Published in final edited form as:

Epidemiology. 2011 May ; 22(3): 400–407. doi:10.1097/EDE.0b013e318212fec6.

Family-based Gene-by-Environment Interaction Studies: Revelations and Remedies

Min Shi¹, David M. Umbach¹, and Clarice R. Weinberg¹

¹Biostatistics Branch, NIEHS, NIH, DHHS, Research Triangle Park NC

Abstract

Bias can arise in case-control studies of genotype effects if the underlying population is structured (genetically stratified or admixed). Nuclear-family-based studies enjoy robustness against such bias, provided that inference conditions properly on the parents. Investigators have extended family-based methods to study gene-by-environment interactions, regarding such extensions as retaining robustness. We demonstrate via simulations that, if population structure involves the exposure, nuclear-family-based analyses of gene-by-exposure interaction remain vulnerable to inflated Type I error rates through subtle dependencies that investigators have failed to appreciate. Motivated by the Two Sister Study, an ongoing study of families affected by young-onset breast cancer, we consider a design that supplements the case-parents design with a sibling who is not genotyped but provides exposure data. If, in the population at large, inheritance is Mendelian and genotypes do not influence propensity for exposure, then this four-person (or tetrad) structure permits the study of genetic effects, exposure effects, and genotype-by-exposure interactions. We show for a dichotomous exposure that, when exposure of an unaffected sibling is available, a modification to the analysis of case-sib or tetrad data re-establishes robustness for tests of multiplicative gene-by-environment interaction. We also use simulations to assess the power for detecting interaction across a range of scenarios, designs and analytic methods.

The etiology of most diseases is not purely genetic but involves both genetic variants and exposures. Consequently, along with genetic effects, investigators need to be able to assess environmental effects and gene-by-environment interactions (GxE). Case-control studies can estimate all these effects but are vulnerable to bias due to population structure, a form of unmeasured confounding due to shared ancestry. Subpopulations that preferentially marry within themselves may differ in their frequencies of a particular allelic variant and also in their baseline risk of disease (the risk in non-carriers of the variant). When such population structure exists, failure to stratify on the ethnic subpopulations produces spurious associations between genotype and risk.

Family-based genetic association studies obviate the population-structure problem for genotype effects. When affected individuals and their parents are genotyped, one can condition on the parental genotypes and base inference on the apparent departures from Mendelian transmission that occur when susceptibility alleles are preferentially passed from

Address for correspondence: Clarice R. Weinberg Biostatistics Branch Mail Drop: A3-03 101/A315 National Institute of Environmental Health Sciences Research Triangle Park 27709 Phone: (919) 541-4927 Fax: (919) 541-4311 weinber2@niehs.nih.gov.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

parents to offspring who later develop the condition of interest. Valid analysis is achieved by either conditioning on the parents' genotypes explicitly, as in the conditional logistic approach,¹⁻³ or implicitly through stratification, as in the log-linear likelihood approach.⁴ Other similar approaches also effectively base inference on transmission.⁵⁻⁸

Family-based approaches are also used to test GxE interactions. With case-parents data and a dichotomous exposure, an analysis for GxE can be carried out by comparing the estimated relative risks for exposed versus unexposed offspring.⁹ Other approaches are available for examining GxE with case-parents data for dichotomous as well as continuous exposures.^{3,10-13} All these approaches rely on an assumption that genotype and exposure are independent within families; that is, that, conditional on the parental genotypes, the inherited genotype does not influence propensity for exposure.

One problem implicit in case-parents data is the inability to estimate exposure effects, hindering the interpretability of any apparent GxE effects. To estimate exposure effects, one needs information about the exposure distribution in the population under study—information that is simply not available with a case-parents design. For a rare disease, the controls in a case-control study provide that information; in a disease-discordant sib-pair (case-sib) design, the unaffected sibling provides it.

