

# Exploiting Gene-Environment Interaction to Detect Genetic Associations

Peter Kraft<sup>a, b</sup> Yu-Chun Yen<sup>a</sup> Daniel O. Stram<sup>c</sup> John Morrison<sup>c</sup>  
W. James Gauderman<sup>c</sup>

Departments of <sup>a</sup>Epidemiology and <sup>b</sup>Biostatistics, Harvard School of Public Health, Boston, Mass., <sup>c</sup>Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, Calif., USA

## Key Words

Gene-environment interaction • Power and sample size calculations • Genome-wide association scans

## Abstract

Complex disease by definition results from the interplay of genetic and environmental factors. However, it is currently unclear how gene-environment interaction can best be used to locate complex disease susceptibility loci, particularly in the context of studies where between 1,000 and 1,000,000 markers are scanned for association with disease. We present a joint test of marginal association and gene-environment interaction for case-control data. We compare the power and sample size requirements of this joint test to other analyses: the marginal test of genetic association, the standard test for gene-environment interaction based on logistic regression, and the case-only test for interaction that exploits gene-environment independence. Although for many penetrance models the joint test of genetic marginal effect and interaction is not the most powerful, it is nearly optimal across all penetrance models we considered. In particular, it generally has better power than the marginal test when the genetic effect is restricted to exposed subjects and much better power than the tests of gene-environment interaction when the genetic effect is not restricted to a particular exposure level. This makes the joint test an attractive tool for large-scale association scans where the true gene-environment interaction model is unknown.

Copyright © 2007 S. Karger AG, Basel

Many common diseases are believed to result from the interplay of genetic and environmental factors. This belief is supported by a wide range of empiric evidence. Ethnic differences in response to exposure to cigarette smoking in lung cancer [1], suggest genetic factors modify the exposure effect [2]. A recent report demonstrated that the association between exposure to large roadways and childhood asthma was modified by family history of asthma [3], again indicating potential importance of gene-environment interaction. Evidence that the effect of specific candidate polymorphisms differ according to environmental exposure has also been found. For example, a non-synonymous single nucleotide polymorphism (SNP) in the DNA repair gene *XRCC1* that results in an Arg to Gln amino acid change was associated with a decreased risk of skin cancer among individuals with five or more severe sunburns during their life [4].

However, it is currently unclear how gene-environment interaction can best be used to locate complex disease susceptibility loci, particularly in the context of studies where between 1,000 and 1,000,000 markers are screened for association with disease. Some argue that even if a disease locus only affects disease risk among those exposed to an environmental factor, the locus will likely still have a detectable marginal association with disease [5], so analysis can proceed in the absence of (or ignoring) data on the environmental exposure. Others propose screening markers for deviations from a multiplicative odds ratio model for gene-environment interac-

tion, for example by using a case-only design [6]. A third approach uses data mining methods to build predictors of disease from a large number of genetic and/or environmental inputs [7–10].

The relative power of these various approaches will depend on the true penetrance model, about which we have little information *a priori* for complex diseases. If a locus truly only affects disease risk among exposed (or unexposed) individuals, then that locus may or may not have a detectable marginal effect, depending on the prevalence of exposure and the magnitude of the genetic effect. Conversely, if the locus affects risk among exposed and unexposed individuals, then tests to detect the marginal effect may be more powerful than tests to detect gene-environment interaction, even when genotype odds ratios differ between exposed and unexposed [11, 12]. The performance of data-mining procedures – including power to detect susceptibility loci, Type I error and false discovery rates, and mean prediction error – is not generally clear. Complex disease loci likely have incomplete penetrance, modest odds ratios, and high phenocopy rates, making them poor predictors of disease. Even with a dominant odds ratio of three – larger than anticipated for most complex disease loci – a common allele can be a poor disease predictor [13]. Hence data-mining procedures geared to identify disease predictors may have low power to detect many complex disease loci, unless each acts in a simple Mendelian fashion in a subset of individuals defined by measured genetic or environmental factors.

In this paper, we present a likelihood ratio test of association between disease and a genetic locus, allowing for the possibility that the genetic effect may be modified by an environmental factor. We focus on case-control data, although similar tests can be developed for non-dichotomous phenotypes and other designs (e.g. family studies [11]). This test is sensitive to a range of alternatives, including situations where the genetic effect is constant across environmental strata, and where the genetic effect is restricted to a particular environmental stratum. We compare the power for this test to the marginal test for genetic association (ignoring exposure information), to the standard test for gene-environment interaction using case-control data, and to the test for gene-environment interaction using case-only data. Any of these tests may be applied in the context of a candidate gene study, or in the search for new genetic loci screening thousands of markers spanning large candidate regions or the whole genome.

## Methods

Letting  $D$  be an indicator of disease, we consider true penetrance models of the form

$$\log \frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)} = b_b + b_g G + b_e E + b_{ge} GE, \quad (1)$$

where  $G$  is some genotype coding and  $E = 1$  or  $0$  for exposed or unexposed subjects. For ease of exposition we consider a dominant genotype coding where  $G = 1$  for carriers of the risk allele and  $G = 0$  for non-carriers. The quantity  $OR_{ge} = \exp(b_{ge})$  is the ratio of the odds ratio comparing exposed carriers and exposed non-carriers to the odds ratio comparing unexposed carriers and unexposed non-carriers. If  $b_{ge} = 0$  the genetic odds ratio is constant across exposure strata, and following convention we say there is no (statistical) interaction – that is, the odds ratio comparing exposed carriers to unexposed noncarriers is the product of the odds ratio comparing exposed noncarriers to unexposed noncarriers ( $OR_e = \exp(b_e)$ ) and the odds ratio comparing unexposed carriers to unexposed noncarriers ( $OR_g = \exp(b_g)$ ). If  $b_{ge} > 0$  ( $< 0$ ) the genetic effect is larger (smaller) in exposed individuals than in unexposed individuals. If  $b_g = 0$  and  $b_{ge} \neq 0$  then the genetic effect is restricted to exposed individuals.

