Testing Gene-Environment Interaction in Large Scale Case-Control Association Studies: Possible Choices and Comparisons

Running title: Tests for Gene-Environment Interaction

BHRAMAR MUKHERJEE, JAEIL AHN, STEPHEN B. GRUBER AND NILANJAN CHATTERJEE*

* Correspondence to Dr. Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, 6120 Executive Blvd, Rockville, MD 20852, (e-mail: chattern@mail.nih.gov).

ABSTRACT

In the era of post genome-wide association studies, many investigators are currently searching for non-multiplicative gene-environment (G x E) interaction effects for studying complex disease phenotypes with established environmental risk factors. Several methods for screening G x E interaction have recently been proposed that address the issue of using gene-environment independence in a data-adaptive way. In this brief report, we present a comparative simulation study of power and Type I error properties of three classes of procedures: (i) The standard one-step case-control method; (ii) The case-only method which requires an assumption of gene-environment independence for the underlying population; (iii) A variety of hybrid methods, including empirical-Bayes, two-step and model averaging, that aim at gaining power by exploiting the assumption of gene-environment independence and yet can protect against false positives when the independence assumption is violated. Our studies suggest that while the case-only method generally has maximum power, it has the potential to create substantial false positives in large scale studies even when a small fraction of markers are associated with the exposure under study in the underlying population. All the hybrid methods perform well in protecting against such false positives and yet can retain substantial power advantages over standard case-control tests. The relative performance of the hybrid methods depend on the true underlying parameters for gene-environment interaction and gene-environment association. We conclude that for future genome-wide scans for G $\times$ E interactions, major power gain is possible by using alternatives to standard case-control analysis. Whether a case-only type scan or one of the hybrid methods should be used, depends on degree and direction of gene-environment association expected, the level of tolerance for false positives and the nature of replication strategies.

KEY WORDS: Case-only; Expected number of false positives; Familywise error rate; Gene-environment independence; Genomewide scan; Hybrid methods; Model averaging; Profile likelihood.

1 Introduction

Risks of most complex traits are influenced by both genetic susceptibility and environmental exposures. Epidemiologic researchers have long anticipated that exploration of gene-environment interactions may hold the key to our understanding of the etiology of chronic diseases and it will ultimately lead to better strategies for disease prevention. In the era of candidate gene studies, studies of gene-environment interactions focused on candidate functional SNPs, tagging SNPs in candidate genes or in whole candidate pathways that are typically chosen a'priori, based on hypothesized mechanisms for the effect of the environmental exposure under study. Unfortunately, such hypotheses-driven studies, although conceptually appealing, have not generally been successful in identifying replicable gene-environment interactions. The widely replicated interaction between NAT2 acetylation activity and smoking on risk of bladder cancer is a rare exception of success from candidate gene studies (1). Many other claims of interactions, however, have often failed to replicate (2).

Genome-wide association studies (GWAS) now provide tremendous opportunities for large-scale exploration of gene-environment interactions. The agnostic approach of searching for genetic associations based on GWAS have been clearly successful in identifying many susceptibility loci for a wide variety of complex traits [http://www.genome.gov/26525384]. However, a large fraction of variation in the different disease phenotypes still remain unknown, with the identified SNPs contributing modestly to prediction of disease risk (3-5). The identified loci are often not within or near genes for which associations could have been expected on an a'priori basis. Thus, there is currently hope that an agnostic genome-wide approach may also lead to detection of gene-environment interactions involving previously unsuspected loci [Gene-Environment Wide Interaction Studies or GEWIS as termed by (6)]. Moreover, as GWAS are now being pooled for further discoveries through meta-analysis, various consortia are now beginning to achieve large enough sample sizes necessary for detection of interactions with high confidence. Thomas (7) presents a detailed review whereas Khoury and Wacholder (6) point out analytical challenges fac-

ing large-scale G x E studies. However, none of the above two papers present numerical results from simulation studies or quantify the comparative performances of the different methods in terms of metrics related to Type I error and Power that are relevant to a GWAS.

Population-based case-control studies are commonly used to study the roles of genes and gene-environment interactions in determining the risks of complex diseases. It is well known that standard case-control analysis often has poor power for detection of multiplicative interaction due to small numbers of cases or controls in cells of crossing genotypes and exposures. In contrast, under the assumption of gene-environment (G-E) independence for the underlying population, one can test for multiplicative interaction in a very powerful fashion based on the genotype-exposure correlation in cases alone (8), but the method can have seriously inflated Type I error when the underlying assumption of gene-environment independence is violated (9). The independence assumption is quite plausible across the genome for exogenous exposures like air-pollution, pesticides, environmental toxins or treatment in a randomized clinical trial. The assumption, however, is expected to be violated for some markers in the genome for behavioral exposures like smoking and alcohol consumption, or anthropometric traits such as height, BMI, which themselves are known to have inherited components.

When gene-environment association is suspected, practitioners often adopt a two-stage procedure where, at first, one formally tests for the adequacy of the gene-environment independence assumption based on the data itself and then uses the outcome of that test to decide whether to choose the powerful case-only or the more robust case-control test. For a given study of modest sample size, however, the power of the tests for gene-environment independence would be typically low and consequently the two-stage procedure, as a whole, could still remain significantly biased (9, 10). The use of independence assumption has been extended to more general analyses that can estimate all the parameters of an association model including main effects and interactions (11, 12). These methods also face the same issue with bias and inflated Type I error when genetic and environmental factors are correlated at the population level.

Several authors recently proposed solutions to the bias vs efficiency dilemma by considering hybrid approaches that combine case-control and case-only analysis (13, 14). Murcray et al. (15) proposed a two-step approach that leverages the independence assumption at an initial screening step. The promising markers are followed-up with a standard case-control analysis at the second step. The purpose of this brief report is to provide a comparative study of these alternative tests for screening gene-environment interactions (G x E effects) with a large number of markers, in terms of Type I error and power. Previous results on Type I error and power comparison for each of these methods with standard case-control and case-only analysis are separately available in each of the above individual papers, but no comparison across methods are available so far. Cornelis et al.(16) apply several of these methods to analyze G x E interactions in a Type 2 diabetes GWAS, but the paper does not contain detailed simulation results. As practitioners are confronting the issue of choosing a method for screening for G x E effects, it is important to realize the advantages and disadvantages associated with each choice. Using simulation studies, in this report, we point out some important operating characteristics of these procedures that could inform/guide such choices.

The report is organized as follows. In Section 2.1, we first describe the different testing procedures that we consider. In Section 2.2, we describe the simulation design followed to evaluate each method. In Section 3, we present results on Type I error and power properties corresponding to these eight tests under different sampling ratio of cases and controls, different number of markers and varying strength of G-E association. Section 4 contains discussion and concluding remarks.

2 Materials and Methods

2.1 Different tests for interaction

We present a guiding summary chart of all methods with glossary and key attributes in Table 1. Following is a more detailed description.

