Testing Haplotype-Environment Interactions Using Case-Parent Triads

Min Shi David M. Umbach Clarice R. Weinberg

Biostatistics Branch, NIEHS, NIH, DHHS, Research Triangle Park, N.C., USA

Key Words

Haplotype-environment interaction • Gene-environment interaction • Case-parent triad • Permutation test • Non-parametric test • Population stratification

Abstract

Objective: Joint analysis of multiple SNP markers can be informative, but studying joint effects of haplotypes and environmental exposures is challenging. Population structure can involve both genes and exposures and a case-control study is susceptible to bias from either source of stratification. We propose a procedure that uses case-parent triad data and, though not fully robust, resists bias from population structure. **Methods:** Our procedure assumes that haplotypes under study have no influence on propensity to exposure. Then, under a no-interaction null hypothesis (multiplicative scale), transmission of a causative haplotype from parents to affected offspring might show distortion from Mendelian proportions but should be independent of exposure. We used this insight to develop a permutation test of no haplotype-by-exposure interaction. **Results:** Simulations showed that our proposed test respects the nominal Type I error rate and provides good power under a variety of scenarios. We illustrate by examining whether SNP variants in *GSTP1* modify the association between maternal smoking and oral clefting. **Conclusion:** Our procedure offers desir-

able features: no need for haplotype estimation, validity under unspecified genetic main effects, tolerance to Hardy-Weinberg disequilibrium, ability to handle missing genotypes and a relatively large number of SNPs. Simulations suggest resistance to bias due to exposure-related population stratification.

Copyright © 2010 S. Karger AG, Basel

Introduction

Gene-environment interactions, where the risk of disease associated with carrying a certain genotype is modified by environmental exposures or, equivalently, where the effects of an exposure are modified by genotype, are likely to be important in the etiology of complex diseases. We consider the assessment of gene-by-environment interaction in the context of a case-parent triad design for a young-onset condition.

Haplotypes can carry more information about interactions than single SNPs. Multiple SNPs together may change protein folding and cause enzyme dysfunction, and the risk conferred by an impaired enzyme may depend on the individual's exposure level. Alternatively, association of disease with untyped causative SNPs may be captured by typed markers that are in linkage disequilibrium with them, and again risk may vary with exposure.

Extensions of single-marker methods to haplotype-based methods are not straightforward, however. Haplotypes cannot be uniquely reconstructed because of phase ambiguity, and the number of possible haplotypes grows exponentially with the number of SNPs. Studying haplotype-environment interaction brings added challenges. For a test of haplotype-by-environment interaction to be valid as broadly as possible, the null model should avoid restrictive specification of genetic main effects, because misspecification of the usually unknown mode of inheritance can invalidate a test for interaction. Another complication is that genetic population stratification can also involve the exposure. Haplotype frequencies may correlate with exposure prevalences across subpopulations with different baseline risks of disease, leading to spurious evidence for haplotype-by-environment interaction in case-control studies. Such bias can afflict even single-SNP family-based tests of interaction if they are not constructed carefully [1].

Family-based studies offer several important advantages over case-control studies. They can provide robustness to population stratification if analyzed properly, and they allow testing of nonstandard genetic effects, such as maternally-mediated genetic effects and imprinting (parent-of-origin) effects. Knowledge of family structure helps resolve phase ambiguity. Moreover, for a gene-environment interaction study involving a single SNP, family-based designs can require fewer genotypes than a case-control design to achieve the same power [2–4]. Self-selection based on exposure levels can bias results for a case-control design, but not for a family-based approach, where the inference is based on a correlation between the exposure and transmission patterns. Family designs also offer practical advantages, because recruiting family members can be much easier than identifying and recruiting appropriate population-based controls.

Other studies have considered testing haplotype-environment interaction using case-parent triads. Cordell et al. [5] generalized the case/pseudo-control approach [6] to haplotype analysis and also incorporated estimation and testing of haplotype-environment interaction effects. The software for implementing this approach allows flexible modeling of haplotype relative risks and interaction effects but discards triads when the haplotype phase of the child cannot be deduced from genotype information – potentially inducing substantial loss of statistical power, especially when some genotype data are missing. Allen and Satten [7] developed a method to cope with missing phase and genotype information when analyzing haplotypes and haplotype-environment interac-

tions. Their approach, projection conditional on parental haplotypes (PCPH), is implemented in software that works well when only a few SNPs are studied. Dudbridge [8] developed a widely applicable likelihood-based association analysis for nuclear families. His software, UNPHASED, accommodates multiple offspring and is capable of testing haplotype-environment interactions in triads. These three approaches are based on parametric models for the relationship between individual haplotypes, exposure, and disease risk. With this reliance on parametric models, one can estimate the relative risk parameters in the model and perform tests for individual parameters. On the other hand, inferences depend to a large extent on the correctness of the postulated model and, with many haplotypes, the number of parameters quickly becomes challenging. Vansteelandt et al. [9] described an adaptation of the FBAT approach of Laird and colleagues [10] for studying haplotype-environment interactions. This approach relies in part on choosing a coding for the haplotypes and in this sense shares some pros and cons of the three fully parametric approaches.

We describe a non-parametric method to test haplotype-environment interaction using case-parent triads under the assumption that haplotypes under study have no influence on propensity to exposure. Then, under a no-interaction null hypothesis (multiplicative scale), transmission of a causative haplotype from parents to affected offspring might show distortion from Mendelian expectations but should be independent of exposure. We exploited this insight to develop a permutation test. Defining strata by sums of unphased parental genotypes, we base our test statistic on an approximation to the average of stratum-specific covariances between exposure and a score reflecting transmission. We generate its null distribution by permuting exposure among families within each stratum.

