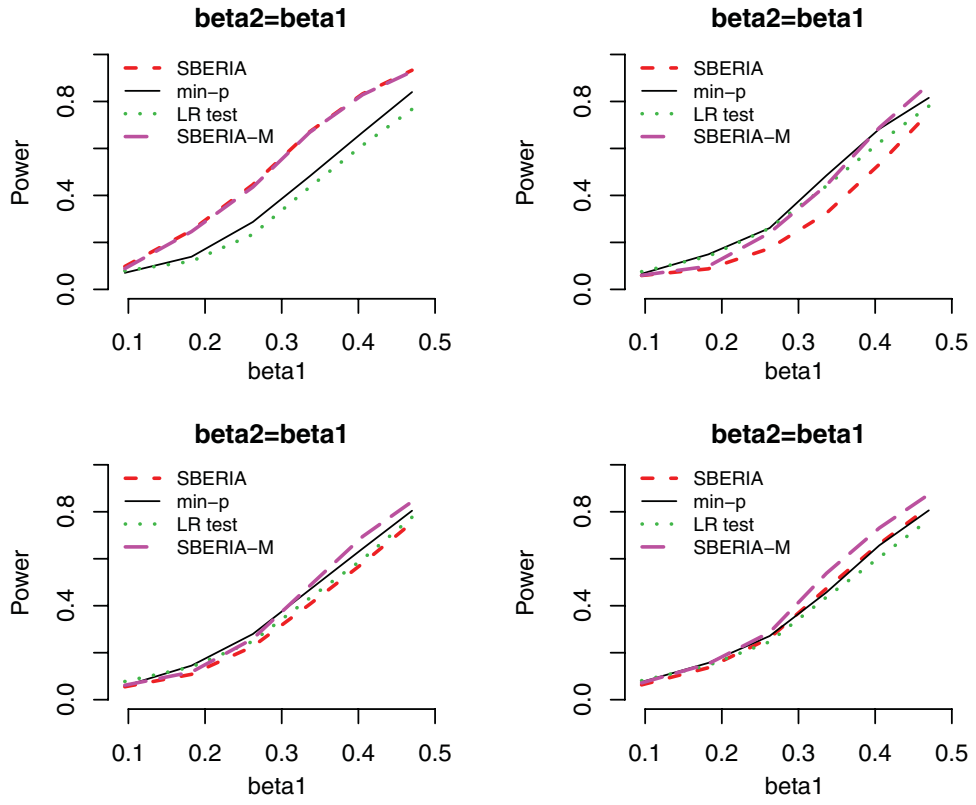

Figure 3. Power comparison between SBERIA, the min-p method, the LR test, and SBERIA-M, the modification to SBERIA (as defined in equation 6, in simulation scenario 3 (E correlated with G) of GWAS settings. The two plots on the top are results when E_i was correlated with two causal SNPs G_{i1} and G_{i2} and the two plots on the bottom are results when E_i was correlated with two randomly selected null SNPs. The plots on the left are for scenarios where E_i is positively with both SNPs and the plots on the right are for scenarios where E_i is positively correlated with one SNP and negatively correlated with the other.

Table 5. Type I error rate (95% CI) for SBERIA and burden G × E in rare variant settings

	SBERIA	Burden G × E	Min-p	LR test
$E_i \sim N(0,1)$	0.045 (0.036 0.054)	0.050 (0.040 0.060)	0.031 (0.023 0.039)	0.085 (0.072 0.097)
$E_i \sim$ Bernoulli(0.3)	0.046 (0.037 0.055)	0.051 (0.041 0.061)	0.031 (0.023 0.039)	0.095 (0.082 0.108)

While applying SBERIA to real data, we found evidence of interaction between genetic risk and smoking status for CRC. rs10936599, the SNP showing the strongest signal, is located at 3q26.2 in the *MYNN* gene. *MYNN* encodes a zinc finger domain-containing protein family, which is involved in the control of gene expression. Given that the function of *MYNN* is largely unknown so far, further functional characterization is needed in order to evaluate and interpret this potential interaction. In the real data application, we included the advanced colorectal adenomas because they are well-known precursor lesions of CRC. As a result, this improves our statistical power to identify G × E that act early in the adenoma-cancer sequence, where adenomas and cancer have a shared etiology. We recognize that the adenoma cases will not show signals for G × E's that act later in the carcinogenic process (i.e., on progression from adenoma to cancer) or G × E's that act through adenoma independent pathways.