The study design of the ongoing Two Sister Study is an alternative design that also provides exposure information. The Two Sister Study is a family-based study of genetic and environmental factors involved in the etiology of young-onset breast cancer. Women with breast cancer diagnosed under age 50 are enrolled along with a cancer-free sister. DNA is collected from the affected sister and their parents. Environmental exposures are ascertained by means of both environmental samples from each sister and extensive computer-assisted telephone interviews. This design augments case-parents data with exposure data from an unaffected sibling; we call it a tetrad design because four family members are studied. By staying within a family-based framework, inference about genetic effects remains robust to population structure. Exposure information from the sibling enables study of exposure effects. We expected analysis of this tetrad design to arise from a straightforward melding of the analyses of a case-parents design with a sibling-pair-matched case-control design, because likelihood inference for each design is possible through conditional logistic regression methods. Unfortunately, we learned that even nuclear-family-based analyses of GxE remain vulnerable to bias due to exposure-related population structure.

What has not been previously appreciated is that, given subpopulations distinct enough to produce preferential mating, many exposures might also vary across subpopulations. Suppose there is exposure-related population structure in the sense that subpopulations differ not only in their allele frequencies but also in their exposure distributions. In such a structured population, one might expect the correlation between a measured marker and any causative locus to differ across subpopulations. When subpopulation-specific exposure prevalence is correlated with the subpopulation-specific linkage disequilibrium (LD) between a marker and a causative locus, assessment of GxE interaction is subject to bias, even in family data. Simply stated, the exposure can act as a surrogate for the LD structure between the marker under study and a causative genetic variant such that the exposure and transmission at the marker may be correlated even when there is no interaction between the exposure and a causative genetic variant. Under such scenarios, typical family-based designs do not automatically protect against bias in assessing GxE interactions unless the locus under study is itself the causative locus and is not in LD with any other locus that is causally related to risk. Moreover, such favorable scenarios will be rare in practice, where SNPs being genotyped are typically markers that are associated with risk indirectly, through LD with nearby causative loci.

We use haplotype-based simulations to demonstrate that these biases occur and can dramatically increase the Type I error rate for existing family-based tests of multiplicative GxE interaction. To compare Type I error rates with the nominal, we consider analyses based on case-parents data, case-sib data, and data from our proposed tetrad design—all derived from a highly structured population with a dichotomous exposure. Although these initial results are discouraging, we go on to show that, when one has ascertained the exposure level for an unaffected sibling, a modification to the analysis of case-sib or tetrad data allows one to achieve the nominal Type I error rate despite exposure-related population structure.

Methods

Type I error rates of existing nuclear-family-based methods

We studied the Type I error rates for tests of GxE involving a single SNP and a dichotomous exposure using three nuclear-family-based designs (case-parents, case-sib, and tetrad) and multiple approaches to their analysis. All the approaches are valid for testing GxE if there is no genetic population structure.

For case-parents data, one enrolls affected offspring and their biological parents, genotyping all three individuals in each family and measuring an exposure for the offspring (Table 1). We studied pseudo-sib analysis using conditional logistic regression.^{1,3} The conditional likelihood is that of a 1:3 matched case-control study where the case is matched to three pseudo-sibling controls, namely, the three possible offspring genotypes (other than that of the case) that could have been produced by the parents. The pseudo-siblings all carry the same exposure as the case (the only exposure measured). For this design, we also studied a polytomous logistic regression approach, QPL,¹³ and a non-parametric approach, FBAT-I.¹¹

For case-sib data, one affected and one unaffected sibling are enrolled; both are genotyped and provide exposure information (Table 1). We studied the usual analysis based on the conditional likelihood for a 1:1 matched case-control study and an alternative analysis proposed by Chatterjee et al.,¹⁴ which enforces assumed within-family gene-by-exposure independence by using the conditional likelihood for a 1:3 matched case-control study. In addition to the case and the control sibling, this likelihood uses two pseudo-sibling controls: one with the case's genotype and the control's exposure and another with the control's genotype and the case's exposure.