Our method offers several desirable features: it does not require investigators to know or to estimate haplotypes or their frequencies; it retains validity under any genetic main effects; it tolerates Hardy-Weinberg disequilibrium; it handles missing genotypes and it can accommodate a relatively large number of SNPs. Although it is not fully robust, we characterize it as ‘resistant’ to bias because our simulations suggest that it does not exceed its nominal Type I error rate under population stratification that could inflate the Type I error rate of a case-control study. We study its performance under scenarios with different numbers of SNPs, with missing genotypes, with a quantitative or a dichotomous exposure, and with or without exposure-involved population stratification.

Methods

Notation

The fundamental genotype data structure is the case-complement genotype-difference matrix $\mathbf{D}$. [11] Consider k di-allelic markers and n families, where families each contribute one case-parent triad. Let M_{ij} , F_{ij} , and C_{ij} (for $i = 1, \dots, n$, $j = 1, \dots, k$) represent the number of copies of the designated allele carried by the mother, father, and affected offspring in triad i at marker locus j (it does not matter which allele is enumerated). We define the *complement* as the hypothetical matched sibling who carries all the alleles not transmitted to the affected child. Thus, the *complement* inherits $M_{ij} + F_{ij} - C_{ij}$. The paired difference in genotype at every locus between each affected offspring and their *complement* produces an $n \times k$ matrix of differences, $\mathbf{D}$, having the ij -th component equal to $D_{ij} = 2C_{ij} - (M_{ij} + F_{ij})$ ($i = 1, \dots, n$, $j = 1, \dots, k$, and $D_{ij} \in [-2, -1, 0, 1, 2]$). We used the observed matrix $\mathbf{D}$ to study apparent transmission distortion of haplotypes from parents to affected offspring through a haplotype-based transmission distortion test called TRIMM [11]. Like the TDT, $\mathbf{D}$ reflects transmissions and, regardless of population stratification, has expected value $\mathbf{0}$ under the null that the set of k markers is not linked or not associated (within families) with the outcome under study. For a single SNP analysis, a case-complement analysis captures most of the information available through a conditional logistic [12] or log-linear [13] model, except that (as for the TDT) some information is lost for triads where both parents are heterozygous. But the case-complement approach (unlike the others) is easily extended to handle multiple linked SNPs.

Let $\mathbf{X}$ represent an n -component column vector with component X_i denoting the exposure for triad i . The exposure can be dichotomous or quantitative; it could be an exposure that was experienced by the mother during her pregnancy with the affected offspring or an exposure experienced post-natally by the affected offspring.

Necessary Assumptions

Any methods that use a case-parent triad design to study haplotype-by-environment interactions require a conditional or within-family independence assumption: we assume that conditional on parental genotypes, the haplotypes transmitted by the parents do not influence the offspring's propensity to exposure. This within-family independence is much weaker than the assumption required for a case-only analysis, where genotype and exposure must be independent unconditionally [14]. Within-family independence is compatible with, for example, exposure-involved population stratification, whereas unconditional independence is not. In addition, when tests rely on LD between unmeasured causative SNPs and other measured SNPs, any method based on transmissions must assume that the exposure is not confounded with the underlying LD structure.

Heuristic

Assuming conditional independence of haplotypes and exposure, we develop a test for the omnibus null hypothesis that any risk-related multi-locus diplotypes act independently of the exposure, in the sense that the joint effects of genotype and exposure on disease risk are multiplicative. Operationally this null means that, even though transmissions of specific haplotypes to affected offspring might violate Mendelian proportions, those

transmissions should nonetheless be statistically independent of the exposure.

Under the no-interaction null, one might expect the covariance between a *transmission score* at locus j , defined by D_{ij} , and the exposure X_i to have expected value 0 for every j . We need, however, to account for the possibility that the exposure participates in bias-inducing population stratification, which can produce an overall correlation between transmissions to affected offspring and exposure, even under this null [1]. Just as when studying a single SNP, protection against bias requires a strategy that groups together families with similar parental genotypes. Our approach calculates covariances within such groups of families.

For the moment imagine that we know phase and can stratify on parental diplotypes. Let g index strata defined by particular pairs of parental diplotypes. Under the no-interaction null, the covariance between D_{ij} and X_i should be 0 within each stratum regardless of exposure-related genetic population stratification; under alternatives, transmission of some SNPs to affected offspring will instead be correlated with exposure. With that background as motivation, we designed our test procedure to estimate the average conditional-on- g (within-stratum) covariance at each component SNP. We make use of the following decomposition (suppressing subscript i) for locus j :

$$\begin{aligned} E_g(\text{Cov}(D_j, X|g)) &= E_g(E(D_j X|g)) - E_g[E(D_j|g)E(X|g)] \\ &= E(D_j X) - E_g[E(D_j|g)E(X|g)] \end{aligned} \quad (1)$$

Here, E_g denotes the expected value taken over g , and $\text{Cov}(\cdot, \cdot|g)$ denotes the covariance between two specified arguments conditional on parental diplotypes. The last term on the right in (1) can be nonzero under exposure-involved population stratification but does not reflect within-stratum correlation between D_{ij} and X_i . This decomposition indicates how to distinguish an interaction signal from population-stratification-induced bias and serves to motivate the development of our testing approach.

The mean of the cross product terms $D_{ij}X_i$ across all triads, namely:

$$c_j = \frac{1}{n} \sum_{i=1}^n D_{ij}X_i$$

estimates the first term on the right in equation (1). We can generate a corresponding permutation distribution for locus j by repeatedly permuting exposure values among families within each g stratum and computing a c_j each time. Let $\bar{c}_j$ denote the average c_j across permutations. This permutation process recapitulates the null hypothesis by allowing transmissions to behave independently of exposure within each stratum. For the permutation distribution, then, the left side of (1) is zero by construction, and the two terms on the right must equal each other. Thus, $\bar{c}_j$, which estimates $E(D_j X)$ for permutations, also estimates $E_g[E(D_j|g)E(X|g)]$ for permutations. Consequently, we propose to regard permutation-based $\bar{c}_j$ as an estimate of $E_g[E(D_j|g)E(X|g)]$ under the null and to use it to adjust for bias in c_j due to exposure-involved population stratification.