I. SIMPLE LOGISTIC REGRESSION BASED ON CASE-CONTROL DATA: The simplest and most com-

monly used test for gene-environment interaction is based on a logistic regression model:

$$\text{logit } P(D|G, E, \boldsymbol{S}) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G * E + \boldsymbol{\beta}^\top \boldsymbol{S}. \tag{1}$$

Where $G = 0, 1, 2$ is the number of alleles present at a bi-allelic locus, E is the environmental exposure and $\boldsymbol{S}$ are a set of other covariates one may adjust for. We will assume that an ensemble of single nucleotide polymorphisms (SNPs) have been genotyped for study participants, leading to data on many such genetic factors G. Instead of assuming a trend or log-additive model as the one described above, one can modify the genetic susceptibility model by binary collapsing of G (dominant, recessive) or by allowing separate log odds-ratios for each category of G compared to the baseline category, leading to a 2 degree of freedom (d.f) test for the main effects of G as well as another 2 d.f. test for the G x E interaction effects. The model for E, when continuous, may involve inclusion of higher order non-linear terms. For categorical E, one may again allow separate log odds-ratios corresponding to each categories of E relative to the baseline category, thus leading to a higher degrees of freedom test for the saturated interaction model. However, we use the above simpler notation with the understanding that appropriate modification of the regression terms can be carried out depending on the nature of G and E. The test for β_{GE} is the standard Wald test for $H_0 : \beta_{GE} = 0$, based on maximum likelihood (ML) estimation, or the corresponding Likelihood ratio (LR) chi-squared test.

II. RETROSPECTIVE LIKELIHOOD BASED TESTS THAT USE GENE-ENVIRONMENT INDEPENDENCE

(II.A) CASE-ONLY ANALYSIS: In its simplest form, this class of methods include the popular case-only analysis. This analytic strategy specifies a regression model for testing association between G and E (conditional on other covariates $\boldsymbol{S}$) among the cases ($D = 1$). This can be achieved through modeling the distribution of $G|E, \boldsymbol{S}$ via a polytomous logistic regression model, namely,

$$\text{logit } P(G = g|E, D = 1) = \gamma_0 + \gamma_{gE}\, E + \gamma_s^\top \boldsymbol{S} \qquad g = 1, 2; \tag{2}$$

Under the assumption of G-E independence conditional on $\boldsymbol{S}$, the likelihood-ratio test for $H_0 : \gamma_{1E} = \gamma_{2E} = 0$ among cases in (2) is a valid test for interaction effects in a corresponding logistic

model. More commonly, when a trend model is assumed with a single coefficient γ_E, such that $\gamma_{1E} = \gamma_E$ and $\gamma_{2E} = 2\gamma_E$ and one employs a Wald test for $H_0 : \gamma_E = 0$, that test is approximately equivalent to testing $H_0 : \beta_{GE} = 0$ in (1) with the log-additive assumption.

A major limitation of the case-only analysis, even when G-E independence assumption is true, is the fact that it can not yield estimates of the main effect parameters β_G and β_E that are essential to evaluate joint effects of G and E or subgroup effects of a genetic factor across exposure strata or effects of an exposure across genotype categories.

(II.B) PROFILE LIKELIHOOD OF CHATTERJEE AND CARROLL (12): One can use log-linear modeling technique for categorical data (11) or a profile likelihood technique more generally (12), and obtain estimates of all model parameters under gene-environment independence using data on **both** cases and controls. The profile likelihood method has complete flexibility of a regression model, and uses gene-environment independence assumption (possibly conditional on covariates $\boldsymbol{S}$, which for example, may include principal components obtained from a large number of genetic markers that can correct for gene-gene and gene-environment dependence that are induced by presence of population stratification). The method considers a retrospective likelihood,

$$P(G, E, \boldsymbol{S}|D) = \frac{P(D|G, E, \boldsymbol{S}) \overbrace{P(G|E, \boldsymbol{S})}^{\text{reduces to P(G}|\boldsymbol{S})} P(E, \boldsymbol{S})}{\sum_{\text{G,E,}\boldsymbol{S}} P(D|G, E, \boldsymbol{S}) P(G|E, \boldsymbol{S}) P(E, \boldsymbol{S})}. \tag{3}$$

The ingredients of the above likelihood are specified as below:

1. A logistic disease incidence model: $P(D|G, E, \boldsymbol{S})$ of the form (1).

2. A model for $P(G|E, \boldsymbol{S})$: This is assumed to be of a multinomial logistic form with three response categories of G and covariates $(E, \boldsymbol{S})$. The assumption of G-E independence, conditional on $\boldsymbol{S}$ implies, $P(G|E, \boldsymbol{S}) = P(G|\boldsymbol{S})$ and the covariate E is simply dropped from this multinomial logit model for $P(G|E, \boldsymbol{S})$ to reflect the assumption. Let this general dependence model be denoted by,

$$\text{logit} P(G = g|E, \boldsymbol{S}) = \theta_0 + \theta_{gE} E + \theta_{gS}^{\top} \boldsymbol{S}, \qquad g = 1, 2. \tag{4}$$

where $\theta_{gE} \equiv 0, g = 1, 2$ under G-E independence assumption. One can assume a log-additive structure and have one single association parameter θ_{GE} as well. The constrained ML estimates of the interaction parameters based on this likelihood with $\theta_{GE} \equiv 0$, retain the same efficiency advantages as a case-only analysis. One can also incorporate assumptions like HWE while modeling the distribution of G to further boost the efficiency advantage of a retrospective likelihood (17, 18).

3. A non-parametric model for $P(E, \boldsymbol{S})$: This renders estimation of a potentially multi-dimensional non-parametric joint distribution and is handled by the profile likelihood technique established in (12). In fact, the likelihood can be treated in an elegant way by establishing equivalence with a simpler pseudo likelihood.

III. EMPIRICAL-BAYES ESTIMATION: To trade-off between bias and efficiency of case-control and case-only analysis, Mukherjee and Chatterjee (13) proposed a shrinkage estimator based on the above retrospective likelihood framework of (12). The estimator is of the following form:

$$\hat{\beta}_{EB} = \Delta^\top \hat{\boldsymbol{\theta}}_{ML} \hat{\boldsymbol{\theta}}_{ML}^\top \Delta [\hat{V}_{\hat{\boldsymbol{\beta}}_{ML}} + \Delta^\top \hat{\boldsymbol{\theta}}_{ML} \hat{\boldsymbol{\theta}}_{ML}^\top \Delta]^{-1} \overbrace{\hat{\boldsymbol{\beta}}_{ML}}^{\text{unconstrained}} + \hat{V}_{\hat{\boldsymbol{\beta}}_{ML}} [\hat{V}_{\hat{\boldsymbol{\beta}}_{ML}} + \Delta^\top \hat{\boldsymbol{\theta}}_{ML} \hat{\boldsymbol{\theta}}_{ML}^\top \Delta]^{-1} \overbrace{\hat{\boldsymbol{\beta}}_{ML}^0}^{\text{constrained}}$$