For tetrad data, one has case-parents data plus the recorded exposure for an unaffected sibling (Table 1). Our proposed analysis uses the conditional likelihood for a 1:7 matched case-control study. Given the parents' genotypes, four offspring genotypes are possible (that of the case and three pseudo-siblings; these genotypes are not necessarily distinct). The seven matched pseudo-sibling controls consist of these four genotypes, each with the unaffected sibling's exposure, and the three non-case genotypes, each with the case's exposure. This 1:7 matching enforces the within-family gene-by-exposure independence.¹⁴ To include PBAT¹⁵ in our comparisons, we augmented the tetrad data with the unaffected sibling's genotype because PBAT requires the unaffected sibling's genotype to test for GxE for a dichotomous exposure.

To avoid bias from model misspecification in assessing GxE interaction, we saturated the model for both genetic and exposure main effects. To reduce the number of parameters, we used a single degree-of-freedom parameterization for interaction. Consequently, for a typical analysis of case-sib (with or without enforcing within-family independence) or an analysis of tetrad data with a dichotomous exposure, we fit a logistic regression model of the form:

$$\text{logit}(D|G, E, \text{family } j) = \alpha_j + \beta_E E + \beta_1 G_1 + \beta_2 G_2 + \gamma_{LIN} G_{LIN} E$$

For case-parents data, the term $\beta_E E$ was omitted as β_E is not estimable. The nuisance parameters, α_j , are neither estimable nor of interest. Here D is the disease indicator; G_i is an indicator that the child carries exactly i copies of the variant; $G_{LIN} = G_1 + 2G_2$ is the number of copies of the variant allele that the child carries, and E is the indicator of the child's exposure. For a rare disease, the β parameters represent log-relative risks for the exposure and genetic main effects. The GxE interaction parameter, γ_{LIN} , models interaction through a single degree-of-freedom log-additive "trend" term. In the absence of population structure, validity for testing interaction is ensured despite possible misspecification of interaction terms if the above model is correct under the null.

Proposed approach to achieve robustness

Suppose that when exposure participates in population structure, the covariance of genotype and offspring exposure (conditioning on parental genotypes) appears as a GxE interaction effect on risk in a naïve analysis. This covariance might be separable into two components, one that represents the actual (within-family) GxE interaction effect and one that represents the spurious association (across families) attributable to the correlation of exposure with subpopulation identity (reflecting sub-population-specific LD between marker and causative locus.) For designs where exposure is available for an unaffected sibling as well as the case, we propose the following logistic regression model. This model accounts for the spurious association and also protects against inflated Type I error rates induced when the exposure participates in population structure (see Appendix):

$$\begin{aligned} \text{logit}(D|G, E, \text{family } j) = & \alpha_j + \beta_E E + \beta_1 G_1 + \beta_2 G_2 + \delta_1 G_1 I_{\left(\bar{E}=0.5\right)} + \delta_2 G_2 I_{\left(\bar{E}=0.5\right)} + \\ & \delta_3 G_1 I_{\left(\bar{E}=1\right)} + \delta_4 G_2 I_{\left(\bar{E}=1\right)} + \gamma_{LIN} G_{LIN} E \end{aligned}$$

Here, $I_{\left(\bar{E}=0.5\right)}$, for example, is an indicator that the average exposure for the case and control is 0.5 (exposed being 1 and unexposed being 0) and δ_1 to δ_4 are parameters that adjust the model for distortions of the genotype main effects due to an across-families association of exposure with sub-population; other symbols are as above.

Simulations

We simulated haplotypes based on HapMap-phased¹⁶ genotype data from the sample with European ancestry, using haplotypes and their frequencies for a 100-kb region around the replication factor C1 gene (*RFC1*). We constructed a haplotype set using 5 LD tagging SNPs for *RFC1*. These SNPs defined 12 haplotypes. We introduced a new locus as a causative SNP, residing on haplotype 1, so we considered 6 SNPs altogether. We simulated a dichotomous exposure assuming a rare-disease model. For each simulated scenario, we generated 1000 datasets, each with 1000 families. We generated families by sampling parental pairs (in scenarios with population stratification, both parents came from the same subpopulation), and then randomly generating two offspring based on Mendelian inheritance. Offspring exposures were randomly assigned according to the exposure prevalence for the corresponding subpopulation. We then generated disease status of a random one of the offspring based on his or her diplotypes and exposure, through the scenario's presumed risk model. Families with an affected offspring were retained until