We assume the joint distribution of  $G$  and  $E$  in the general population is  $\Pr(G,E) = \Pr(G) \Pr(E)$ , i.e. the genetic and environmental factors are independent. We return to consequences of departures from this assumption in the discussion. We denote the population risk allele frequency as  $q_g$  and assume genotypes are in Hardy-Weinberg proportions (e.g. under the dominant model  $\Pr(G=1) = 1 - (1 - q_g)^2$ ). We define  $q_e$  as the population prevalence of exposure  $\Pr(E=1)$ . The population prevalence of disease is then given by  $k_p = \sum_{G,E} \Pr(D=1|G,E) \Pr(G) \Pr(E)$ .

We evaluate four tests of association, all of which involve likelihood ratio statistics of the form

$$2 \left[ \log L(\hat{\beta}_1) - \log L(\hat{\beta}_0) \right] \quad (2)$$

where  $\hat{\beta}_1$  maximizes the likelihood under the alternative hypothesis and  $\hat{\beta}_0$  maximizes the likelihood under the (constrained) null hypothesis. As described below, the statistics differ in the likelihood  $L$  and the constraints placed on  $\beta_0$ . The statistic (2) has an asymptotic chi-squared distribution with  $j$  degrees of freedom, where  $j$  is the number of parameters constrained under the null.

The non-centrality parameter, which can be used to determine power or sample size, has the form

$$\begin{aligned} \delta &= 2 E \left[ \log L(\tilde{\beta}_1) - \log L(\tilde{\beta}_0) \right] \\ &= 2 n \sum_{D,G,E} \left[ \ell(\tilde{\beta}_1; D, G, E) - \ell(\tilde{\beta}_0; D, G, E) \right] f_{b,q}(D, G, E). \end{aligned}$$

Here  $\tilde{\beta}_1$  and  $\tilde{\beta}_0$  maximize the expected log likelihood, and  $n$  is the total sample size (including both cases and controls, where applicable). The quantity  $\ell(\beta; D, G, E)$  is the log-likelihood for an individual subject, and  $f_{b,q}(D, G, E)$  is the expected fraction of the data with disease status  $D$ , genotype coding  $G$  and exposure  $E$ , given the ascertainment scheme. For example, for case-control data,

$$\begin{aligned} f_{b,q}(D=1, G, E) &= (n_{\text{cases}}/n) \Pr(G, E | D=1) = \\ &= \frac{\Pr(D=1|G,E) \Pr(G) \Pr(E)}{k_p}, \end{aligned}$$

and

$$f_{b,q}(D = 0, G, E) = (n_{\text{controls}}/n) \Pr(G, E | D = 0) = \Pr(D = 0 | G, E) \Pr(G) \Pr(E) / (1 - k_p),$$

where  $\Pr(D|G,E)$  is based on the true model (1) and the ratio  $(n_{\text{cases}}/n)$  is fixed by design. For case-control designs, we will assume  $n_{\text{cases}}/n = n_{\text{controls}}/n = 1/2$ , while in the case-only design  $n_{\text{cases}}/n = 1$ . We calculate  $\delta$  using the exemplary data procedure described in [14].

Given  $\delta$ , we calculate the power for a test with significance level  $\alpha$  as  $1 - X_{j,\delta}(\chi^2_{1-\alpha;j,0})$ , where  $X_{j,\delta}$  is the cumulative density function for a chi-squared random variable with  $j$  degrees of freedom (d.f.) and non-centrality parameter  $\delta$ , and  $\chi^2_{1-\alpha;j,0}$  is the  $1 - \alpha$ -th quantile of the chi-squared distribution with  $j$  d.f. and non-centrality parameter 0. Alternatively, the sample size needed to reject the null hypothesis with power  $1 - \beta$  can be estimated as

$$n = \frac{\delta_{j;\alpha,\beta}^*}{2 \sum_{D,G,E} \left[ \ell(\tilde{\beta}_1; D, G, E) - \ell(\tilde{\beta}_0; D, G, E) \right] f_{b,q}(D, G, E)}. \quad (3)$$

where  $\delta_{j;\alpha,\beta}^*$  solves  $1 - X_{j,\delta}(\chi^2_{1-\alpha;j,0}) = 1 - \beta$ .

We now describe the likelihoods for the four tests that we compare. Note that the null hypotheses for the tests differ.

#### Case-Control Test for Marginal Genetic Effect (G Test)

This is a test of overall association between  $G$  and  $D$ , and does not use any environmental data. The likelihood for this test has the form:

$$\ell^{(1)}(\beta^{(1)}; D, G, E) = \frac{\exp[\beta_b^{(1)} + \beta_g^{(1)}G]^D}{1 + \exp[\beta_b^{(1)} + \beta_g^{(1)}G]}.$$

The test constrains  $\beta_{ge}^{(1)} \equiv 0$  under the null hypothesis of no genetic association, so  $j = 1$ .