where $\Delta = \left. \frac{\partial \hat{\beta}_{ML}^\top(\theta)}{\partial \theta} \right|_{\theta=0}$. Here $\hat{\boldsymbol{\theta}}_{ML}^\top$ denotes estimates of the parameters from (4) whereas $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{V}_{\hat{\boldsymbol{\beta}}_{ML}}$ are parameter estimates and variance estimates obtained from (3) that allows for gene-environment dependence. Finally, $\hat{\boldsymbol{\beta}}_{ML}^0$ are the constrained ML estimates of the disease-risk parameters from (3) while using gene-environment independence by setting $\theta_{gE} = 0$ in (4), $g = 1, 2$. Variance approximations for $\hat{\beta}_{EB}$ are provided by Delta approximations in (13). Wald tests based on approximate asymptotic normality is used. The authors in fact consider another form of shrinkage weight in (10), where in the above expression, $\hat{V}_{\hat{\boldsymbol{\beta}}_{ML}}$ is replaced by $\hat{V}(\hat{\boldsymbol{\beta}}_{ML} - \hat{\boldsymbol{\beta}}_{ML}^0)$. The resultant estimator will be denoted by $\hat{\beta}_{EB2}$ from here onwards. The specific form of the "shrinkage" weights is obtained by the expression for the posterior mean obtained in a conjugate analysis under a normal-normal model (19, p. 131), with the prior variance substituted by an

estimate obtained using a method of moments approach (20).

The retrospective profile-likelihood methods, including the unconstrained maximum likelihood, constrained maximum likelihood and EB are now readily implementable in the R-package CGEN (dceg.cancer.gov/bb/tools/genetanalcasecontdata), also available as a part of the R-bioconductor computing repository.

IV. Model Averaging

IV.A: Bayesian Model Averaging:

Li and Conti (14) present the following approach that uses case-control data. The method combines the case-only model and the case-control model via the following model-averaging framework, assuming all *categorical* covariates (G, E).

- Note that one can represent the logistic model in (1) (without $\boldsymbol{S}$) by an equivalent log-linear model (M_1), where μ represents the cell count of a particular (D, G, E) configuration:

$$\log(\mu|D, G, E) = \alpha_0 + \alpha_G\, G + \alpha_E\, E + \alpha_{\mathbf{GE}}\, \mathbf{GE} + \beta_0\, D + \beta_G\, GD + \beta_E\, ED + \beta_{GE}\, GED.$$

  Under G-E independence assumption, the above model is fitted under constraints which sets $\alpha_{GE} \equiv 0$. Let this reduced model be termed as M_2. The estimator for β_{GE} under M_2 is approximately equivalent to the case-only estimator $\boldsymbol{\gamma}$ in (2) for binary G. For a trinary G, under the log-additive trend model, one can assign scores of 0,1,2 to the three categories of G, that are equivalent to introducing a linear by linear association term in the log-linear model for ordinal data (21).

- By specifying the prior odds $W = P(M_1)/P(M_2)$, one can get the posterior distribution of β_{GE} and the resultant BMA estimator as

$$P(\beta_{GE}|\text{Data}) = \sum_{k=1}^{2} P(\beta_{GE}|\text{Data}, M_k)P(M_k|\text{Data});$$

$$\hat{\beta}_{BMA} = E(\beta_{GE}|\text{Data}) = \sum_{k=1}^{2} \hat{\beta}_{GE}^{k} P(M_k|\text{Data}).$$

The variance expression of the BMA estimator is provided in (14) and Wald tests based on the ratio of estimate to standard error is recommended for being used as a test statistic. One can note that the BMA estimator may be computed by using the profile likelihood in (3) with and without the independence assumption following the same recipe. Using the profile likelihood allows more flexibility than the construct of a log-linear model which limits the analysis to categorical covariates. Li and Conti (14) vary the prior odds-ratio W over a range of values to assess the properties of their method.

IV.B FREQUENTIST AIC MODEL AVERAGING: The notion of combining two models is fairly general (22). In addition to the BMA and EB approach, we consider another model-averaged estimator based on the Akaike information criterion (AIC). Note that, the BMA approach is approximately equivalent to weighting the two models by their respective Bayes Information Criterion [BIC, (23)]. In averaging a full model estimate and a reduced model estimate, the EB estimators are also implicitly performing model averaging with a particular choice of model weights (22). One can also use model AICs as alternative weights for the two models. For AIC, define $w(v) = \{1 + \exp(v/2 - d)\}^{-1}$, where d is the dimension of θ_{GE} in (4), or the difference in dimension between the full (dependence model/unconstrained model) and the reduced model (independence model/constrained model). Let $\mathcal{L} = 2\{L_{dep} - L_{indep}\}$ be the likelihood ratio test statistic comparing the two models. The weight assigned to the reduced (independence) model by AIC is $w(\mathcal{L})$. All models are assumed to be fitted in the profile likelihood framework of (12). An approximate variance expression for the AIC averaged estimator is obtained by following an exactly analogous formula to the variance expression in BMA as presented in (14) where the BIC model weights are replaced by AIC weights.

V. THE TWO-STEP SCREENING STRATEGY: Murcray et al. (15) proposed a simple but very useful two-step approach to again leverage the efficiency advantage of case-only type methods for screening, without compromising on the robustness properties of the final tests for gene-environment interaction. Their method can be described through the following steps.

- Step 1. A first step screening test: A likelihood ratio test of association between G and E in a combined sample of cases and controls is carried out. For trinary G, test $H_0 : \gamma_{gE} = 0$, $g = 1, 2$ in the following association model:

$$\text{logit } P(G = g|E, \boldsymbol{S}) = \gamma_0 + \gamma_{gE}\, E + \gamma_s^\top \boldsymbol{S}, \qquad g = 1, 2$$

  with $\gamma_{2E} = 2\gamma_{1E}$ under a log-additive model.

- A subset of m SNPs will exceed a first step threshold of significance α_1, with $P < \alpha_1$.

- Step 2. For the m SNPs passing through Step 1, test $H_0 : \beta_{GE} = 0$ in the logistic model (1). Significance at the second stage assessed by $P < \alpha/m$, where α the overall Type I error rate.

The first-step test of the two-step approach exploits the fact that under the gene-environment independence assumption in the underlying population, the presence of G-E correlation in the case-enriched case-control sample indicates presence of gene-environment interaction on the risk of the disease. It is important to note that the first step test is done in the *combined* sample of cases and controls, and not cases only. The resulting test is less powerful than a case-only test under gene-environment independence, but it being independent of the second step case-control test ensures that it can be used as a pure "screening" method. Consequently, the two-step method maintains nominal Type I error level as the ultimate second step test is the model-robust case-control test for G x E interaction. The power advantage of the two-step method comes from reducing the multiplicity burden by decreasing the number of SNPs that are being carried forward to Step 2. The amount of power gain of the two-step procedure depends on the choice of the first step threshold α_1. The authors use α_1 as 0.05 in the original paper but follow-up empirical studies suggest α_1 can be chosen adaptively depending on study size and other parameter guesses for enhanced power.