There are several possible improvements that can be made to SBERIA. First, we chose θ_N such that it corresponds to *P*-value cutoff 0.1. We also tried other *P*-value cutoffs such as 0.05 and 0.2 in the simulation and the power of SBERIA does not change substantially (results not shown). However, it should be noted that the minor allele frequency affects the power of the correlation screening, and the SNPs with larger MAF will be more likely to pass the screening compared to less common SNPs. Hence, it is of interest to let the threshold vary with MAF. More work should be done to find an optimal θ_N. In addition, the current weighting of SBERIA is either 1, –1, or 0. Further work should explore whether the use of more advanced weight, such as the effect size of the correlation screening or the main effect, would increase power. In SBERIA, the main effect is modeled separately for each SNP in the set. It would be interesting to model main effects also in a set-based manner, which could potentially increase power.

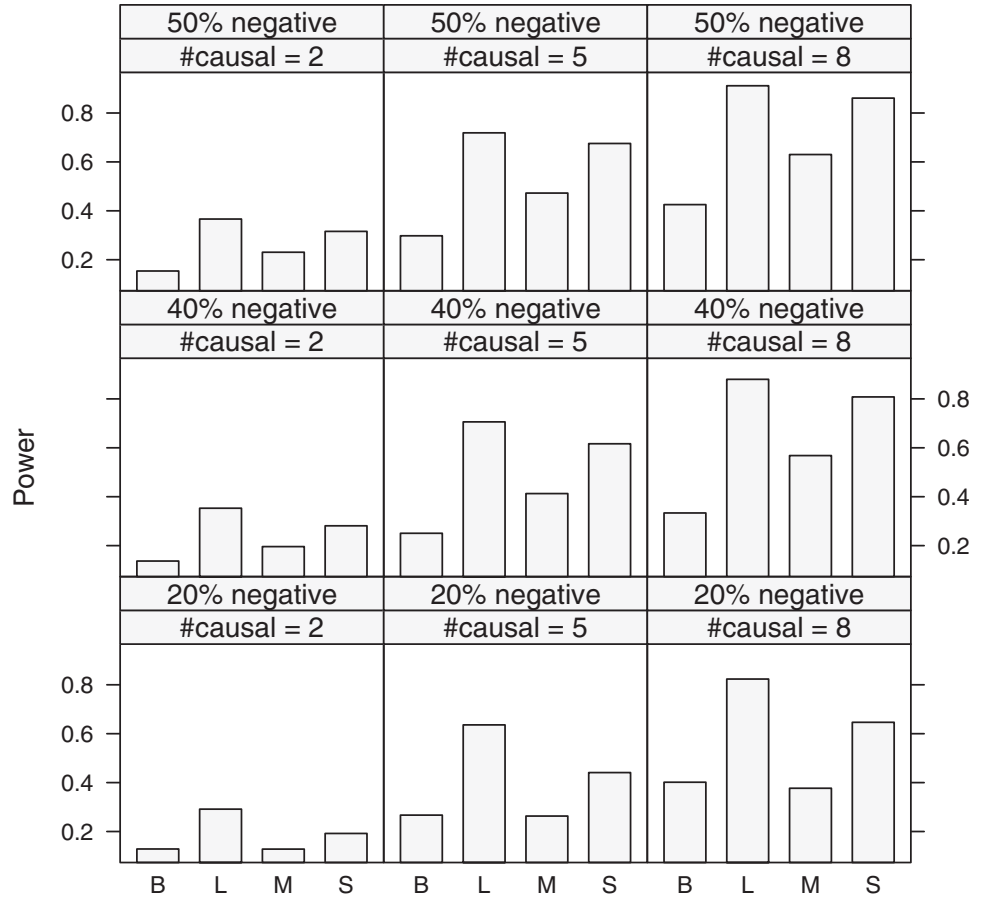

Figure 4. Power comparisons between SBERIA (S), min-p (M), LR test (L), and the burden G × E method (B) for different simulation scenarios in rare variant settings. The results are categorized by combinations of the number of causal variants and the proportion that a causal variant has a negative interaction effect.

Table 6. The results for testing interaction between the known CRC loci marker set and smoking status using different methods

	SBERIA	Min-p	GRS
P-value	5.92×10^{-3}	0.28	5.41×10^{-2}

Furthermore, more sophisticated methods can be built upon the framework of our method. For example, SBERIA drops the markers that are not selected based on screening. However, as the screening is not perfect, those SNPs can still contain useful information. Hence, it could potentially increase power to apply the traditional method (i.e., variance component based method) to the unselected SNPs and combine the results from the selected and unselected SNPs. SBERIA uses the correlation screening to combine SNPs in case-control studies. The strength of the correlation screening is mainly driven by the correlation between G and E in cases when there is G × E interaction. Hence, it is expected that if there are much more controls than cases, the correlation signal will be weakened and the power of correlation screening will be reduced.