Proposed Tests

To construct a multi-locus test of no haplotype-by-exposure interaction, we consider a vector with component c_j for each locus j ($j = 1, 2, 3, \dots, k$). Because the bias-related second term on the

right in (1) will differ for different loci, from each component c_j we subtract $\bar{c}_j$, its permutation-based mean under the null, to form v_j , $v_j = c_j - \bar{c}_j$. We then form a test statistic $T_{\Sigma v^2}$ as the sum of squared v_j , that is,

$$T_{\Sigma v^2} = \sum_{j=1}^k v_j^2.$$

We use the within-stratum permutation procedure described above to generate the permutation distribution for $T_{\Sigma v^2}$.

A modification of the statistic may enhance power under certain circumstances. The k SNPs may contribute unequally to the genetic risks, hence to the interaction, especially when a disease-susceptibility mutation is one of the typed SNPs and that SNP not only confers risk among the unexposed but also interacts with the exposure. In that circumstance one could potentially increase the power by combining the v_j in a weighted fashion, using the estimated marginal genetic main effects of marker SNPs as weights. Accordingly, in a second procedure we used z_j , the j -th component of the $\mathbf{Z}$ statistic from TRIMM [11] (the standardized mean of the j -th column of $\mathbf{D}$) as a weight for SNP j and formed an alternate test statistic

$$T_{\max_ZV} = \max_j \left([z_j v_j]^2 \right).$$

The use of $T_{\max_ZV}$ should also tend to favor finding interactions when the genotype main effect and a genotype-by-exposure interaction both involve the same haplotype.

In addition to the two test statistics $T_{\Sigma v^2}$ and $T_{\max_ZV}$, we also considered a combined test, T_{combined} , that takes advantage of the complementary strengths of both. To do this, we used the sum_log(P) approach of Shi et al. [11]. If the interaction effect is due to a single causative SNP, a SNP that was fortuitously included in our measured set of k markers, then we expect $T_{\max_ZV}$ to do better than $T_{\Sigma v^2}$. If instead a susceptibility haplotype is related to risk, or the SNPs studied behave collectively as markers for an untyped causative SNP or haplotype, then the advantage should go to $T_{\Sigma v^2}$. Consequently, the combination of the two might perform well under either scenario. We refer to the test based on T_{combined} as GEI-TRIMM (gene-environment interaction triad multi-marker method).

With unphased genotype data, parental diplotypes are typically unknown, so instead of strata defined by parental diplotype pairs, we used strata defined by the vector of sums of parental genotypes ($M_{i,1} + F_{i,1}$, $M_{i,2} + F_{i,1}$, L , $M_{i,k} + F_{i,k}$). Strata defined this way represent a coarsening of strata based on parental diplotypes. Consequently, tests based on those strata are less than fully robust but remain resistant to bias from population structure. Other than this re-definition of strata, one generates the permutation distribution as described above. The empirical p value is the proportion of permuted-data-based test statistics that are larger than or equal to the observed-data-based test statistic.

Tests in the Presence of Missing Genotypes

Several available programs, such as Haplore [15] infer haplotypes based on genotypes from a pedigree even with incomplete genotype data. These programs typically provide a list for each family of the haplotype configurations compatible with the observed genotypes and corresponding estimated probabilities. These lists provide an efficient way to include families with miss-

ing genotypes. Using such a list of haplotype configurations, we calculated expected allele counts for each person with a missing SNP genotype and used those expected counts to replace any missing values. In this way, we imputed missing genotypes before calculating $\mathbf{D}$ and the observed-data c_j .

Depending on a family's list of possible data-compatible haplotype configurations, the family may be assignable to several different strata based on the vector of sums of parental genotypes. For permutations, we handled the assignment to strata stochastically, by sampling. For each family with more than one possible triad-haplotype configuration, we first sampled a configuration from the list with probability proportional to the estimated probability from Haplore. Each family then had a provisional haplotype configuration and corresponding assigned stratum. We next generated a permuted data set by permuting exposure values among families within strata defined by summing the two corresponding parental genotypes. Note that, when calculating test statistics for permuted data, we use the same set of (possibly imputed) genotypes as those used for calculating observed-data statistics. The sampling of a triad-haplotype configuration thus helps define the permutation strata (which determine how to permute exposure), but it does not affect $\mathbf{D}$. We then repeated the process by sampling (with replacement) a new configuration for each family from all possible triad-haplotype configurations and then carrying out exposure permutations within the parental-genotype-based strata. For improved computational efficiency, we generated five exposure permutes for each set of sampled triad-haplotype configurations, but one could also use one permute per set.

Simulation Studies

We used simulations to evaluate the performance of our proposed tests in a realistic context. We simulated haplotypes based on HapMap-phased 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*). This haplotype set is defined by 12 SNPs, which specify 17 different haplotypes (Appendix). We then created a thirteenth SNP, representing a causative mutation, whose variant allele resided on haplotype 1. We also constructed a second haplotype set with fewer SNPs by using 5 LD tagging SNPs for *RFC1*. We again introduced a causative SNP, residing on the same haplotype. This second set contained 12 haplotypes (Appendix). In both these haplotype sets, only one haplotype was associated with risk and its frequency was typically 0.28, although we also examined the influence of risk-haplotype frequency on the performance of our tests. Our simulations always involved only a single risk-associated haplotype; in contrast, our test procedures neither presuppose a particular number of risk haplotypes nor designate any haplotypes as candidates.

We simulated both dichotomous and continuous exposures. Dichotomous exposures were generated via Bernoulli trials with outcome probability equal to the designated exposure prevalence. We also simulated continuous exposures from a zero-enriched truncated lognormal distribution. A proportion of individuals had no exposure (again generated via Bernoulli trials). For the remaining individuals, we generated exposure levels by exponentiating a standard normal random number and rejecting values above 6.