REMARK 1: Like the case-only approach, one major limitation of the two-step approach is that one can not screen for joint effects through this approach as only the SNPs with significant interaction

and not necessarily main effects are carried over to Step 2, for the final case-control analysis. Thus it is very much a targeted interaction searching procedure with the first step screening test only filtering markers with significant interactions. Moreover, the first step test statistic is not independent of the second step test statistic for testing main effects, the independence holds only for testing interaction effects. Thus, for detecting joint effects, this approach is not optimal.

2.2 Simulation Setting

We first describe the simulation mechanism for any given marker. For simulation purpose, we consider the simple set-up of an unmatched case-control study with a binary genetic factor G and a binary environmental exposure E. Let $E = 1$ ($E = 0$) denote an exposed (unexposed) individual and $G = 1$ ($G = 0$) denote whether an individual is a carrier (non-carrier) of the susceptible genotype. Let D denote disease status, where $D = 1$ ($D = 0$) stands for an affected (unaffected) individual. Let n_0 and n_1 be the number of selected controls and cases, respectively. The data can be represented in the form of a 2×4 table as displayed below.

A binary genetic factor and a binary environmental exposure

	$G=0$		$G=1$	
	$E=0$	$E=1$	$E=0$	$E=1$
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}

Let $\boldsymbol{r}_0 = (r_{000}, r_{001}, r_{010}, r_{011})$ and $\boldsymbol{r}_1 = (r_{100}, r_{101}, r_{110}, r_{111})$ denote the vector of observed cell frequencies in the controls and cases respectively. The population parameters, namely, the cell probabilities corresponding to a particular G-E configuration in the underlying control and case populations are denoted as $\boldsymbol{p}_0 = (p_{000}, p_{001}, p_{010}, p_{011} = 1 - p_{000} - p_{001} - p_{010})$ and $\boldsymbol{p}_1 = (p_{100}, p_{101}, p_{110}, p_{111} = 1 - p_{100} - p_{101} - p_{110})$, respectively. The observed vectors of cell counts can be viewed as realizations from two independent multinomial distributions, namely, $\boldsymbol{r}_0 \sim$ Multinomial$(n_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \sim$ Multinomial$(n_1, \boldsymbol{p}_1)$. Let $OR_{10} = p_{000}p_{101}/p_{001}p_{100}$ denote the odds-ratio associated with E for nonsusceptible subjects ($G = 0$), $OR_{01} = p_{000}p_{110}/p_{010}p_{100}$ denote the odds-ratio associated with G for unexposed subjects ($E = 0$) and $OR_{11} = p_{000}p_{111}/p_{011}p_{100}$

denote the odds-ratio associated with $G = 1$ and $E = 1$ compared to the baseline category $G = 0$ and $E = 0$. Therefore,

$$\psi = OR_{11}/(OR_{10}OR_{01}) = (p_{001}p_{010}p_{100}p_{111}) \,/\, (p_{000}p_{011}p_{101}p_{110})$$

is the multiplicative interaction parameter of interest. The interaction log OR parameter is $\beta_{GE} = \log(\psi)$.

For each marker, given the values for the prevalence of G and E, namely P_G and P_E, and the value of the odds-ratio θ_{GE} in the control population, one is able to obtain the control probability vector $\boldsymbol{p}_0$ by solving the following system of equations.

$$\begin{aligned} \theta_{GE} &= \frac{p_{000}(p_{000} - (1 - P_G - P_E))}{(1 - P_G - p_{000})(1 - P_E - p_{000})}, \\ p_{001} &= 1 - P_G - p_{000},\ p_{010} = 1 - P_E - p_{000}. \end{aligned}$$

We then set the values of OR_{10}, OR_{01} and ψ, which together with $\boldsymbol{p}_0$, defines the case-probability vector (24). For each marker, we generate data independently from the two multinomial distributions corresponding to the case and control populations.

To mimic a large-scale study, we generated data on M markers independently distributed across the genome. For Type I error evaluation we consider the situation with 2000 cases and 2000 controls with $M = 100,000$. For evaluating power characteristics we consider $M = 100,000$ with $n_1 = 2000$, n_0=2000 and 4000; and a larger study with $n_1 = 10000$, n_0=10000 and 15000. Simulation results for some additional settings with a smaller number of markers $M = 10,000$ and intermediate sample sizes $n_1 = 7000$, $n_0 = 7000, 14000$ are presented in the online supplementary material. Throughout, we consider E as a binary environmental covariate with $P(E) = 0.5$, reflecting a common situation with dichotomization of a continuous covariate at the sample median. All main effect parameters are assumed to be unity, namely, $OR_{10} = OR_{01} = 1.0$ across all scenarios. The trend of results remain unchanged with non-null main effects.

We assume a situation with only 1 causal locus having true interaction with E, others null with no interaction effect. At the causal locus, the minor allele frequency is set at $q_A = 0.2$, and

we assume a dominant genetic susceptibility model with $G = 1(AA, Aa)$, $G = 0(aa)$, yielding $P(G = 1) = 0.36$. The G-E OR among controls for the causal locus is set at three values, namely, $\exp(\theta_{GE}) = 1.0, 0.8, 1.1$ corresponding to independence, negative and positive dependence. The interaction parameter at the causal locus $\exp(\beta_{GE})$ is varied from 1.1 to 2.0 for 2000 cases and from 1.1 to 1.5 for the situation with 10000 cases.

Among the $M - 1$ null loci, without any interaction effects with E, the allele frequency distribution is assumed to be uniform $q_A \sim Uniform(0.1, 0.3)$. The population level G-E association structure among null loci is assumed to be of the form of a mixture distribution reflecting that a large fraction, say p_{ind}, of the SNPs, indeed are independent of E in the population, whereas the remaining SNPs show some departures from the independence assumption. We generated the log OR of GE association in controls corresponding to null loci as $\theta_{G^0E} \sim p_{ind}\ \delta_0 + (1 - p_{ind})\ N(0, \mathrm{sd} = \log(1.5)/2)$. Here δ_0 is a point mass at 0 reflecting G-E independence. The standard deviation parameter of the normal distribution part of the mixture is chosen such that of the θ_{GE} values that depart from independence, 95% fall within $\pm \log(1.5)$. We vary the simulation parameter p_{ind} to create G and E dependence among more (less) null markers.

For the generated ensemble of markers, we then carry out the eight tests described in Section 2: case-control (CC), case-only (CO), Empirical Bayes (EB, EB2), Two-step (TS with two choices of α_1, 0.05 and 5×10^{-4}), BMA ($W = 1$), AIC model averaging (AIC). In our simulation, we investigate the family-wise Type I error rate (FWER), expected number of false positives and power of the above testing procedures. All family-wise Type I error rates are estimated by the empirical proportion of simulated datasets that declare false significance for $H_0 : \beta_{GE} = 0$ corresponding to at least one null marker. We also collect the number of false rejections in each simulation run and present average of that count over all simulation runs. This estimate of the expected number of false positives may be a more reasonable quantity to examine in a GWAS setting instead of the more conservative FWER which counts the proportion of *at least* one false rejection. The power values are estimated by empirical proportion of rejection of $H_0 : \beta_{GE} = 0$ corresponding

to the causal marker. The results are based on 5000 simulated datasets. The margin of error in the estimated proportions with nominal level $\alpha = 0.05$ with 5000 simulated datasets is approximately 0.003.