In summary, SBERIA shows a promising performance both in simulation and real data application. With its easy implementation and fast computation time, SBERIA provides an attractive approach to detecting set-based gene-environment interactions.

Acknowledgments

GECCO: National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088; R01 CA059045).

ASTERISK: a Hospital Clinical Research Program (PHRC) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC).

COLO2&3: National Institutes of Health (R01 CA60987).

DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6–1, BR 1704/6–3, BR 1704/6–4, and CH 117/1–1), and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814).

DALS: National Institutes of Health (R01 CA48998 to M.L.S).

Guangzhou-1: National Key Scientific and Technological Project—2011ZX09307–001–04 and the National Basic Research Program—2011CB504303, People's Republic of China.

HPFS is supported by the National Institutes of Health (P01 CA 055075, UM1 CA167552, R01 137178, and P50 CA 127003), NHS by the National

Institutes of Health (R01 137178, P01 CA 087969, and P50 CA 127003), and PHS by the National Institutes of Health (CA42182).

MEC: National Institutes of Health (R37 CA54281, P01 CA033619, and R01 CA63464).

OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783); see CCFR section below. OFCCR is supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research, and the Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute. Thomas J. Hudson and Brent W. Zanke are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Economic Development and Innovation.

PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Control samples were genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan, supported by the Intramural Research Program of the National Cancer Institute. The datasets used in this analysis were accessed with appropriate approval through the dbGaP online resource (http://www.cgems.cancer.gov/data_acess.html) through dbGaP accession number 000207v.1p1.c1 (National Cancer Institute (2009) Cancer Genetic Markers of Susceptibility (CGEMS) data website http://cgems.cancer.gov/data_access.html; Yeager et al., 2007). Control samples were also genotyped as part of the GWAS of Lung Cancer and Smoking [Landi et al., 2009]. Funding for this work was provided through the National Institutes of Health, Genes, Environment and Health Initiative [NIH GEI] (Z01 CP 010200). The human subjects participating in the GWAS are derived from the Prostate, Lung, Colon and Ovarian Screening Trial and the study is supported by intramural resources of the National Cancer Institute. Assistance with genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01 HG 004438). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000093.

PMH: National Institutes of Health (R01 CA076366 to P.A.N.).

VITAL: National Institutes of Health (K05 CA154337).

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

ASTERISK: We are very grateful to Dr. Bruno Buecher without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians, and students.

DACHS: We thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benscheid, Muhabbet Celik, and Ursula Eilber for excellent technical assistance.

GECCO: The authors would like to thank all those at the GECCO Coordinating Center for helping bring together the data and people that made this project possible.

HPFS, NHS, and PHS: We would like to acknowledge Patrice Soule and Hardeep Ranu of the Dana Farber Harvard Cancer Center High-Throughput Polymorphism Core who assisted in the genotyping for NHS, HPFS, and PHS under the supervision of Dr. Immaculata Devivo and Dr. David Hunter, Qin (Carolyn) Guo and Lixue Zhu who assisted in programming for NHS and HPFS, and Haiyan Zhang who assisted in programming for the PHS. We would like to thank the participants and staff of the Nurses' Health Study and the Health Professionals Follow-Up Study, for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY.

PLCO: The authors thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff or the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, Mr. Tom Riley and staff, Information Management Services, Inc., Ms. Barbara O'Brien and staff, Westat, Inc., and Drs. Bill Kopp, Wen Shao, and staff, SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions to making this study possible.

PMH: The authors would like to thank the study participants and staff of the Hormones and Colon Cancer study.

WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: https://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI% 20Investigator%20Short%20List.pdf

References

Beckmann L, Thomas DC, Fischer C, Chang-Claude J. 2005. Haplotype sharing analysis using mantel statistics. *Hum Hered* 59:67–78.

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, Lubbe S, Spain S, Sullivan K, Fielding S and others. 2007. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* 39:1315–1317.

Cai T, Lin X, Carroll RJ. 2012. Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics (Oxford, England)* 13:776–790.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92:399–418.

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* 79:1002–1016.

Cornelis MC, Tchetgen EJT, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. 2012. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol* 175:191–202.

Dai JY, Kooperberg C, Leblanc M, Prentice RL. 2012. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* 99:929–944.

Dempfle A, Hein R, Beckmann L, Scherag A, Nguyen TT, Schäfer H, Chang-Claude J. 2007. Comparison of the power of haplotype-based versus single- and multilocus association methods for gene x environment (gene x sex) interactions and application to gene x smoking and gene x sex interactions in rheumatoid arthritis. *BMC Proc* 1(Suppl. 1):S73.

Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, Ooi L-Y and others. 2012. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 44:770–776.