Besides specifying distributions of haplotypes and exposure for our simulations, we need a risk model to generate case-parent triads. Our simulations allowed the risk associated with haplotypes to depend on exposure level; but the risk model that we used, though convenient for simulations, is neither required for the validity of nor imposed by our tests. Because the case-parent triad design precludes estimation of effects of exposures, our simulations assumed for simplicity that exposure was unrelated to risk among offspring with no copies of the risk-associated haplotype. Suppose X_0 is a designated reference level of the exposure and let H denote the number of offspring-inherited copies of the risk-associated haplotype. In our simulations, X_0 was always set to zero. We specify the genetic main effects parameters R_h for $h = 1$ or 2 as:

$$R_h = \frac{\Pr[D|H = h, X_0, \text{parental diplotypes}]}{\Pr[D|H = 0, X_0, \text{parental diplotypes}]}.$$

Thus, haplotype relative risks at the reference exposure level depend on the inherited number of copies, H , conditional on parental genotypes. For $\delta = X - X_0$, we define interaction parameters for the risk-associated haplotype as $I_h(\delta)$ for $h = 1$ or 2 as:

$$I_h(\delta) = \frac{\Pr[D|H = h, X_0 + \delta, \text{parental diplotypes}]}{R_h \Pr[D|H = 0, X_0 + \delta, \text{parental diplotypes}]}.$$

$I_h(\delta)$ could be any function of δ , but for simulations we used a log-linear function, so that the $\log I_h(\delta)$ was proportional to δ , a function we can characterize by specifying choices of $I_h(1)$ (simply denoted I_h) for $h = 1$ or 2. Consequently, risk models in our simulations, whether exposure is dichotomous or continuous, can be represented by the risk vector (R_1, R_2, I_1, I_2) for the designated risk-associated haplotype. Implicitly, all other simulated haplotypes have the risk vector $(1, 1, 1, 1)$.

Extending the notation established above from a single risk haplotype to include separate functions $I_h(\delta)$ for each haplotype under study allows a more precise statement of the null hypothesis of interest, namely, $H_0: I_1(\delta) = 1 = I_2(\delta)$ for all δ and for all haplotypes studied. This null asserts that, conditional on parental diplotypes, the relative risks associated with having inherited copies of any haplotype do not vary with the level of the exposure. This null is equivalent to one where the effect on risk of a change in exposure level does not depend on how many copies of any haplotype were inherited. Alternatives would include situations where inheritance of one specific haplotype influences risk in the unexposed while a different haplotype may interact with the exposure. For simplicity, however, we simulate only scenarios where the same haplotype that contributes to risk in the unexposed might also interact with the exposure.

For each simulated scenario, we generated 1,000 data sets, each with 1,000 triads. We generated triads by sampling parental pairs (with population stratification, both parents came from the same subpopulation), randomly generating the offspring haplotypes based on Mendelian inheritance and an offspring exposure from the scenario's exposure model, and then generating disease status randomly from diplotype and exposure through the scenario's presumed risk model. Cases were retained until 1,000 case-parent triads were generated. For each risk scenario, we considered four situations: the mutation SNP was either typed or not and geno-

types had either 0 or 20% of loci randomly missing. We applied our three proposed tests to each simulated data set. This tactic enabled us to examine whether $T_{combined}$ (GEI-TRIMM) had power superior to $T_{\max, ZV}$ when the causative SNP was genotyped and power superior to $T_{\Sigma V^2}$ when the causative SNP was not genotyped. We used 1,000 permutations with each simulated data set to estimate p values.

To examine validity of our tests under population stratification, we simulated a no-interaction null scenario with a dichotomous exposure and two equal-sized subpopulations, each subpopulation having haplotypes in Hardy-Weinberg equilibrium (HWE). We used only the second haplotype set with 6 SNPs (Appendix). The disease ratio between subpopulations was 1:1.5 (ratio of risks for unexposed individuals with no copies of the risk haplotype). Risk-haplotype frequency and exposure prevalence were both 0.2 in one subpopulation and both 0.4 in the other. Thus this no-interaction scenario had both genetic- and exposure-related population stratification. We set $(R_1, R_2, I_1, I_2) = (1.73, 3, 1, 1)$ in each subpopulation.

To further examine the Type I error behavior under population stratification, we used *RFC1* with 6 SNPs again; but we reduced the risk haplotype frequencies and varied the proportional representation of the remaining haplotypes between the two subpopulations. We set the risk haplotype frequencies at 0.1 and 0.05 in the two subpopulations, respectively. In the first subpopulation, all non-risk haplotypes retained their original HapMap frequencies rescaled to total 1–0.1; in the second subpopulation, the non-risk haplotype frequencies were rescaled to total 1–0.05 and then re-assigned at random so that the frequency ranking of haplotypes differed between the two subpopulations. The exposure prevalences, disease ratio, and relative risk parameters were as in the previous paragraph. We constructed 12 distinct simulation scenarios by letting each haplotype in turn be the risk haplotype, and we simulated 1,000 data sets for each scenario.

We examined power for detecting interaction with a dichotomous exposure where either 13 or 6 SNPs (Appendix) were genotyped. Samples came from a homogeneous population, where the exposure prevalence was 0.2. For simplicity, we simulated genotypes in HWE (not required for test validity). We simulated data with four (R_1, R_2, I_1, I_2) settings: $(1, 1, 2, 4)$, $(1, 1, 1, 4)$, $(1, 2, 2, 2)$, or $(1.414, 2, 1.414, 2)$.

We also studied a quantitative exposure where the risk haplotype and exposure had a pure interaction, that is, haplotype was unrelated to risk among the unexposed. We drew triads from a population consisting of a mixture of two equal-sized subpopulations, each with haplotypes in HWE and with disease ratio 1.65. The proportions of unexposed individuals were 0.2 and 0.4, and the haplotype frequencies were 0.5 and 0.2 for the two subpopulations, respectively. We simulated scenarios with four (R_1, R_2, I_1, I_2) settings: a null scenario $(1, 1, 1, 1)$ and three alternative scenarios $(1, 1, 1.2, 1.44)$, $(1, 1, 2.4, 2.4)$ and $(1, 1, 1, 1.36)$.