REMARK 2: Under departure from the independence assumption, methods except for the two-step and the standard case-control analysis may not adhere to strict nominal FWER for all parameter settings considered. We present two metrics that combine Power and Type I error (25) in the online supplementary material. The metrics are accuracy (ACC) and positive predictive value (PPV), given by ACC=(1-Type 1 error+Power)/2 and PPV=Power/(Power+Type I error). We also present mean-squared error (MSE) corresponding to each method except two-step, which is a pure screening procedure. The MSE provides a combined metric of bias and variance from an estimation standpoint.

3 Results

We now summarize the main findings of the simulation study. Table 2 presents the FWER and expected number of false positives corresponding to each method with $M = 100,000$ for varying values of p_{ind} from 0.95 to 1.00, the proportion of markers that follow gene-environment independence. The simulation is carried out with $n_1 = n_0 = 2000$. One can note that the two-step and the case-control procedures always maintain FWER, whereas the FWER for the case-only method is at 0.80 even when 99.95% of the SNPs are independent of E. The FWER control of EB-type procedures and model averaging procedures is much superior than case-only method and FWER is maintained if the fraction of SNPs that actually follow the G-E independence assumption is 99.0% or more. The model averaging procedures BMA and AIC offer better control of FWER compared to EB-Type procedures when p_{ind} is lower, say for example, 0.95, i.e., when more than 5% SNPs are associated with E. One may note that for higher values of p_{ind}, closer to 1.00, which is likely to be realistic in practice, the EB-type as well as model averaging procedures can maintain strict FWER and even be conservative.

In terms of expected number of false positives in Table 2, the case-only analysis is still worse among all methods but does not appear to be an unreasonable strategy with expected number of false positives less than one when $p_{ind} = 0.9995$, around 7 when $p_{ind} = 0.9975$, which rises to around 158 with $p_{ind} = 0.95$. This may be a more rational metric to examine in GWAS instead of FWER which only considers the more conservative criterion of probability of at least one false rejection under the global null hypotheses.

Figures 1 and 2 represent the power values for testing $H_0 : \beta_{GE} = 0$ at the causal locus for the eight methods with $M = 100,000$. The exact numerical values corresponding to the graphs are contained in the online supplementary material. The fraction of null SNPs that satisfy the independence assumption is set at 0.995 in each case.

We first discuss the main features in Figure 1 with $n_1 = n_0 = 2000$ when the independence assumption holds at the causal locus ($\exp(\theta_{GE}) = 1.0$). As expected, case-only analysis has the maximum power compared to all other contenders. Among hybrid methods, two-step and EB perform similarly and these two methods generally have higher power than BMA or AIC. The two-step approach with $\alpha_1 = 5 \times 10^{-4}$ has slightly higher power than EB for interaction OR exceeding 1.6. For example, with $n_1 = n_0 = 2000$, at interaction OR=1.8, the power of EB is 0.68, of EB2 is 0.58, BMA and AIC at 0.59 whereas two-step method with $\alpha_1 = 5 \times 10^{-4}$ has power 0.74. In this setting, two-step with $\alpha_1 = 0.05$ attains a power of 0.53. The case-only method has power 0.92 whereas the case-control analysis has a low power of 0.31. For lower values of the interaction OR (≤ 1.6), the EB and TS have very similar performance and they outperform the model averaging approaches. For example, at interaction OR=1.6, both EB and TS ($\alpha_1 = 5 \times 10^{-4}$) have a power 0.30, whereas the power of BMA and AIC is 0.18. The EB2 power is at 0.23 whereas TS with $\alpha_1 = 0.05$ has power 0.20. The case-control and case-only analyses have power values of 0.08 and 0.54 respectively. For interaction OR less than 1.3, the EB procedure appeared to have slightly greater power than TS but both of the power values were very low under this sample size.

The bottom panel of Figure 1 with $n_1 = 2000$ and $n_0 = 4000$ show similar trend but weaker

performance by the TS method with both choices of α_1. With case:control ratio being tilted towards having more controls than cases, the first step screening test in combined sample of cases and controls loses the power advantage of a case-only approach, when compared to 1:1 case:control ratio. Thus we note a surprising and counter-intuitive finding that under identical simulation settings, the power of two-step procedure actually decreases with increasing the number of controls, if the number of cases are held fixed, as the power depends on the case:control ratio. For example, under the independence assumption, with interaction OR=1.9, with $n_0 = 2000$, TS ($\alpha_1 = 5 \times 10^{-4}$) has power 0.87 that reduces to 0.82 as n_0 increases to 4000. On the other hand, under the same setting the power of TS ($\alpha_1 = 0.05$) is 0.69 with $n_0 = 2000$ and increases to 0.89 with $n_0 = 4000$. This indicates that the optimal screening threshold α_1 in two-step procedure does heavily depend on case:control ratio, everything else remaining the same.

Under departures from the independence assumption at the causal locus, we consider two situations: one with positive and the other with negative association between G and E in the controls. With ($\exp(\theta_{GE}) = 1.1$), again the case-only method has the highest power, two-step with $\alpha_1 = 5 \times 10^{-4}$ has the second highest power and a clear dominance over other hybrid methods. In contrast, under negative dependence at the causal locus ($\exp(\theta_{GE}) = 0.8$), case-control analysis is the most powerful analysis and case-only analysis performs quite poorly [see also (14)]. In this situation, where β_{GE} is positive and θ_{GE} is negative, the G-E log odds-ratio in cases (which is simply $\beta_{GE} + \theta_{GE}$ for a 2×4 table) is close to null, explaining the loss of power. The two-step approach also performs quite poorly in this setting, especially with the more stringent choice of α_1=5×10^{-4}. The BMA, EB, EB2 and AIC perform comparably among the hybrid methods with BMA/AIC having an edge over the EB-type methods in this scenario. The loss of power in TS under a study design with more controls than cases becomes quite drastic with negative G-E association as one can notice in the left most panels in Figure 1. TS with both choices of α_1 loses power as n_0 increases under such negative dependence.

In order to understand the phenomenon of better power property of EB over TS at smaller

values of interaction OR under independence, we increased the sample size to $n_1 = 10,000$ and repeated the same simulation over a more modest range of interaction OR from 1.1 to 1.5. Figure 2 essentially captures the same features of the different methods as discussed for Figure 1. Under independence, EB has power advantages over TS for smaller values of interaction, especially for unequal case:control ratio (bottom panel, center graph in Figure 2). Under positive dependence, TS has a clear dominance and under negative dependence BMA/AIC has advantage over EB, whereas TS performs quite poorly. This larger sample size setting is more reflective of current post-GWAS consortium studies exploring G x E effects. Results for several other simulation settings are presented in the online supplementary material.