Gao X, Starmer J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 32:361–369.

García-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein DW, Tardón A, Serra C, Carrato A, García-Closas R and others. 2005. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet* 366:649–659.

Gauderman JW, Zhang P, Lewinger PJ. 2012. Finding GWAS signals in the lower Manhattan by testing GxE interactions. In *International Genetic Epidemiology Society Annual Conference*. Stevenson, WA.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. 2007. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 31:383–395.

Goeman JJ, Van de Geer SA, De Kort F, Van Houwelingen HC. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics (Oxford, England)* 20:93–99.

Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, Tenesa A, Kusel VI, Sheehan P, Eaaswarkhanth M and others. 2011. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet* 7:e1002237.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70:42–54.

Hashibe M, McKay JD, Curado MP, Oliveira JC, Koifman S, Koifman R, Zaridze D, Shangina O, Wünsch-Filho V, Eluf-Neto J and others. 2008. Multiple ADH genes are associated with upper aerodigestive cancers. *Nat Genet* 40:707–709.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106:9362–9367.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S and others. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 40:1426–1435.

Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S and others. 2010. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 42:973–977.

Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. 2012. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol* 36:183–194.

Huang H, Chanda P, Alonso A, Bader JS, Arking DE. 2011. Gene-based tests of association. *PLoS Genet* 7:e1002177.

Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, Walther A, Spain S, Pittman A, Kemp Z and others. 2008. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* 40:26–28.

Kooperberg C, Leblanc M. 2008. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* 32: 255–263.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82:386–397.

Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M and others. 2009. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* 85:679–691.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

Li B, Leal SM. 2009. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 5:e1000481.

Li D, Conti DV. 2009. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol* 169:497–504.

Li M, Wang K, Grant SFA, Hakonarson H, Li C. 2009. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics (Oxford, England)* 25:497–503.

Lin D.-Y, Tang Z.-Z. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367.

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG and others. 2010. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 87:139–145.

Luan Y, Li H. 2008. Group additive regression models for genomic data analysis. *Biostatistics (Oxford, England)* 9:100–113.

Lubbe SJ, Di Bernardo MC, Broderick P, Chandler I, Houlston RS. 2012. Comprehensive evaluation of the impact of 14 genetic variants on colorectal cancer phenotype and risk. *Am J Epidemiol* 175:1–10.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5:e1000384.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615:28–56.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34:188–193.

Moskvina V, Schmidt KM. 2008. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* 32:567–573.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 64:685–694.

Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. 2010. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 34:213–221.

Murcray CE, Lewinger JP, Gauderman WJ. 2009. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 169:219–226.

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. 2011. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol* 35:201–210.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Peters U, Hutter CM, Hsu L, Schumacher FR, Conti DV, Carlson CS, Edlund CK, Haile RW, Gallinger S, Zanke BW and others. 2011. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* 131:217–234.

Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, Berndt SI, Bézieau S, Brenner H, Butterbach K and others. 2013. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* 144:799–807.

Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 13:153–162.

Price AL, Kryukov GV, De Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86:832–838.

Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D and others. 2010. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* 42:978–984.

Schaid DJ. 2010. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum Hered* 70:109–131.

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. 2005. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* 76:780–793.

Smith PG, Day NE. 1984. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 13:356–365.

Tenesa A, Farrington SM, Prendergast JGD, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N and others. 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 40:631–637.

The International HapMap Project. 2003. *Nature* 426:789–796.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W and others. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 39:984–988.

Tomlinson IPM, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Penegar S, Chandler I, Gorman M, Wood W and others. 2008. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 40:623–630.

Tomlinson IPM, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, Howarth K, Palles C, Broderick P, Jaeger EEM, Farrington S and others. 2011. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* 7:e1002105.

Tzeng J-Y, Zhang D. 2007. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* 81:927–938.

Tzeng J-Y, Devlin B, Wasserman L, Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 72:891–902.

Tzeng J-Y, Zhang D, Chang S.-M, Thomas DC, Davidian M. 2009. Gene-trait similarity regression for multimarker-based association analysis. *Biometrics* 65:822–832.

Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu F-C, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89:277–288.

Wang K, Abbott D. 2008. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol* 32:108–118.

Wang T, Elston RC. 2007. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80:353–360.

Wei Z, Li M, Rebbeck T, Li H. 2008. U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet* 72:821–833.

Wessel J, Schork NJ. 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* 79:792–806.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86:929–942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.

Zanke BW, Greenwood CMT, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E and others. 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 39:989–994.

Zhao J, Boerwinkle E, Xiong M. 2005. An entropy-based statistic for genomewide association studies. *Am J Hum Genet* 77:27–40.