To study the influence of risk-haplotype frequency on power, we generated triads either from a homogeneous population or from a stratified population formed from two equal-sized subpopulations each in HWE and with disease ratio 1.65. We used a quantitative exposure with unexposed prevalence of 0.3 for the homogeneous scenarios and of 0.2 and 0.4, respectively, for each subpopulation in the stratified scenarios. In these simulations, we set (R_1, R_2, I_1, I_2) at $(1.2, 1.5, 1.2, 1.44)$. For the homogeneous sce-

narios, risk-haplotype frequencies ranged from 0.05 to 0.95. For the stratified scenarios, risk-haplotype frequencies (f_1 and f_2 , respectively) were set so that the odds ratio:

$$f_1/(1-f_1)/f_2/(1-f_2)=4$$

and the overall haplotype frequency, namely, $(f_1 + f_2)/2$, ranged from 0.05 to 0.95. For example, for an overall haplotype frequency of 0.3, the frequencies in the two subpopulations are 0.5 and 0.2 respectively, which meets the odds ratio specification with $(f_1 + f_2)/2 = 0.3$.

We compared the performance of our method on complete genotype data with the pseudocontrol approach [16]. Because the current implementations of UNPHASED [8] and PCPH [7] do not provide for saturating the genetic main effects (i.e. for fitting a separate parameter for each diplotype), we excluded them from our comparison. We evaluated Type I error rates under a null scenario with population stratification in our second haplotype set with 5 SNPs (causative SNP not typed). This scenario used a quantitative exposure with unexposed prevalences 0.1 and 0.9 and risk haplotype frequencies 0.9 and 0.1 for the two equal-sized subpopulations, respectively. The disease ratio was set at 1.65 and (R_1, R_2, I_1, I_2) at (1.73, 3, 1, 1). We also simulated populations with three haplotypes, each with two SNPs. We simulated a no-interaction null scenario with population stratification: a quantitative exposure with unexposed prevalences 0.2 and 0.4 and risk haplotype frequencies 0.5 and 0.2 for the two equal-sized subpopulations, respectively, with disease ratio 1.65. (R_1, R_2, I_1, I_2) was set at (1, 3, 1, 1). We also simulated a non-null scenario with a homogeneous population: a quantitative exposure with unexposed prevalence 0.3, risk haplotype frequency 0.3, and (R_1, R_2, I_1, I_2) set at (1, 1, 1.2, 1.44). For the pseudocontrol approach [16], we used 'genotype' coding, as described in the program's R-package documentation, to saturate main effects by fitting a parameter for each diplotype and tested for log-additive interaction with a likelihood ratio test.

Results

Simulations

With our second haplotype set (5–6 SNPs, Appendix) and a dichotomous exposure, our tests maintained nominal Type I error rates in the presence of exposure-involved population structure. The observed Type I error rates for all three tests were always statistically compatible with the nominal 0.05 with complete genotypes but fell slightly below nominal with 20% of SNPs missing (data not shown). In the set of simulations where each haplotype in turn was the risk haplotype (with frequency <10%), the empirical Type I error rates across the 12 complete-data scenarios ranged from 0.037 to 0.059.

With a homogeneous population, the power of our three tests with 1,000 triads was reasonably good for a dichotomous exposure with prevalence 0.20 and a risk-haplotype frequency of 0.28 in the pure interaction sce-

narios (1, 1, 2, 4) and (1, 1, 1, 4), regardless of the number of SNPs (table 1). Power was much reduced, however, for scenarios (1, 2, 2, 2) and (1.4, 2, 1.4, 2), where the haplotype influenced risk among both unexposed and exposed. Compared to complete data scenarios, power dropped somewhat under all four risk scenarios when 20% of the SNP genotypes were sporadically missing. When all other factors were equal, 6-SNP scenarios had somewhat higher power than 13-SNP scenarios. This observation could in part reflect the LD structures for the particular haplotypes used in the simulations; or it could reflect the increase in the number of strata that need to be accommodated when the lists of data-compatible haplotype configurations become long.

The drop in power from the pure interaction risk scenarios with a dichotomous exposure can be understood by considering I_2/I_1 . In families where one parent is heterozygous for the risk haplotype and the other is homozygous, for example, the scenario with (1, 2, 2, 2) produces identical transmission patterns to affected offspring regardless of exposure (because $I_1 = I_2 = 2$). Hence, those families become noninformative for interaction under this scenario.

For a quantitative exposure with exposure-related population stratification, the Type I error rates for our tests were statistically consistent with 0.05 when no genotypes were missing (table 2). Both the 6- and 13-SNP scenarios appeared to maintain the nominal rate whether the causative SNP was typed or not, although the rate was slightly elevated for the scenario with 6 SNPs when the causative SNP was not typed. With 20% of genotypes missing, the Type I error rates were slightly below nominal for all scenarios. This conservatism can perhaps be understood by recognizing that our permutation distribution is constructed as a mixture of possible permutation distributions and as such has a larger variance than the idealized permutation distribution that would have been generated with no phase ambiguity.

As we predicted, for both dichotomous and quantitative exposures, the power for $T_{\Sigma, V}^2$ was superior to that of $T_{\max, ZV}$ when the causative SNP was not genotyped, while the reverse was true when the causative SNP was genotyped. In both situations the combined test, GEI-TRIMM, achieved power that was close to the better of the procedures being combined.