Figure 3 presents estimated relative MSE corresponding to the log odds-ratio parameter at the causal locus under the simulation setting of Figure 1 for all the methods except the two-step method (which is more of a screening tool and not an estimation method). The MSE for each method is divided by the MSE of standard case-control analysis. One can notice the advantage of EB type methods in terms of this metric as one tries to balance between bias and efficiency in a data-adaptive way. The case-only method is best only when the independence assumption is true (the central block) and performs worse under any departures from the independence assumption.

4 Discussion

In summary, our study indicates that the data-adaptive hybrid methods like EB, TS, BMA or AIC model averaging can achieve balance between power gain and Type 1 error rate control for testing G x E effects in large-scale association studies. There is no uniform dominance of one method versus the other in terms of their operating characteristics across all simulation scenarios. The performance of the methods differ according to magnitude/direction of the G-E association and interaction OR. All the new hybrid methods offer power gain over standard case-control analysis and better control of Type I error rate compared to a case-only analysis. We summarize and conclude with some observations that merit further discussion.

Type I error control: We note that even if only a very small fraction of the SNPs actually depart from independence, say 0.05% (p_{ind}=0.9995 in Table 1), using a case-only method will still not offer nearly adequate control of Type I error rates to prioritize lead G x E candidates whereas the hybrid approaches offer protection from false positives. An attractive feature of the two-step method is that it always maintains the desired level of FWER. The EB-type methods have worse FWER control compared to model averaging when the fraction of SNPs truly associated with E is more than 1% (under the G-E association distribution we assumed among the null markers). However, in a GWAS study to expect the fraction of SNPs departing from independence assumption to be much less than 1% may be a quite realistic assumption, and in that range of p_{ind}, all the weighted methods maintain nominal FWER. We note that if the prior probability for the case-only model is increased in BMA, it boosts the power but inflates the FWER as one would expect. However, one can always postulate alternative distributions for G-E association parameters among the null loci, instead of the mixture distribution we assumed, and the operating characteristics of the methods in terms of FWER will change substantially based on that distribution. Note that for the small fraction of markers that depart from independence, we assumed a $N(0, \text{sd} = \log(1.5)/2)$ distribution. Shrinking the variance of this distribution further, leads to an improvement in the FWER properties of case-only and all hybrid methods.

Another interesting observation is the fact that for p_{ind} exceeding 0.99, the weighted methods have a conservative FWER falling well below 0.05. Thus there is some scope of employing a more aggressive form of shrinkage and enhance the power of these methods further.

We observe that if one desires to control the number of false positives instead of FWER, a case-only analysis appears to perform quite reasonably in a range of scenario for gene-environment association that is likely to arise in practice (p_{ind} between 0.995-1.00). If in a GWAS study, discoveries are going to be followed by replication, it may be reasonable to accept a few false positives if a significant boost in power occurs using a specific method. Still caution is needed to avoid large number of false positives as they could infiltrate the limited number of top ranked SNPs that may

be followed up for replication.

Power Comparison: In general, under the independence assumption and for positive G-E association, the case-only method has the highest power. The two-step approach is preferred over the case-only approach for positive association scenarios as it maintains FWER, which the case-only method does not. The power properties of the methods when the G x E interaction is positive and the G-E association is negative is worth noting as this may very likely occur for a fraction of SNPs in a large-scale association study. Case-only analysis is not the most powerful analysis in such situations and a standard case-control analysis can be more powerful. The weighted methods strike a compromise in this reverse situation as well. The performance of the two-step method in such situation is concerning as it suffers severely from the lack of power of the first step case-only type screening procedure, especially with a more stringent choice of the first step threshold α_1.

The power performance of the two-step method in terms of study designs where control:case ratio is larger than 1:1 is also noteworthy and have not been previously pointed out. The first step screening test for interaction in the two-step method can be viewed as a weighted test of G-E association in cases and in controls. When the weight corresponding to controls increase, there is an attenuation of the test statistic, leading to a loss of power. In fact, because of such phenomenon, there could be situations where the power of the two-step method as a whole, may decrease, as the number of controls are increased, everything else remaining fixed. The power loss of two-step method is more pronounced for negative G-E association in controls and positive interaction (or vice versa). In this situation, very few of the "promising" SNPs filter through the screening step causing this behavior. One may attempt to correct this drawback by attaching differential sampling weights to case and control observations in the first step screening procedure, but that will destroy the desirable independence property of the first step screening test with the second step case-control test.

Given the sample-size and G-E association configurations, it appears that the weighted methods EB, EB2, BMA, AIC have robustness advantage of performing reasonably well across a spec-

trum of alternatives in terms of their power properties. Among the weighted methods EB has advantage over BMA/AIC in the situation with positive G-E association, whereas BMA has power advantage in the negative G-E association scenario.

Combined metrics of power and Type 1 error: Since five of the seven methods may not adhere to nominal Type I error levels (in the sense of FWER), in Web Figures 4-9 of the online supplementary material we present two combined metrics ACC and PPV as described before. One can notice that the case-only method is least desirable in terms of these metrics even when $p_{ind} = 0.9995$. The performance of the hybrid methods across a spectrum of G-E association scenarios are indeed encouraging.

Estimation and testing for effects other than multiplicative interaction: In this article, we have focused on tests for multiplicative interactions only. It is, however, important to recognize that the value of studying genetic and environmental exposures together does not necessarily stem from the ability to test for statistical interactions. Various alternative parameters, such as the joint effect of two exposures or the sub-group effects of one exposure within strata defined by the other exposure, may be useful for developing powerful test of association, understanding the public health impact of the exposures, targeting intervention and risk prediction. The hybrid procedures can be extended to carry out inference regarding such alternative parameters of interest. In recent years, for example, omnibus tests, that can simultaneously account for genetic main effects and gene-environment/gene-gene interactions have received attention as a powerful approach for detection of disease of susceptibility loci (26, 27). A major limitation of the two-step method is that it is targeted towards only testing for multiplicative interactions and cannot be easily generalized to alternative tests that may involve main effect parameters (see Remark 1).

Sensitivity to user defined choices: The choice of the first step threshold α_1 can largely determine the power properties of the two-step approach. It is hard to optimize this choice for a large-scale study as it depends on the number of markers, disease prevalence, case:control ratio, distribution of unknown G-E association parameters and interaction effect sizes. In a more recent manuscript, an

optimal choice of α_1 has been proposed (28) and α_1=0.0005 that we have used is found to be nearly optimal under most of our simulation configuration. The performance of the BMA procedure can also change by varying the ratio of prior weights W. We used $W = 1$ in our study but it may be more reasonable to assign a larger prior mass to the case-only model or to the assumption of gene-environment independence. On the other hand, the EB procedures and AIC averaging does not require any prior or tuning parameter specification and is completely data adaptive.

