Published in final edited form as:

Stat Med. 2014 January 30; 33(2): 304–318. doi:10.1002/sim.5930.

Testing gene-environment interactions in family-based association studies using trait-based ascertained samples

Weiming Zhang¹, Carl D. Langefeld², Gary K. Grunwald¹, and Tasha E. Fingerlin^{1,3}

¹Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus

²Department of Biostatistical Sciences, Wake Forest University Health Sciences

³Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus

Abstract

The study of gene-environment interactions is an increasingly important aspect of genetic epidemiological investigation. Historically, it has been difficult to study gene-environment interactions using a family-based design for quantitative traits or when parent-offspring trios were incomplete. The QBAT-I[1] provides researchers a tool to estimate and test for a gene-environment interaction in families of arbitrary structure that are sampled without regard to the phenotype of interest, but is vulnerable to inflated type I error if families are ascertained based on the phenotype. In this study, we verified the potential for type I error of the QBAT-I when applied to samples ascertained on a trait of interest. The magnitude of the inflation increases as the main genetic effect increases and as the ascertainment becomes more extreme. We propose an ascertainment-corrected score test that allows use of the QBAT-I to test for gene-environment interactions in ascertained samples. Our results indicate that the score test and an ad-hoc method we propose can often restore the nominal type I error rate, and in cases where complete restoration is not possible, dramatically reduce the inflation of the type I error rate in ascertained samples.

Keywords

Gene-environment interaction; QBAT-I; ascertainment; family-based association study; quantitative trait

1 Introduction

The personal and public health burden from complex diseases such as type 2 diabetes, heart disease and many forms of cancer is great. The current paradigm is that the etiology of most complex genetic traits involves genetic, epigenetic and environmental factors and their interaction. Even though the exact natures of the interactions are not known, there are

Corresponding Authors: Weiming Zhang, PhD, 13001 East 17th Avenue, Box B-119, Aurora, CO 80045, Ph: 303-724-7548, FAX: 303-724-4476, weiming.zhang@ucdenver.edu, Tasha E. Fingerlin, PhD, 13001 East 17th Avenue, Box B-119, Aurora, CO 80045, Ph: 303-724-4416, FAX: 303-724-4489, tasha.fingerlin@ucdenver.edu.

The authors declare no conflicts of interest and thank the families who volunteered to participate in the AGRE and the IRASFS.

several examples of complex diseases that are likely due to complex interactions between genetic susceptibility and environmental factors. For example, the ratio of female to male patients with lupus is 9:1 and many hypothesize that hormonal reproductive factors interact with genetic factors to influence disease risk [2]. It is clear that understanding the interplay between a gene and the environment has the potential to aid in the development of prevention and treatment strategies. As such, the ability to incorporate tests of interactions between genetic variants and environmental covariates is of great interest and potential utility in both identifying susceptibility variants and in understanding their potential functional role.

One of the major concerns in the implementation of genetic association studies is the potential for population substructure that may lead to confounding by ancestry of the relationship between a genetic factor and the outcome of interest. Family-based association tests were developed specifically to address the issue of population stratification and do not require additional genotyping of ancestry informative markers. Coupling these characteristics with the value of incorporating linkage information and family-based imputation makes methods for family-based testing a valuable approach. The form of the tests for general pedigrees often makes conventional linear modeling and testing strategies inappropriate, such that including environmental covariates either as main effects or in interaction terms is difficult. The transmission disequilibrium test (TDT) [3] uses a trio of parents and their affected child to test for excess transmission of an allele to an affected child from parents who carry two different alleles at a genetic location. The TDT, which is McNemar's Test[3], takes advantage of the perfectly matched nature of transmitted and untransmitted alleles from a single parent. Many extensions to the TDT have been described and implemented, including the ability to consider incomplete trios, nuclear families with more than one child, and extended pedigrees [4–9]. The core of all of these extensions is the ability to condition on the observed parental genotypes of the affected individual(s) if available, or the expected distribution or sufficient statistic for those parental genotypes if they are not observed. Importantly, until recently, none of the extensions allowed for a direct test of the interaction between a genetic covariate and an environmental (nongenetic) covariate due to the complexities of maintaining the ability to condition on the parental genotype configuration in the presence of interaction effects.