Power for our tests with 1,000 case-parent triads for our quantitative exposure scenarios under population stratification was reasonably good under most risk scenarios that we examined (table 2). We saw the lowest power with risk scenario (1, 1, 2.4, 2.4) with 13 SNPs. Gener-

Table 1. Power of tests with 1,000 case-parent triads from a homogeneous population with risk-haplotype prevalence 0.28 and a dichotomous exposure with prevalence 0.20

Causative SNP	Missing-ness	Tests	6 (5) SNPs				13 (12) SNPs			
			(1, 1, 2, 4)	(1, 1, 1, 4)	(1, 2, 2, 2)	(1.4, 2, 1.4, 2)	(1, 1, 2, 4)	(1, 1, 1, 4)	(1, 2, 2, 2)	(1.4, 2, 1.4, 2)
Typed	none	$T_{\Sigma v^2}$	957	869	502	451	824	808	271	289
		$T_{\max_ZV}$	971	862	565	512	895	847	335	368
		$T_{combined}$	967	886	558	502	879	847	317	345
	20%	$T_{\Sigma v^2}$	952	860	473	366	784	760	202	211
		$T_{\max_ZV}$	966	861	547	434	868	808	262	286
		$T_{combined}$	970	879	518	401	853	819	236	263
Not typed	none	$T_{\Sigma v^2}$	944	804	464	381	819	789	293	292
		$T_{\max_ZV}$	882	693	420	361	775	708	294	276
		$T_{combined}$	937	779	469	400	831	787	306	295
	20%	$T_{\Sigma v^2}$	908	753	394	319	762	730	206	214
		$T_{\max_ZV}$	833	639	351	292	715	640	215	215
		$T_{combined}$	889	724	380	319	762	716	214	218

Entry is the number rejecting the null out of 1,000 simulated data sets.

Table 2. Power of tests with 1,000 case-parent triads from a stratified population with a quantitative exposure

Causative SNP	Missing-ness	Tests	6 (5) SNPs				13 (12) SNPs			
			(1, 1, 1, 1)	(1, 1, 1.2, 1.44)	(1, 1, 2.4, 2.4)	(1, 1, 1, 1.36)	(1, 1, 1, 1)	(1, 1, 1.2, 1.44)	(1, 1, 2.4, 2.4)	(1, 1, 1, 1.36)
Typed	none	$T_{\Sigma v^2}$	54	880	873	826	48	762	464	795
		$T_{\max_ZV}$	53	909	919	843	40	830	587	832
		$T_{combined}$	58	910	906	857	50	816	561	840
	20%	$T_{\Sigma v^2}$	27	868	886	809	23	747	428	761
		$T_{\max_ZV}$	29	892	923	827	30	822	562	798
		$T_{combined}$	24	889	916	830	25	814	533	798
Not typed	none	$T_{\Sigma v^2}$	65	847	897	778	57	746	540	748
		$T_{\max_ZV}$	57	799	872	703	54	741	540	701
		$T_{combined}$	64	845	897	769	57	762	555	749
	20%	$T_{\Sigma v^2}$	36	807	845	694	20	702	433	727
		$T_{\max_ZV}$	35	744	827	639	21	688	453	670
		$T_{combined}$	36	796	856	691	20	715	465	718

Entry is the number rejecting the null out of 1,000 simulated data sets.

ally speaking, power was larger when the causative SNP was typed compared to not typed, when data were complete compared to missing, and when 6 tagging SNPs were studied compared to 13.

We examined the power of our combined test GEI-TRIMM as a function of the risk-haplotype frequency for homogeneous and stratified populations, with the causative SNP either typed or not, and with 6 SNPs (fig. 1) or

13 SNPs (fig. 2). For 1,000 case-parent triads, power exceeded ~70% for risk-haplotype frequencies between 0.25 and 0.75, with much lower power for more extreme frequencies.

With our 5-SNP haplotype set (causative SNP not typed) and strong exposure-related population stratification, the pseudocontrol method showed inflated Type I error rates of 0.20, whereas GEI-TRIMM showed a

Fig. 1. Power of $T_{combined}$ (GEI-TRIMM) with 1,000 triads at different haplotype frequencies for scenarios with six (causative SNP typed) or five (causative SNP not typed) SNPs where samples are from either a single homogeneous population or a stratified population (see Methods for details). Symbols: circles with solid curve for causative SNP typed homogeneous population; diamonds with long-dashed curve for causative SNP not typed, homogeneous population; squares with short-dashed curve for causative SNP typed, stratified population; triangles with dotted curve for causative SNP not typed, stratified population.

Fig. 2. Power of $T_{combined}$ (GEI-TRIMM) with 1,000 triads at different haplotype frequencies for scenarios with 13 (causative SNP typed) or 12 (causative SNP not typed) SNPs where samples are from either a single homogeneous population or a stratified population (see Methods for details). Symbols: circles with solid curve for causative SNP typed homogeneous population; diamonds with long-dashed curve for causative SNP not typed, homogeneous population; squares with short-dashed curve for causative SNP typed, stratified population; triangles with dotted curve for causative SNP not typed, stratified population.

rate of 0.04, which was statistically consistent with 0.05. In a 5-SNP 12-haplotype null scenario with no exposure-related population stratification, the pseudocontrol method still showed inflated Type I error rates (data not shown). These results demonstrate that GEI-TRIMM can cope with larger numbers of SNPs or haplotypes better than can the otherwise sound parametric method, reflecting the difficulties of asymptotic inference in highly parameterized models. For the 2-SNP null scenario, pseudocontrol, and GEI-TRIMM both yielded nominal Type I error rates near 0.05. For the 2-SNP non-null scenario, GEI-TRIMM (power = 0.948) performed better than did the pseudocontrol approach (power = 0.785).