#### Case-Control Test for Gene-Environment Interaction (GE Test)

The likelihood for this test includes genetic and environmental 'main effects' and a product 'interaction' term:

$$\ell^{(2)}(\beta^{(2)}; D, G, E) = \frac{\exp[\beta_b^{(2)} + \beta_g^{(2)}G + \beta_e^{(2)}E + \beta_{ge}^{(2)}GE]^D}{1 + \exp[\beta_b^{(2)} + \beta_g^{(2)}G + \beta_e^{(2)}E + \beta_{ge}^{(2)}GE]}. \quad (4)$$

The test constrains the interaction parameter  $\beta_{ge}^{(2)} \equiv 0$  under the null hypothesis of no interaction, so  $j = 1$ .

#### Case-Control Test for Genetic Association, Allowing for Heterogeneity in Genetic Effect Across Exposure Strata (G-GE Test)

This is a joint test that combines information about genetic marginal effect and gene-environment interaction. The test uses the same likelihood as the test for gene environment interaction, but constrains both  $\beta_{ge}^{(2)} \equiv 0$  and  $\beta_g^{(2)} \equiv 0$  under the null, so that  $j = 2$ . We refer to this test throughout the rest of this paper as the 'joint 2-d.f. test.'

#### Case-Only Test for Gene-Environment Interaction (GE<sub>ca</sub> Test)

Like the GE test, this is a test of interaction only. Using only the cases, this likelihood models the relationship between  $E$  and  $G$  and has the form:

$$\ell^{(3)}(\beta^{(3)}; G, E) = \frac{\exp[\beta_b^{(3)} + \beta_{ge}^{(3)}G]^E}{1 + \exp[\beta_b^{(3)} + \beta_{ge}^{(3)}G]}.$$

If  $G$  and  $E$  are independent among the cases,  $\hat{\beta}_{ge}^{(3)}$  is a consistent estimator of the log of the genetic *relative-risk* ratio between exposed and unexposed individuals, i.e. of the parameter  $b_{ge}$  from a log-linear model rather than the logistic model in (1) [15, 16]. However, if the disease is rare (i.e.  $(1 + \exp[b_0 + b_gG + b_eE + b_{ge}GE])^{-1}$  is  $\approx 1$  for all  $G$  and  $E$ ) then  $\hat{\beta}_{ge}^{(3)}$  is also a good estimator of  $b_{ge}$  from the logistic model. Under the null constraint that  $\beta_{ge}^{(3)} \equiv 0$ , this test has  $j = 1$  d.f.

To compare the above tests, we calculated power for studies with 250 cases (and 250 controls, where applicable) and computed sample sizes needed to ensure 80% power. We considered a range of allele frequencies (0.10, 0.25), exposure prevalences (0.10, 0.25), genetic main effect odds ratios ( $e^{bg} = 1.00, 1.75$ ), environmental main effect odds ratios ( $e^{be} = 1.00$  to  $4.00$  by  $0.50$ ), and gene-environment interaction effects ( $e^{bge} = 1.00$  to  $2.50$  by  $0.25$ ). In all calculations, we assumed a significance level of  $\alpha = 0.01$  and a 2-sided alternative hypothesis.

## Results

Table 1 shows the power for allele frequency  $q_g = 0.10$ , exposure frequency  $q_e = 0.25$ , and several settings of the disease model parameters. Not surprisingly, in the absence of interaction ( $OR_{ge} = 1.0$ ), the pure interaction tests (GE and GE<sub>ca</sub>) have no power (beyond the Type I error rate). In the presence of any interaction effect ( $OR_{ge} > 1.0$ ), the case-only test is always more powerful than the GE test, a result that has been shown by others [15, 17, 18]. When  $OR_{ge} = 1.0$ , the G-GE test is always less powerful than the G test. In this case the non-centrality parameters for the two tests are identical; the (often slight) drop in power for the G-GE test follows from the fact that it has one extra degree of freedom.

In the presence of an interaction ( $OR_{ge} \geq 1.5$ ), the relative power of the marginal (G) and flexible (G-GE) tests depends on the main effects of both the genetic ( $OR_g$ ) and environmental ( $OR_e$ ) factors. In the absence of main effects ( $OR_g = OR_e = 1.0$ ) or for small main effects, the G-GE test has greater power than the marginal (G) test. On the other hand, the G test is more powerful than the G-GE test when the genetic and environmental main effects are of modest to large magnitude, even when  $OR_{ge}$  is large. In most situations, however, the difference in power between the G and G-GE tests is small. When there was any genetic main effect ( $OR_g \geq 1.5$ ), both the G and G-GE tests have greater power than the pure interaction (GE and GE<sub>ca</sub>) tests. Interestingly, the 2 d.f. G-GE test is more powerful than the 1 d.f. GE test in all situations shown here, although for a rarer exposure ( $q_e = 0.10$ ) the latter test can be slightly more powerful (see fig. 1).