Analytical power calculation is intractable in closed form for the hybrid methods and we resort to simulation studies to evaluate the current ensemble of methods. We have presented results under one particular simulation scheme, the trend in the results remain similar for changes in simulation parameters like the allele frequency, exposure prevalence, number of cases and controls with a given case:control ratio and number of markers. However, if one changes the parameters of the mixture distribution for $\log(\theta_{G^0E})$, or uses an alternate form of distribution as elicited in (5), the FWER comparison may change appreciably.

Issues with lack of coherence and the violation of the likelihood principle: A reviewer has raised an important point that some of these methods (case-only, two-step) ignore data and still gain power. This may appear to be counterintuitive to foundational statistical principles and raises the question: "how can ignoring data lead to better performances than using the entire information content of a dataset?" The case-only approach makes a strong assumption to gain efficiency and provides unbiased estimates only under the assumption of gene-environment independence. But the method is "coherent"in the sense that it can be justified via a proper "likelihood" of the entire data as long as the independence assumption is valid (11, 12). Note in Figures 1 and 2 (central block) that when the independence assumption is true, the case-only method that yields the constrained MLE is indeed optimal in terms of power. However, as our simulation study has shown that the gain in power by making this assumption comes at a price of inflated FWER, under departures from this assumption. If instead of controlling the FWER, one is willing to accept a limited number of false positives, the case-only type approach may be a reasonable strategy, if we believe

that only a handful of SNPs in a GWAS may be truly associated with the environmental exposure under consideration (the situations corresponding to $p_{ind} \geq 0.9975$ in Table 2).

In contrast, the two-step method essentially divides information in the total likelihood into two independent components, one used for screening, the other for validation. In general, these types of two-step screening strategies that partition the total information in the data for a clever work-around the multiple testing problem, can not be justified based on a likelihood principle and thus sometimes can face "incoherence" issues. For example, we have noted earlier that in some situations the power of the two-step method may decrease as the sample size for controls increase in a study, everything else remaining fixed. Future research is merited to explore methods that can combine the two independent sources of information used in the two-step procedure in a more coherent fashion, while retaining the desirable unbiased property that the Type-I error rate for the procedure overall is not influenced by the underlying gene-environment independence assumption. Analogous developments have recently taken place for combining within and between family information in family-based association studies (29).

The current study is certainly not exhaustive. For example, Kooperberg and LeBlanc (30) proposed another two-step approach to screen for G x G effects by filtering the marginal genetic associations and restricting interaction testing to this subset. Similar strategies can be adapted to G x E screening. Murcray et al. (28) propose a hybrid approach that combines the above Kooperberg-Leblanc marginal screening and the two-step screening of Murcray et al. (15) to improve upon the original two-step procedure. The new paper (28) addresses some of the critiques pointed in the discussion of the original paper by Chatterjee and Wacholder (31), specifically the issue of choosing the optimal α_1. These new and improved methods may be included in the future to expand on the findings of the current study. We primarily considered case-control sampling and did not consider family-based designs. Gauderman et al. (32) extended the two-step method for G x E interactions in case-parent trios. It remains an interesting open question to compare the hybrid methods under studies that include related individuals.

To conclude, we find it encouraging that under realistic violations of gene-environment independence, the hybrid procedures can protect against false positives due to gene-environment association and yet can gain substantial power over the standard case-control analysis. Moreover, under a range of realistic scenarios, the hybrid methods are likely to be conservative and further power gain is possible by using case-only type methods, assuming a moderate number of false positives could be ruled out in further replication studies. Thus, future analysis of gene-environment interactions in GWAS is likely to benefit by using the new alternatives.

Software: The R-codes for simulating power for the different tests of interaction is available at http://www.sph.umich.edu/$\sim$ bhramar/public_html. An R-package CGEN for semiparametric ML and EB procedure is available at http://dceg.cancer.gov/bb/tools/genetanalcasecontdata.

ACKNOWLEDGEMENTS:

Author affiliations: Department of Biostatistics, School of Public Health, The University of Michigan, Ann Arbor, Michigan (Bhramar Mukherjee, Jaeil Ahn); Department of Internal Medicine, Human Genetics and Epidemiology, The University of Michigan, Ann Arbor, Michigan (Stephen B. Gruber); Division of Cancer Epidemiology and Genetics, National Cancer Institute (Nilanjan Chatterjee).

The research of Bhramar Mukherjee was partially supported by NIH grant R03 CA130045-01 and NSF grant DMS-076935. The research of Stephen B. Gruber and Bhramar Mukherjee was supported by NIH grant U19 NCI-895700. Nilanjan Chatterjee's research was supported by the Intramural Research Program of the National Cancer Institute. The authors will like to thank Duncan Thomas and an anonymous reviewer for their valuable comments on the originally submitted manuscript. The authors are grateful to David V Conti and Dailin Li for sharing their R codes for the BMA method.

References

1. Garcia-Closas M, Malts N, Silverman D, et al. NAT2 slow acetylation and GSTM1 null geno-

types increase bladder cancer risk: Results from Spanish Bladder Cancer Study and meta-analysis. *Lancet.* 2005;366:649-659.

2. Risch N, Herrell R, Lehner T, et al. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *JAMA*. 2009;301:2462-2471.

3. Gail M. Discriminatory Accuracy from Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. *J Natl Cancer Inst*. 2008;100:1037-1041.

4. Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med.* 2010;362(11):1043-1045.

5. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Gen.* 2010;42:570-575.

6. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. *American J of Epid.* 2009;169:219-226.

7. Thomas DC. Gene-environment-wide association studies: emerging approaches. *Nature Reviews, Genetics*. 2010;11:259-272.

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

9. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Amer Jour of Epid.* 2001;154:687-693.

10. Mukherjee B, Ahn J, Rennert G, et al. Testing gene-environment interaction from case-control data: A novel study of Type-1 error, power and designs. *Gen Epid.* 2008;32:615-626.

11. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat in Med.* 1997;16:1731-1743.

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

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

14. Li D, Conti DV. Detecting Gene-Environment Interactions Using a Combined Case-Only and Case-Control Approach. *Am J Epidemiol*. 2009; 169:497-504.

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

16. Cornelis M, Tchetgen E, Liang L, et al. Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. (*Unpublished Manuscript i Submission*).

17. Chen J, Chatterjee N. Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. *Hum Hered.* 2007; 63:196-204.

18. Luo S, Mukherjee B, Chen J, Chatterjee N. Shrinkage estimation for robust and efficient screening of single-SNP association from case-control genome-wide association studies. *Genet Epidemiol.* 2009; 33:740-750.

19. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. New York, Springer Verlag;1985.

20. Greenland S. Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum likelihood, preliminary-testing, and empirical Bayes regression. *Stat in Med.* 1993; 12:717-736.

21. Agresti A. Categorical data analysis (2nd ed.). New York: John Wiley and Sons;2002.

22. Hjort NL, Claeskens G. Frequentist model average estimators (with discussion). *Jour of the Amer Stat Assoc.* 2003; 98:879-99.