1000 families were accrued. Imposing the rare disease assumption, the other sibling was taken to be unaffected. In the analysis we fit the above models to each SNP separately and report GxE test results for individual SNPs. We also computed a multi-SNP test of GxE interaction by combining the single correlated SNP tests using Simes' procedure.¹⁷

To examine validity of the tests under population stratification, we simulated a no-interaction null scenario with a dichotomous exposure and two equal-sized subpopulations; each subpopulation had all haplotypes in Hardy-Weinberg equilibrium and the same baseline risk of disease. Risk-haplotype frequency and exposure prevalence were 0.1, 0.05, respectively, in one subpopulation and 0.9, 0.5, respectively, in the other. The haplotypes that were not associated with the risk allele occurred in the same relative frequencies as in HapMap. We set $(R_1, R_2, I_1, I_2, R_e) = (1, 3, 1, 1, 2)$ in each subpopulation, where R_i is the relative risk among the unexposed associated with inheritance of i copies of the causative allele, R_e is the relative risk associated with having the exposure and I_i is the interaction parameter defined as the ratio of the i -copy within-family relative risk among the exposed to that among the unexposed.

To study the influence of risk-allele frequency on power, we generated families either from a homogeneous population or from a stratified population formed from two equal-sized subpopulations, each in Hardy-Weinberg equilibrium. For the homogeneous scenarios, the exposure prevalence was set at 0.3 and the risk-haplotype frequency ranged from 0.1 to 0.5. For population-structured scenarios, the exposures in the two subpopulations were 0.05 and 0.4, respectively, and risk haplotype frequencies ranged from 0.1 up to 0.5 in population 1 and, correspondingly, from 0.9 down to 0.5 in population 2. (The scenario where both are 0.5 corresponds to an unstructured population.) In these simulations, we set $(R_1, R_2, I_1, I_2, R_e)$ at $(1, 3, 1.5, 2.25, 2)$. Thus, the risk model had interaction parameters where $I_2 = I_1^2$, which corresponded to the interaction parameterization that we used in fitting the simulated data. In general, of course, one need not know the proper interaction model. We did not show power for PBAT because, as currently implemented, it requires log-additive genetic main effects for validity and so is invalid under our simulation scenario.

Results

Simulations under a no-interaction null for a highly structured population where exposure participated in the structure revealed extreme inflation of the Type I error rate for all family-based designs and existing single-SNP analytic methods (Figure 1A). The log-linear⁹ results are not shown because they were the same as those based on the case-parent pseudo-sib analysis. Type I error rates for polytomous logistic and FBAT-I were also inflated, but less so than for the other tests. Only for SNP 1, the causal locus, was Type I error consistent with the nominal 0.05. The exception was PBAT; it showed inflated Type I error even at the causal locus, a feature attributable to its failure to saturate the genetic main effects (at least in the implementation available to us). All other individual SNPs as well as the multi-SNP test based on Simes' procedure showed a strong tendency to reject too often under the no-interaction null.

When we analyzed the same data (either the tetrads or the sibling case-control) using our proposed alternative regression model, which included four additional covariates that allow the siblings' set of exposures to influence the main effects of genotype, the Type I error rates were consistent with the nominal 0.05 level for each SNP individually as well as for the multi-SNP test (Figure 1B). With case-parents data, this kind of alternative analysis is precluded by lack of supplementary exposure data

When exposure does not participate in population structure, the usual family-based analyses of GxE interaction are valid, and inclusion of the terms involving the family-based $\bar{E}$ is unnecessary. In this situation, for the uncorrected multi-SNP test, the tetrad design had the best power (but required the most genotyping, 3 per family); the case-sib analysis via Chatterjee's method had intermediate power; the usual case-sib analysis had the lowest power. The power of case-parents analysis was slightly higher than that of the usual case-sib analysis but still much lower than that of the other designs (Figure 2A). When terms involving $\bar{E}$ were included in these models, the power of each approach fell, as expected (Figure 2B). In this scenario, the magnitude of the power loss was about the same for the two case-sib methods and was larger for the tetrad design. After adjustment, the tetrad analysis and the case-sib analysis via Chatterjee's method exhibited similar power.