**Table 1.** Power for four association tests across a range of environmental and genetic main effects and gene-interaction effects

q <sub>e</sub>	OR <sub>e</sub>	OR <sub>g</sub>	OR <sub>ge</sub>	Study with 250 cases				Study with 1,000 cases				
				G	GE	G-GE	G <sub>ca</sub>	G	GE	G-GE	G <sub>ca</sub>	
0.1	1.00	1.00	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
			1.50	0.01	0.02	0.02	0.04	0.02	0.08	0.06	0.19	
			2.00	0.02	0.06	0.05	0.15	0.04	0.30	0.26	0.70	
			2.50	0.03	0.12	0.10	0.34	0.09	0.58	0.55	0.96	
		1.25	1.00	0.06	0.01	0.04	0.01	0.29	0.01	0.21	0.01	
		1.50	0.09	0.02	0.08	0.05	0.46	0.08	0.45	0.22		
		2.00	0.13	0.06	0.15	0.18	0.64	0.31	0.74	0.76		
		2.50	0.18	0.12	0.24	0.39	0.78	0.60	0.91	0.98		
		1.50	1.00	0.24	0.01	0.17	0.01	0.88	0.01	0.81	0.01	
		1.50	0.33	0.02	0.26	0.05	0.95	0.08	0.93	0.24		
		2.00	0.41	0.06	0.38	0.19	0.98	0.32	0.99	0.80		
		2.50	0.50	0.13	0.50	0.42	>0.99	0.61	>0.99	0.99		
	2.00	1.00	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
			1.50	0.02	0.03	0.02	0.07	0.04	0.09	0.08	0.33	
			2.00	0.03	0.07	0.06	0.27	0.14	0.36	0.34	0.91	
			2.50	0.07	0.14	0.12	0.55	0.35	0.66	0.67	>0.99	
		1.25	1.00	0.06	0.01	0.04	0.01	0.29	0.01	0.21	0.01	
		1.50	0.12	0.03	0.09	0.08	0.61	0.10	0.50	0.38		
		2.00	0.22	0.07	0.18	0.30	0.85	0.37	0.81	0.94		
		2.50	0.34	0.15	0.29	0.61	0.96	0.68	0.95	>0.99		
		1.50	1.00	0.24	0.01	0.17	0.01	0.88	0.01	0.80	0.01	
		1.50	0.40	0.03	0.28	0.08	0.98	0.10	0.94	0.42		
		2.00	0.56	0.08	0.42	0.33	>0.99	0.38	0.99	0.96		
		2.50	0.70	0.15	0.56	0.65	>0.99	0.69	>0.99	>0.99		
0.25	1.00	1.00	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
			1.50	0.02	0.04	0.04	0.08	0.06	0.17	0.18	0.42	
			2.00	0.06	0.13	0.14	0.34	0.29	0.61	0.71	0.96	
			2.50	0.13	0.26	0.31	0.65	0.64	0.90	0.96	>0.99	
		1.25	1.00	0.06	0.01	0.04	0.01	0.29	0.01	0.21	0.01	
		1.50	0.16	0.04	0.15	0.09	0.71	0.18	0.74	0.48		
		2.00	0.31	0.13	0.36	0.38	0.94	0.64	0.98	0.98		
		2.50	0.49	0.27	0.59	0.70	0.99	0.91	>0.99	>0.99		
		1.50	1.00	0.24	0.01	0.17	0.01	0.88	0.01	0.81	0.01	
		1.50	0.46	0.04	0.39	0.10	0.99	0.19	0.99	0.52		
		2.00	0.66	0.14	0.65	0.41	>0.99	0.65	>0.99	0.98		
		2.50	0.82	0.28	0.84	0.74	>0.99	0.92	>0.99	>0.99		
	2.00	1.00	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
			1.50	0.04	0.04	0.04	0.10	0.18	0.19	0.23	0.53	
			2.00	0.15	0.14	0.17	0.41	0.70	0.65	0.80	0.98	
			2.50	0.35	0.27	0.38	0.73	0.97	0.91	0.99	>0.99	
		1.25	1.00	0.06	0.01	0.04	0.01	0.29	0.01	0.20	0.01	
		1.50	0.24	0.04	0.17	0.12	0.88	0.20	0.80	0.59		
		2.00	0.53	0.14	0.42	0.46	>0.99	0.67	0.99	0.99		
		2.50	0.78	0.28	0.67	0.77	>0.99	0.92	>0.99	>0.99		
		1.50	1.00	0.24	0.01	0.17	0.01	0.88	0.01	0.80	0.01	
		1.50	0.59	0.05	0.43	0.13	>0.99	0.21	0.99	0.63		
		2.00	0.84	0.15	0.71	0.49	>0.99	0.68	>0.99	0.99		
		2.50	0.96	0.29	0.88	0.80	>0.99	0.93	>0.99	>0.99		

Based on a study with N (250 or 1,000) cases and N controls and Type I error rate  $\alpha = 0.01$  with 2-sided alternative hypothesis. Dominant penetrance model with allele frequency q<sub>g</sub>; population prevalence of E is 0.25; baseline disease prevalence is 0.0001. G is the marginal genetic test; GE is the standard case-control test for gene environment interaction; G-GE is the flexible two-degree-of-freedom test; GE<sub>ca</sub> is the case-only test for gene-environment interaction.

**Fig. 1.** Log<sub>10</sub> sample sizes to achieve 80% power (y-axis) versus the prevalence of the environmental exposure (x-axis) for a range of genetic (OR<sub>g</sub>) main effects and allele frequencies. Type I error rate  $\alpha = 0.01$ ; dominant penetrance model with allele frequency 0.10; gene-environment interaction parameter  $b_{ge} = \log 1.5$ . Sample sizes range from 524 ( $y = 2.72$ ) to 62,814 ( $y = 4.80$ ).