The QBAT-I[1] was developed to estimate parameters for and test gene-environment interaction effects in a family-based genetic study. The QBAT-I uses an E-estimation procedure [10] to estimate genetic effects that are adjusted for ancestry. One of the assumptions of using E-estimation in this con is that the sampling scheme is not dependent on the genotypes at the genetic locus under study. If this assumption is not met, bias in the parameter estimates may result, which can lead to inflated type I error rates of the test for the gene-environment interaction. This failure to maintain the appropriate type I error rate if sampling is not independent of genotype is important because for most family-based genetic studies, individuals and families were recruited based on traits (outcomes) of family members. If the locus under study contributes to the trait values that were used for selection (e.g. there is a main effect of the genetic locus on the trait), the assumption cannot be met. Thus, the QBAT-I is not appropriate to test for a gene-environment interaction when

individuals or family members have been ascertained based on the value of a trait under study.

In this paper, we focus on the study of a single marker and a quantitative trait in nuclear families. We quantify the inflation in type I error generated by sampling based on phenotype (ascertainment) and propose two methods to eliminate or reduce this inflation for a quantitative trait. We show that an ascertainment-adjusted score statistic for testing an interaction in parent-offspring trios is robust to ascertainment. We also investigate an ad-hoc approach for larger pedigrees and show that ignoring the phenotype of the proband (the individual through whom the family is recruited) can dramatically reduce the type I error rate in larger pedigrees.

2. Methods

We consider a sample of n families, each with m_i offspring. For the j^{th} offspring in the i^{th} family, we consider the case of a quantitative phenotype Y_{ij} with distribution $N(E[Y], \sigma^2)$. Let X_{ij} be a function of the offspring genotype of interest and Z_{ij} a covariate measuring an environmental exposure.

2.1 Confounding by ancestry, E-estimation and effects of ascertainment

A conventional approach to address the potential for confounding is to adjust for confounders by including covariates for those confounders in whatever statistical model is being used. This approach essentially assumes a specific form of the relationship between the outcome and the potential confounders. When the relationship between the potential confounders and the exposure of interest is better understood than that between the outcome and the potential confounders, an E-estimator [10] for the effect of the exposure is obtained by replacing the observed exposure value with the estimated expected value of the exposure, conditional on the confounders. For family-based genetic association studies, E-estimation provides a simple method to control for ancestry information as described below.

Consider estimating the genetic main effect in a linear model. In order to adjust for confounding by ancestry, we can include some function, $h(S)$, of the parental genotypes, S , which hold all the information about genetic ancestry, in the model

$$Y_{ij} = \beta_0 + \beta_x X_{ij} + h(S_i) + \varepsilon_{ij} \quad (2.1)$$

where $E(\varepsilon_{ij} | S_i, X_{ij}) = 0$. The relationship h between Y and the parental genotypes, S , is not known. However, the distribution of offspring genotypes conditional on parental genotypes is known and follows simple rules of Mendelian inheritance. Thus, $E[X_{ij} | S_i]$ can be easily calculated, and β_x can be estimated by using the E-estimator β_{E-x} :

$$\beta_{E-x} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij} [X_{ij} - E(X_{ij} | S_i)]}{\sum_{i=1}^n \sum_{j=1}^{m_i} X_{ij} [X_{ij} - E(X_{ij} | S_i)]}. \quad (2.2)$$

Valid inference based on the application of E-estimation requires that the model for $E[X_{ij} | S_i]$ is correct and that the ε_{ij} from the model (2.1) are $N(0, \sigma^2)$.

Vansteelandt et al. use this E-estimation approach to develop a family-based test for gene-environment interaction [1]. Consider the following interaction model for the j^{th} offspring in the i^{th} family:

$$Y_{ij} = \beta_0 + \beta_z Z_{ij} + \beta_x X_{ij} + \beta_{xz} X_{ij} Z_{ij} + \varepsilon_{ij} \quad (2.3)$$

Tests of $H_0 : \beta_{xz} = 0$ vs. $H_A : \beta_{xz} \neq 0$ serve as a test for the presence of a gene-environment interaction on Y . Meeting the assumptions of E-estimation in the context of tests based on this model requires that X and Z are conditionally independent given S [1].