When exposure does participate in population structure, the usual family-based analyses of GxE interaction have inflated Type I error rates, as demonstrated. Adjustment by terms involving $\bar{E}$ is necessary for tests to have proper size. As expected, tests of interaction that did not adjust for $\bar{E}$ showed higher apparent power for the multi-SNP test than those that did (data not shown)—but in this situation such tests are invalid and would reject too often even under the null. Among the valid tests in this situation, the tetrad analysis exhibited the highest power, and its adjusted power was similar to that for Chatterjee's method (Figure 3).

Additional results for simulations under a broader range of scenarios are available in the eAppendix (<http://links.lww.com>) and the authors' Web site¹⁸. The general ranking persists in these additional simulation studies.

Discussion

Testing GxE interactions with family-based data has a bit of a checkered history. For a dichotomous exposure, interest properly centers on whether the relative risk associated with carrying a variant genotype differs between exposed and unexposed individuals. While Mendelian inheritance guarantees that family-based, *i.e.*, transmission-based, inferences about genetic effects are protected against inflated Type I error rates due to genetic population structure, our simulations document that this protection does not extend to inference related to gene-by-environment interaction.

One early method¹⁹ treated transmission of the variant allele as the dichotomous event of interest, and used logistic regression to compare transmission rates to unexposed versus exposed affected offspring. A similar and seductively simple method creates a two-by-two table based on categorizing all the heterozygous parents, with transmission/nontransmission of the designated allele forming the columns and exposed/unexposed (offspring) forming the rows.¹⁰ One simply carries out a chi-squared test for independence. Unfortunately, while still used, such approaches that directly compare allelic transmission rates are invalid.^{9,20} First, they do not account for the induced dependency, present even in a homogeneous population, between transmissions from the mother and the father to an affected offspring. Second, and more importantly, in stratified populations transmission rates can differ between exposed and unexposed offspring even when the relative risks for carrying a variant genotype do not.

Other existing methods for testing interaction with case-parents data also in effect compare transmissions to exposed versus unexposed affected offspring, but do not use allelic transmission rates directly. These methods avoid the problems of transmission dependency by treating the family rather than each allelic transmission as the unit of analysis.

Many were developed with candidate SNPs in mind, that is, under the strong assumption that the SNP under study is causative and not in LD with another causative SNP. These

methods are valid under that unrealistically narrow assumption, as shown in our simulations. Our simulations further demonstrate that these approaches are invalid generally for structured populations when the structure is exposure-related. The knotty problem is that population structure tends to produce heterogeneity in marker transmission rates even when genotype relative risks for the causative allele do not differ between exposed and unexposed individuals within families.

To demonstrate the potential for inflated Type I error rates when testing GxE interactions, we used an extreme scenario in our simulations. For less extreme scenarios, the inflation will be less. Of course, an investigator will generally not know how extreme exposure-related population structure may be in a targeted population. Other approaches to alleviate such bias include stratification on reported ethnicity or on strata derived from a large genome-wide panel of SNPs.²¹ The ability of such methods to overcome bias depends heavily on how well the assigned strata can identify sufficiently homogeneous subpopulations. If families can be assigned unambiguously to their truly relevant subpopulation, then stratification will correct the inflation of Type I error rates for GxE; however, in most settings this expectation would be unrealistic.

Tests of interaction depend heavily on how one specifies the null model. The choice between the multiplicative versus the additive interaction null models has been a long-debated subject.²² The multiplicative model is widely used mainly due to the mathematical convenience of logistic regression, whereas the additive model has been argued to be more biologically relevant. We focused on testing a multiplicative null in this paper, but the tetrad design and the case-sib design also allow testing of an additive model.