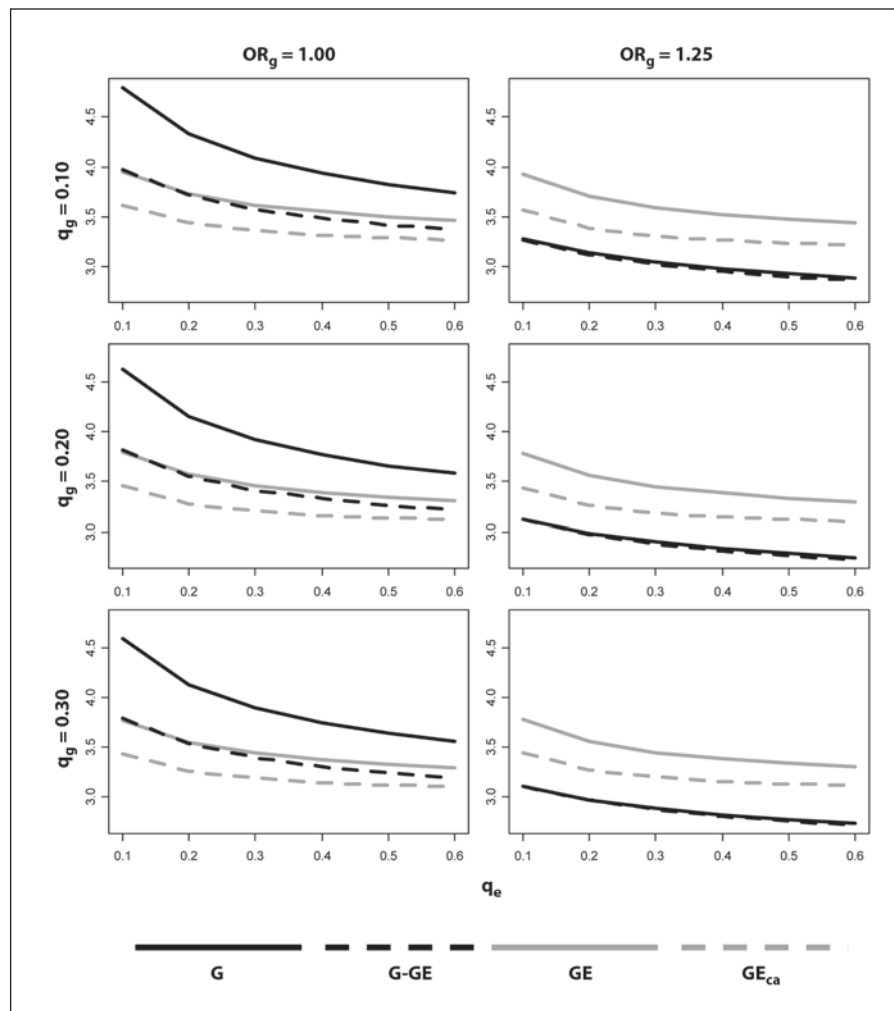
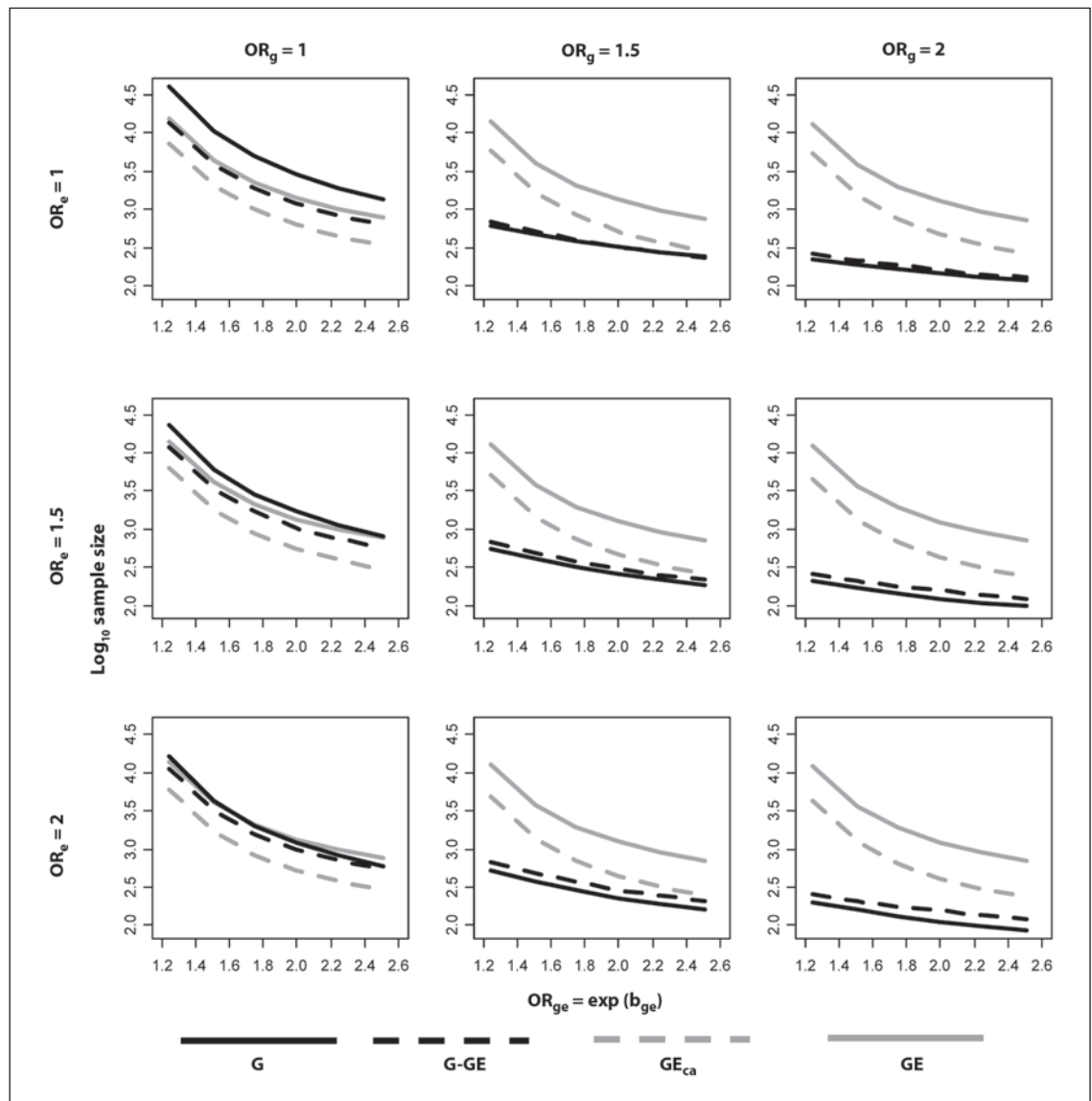