Example: Orofacial Clefts, *GSTP1*, Smoking

GSTP1, one of the major enzymes contributing to inactivation of metabolites from cigarette smoke, is widely expressed in normal human epithelial tissue and is particularly abundant in the lung, esophagus and placenta. *GSTP1* is the most highly expressed *GST* gene in all tissues of both human embryo and fetus [17]. When treated

with carcinogens, mice with deletions in the *GST-π* gene cluster (*Gst-π1*, and *Gst-π2*) showed a significantly increased incidence of papillomas [18]. Thus, variable *GSTP1* expression may be an important determinant of susceptibility to environmental chemicals. Shi et al. [19], using the method of Kistner et al. [4], reported a significant interaction between SNP rs947894 in *GSTP1* and maternal smoking during pregnancy on the risk of cleft lip with or without cleft palate (CL/P) in a Danish sample. We reanalyzed those data to explore possible haplotype-by-smoking interaction using three SNPs (rs947894, rs1799811, rs762803) in *GSTP1* genotyped in 498 Danish triads with CL/P.

We observed $p = 0.062$ for $T_{\Sigma V^2}$, $p = 0.059$ for $T_{\max, ZV}$, and $p = 0.050$ for GEI-TRIMM, suggesting *GSTP1* haplotypes modulate the effect of maternal smoking on risk of CL/P. The corrected mean cross product vector was (0.026, 0.036, 0.013), where the positive signs indicate increased transmission of minor alleles to affected offspring born to smoking mothers compared to affected offspring born to nonsmokers.

Such interactions are, however, hard to interpret without examining main effects. In this example, when we applied TRIMM to the data without considering maternal smoking status, the $\mathbf{Z}$ vector was $(-0.5, -1.3, -1.7)$ with $p = 0.25$. Thus, we saw no significant overall main effect of haplotypes, with possibly some under-transmission of the interaction-identified haplotype to affected offspring. When smokers and nonsmokers were considered separately, the $\mathbf{Z}$ statistic among nonsmokers was $(-1.6, -1.9, -2.45)$, with a TRIMM $p = 0.05$, while that among smokers was $(1.8, 1.3, 0.1)$ with a TRIMM $p = 0.21$. This analysis suggests that a haplotype marked by the minor alleles (the ones being enumerated) at these three loci confers protection among nonsmokers (since all the signs are negative for the $\mathbf{Z}$ vector) but that the protection is attenuated by maternal smoking.

Discussion

Multiplicative interaction between an inherited set of SNP markers and an exposure can be assessed through our combined GEI-TRIMM test. Stratifying on the vector of the sum of unphased parental genotypes allows permutation of exposure measures among genetically similar families, a strategy that confers resistance to exposure-related population stratification and its attendant biases, but does not guarantee complete robustness. Our simulations suggest that, in a realistic setting with a moderate sample size, GEI-TRIMM applied to the offspring-inherited genotype resists bias due to fairly strong exposure-related population stratification and is also reasonably powerful for detecting haplotype-by-exposure interactions – even though Type I error rates sometimes fall below the nominal level (table 2). It continues to perform well when 20% of SNPs are sporadically missing, or when certain individuals (e.g. 20% of the fathers) are missing altogether (data not shown). It requires no parametric assumptions about the exposure distribution, and the investigator need not assume a particular underlying mode of inheritance. Its construction, by building on a combination of tests, allows GEI-TRIMM to perform well if there is a single causative SNP, whether or not that SNP was directly typed. Our simulations were, however, limited to scenarios where at most one haplotype was involved in main effects and in interactions with the exposure. Software for implementing GEI-TRIMM is freely available at <http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads>.

Although GEI-TRIMM can cope more successfully with larger numbers of SNPs than its competitors can in testing haplotype-environment interaction, increasing numbers of SNPs do impose costs. As the number of SNPs increases, the number of possible strata based on the vectors of sums of parental genotypes increases rapidly. Large numbers of strata will reduce power because many strata will then have only one triad, or possibly contain only triads with tied exposure values, and such strata will not contribute to the test statistic. In one extreme example among our simulations with a dichotomous exposure, when 13 SNPs were included in the analysis, we lost 60% of triads in this way. Also, with missing genotype data, GEI-TRIMM and its competitors assume that no recombinations have occurred between parents and offspring. With sets of SNPs that span long genomic regions, the likelihood of recombination grows; and the performance of GEI-TRIMM and its competitors will suffer.

Because GEI-TRIMM does not fully stratify on parental diplotypes, residual bias due to population stratification can potentially persist. One can construct extreme scenarios where a stratum defined by $\mathbf{M} + \mathbf{F}$ still contains subpopulations that differ in both their exposure and risk haplotype distributions. In such a setting the Type I error rate for GEI-TRIMM can be inflated. Consequently, the proposed method can claim only to be resistant to population stratification and not fully robust. Nevertheless, our simulations revealed a Type I error rate that respected the nominal level of the test. Moreover, if many linked SNPs were included in the analysis, fully stratifying on the parental diplotypes, even if possible, would require more strata than would stratification on $\mathbf{M} + \mathbf{F}$, rendering the number of strata so large that the power of the test would be compromised. From this perspective, GEI-TRIMM represents a pragmatic compromise between bias and efficiency.

Just as our TRIMM method [11] can be adapted to detect effects either of an offspring-inherited haplotype or of a haplotype carried by the mother that acts on the offspring via effects on the prenatal environment, the GEI-TRIMM method can also be adapted to detect interactions between exposure and maternal genotype. One simply needs to replace the case-complement difference with the mother-father difference. Conveniently, when no genotypes are missing, GEI-TRIMM tests for interactions with maternal haplotypes are statistically independent of their transmission-based counterparts for interactions with offspring-inherited haplotypes. The user should be cautious, however, when employing these methods to study gene-environment interactions related to maternal genetic effects. The conditional independence assumption

tion needed for the validity of tests for interaction with offspring-inherited genes is a weak one, specifying that inheritance of the haplotype does not influence the offspring's propensity to be exposed, conditional on the parental genotypes. The corresponding assumption needed to assess maternal haplotype-by-exposure interaction is stronger: one must assume that when the parents differ in their diplotypes, that difference is not predictive of the propensity for the mother to be exposed (presumably during or before the pregnancy). In a population with exposure-related population stratification, such an assumption might often fail, for example, if maternal ancestry is related to both maternal genotype and maternal exposures, and ethnically mixed marriages are common.