While results shown here have been restricted to a dichotomous exposure and a dichotomous phenotype, the same sorts of biases occur in the more general context where the exposure is continuous or even where the phenotype is quantitative. Biases can also occur in haplotype-based analyses when haplotypes under study are in LD with a causative SNP. Our simulations indicate that the Type I error rates are inflated for several haplotype-based approaches such as GEI-TRIMM²³ PCPH,²⁴ Unphased²⁵ Pseudocontrol²⁶ (see Shi et al¹⁸). Thus, great care must be taken with inferences about GxE interactions when using family data. The usual analyses suggesting causative multiplicative interactions between an exposure and a genotype may simply be showing the tendency for exposure to serve as a marker for the LD relationship between the measured marker (even if a haplotype) and an unmeasured causative variant. It is worth noting that bias can occur even without differential LD in the subpopulations. Consider a scenario where a causative locus A is typed, but there is also another causative locus B. Bias can occur whenever both the haplotype frequencies and exposure prevalence differ in the two subpopulations even if the LD between loci A and B remains the same across subpopulations.²⁷

Our proposed remedy is to adjust for a family-based measure of the exposure distribution, $\bar{E}$, multiplied by genotype. This remedy works extremely well for a dichotomous exposure. If the exposure is continuous, then correcting for exposure-related population structure bias in assessing gene-by-environment interaction becomes more complex, and this problem is the subject of ongoing research.

Of the interaction analyses that used the $G\bar{E}$ -adjustment, the tetrad analysis was virtually identical to the sibpair analysis that imposed within-family G-by-E independence. The other case-sib analysis was consistently less powerful. The within-family independence assumption used here to good advantage is far less stringent than independence of genotype and exposure in the general population (the assumption required for case-only analyses) and should often be plausible. While this supports use of the case-sib method for assessing GxE,

genotyping parents provides more power for testing genetic main effects and it permits additional questions to be addressed, *e.g.*, whether there are prenatal maternally-mediated effects on risk, and whether the risk associated with a variant allele depends on the parent of origin.

Although GE -adjustment removes bias from the assessment of GxE interaction, it also costs power in situations where adjustment is not needed. One could perform a 4-df (degree-of-freedom) likelihood ratio test by comparing the base genotype/exposure models with and without GE -adjustment to investigate whether inference in a particular data set will likely be subject to bias from population structure. To reduce the number of degrees of freedom, one can fit both G and E as linear, resulting in a 1-df likelihood ratio test. Unfortunately even this 1-df test is not very sensitive. One can nevertheless set a liberal α -level (*e.g.* $\alpha=0.2$) and use the model without GE -adjustment to achieve more power when the 1-df likelihood ratio test is not rejected. Empirical Bayes approaches could also potentially help the investigator to negotiate a compromise between bias and efficiency.²⁸

The non-robustness problem highlighted here for family-based analyses of genotype-by-exposure interaction also will plague family-based analyses of genotype-by-genotype (GxG) interactions aimed at elucidating epistatic effects. Even for loci that are unlinked (*e.g.*, on different chromosomes), analyses can generate spurious evidence for epistasis if there is genetic population structure. A robust analysis for GxG interactions could be accomplished through stratification on parental genotypes at both loci (*i.e.*, 36 mating type strata), but this strategy would require a large sample size.

Most researchers, whether involved in the development or application of GxE methods, have mistakenly presumed that family-based methods must be robust to population structure. Under that mistaken assumption, investigators have scanned the genome SNP by SNP looking for GxE. We have shown that when the exposure participates in the population structure, the usual analyses of markers do not guarantee robust tests for GxE interaction effects. Recognition of this potential source of bias is particularly important for SNP-by-SNP analyses of family-based genome-wide association studies. Our proposed method provides one strategy for ameliorating the problem, at least for dichotomous exposures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Susan G. Komen for the Cure (FAS 0703856) for the support of the Two Sister Study and Chia-Ling Kuo and Dmitri Zaykin for their careful review and valuable comments.

Financial Support: Supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES040007; Z01-ES45002).