Figure 2 shows the sample sizes (number of cases) needed to achieve 80% power for a range of genetic and environmental main effects and gene-environment interaction effects when the prevalence of exposure is 0.25. The patterns are similar to those seen in table 1. When there is no genetic main effect, the case-only design is most efficient, i.e. requires the lowest sample size to achieve 80% power, while the G-GE test is slightly more efficient than the GE test (with the efficiency advantage increasing as the environmental main effect increases). When there is a genetic main effect ( $OR_g \geq 1.5$ ) the G-GE and G tests are more efficient than both the GE<sub>ca</sub> and GE tests. The marginal test requires a smaller sample size than the G-GE test in these situations, although the differences are relatively modest.

The relative performance of the tests as depicted in figure 1 is identical or nearly so for the more stringent

significance levels needed to control false positives for large-scale association scans involving between 1,000 and 1,000,000 markers. The efficiency of test A relative to test B is the ratio of the sample sizes needed to achieve a given power for a given type I error rate, i.e.  $(\delta_{j_A;\alpha,\beta}^* L_B) / (\delta_{j_B;\alpha,\beta}^* L_A)$ , where  $L_A$  and  $L_B$  are twice the difference of the expected log-likelihoods (the denominator in expression (3)) for tests A and B, respectively. Hence for tests with the same number of degrees of freedom ( $j_A = j_B$ ) – such as the G, GE, and GE<sub>ca</sub> tests – the relative efficiency does not depend on the significance level or desired power. When  $j_A = 2$  and  $j_B = 1$ , the ratio  $\delta_{2;\alpha,\beta}^* : \delta_{1;\alpha,\beta}^*$  decreases as either  $\alpha$  or  $\beta$  decrease, from about 1.25 for  $\alpha = 0.01$  and  $\beta = 0.2$  to 1.10 for  $\alpha = 10^{-6}$  and  $\beta = 0.2$ . This means the efficiency of the G-GE test relative to the others actually increases as the significance level becomes more stringent. Of course the absolute numbers necessary to achieve



**Fig. 2.** Log<sub>10</sub> sample sizes to achieve 80% power (y-axis) versus gene-environment interaction parameter  $\exp(b_{ge}) = OR_{ge}$  (X-axis) for a range of environmental ( $OR_e$ ) and genetic ( $OR_g$ ) main effects. Type I error rate  $\alpha = 0.01$ ; dominant penetrance model with allele frequency 0.10; population prevalence of E is 0.25; baseline disease prevalence is 0.0001. Sample sizes range from 86 ( $y = 1.93$ ) to 40,537 ( $y = 4.61$ ).

80% power will change with more stringent significance levels: about 1.9 times as many subjects for  $\alpha = 10^{-4}$  and 2.8 times as many for  $\alpha = 10^{-6}$ .

### Discussion

We have demonstrated that for case-control designs the flexible two-degree-of-freedom joint test of marginal genetic effects and gene-environment interaction pro-

vides good power for detecting a gene across a wide range of underlying models. The flexible joint test generally has greater power than a simple marginal test when the genetic effect is only expressed in exposed individuals, and it almost always has greater power than a standard case-control based test of interaction. Even in situations for which the marginal test provides the greatest power, the flexible joint test is not substantially less powerful. This makes the flexible test attractive for screening a large number of markers for association with disease, where

the association may be restricted to those exposed to a particular environmental factor, or differ between exposed and unexposed, or alternatively not vary with exposure at all. Similar joint tests of genetic marginal effect and gene-environment interaction could also be implemented for non-dichotomous (e.g. quantitative) phenotypes and/or other designs (e.g. case-parent trios), and we would expect similar trends in relative efficiency compared to other tests.

Large-scale genetic association studies will likely uncover multiple disease susceptibility loci, with distinct penetrance models. Often researchers will have strong *a priori* suspicions that a particular environmental exposure will modify genetic risk, although they will not know whether that exposure modifies the effect of all risk genes, or how it changes the effect of any particular risk gene [19]. Rather than screen thousands of markers using multiple tests, each sensitive to a particular alternative, the flexible two-degree-of-freedom test allows researchers to screen markers using a single test that is sensitive to a range of alternatives. This avoids some multiple-testing complications and difficulties interpreting results (although some adjustment for testing many markers will still be needed). In the context of genome-wide association scans, researchers will need to prioritize a set of markers to genotype in additional studies [20, 21]. How should markers that yield strong evidence for a marginal effect but moderate or no evidence for gene-environment interaction be ranked relative to markers that yield strong evidence for gene-environment interaction but little evidence of a marginal effect? The two-degree-of-freedom test avoids this dilemma by providing a single measure of association between a marker and disease, allowing for heterogeneity in genetic effect across environmental strata.