To our knowledge, GEI-TRIMM is the first method that can validly study the possible multiplicative interaction of many SNPs simultaneously in relation to an exposure, without requiring prior knowledge of candidate haplotypes, or identification of phased diplotypes, or specification of a genetic inheritance model. It can handle sporadically missing SNPs and can fully use data from triads having only one parent available for genotyping. Though GEI-TRIMM evidently can become conservative when

many SNPs are missing, it was designed to be approximately robust to exposure-involved genetic population stratification, and our simulations support that robustness.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES040007; Z01-ES45002). We thank Dr. Jeff Murray for sharing the clefting genotype data and Drs. Dana Hancock and Dmitri Zaykin for helpful comments on the manuscript.

Electronic Resources

Software: <http://dir.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads>.

Appendix

Haplotype structures and frequencies used in the simulation studies were based on HapMap phased genotype data for the *RFC1* gene in the sample with European ancestry.

Appendix 1

First haplotype set with 13 SNPs

	SNP index													Haplotype frequency
	1	2	3	4	5	6	7	8	9	10	11	12	13	
Haplotype														
1	2	1	2	2	2	2	1	1	1	1	2	1	2	0.284
2	1	1	2	1	1	2	2	2	2	2	2	2	1	0.108
3	2	1	2	1	2	2	1	1	1	1	2	1	1	0.095
4	2	1	2	2	2	2	1	1	2	1	1	2	1	0.068
5	1	2	1	1	1	1	1	2	2	1	2	2	1	0.054
6	2	1	2	1	1	1	1	2	2	1	1	2	1	0.054
7	1	2	1	1	1	1	1	2	2	1	1	2	1	0.041
8	1	2	1	2	1	1	1	2	2	1	1	2	1	0.041
9	1	2	2	1	1	1	1	2	2	1	1	2	1	0.041
10	1	2	1	1	1	1	1	2	2	1	2	1	1	0.027
11	1	2	1	2	1	1	1	2	2	1	2	2	1	0.027
12	1	2	1	2	2	2	1	1	1	1	2	1	1	0.027
13	1	2	2	2	1	1	1	2	2	1	1	2	1	0.027
14	1	2	2	2	1	2	2	2	2	2	2	2	1	0.027
15	2	1	2	1	2	2	1	1	1	2	2	2	1	0.027
16	2	1	2	1	2	2	1	1	2	1	1	2	1	0.027
17	2	1	2	2	2	2	2	2	2	2	2	2	1	0.027

SNP 13 is the hypothetical causative SNP that we created (when SNP 13 is not typed, we refer to a 12-SNP scenario); haplotype 1 is the risk-associated haplotype.

Appendix 2

Second haplotype set with 6 SNPs

	SNP index						Haplotype frequency
	2	4	5	7	12	13	
Haplotype							
1	1	2	2	1	1	2	0.284
2	1	1	1	2	2	1	0.108
3	1	1	2	1	1	1	0.095
4	1	1	2	1	2	1	0.054
5	1	1	1	1	2	1	0.054
6	1	2	2	1	2	1	0.068
7	1	2	2	2	2	1	0.027
8	2	1	1	1	1	1	0.027
9	2	1	1	1	2	1	0.135
10	2	2	1	1	2	1	0.095
11	2	2	1	2	2	1	0.027
12	2	2	2	1	1	1	0.027

Five LD-tagging SNPs from the previous set plus SNP 13, which is the hypothetical causative SNP that we created (when SNP 13 is not typed, we refer to a 5-SNP scenario); haplotype 1 is the risk-associated haplotype.

References

- 1 Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000; 66:251–261.
- 2 Gauderman WJ: Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* 2002;21: 35–50.
- 3 Schaid DJ: Case-parents design for gene-environment interaction. *Genet Epidemiol* 1999;16:261–273.
- 4 Kistner EO, Shi M, Weinberg CR: Using cases and parents to study multiplicative gene-by-environment interaction. *Am J Epidemiol* 2009;170:393–400.
- 5 Cordell HJ: Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genet Epidemiol* 2004;26:186–205.
- 6 Self SG, Longton G, Kopecky KJ, Liang KY: On estimating hla/disease association with application to a study of aplastic anemia. *Biometrics* 1991;47:53–61.
- 7 Allen AS, Satten GA: Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. *Genet Epidemiol* 2007;31:211–223.
- 8 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* 2008;66:87–98.
- 9 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: 458–467.
- 10 Laird NM, Horvath S, Xu X: Implementing a unified approach to family-based tests of association. *Genet Epidemiol* 2000;19(suppl 1):S36–S42.
- 11 Shi M, Umbach DM, Weinberg CR: Identification of risk-related haplotypes with the use of multiple snps from nuclear families. *Am J Hum Genet* 2007;81:53–66.
- 12 Schaid DJ: General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996;13:423–449.
- 13 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;62:969–978.
- 14 Piegorsch WW, Weinberg CR, Taylor JA: Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153–162.
- 15 Zhang K, Sun F, Zhao H: Haplore: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics* 2005;21:90–103.
- 16 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;26:167–185.
- 17 Raijmakers MT, Steegers EA, Peters WH: Glutathione s-transferases and thiol concentrations in embryonic and early fetal tissues. *Hum Reprod* 2001;16:2445–2450.
- 18 Henderson CJ, Smith AG, Ure J, Brown K, Bacon EJ, Wolf CR: Increased skin tumorigenesis in mice lacking pi class glutathione s-transferases. *Proc Natl Acad Sci USA* 1998; 95:5275–5280.
- 19 Shi M, Christensen K, Weinberg CR, Romitti P, Bathum L, Lozada A, Morris RW, Lovett M, Murray JC: Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *Am J Hum Genet* 2007;80:76–90.