Appendix: Adjustment for bias in testing within-family GxE interaction for a categorical exposure and a marker SNP

Our simulations demonstrated non-robustness of the GxE analysis using any of a number of methods. This issue arises for case-sib analyses because even after conditioning on both the exposure set for the sibling pair, $\{E_a, E_u\}$, and the genotype set for the sibling pair, $\{G_a, G_u\}$, if the exposure participates in population stratification, the product EG may be predictive of disease even in the absence of within-family causal multiplicative interaction. Spurious interaction arises because E can act as a marker for the subpopulation (*i.e.*,

ancestry), hence, for the LD structure between the SNP marker under study and the causative SNP(s).

EG is predictive, however, only because the conditioning set $\{E_a, E_u\}$ is itself predictive of the LD structure, hence of the main effects of the marker genotype. Let E denote a dichotomous exposure, which is either absent ($E=0$) or present ($E=1$). Let C denote the number of copies of the variant allele carried by the offspring at the marker. A realistic model is:

$$\text{logit}(\text{risk}|\text{family } j, \{E_a, E_u\}, \{G_a, G_u\}) = \alpha_j + \delta E + I_{(C=1)} [\beta_1(\{E_a, E_u\}) + \omega_1 E] + I_{(C=2)} [\beta_2(\{E_a, E_u\}) + \omega_2 E].$$

Here $\beta_1(\{E_a, E_u\})$, for example, denotes a parameter whose value is a function of a set to represent that the relative risk associated with inheriting a single copy of the variant allele can be a function of the observed set of exposures. If the analysis is structured so that any possible dependence of the main effect of G on the set $\{E_a, E_u\}$ is accounted for, then the interaction parameters ω_1 and ω_2 will be 0 unless there is a causal multiplicative $G \times E$ interaction within families. If E is dichotomous, the argument for each of the two $\beta_c(\{E_a, E_u\})$ functions has three possible values, so we can saturate the main effects of G in a way that allows for heterogeneity of LD, by allowing three distinct values for each β coefficient. Thus, an analysis that fully stratifies in this way, by including four additional adjustment parameters, provides a robust test for within-family $G \times E$ interaction. The adjustment has an obvious extension for a multi-level categorical E , but how to saturate the G effects when E is continuous remains problematic.

References

1. Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics*. 1991; 47(1):53–61. [PubMed: 2049513]
2. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*. 1993; 53(5):1114–26. [PubMed: 8213835]
3. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet*. 2002; 70(1):124–41. [PubMed: 11719900]
4. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet*. 1998; 62(4):969–78. [PubMed: 9529360]
5. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. *Genet Epidemiol*. 2000; 19(Suppl 1):S36–42. [PubMed: 11055368]
6. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*. 1993; 52(3):506–16. [PubMed: 8447318]
7. Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet*. 2000; 67(1):146–54. [PubMed: 10825280]
8. Martin ER, Bass MP, Hauser ER, Kaplan NL. Accounting for linkage in family-based tests of association with missing parental genotypes. *Am J Hum Genet*. 2003; 73(5):1016–26. [PubMed: 14551902]
9. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet*. 2000; 66(1):251–61. [PubMed: 10631155]
10. Schaid DJ. Case-parents design for gene-environment interaction. *Genet Epidemiol*. 1999; 16(3):261–73. [PubMed: 10096689]