We have focused on a simple dichotomous genetic coding and dichotomous environmental exposure. The logistic regression framework (4) can be naturally extended to other genetic codings and other kinds of environmental exposures (general categorical, ordinal, continuous). It can also accommodate multiple environmental exposures, which is particularly relevant as many complex diseases have more than one known environmental risk factor that may act as a genetic effect modifier. For example, comparing the saturated model involving a dichotomous genetic and  $k$  dichotomous environmental exposures to the model with genetic and environmental main effects leads to a  $2^k - 1$  degree-of-freedom test. To reduce test degrees of freedom and avoid sparse-data problems, three-way and higher-order inter-

actions could be excluded from the alternative model. However, depending on sample size, allele frequency, exposure frequency and number of exposures  $k$ , even this reduced model may have low power or increased Type I error due to high number of degrees of freedom, large number of nuisance parameters (environmental main effects), and small numbers of subjects with rare exposure profiles. Alternatively, multiple tests involving a single environmental exposure could be conducted for each SNP and a permutation procedure (or other multiple-testing correction) could be used to control the Type I error rate. The relative performance of these approaches to multiple potential effect-modifiers (including their efficiency relative to simpler tests such as the marginal test for association) is a subject of ongoing investigation.

Consistent with earlier work [15, 17] we found that the case-only analysis is more powerful than any of the case-control-based tests we considered when the genetic effect is expressed only in exposed individuals (i.e. there is no genetic main effect). Several authors have proposed methods that use data on both cases and controls under the assumption of that genes and environment are independently distributed in the underlying population [22, 23]. Unlike the case-only analysis, these methods can estimate genetic and environmental main effects. In fact, these main-effect estimators can be more efficient than the standard main-effect estimators that do not assume gene-environment independence [22]. This suggests that a flexible joint test that exploits the gene-environment independence assumption could be nearly as powerful as (or more powerful than) the case-only test in all of the situations we considered. Further, these tests could be more powerful than the flexible two-degree-of-freedom test we discussed, which is based on a standard logistic regression model.

However, if genetic and environmental factors are not independently distributed in controls, tests that assume gene-environment independence will have inflated Type I error rates and can lose power relative to standard methods which do not assume gene-environment independence. There are several reasons variation at a single locus might not be independent of environmental exposure, including correlated differences in exposure rates and allele frequencies across latent subpopulations and potential direct effects of the gene on exposure (e.g. genes that affect behavior). Although the probability that any particular locus is not independent of environmental exposure may be small, it is unclear what proportion of the thousands of markers in a screening study will be associated with exposure in the general population. A recent review

of several candidate gene studies found that although statistically significant correlations between genetic and environmental factors were rare, subtle undetected gene-environment correlations could greatly bias case-only interaction estimates [24]. Newer methods that only require genes-environment independence in observed strata [23, 25] may be more robust, but the trade-off between increased power when the gene-environment assumption holds and increased Type I error rate (or decreased power) when it does not is currently not well understood. We emphasize that although we assumed gene-environment independence for the power calculations presented here, the standard gene-environment interaction (GE) and joint tests (G-GE) remain valid (have appropriate Type I error rates) when genotypes and the environmental exposure are correlated. The marginal test will have increased Type I error rate if the exposure has an effect, since the environmental exposure will confound the gene-disease association.

Relative to prospective studies, retrospective case-control studies are generally faster to conduct and less expensive, but the retrospective design is also more susceptible to recall and selection bias [26, 27]. Differential misclassification of exposure between cases and controls may lead to biased estimates of environmental and genetic main effects, although it does not bias estimates of the gene-environment interaction parameter  $b_{ge}$  unless misclassification probabilities depend on genotypes independent of case-control status (an unlikely situation) [28, 29]. Thus recall bias is unlikely to increase the Type I error rate of the interaction tests (GE and  $GE_{ca}$ ), while it may increase the Type I error rate of the joint test (G-GE). Both differential and non-differential misclassification can reduce the power of all tests that use exposure data [30]. Differential participation can also lead to biased estimates of environmental and genetic main effects. If participation rates do not differ by genotype conditional on disease and exposure status, then differential participation does not lead to biased estimates of gene-environment interaction [29]. However, if participation varies by ethnicity, then participation rates may very well differ by genotype conditional on disease and exposure status, leading to selection bias in the estimation of the gene-environment interaction parameter as well.

We emphasize that we have considered these tests in the context of screening markers for association with disease, rather than making etiologic arguments about biologic interaction between a locus and an environmental exposure. The latter is difficult to do from epidemiologic data as the relevant scale of risk measurement is typically

unknown [31]. These tests should also not be used for assessing the public health impact of a particular locus or environmental factor. While etiologic inference and public health assessment are essential tasks, they belong to the process of gene characterization [32], not gene discovery. Certainly any marker-disease associations discovered from genome-wide association scans will require replication in distinct populations, further fine-mapping and laboratory work to pinpoint and understand the functional variants. Once these variants are known, work can proceed on characterizing their effect in human populations.