23. Schwarz, G. Estimating the dimension of a model. *The Annal of Stat.* 1978; 6:461-464.

24. Satten GA, Kupper LL.Inferences about exposure-disease associations using probability-of-exposure information. *Jour of the Amer Stat Assoc.* 1993; 88:200-208.

25. Chatterjee N, Kalaylioglu Z, Moslehi, R, et al. Powerful Multilocus Tests of Genetic Associ-

ation in the Presence of Gene-Gene and Gene-Environment Interactions. *Am J Human Genet.* 2006; 79:1002-1016.

26. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered.* 2007; 63:111-119.

27. Mirea L, Sun L, Staf ford JE, Bull SB. 2010. Using evidence for population stratification bias in combined individual- and family- level genetic association analyses of quantitative traits. Genet Epidemiol 34:502511.

28. Murcray CE, Lewinger JP, Conti DV, Thomas, DC and Gauderman, WJ. Sample Size Requirements to Detect Gene-Environment Interactions in Genome-wide Association Studies. *Genetic Epidemiology*, 2011, E-pub ahead of print.

29 Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, O'Connor GT, Weiss ST, Lange C. On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet.*, 2009 Nov;5(11):e1000741.

30. Kooperberg C, Leblanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* , 2008; 32:255-63.

31. Chatterjee N. and Wacholder S. Efficient testing of gene-environment interaction. *Am J Epidemiol.*, 2009; 169: 231-233.

32. Gauderman WJ,Thomas DC,Murcray CE,Conti D,Li D,Lewinger JP - Efficient genome-wide association testing of gene-environment interaction in case-parent trios. - *Am J Epidemiol*, 2010, 172:116-22.

Table 1. Glossary of Methods with Summary Features and Key Attributes

Method	Key Advantages	Key Limitations
Case-Control (CC)	**Always maintains Type 1 error rate irrespective of gene-environment association. Robust. Makes no assumption about gene-environment independence/dependence. Can provide tests for joint effects.**	**Lacks power as a screening tool for discovery of G x E interaction.**
Case-Only (CO) (Piegorsch et al, 1994)	**Valid under G-E independence. Provides substantial gain in terms of power and efficiency under G-E independence. Does not require use of control data and tests association of G and E in cases. Reasonable control of expected number of false positives in large-scale testing even with G-E association being present for a fraction of SNPs.**	**Severely inflated Type I error rates even with a very small fraction of SNPs showing association with environmental exposure. Loses power advantages if G-E association at the causal locus is negative and interaction log OR is positive (or vice versa). Cannot yield tests for joint effects and only provides test for interaction.**
Empirical Bayes (EB) Empirical Bayes ver 2 (EB2) (Mukherjee and Chatterjee, 2008)	**A data adaptive shrinkage approach that provides increase in power compared to CC and superior control of Type 1 error compared to CO. Completely data driven and not reliant on user-defined choices. Works well across all G-E association scenarios. Converges to CC in large sample. EB is preferred over EB2 in terms of MSE. Can test for joint effects.**	**Does not strictly adhere to nominal Type 1 error level under violation of the G-E independence assumption and moderate sample sizes. Conservative under the independence assumption with lower than nominal Type-1 error levels. Further power-improvements can possibly be achieved with more aggressive shrinkage weights.**
Model Averaging **Bayesian (BMA) (Li and Conti, 2009)** **Frequentist (AIC) (Introduced in this paper for G x E studies)**	**Data-adaptive compromise estimators that trade-off between bias and efficiency and combine CC and CO analysis. Similar to the EB approaches in terms of operating characteristics. AIC uses the normalized model AICs as weights and is completely data driven. Works well across all G-E association scenarios. Converges to CC analysis in large samples. Can test joint effects.**	**Not guaranteed to maintain nominal Type I error levels with moderate sample sizes and under violation of G-E independence. Conservative under G-E independence. BMA depends on prior weight on the case-control vs. case-only analysis and the results are dependent on this choice.**
Two-Step Procedure (TS) (Murcray et al,2009)	**Uses the independence assumption at the first step scan by testing association of G and E in cases and controls. Filtered markers are followed up by case-control test. Provides power gain by using a powerful first-step scan and reduction in multiple testing burden at second step. Always maintains nominal Type 1 error level and provides power gain in most settings.**	**Loses power advantages when G-E association is negative and interaction positive (or vice versa), a situation where case-only also suffers. With more controls than cases, the power-advantage over one-step case-control decreases. The power advantages depend on the choice of the first-step significance threshold. Cannot provide tests for joint effects.**

Table 2. Family-wise Type 1 error rate (Expected number of false positives) corresponding to the 8 testing procedures when p_{ind}, the fraction of SNPs being independent of E, varies from 0.95 to 1.00. Number of Markers considered $M = 100,000$.

	Method							
p_{ind}	CC	CO	EB	EB2	TS(α_1=0.0005)	TS(α_1=0.05)	AIC	BMA
0.9500	0.050 (0.051)	1.000 (158.444)	0.186 (0.205)	0.138 (0.148)	0.058 (0.059)	0.048 (0.048)	0.064 (0.065)	0.068 (0.068)
0.9900	0.047 (0.047)	1.000 (30.795)	0.057 (0.058)	0.042 (0.043)	0.051 (0.051)	0.037 (0.038)	0.033 (0.034)	0.033 (0.034)
0.9950	0.050 (0.051)	1.000 (14.772)	0.036 (0.036)	0.031 (0.032)	0.054 (0.056)	0.049 (0.050)	0.031 (0.032)	0.032 (0.033)
0.9975	0.046 (0.047)	1.000 (7.038)	0.026 (0.036)	0.025 (0.025)	0.052 (0.053)	0.050 (0.052)	0.033 (0.033)	0.029 (0.030)
0.9995	0.046 (0.048)	0.802 (0.830)	0.022 (0.023)	0.022 (0.023)	0.047 (0.049)	0.048 (0.050)	0.030 (0.031)	0.030 (0.031)
1.0000	0.047 (0.049)	0.041 (0.042)	0.016 (0.016)	0.016 (0.017)	0.047 (0.048)	0.043 (0.044)	0.030 (0.030)	0.031 (0.031)

†For the simulation 5,000 replicate data sets including 2,000 cases and 2,000 controls, each with genotype information on M=100,000 markers are used. This ensures simulation accuracy for Type I error rates of the order of 0.003 at $\alpha_1 = 0.05$. The prevalence of exposure (E) is 0.5 and the allele frequency of genotype (G) at the causal locus is 0.2 and allele frequencies among null markers are uniformly distributed in [0.1, 0.3]. The odds ratios for E (OR_{10}) and G (OR_{01}) at all loci are fixed as 1.0 and the interaction effect size ($\exp(\beta_{GE})$) at the causal locus is fixed at the null value 1.00. The probability of a null SNP being independent of E, p_{ind} varies in (0.95, 1.0). Family-wise Type 1 error rate is estimated as the empirical proportion of datasets declaring *at least* one null marker to be significant. The expected number of false positives are estimated as the average number of falsely rejected null hypotheses, averaged over 5000 datsets.