11. Lake SL, Laird NM. Tests of gene-environment interaction for case-parent triads with general environmental exposures. *Ann Hum Genet.* 2004; 68(Pt 1):55–64. [PubMed: 14748830]
12. Lim S, Beyene J, Greenwood CM. Continuous covariates in genetic association studies of case-parent triads: gene and gene-environment interaction effects, population stratification, and power analysis. *Stat Appl Genet Mol Biol.* 2005; 4 Article20.
13. Kistner EO, Shi M, Weinberg CR. Using Cases and Parents to Study Multiplicative Gene-by-Environment Interaction. *Am J Epidemiol.* 2009
14. Chatterjee N, Kalaylioglu Z, Carroll RJ. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genet Epidemiol.* 2005; 28(2):138–56. [PubMed: 15593088]
15. Vansteelandt S, Demeo DL, Lasky-Su J, Smoller JW, Murphy AJ, McQueen M, Schneiter K, Celedon JC, Weiss ST, Silverman EK, Lange C. Testing and estimating gene-environment interactions in family-based association studies. *Biometrics.* 2008; 64(2):458–67. [PubMed: 17970814]
16. International HapMap Consortium. The International HapMap Project. *Nature.* 2003; 426(6968): 789–96. [PubMed: 14685227]
17. Simes R. An improved Bonferroni procedure for multiple tests of significance. *Biometrika.* 1986; 73:751–754.
18. Shi, M.; Umbach, DM.; Weinberg, CR. Online supplementary materials: <http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads>
19. Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang KY, Duffy DL, VanderKolk C. Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. *Am J Med Genet.* 1997; 73(3):337–44. [PubMed: 9415696]
20. Shin JH, McNeney B, Graham J. On the Use of Allelic Transmission Rates for Assessing Gene-by-Environment Interaction in Case-Parent Trios. *Ann Hum Genet.* 2010
21. Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N. Using principal components of genetic variation for robust and powerful detection of gene-gene interactions in case-control and case-only studies. *Am J Hum Genet.* 2010; 86(3):331–42. [PubMed: 20206333]
22. Weinberg CR. Less is more, except when less is less: Studying joint effects. *Genomics.* 2009; 93(1):10–2. [PubMed: 18598750]
23. Shi M, Umbach DM, Weinberg CR. Testing Haplotype-Environment Interactions Using Case-Parent Triads. *Hum Hered.* 2010; 70(1):23–33. [PubMed: 20413979]
24. Allen AS, Satten GA. Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. *Genet Epidemiol.* 2007; 31(3):211–223. [PubMed: 17266114]
25. Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered.* 2008; 66(2):87–98. [PubMed: 18382088]
26. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol.* 2004; 26(3):167–85. [PubMed: 15022205]
27. Zaykin DV, Shibata K. Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am J Hum Genet.* 2008; 82(3):794–6. author reply 796–7. [PubMed: 18319078]
28. 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(3):685–94. [PubMed: 18162111]

Figure 1. Simulation results on Type I error rate for tests of GxE interaction in a population with strong exposure-related stratification: A. commonly used methods; B. methods using proposed GE -adjustment. The abscissa indexes single SNP tests for SNPs 1 to 6, followed by a Simes' test for the 6 SNPs together. Symbols: analysis of case-parents data via the pseudo-sib approach (open triangle), via FBAT-I (filled square), and via QPL (open square); analysis of case-sib data via conditional logistic regression (filled circle) and via Chatterjee's method (open circle); analysis of tetrad data (filled triangle); analysis of tetrad data with unaffected sib genotyped via PBAT ($\times$).

Figure 2. Power of tests of GxE interaction for a homogeneous population under the risk scenario (R_1, R_2, I_1, I_2, R_e) = (1, 3, 1.5, 2.25, 2): A. for unadjusted models; B. for models with (dashed

line) and without (solid line) $G\bar{E}$ -adjustment (dashed line with filled triangles represents both the tetrad design and Chatterjee's method for the case-sib design because after $G\bar{E}$ -adjustment their power was the same so the curves coincide). Symbols: analysis of case-parents data via the pseudo-sibling approach (open triangle) and via FBAT-I (filled square); analysis of case-sib data via conditional logistic regression (filled circle) and via Chatterjee's method (open circle); analysis of tetrad data (filled triangle).

Figure 3. Power of valid tests of GxE interaction (those from GE -adjusted models) under a scenario with exposure-related stratification. After GE -adjustment, the power of the tetrad design coincides with that of the Chatterjee's method; therefore the dashed line with filled triangles represents both methods. Symbols: analysis of case-sib data via conditional logistic regression (filled circle); analysis of tetrad data (filled triangle).

Table 1

Information collected for different designs

Study Design	Parents		Cases		Sibling	
	Genotype	Exposure	Genotype	Exposure	Genotype	Exposure
Triad	✓		✓	✓		
Case-sibling			✓	✓	✓	✓
Tetrad	✓		✓	✓		✓