We have demonstrated that the specific test that will have the greatest power for detecting a genetic association depends on the underlying relationship of the gene and exposure factor to disease. Across a wide range of parameter combinations, we have shown that a joint test of genetic marginal effect and gene-environment interaction provides good power relative to other alternatives. Additionally, the incorporation of environmental data into tests of genetic association has the potential to uncover a specific subgroup (e.g. exposed subjects) for which the genetic effect is largest, and thus might suggest that future studies might best be targeted at that subgroup. We recognize that the best testing approach is highly dependent on the specific disease, the a priori information known at the time of study planning, availability of resources for exposure assessment, etc. To assist investigators in planning a study, we distribute the Windows-based software program QUANTO [33], freely available at <http://hydra.usc.edu/gxe>. A SAS macro is also available at <http://www.hsph.harvard.edu/faculty/kraft/soft.htm>. These programs provide power or required sample size for all of the tests considered in this paper, and thus allow investigators to better understand the relative tradeoffs of each approach for their given situation.

### Acknowledgements

PK was supported by NCI grants U01 CA098233 and P01 CA08796. JG was supported by NIEHS grants P01 ES07048 and U01 ES015090 and by NCI grant RO1 CA52862.

## References

- 1 Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, Le Marchand L: Ethnic and racial differences in the smoking-related risk of lung cancer. *N Engl J Med* 2006;354:333–342.
- 2 Risch N: Dissecting racial and ethnic differences. *N Engl J Med* 2006;354:408–411.
- 3 McConnell R, Berhane K, Yao L, Jerrett M, Lurmann F, Gilliland F, Kunzli N, Gauderman J, Avol E, Thomas D, Peters J: Traffic, susceptibility, and childhood asthma. *Environ Health Perspect* 2006;114:766–772.
- 4 Han J, Hankinson SE, Colditz GA, Hunter DJ: Genetic variation in XRCC1, sun exposure, and risk of skin cancer. *Br J Cancer* 2004;91:1604–1609.
- 5 Clayton D, McKeigue PM: Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;358:1356–1360.
- 6 Botto L, Khoury M: Facing the challenge of complex genotypes and gene-environment interaction: the basic epidemiologic units in case-control and case-only designs; in Khoury M, Little J, Burke W (eds): *Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. Oxford, Oxford University Press, 2004.
- 7 Cupples LA, Bailey J, Cartier KC, Falk CT, Liu KY, Ye Y, Yu R, Zhang H, Zhao H: Data mining. *Genet Epidemiol* 2005;29(suppl 1):S103–109.
- 8 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003;4:701–709.
- 9 Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147.
- 10 Kooperberg C, Ruczinski I: Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005;28:157–170.
- 11 Selinger-Leneman H, Genin E, Norris J, Khlat M: Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? *Genet Epidemiol* 2003;24:200–207.
- 12 Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006;78:15–27.
- 13 Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P: Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;159:882–890.
- 14 Longmate J: Complexity and power in case-control association studies. *Am J Hum Genet* 2001;68:1229–1237.
- 15 Piegorsch W, Weinberg C, Taylor J: Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153–162.
- 16 Yang Q, Khoury MJ, Sun F, Flanders WD: Case-only design to measure gene-gene interaction. *Epidemiology* 1999;10:167–170.
- 17 Yang Q, Khoury MJ, Flanders WD: Sample size requirements in case-only designs to detect gene-environment interaction. *Am J Epidemiol* 1997;146:713–720.
- 18 Gauderman W: Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002;155:478–484.
- 19 Thomas DC: Are we ready for genome-wide association studies? *Cancer Epidemiol Biomarkers Prev* 2006;15:595–598.
- 20 Kraft P: Efficient two-stage genome-wide association designs based on false positive report probabilities. *Pac Symp Biocomput* 2006, pp 523–534.
- 21 Wang H, Thomas DC, Pe'er I, Stram DO: Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol* 2006;30:356–368.
- 22 Umbach D, Weinberg C: Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med* 1997;16:1731–1743.
- 23 Chatterjee N, Carroll R: Semiparametric maximum likelihood estimation exploiting gene-environment independence. *Biometrika* 2005;92:399–418.
- 24 Liu X, Fallin MD, Kao WH: Genetic dissection methods: designs used for tests of gene-environment interaction. *Curr Opin Genet Dev* 2004;14:241–245.
- 25 Chatterjee N, Kalaylioglu Z, Carroll R: Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. *Genet Epidemiol* 2005.
- 26 Langholz B, Rothman N, Wacholder S, Thomas D: Cohort studies for characterizing measured genes. *Monogr Natl Cancer Inst* 1999;26:39–42.
- 27 Garcia-Closas M, Wacholder S, Caporaso N, Rothman N: Inference issues in cohort and case-control studies of genetic effects and gene-environment interactions; in Khoury M, Little J, Burke W (eds): *Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. Oxford, Oxford University Press, 2004.
- 28 Garcia-Closas M, Thompson WD, Robins JM: Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol* 1998;147:426–433.
- 29 Morimoto LM, White E, Newcomb PA: Selection bias in the assessment of gene-environment interaction in case-control studies. *Am J Epidemiol* 2003;158:259–263.
- 30 Garcia-Closas M, Rothman N, Lubin J: Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev* 1999;8:1043–1050.
- 31 Thompson W: Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;44:221–232.
- 32 Thomas D: Gene characterization studies: an overview. *Monogr Natl Cancer Inst* 1999;26:17–23.
- 33 Gauderman WJ: Sample size calculations for matched case-control studies of gene-environment interaction. *Stat Med* 2002;21:35–50.