There are at least three characteristics of samples obtained from trait-based ascertainment that can result in violation of the assumptions of the QBAT-I when there are main effects of X and Z . Assume a family is selected only when the proband's trait value is larger than some predefined threshold, C (e.g., body mass index ($BMI > 27$)). First, X and Z are dependent conditional on S due to the fact that the proband's trait has to be $> C$ to be included in the sample (Figure 1). Second, the variance of Y depends on X ; that is, there is unequal variance across proband genotype groups. Since the proband genotype, X , and environmental covariate, Z , are dependent under ascertainment, the variance of Y also depends on Z . Third, due to ascertainment $> C$, the proband trait distribution is truncated conditional on X and Z such that the ε_{ij} from model (2.1) are not $iid N(0, \sigma^2)$ even when conditioned on X and Z . Even if there is no main effect of Z , the assumptions of the QBAT-I may not be met under trait-based ascertainment in the presence of a main effect of X . First, the third characteristic of trait-based ascertained samples noted above is true even if there is no main effect of Z on Y . In addition, the QBAT-I assumes that $E[X_{ij} | S_i]$ can be calculated based on Mendel's Laws of Segregation and Independent Assortment; however, those simple proportions do not hold under trait-based ascertainment when there is a main effect of X (Figure 2) such that an alternative estimate of $E[X_{ij} | S_i]$ is required.

2.2 Ascertainment-corrected score test

The QBAT-I is a score test derived for likelihood functions of the form $f(y | \theta)$, where θ is a vector of parameters. Ascertained samples can be considered as being generated from a

truncated distribution, $\frac{f(y|\theta)}{P(A)}$, where A is the selected subset of the original sample space for the trait and $f(y | \theta)$ is defined to be 0 outside A . We consider the case when A is determined by a predefined constant trait value C . When $f(y | \theta)$ is a member of the exponential family

of distributions and C is a constant, then $\frac{f(y|\theta)}{P(A)}$ is a member of an exponential family (Appendix A). Therefore, the appropriate score test for the ascertained samples should be derived from the truncated distribution. Assume our ascertainment criteria is that the

proband's trait value is larger than C so that the truncated trait distribution is $\frac{f(y|\theta)}{P(Y > C)}$. In our further development and simulations, we adopt this ascertainment criterion and assume $f(y | \theta)$ is a normal distribution, but our rationale and methods apply to other ascertainment criteria with fixed boundaries for distributions that are members of the exponential family of distributions (Appendix A).

Assume ascertainment of n independent nuclear families based on the trait value of the proband and that the i^{th} family has m_i offspring. Let Y have probability density function $f(y | x, z, s)$ and cumulative distribution function $F(y | x, z, s)$, with

$$E(Y_{ij}|X_{ij}, Z_{ij}, S_i) = \mu_{ij} + \beta_x X_{ij} + \beta_{xz} X_{ij} Z_{ij} \text{ and } \text{Var}(Y_{ij}|X_{ij}, Z_{ij}, S_i) = \sigma^2,$$

where μ_{ij} is a parameter that accounts for the main effects of Z and S . When $f(y | x, z, s)$ is a normal distribution, the ascertainment-adjusted score function under the null hypothesis of

no interaction (i.e. $H_0: \beta_{xz} = 0$) is $U = \sum_{i=1}^n U_i$, with

$$U_i = \sum_{j=1}^{m_i} (x_{ij} - E(X_{ij}|Z_{ij}, S_i))(z_{ij} - \mu_z) \left(\frac{y_{ij} - \mu_{ij} - \beta_x x_{ij}}{\sigma^2} - \frac{1}{\sigma} \frac{\phi\left(\frac{C - \mu_{ij} - \beta_x x_{ij}}{\sigma}\right)}{1 - \Phi\left(\frac{C - \mu_{ij} - \beta_x x_{ij}}{\sigma}\right)} \right), \quad (2.4)$$

where ϕ and Φ are the probability density function and cumulative distribution function of the standard normal distribution, respectively. This score function is similar to the original QBAT-I score function with the

$$\sum_{j=1}^{m_i} \left[(x_{ij} - E(X_{ij}|Z_{ij}, S_i))(z_{ij} - \mu_z) \left(\underbrace{\frac{1}{\sigma} \frac{\phi\left(\frac{C - \mu_{ij} - \beta_x x_{ij}}{\sigma}\right)}{1 - \Phi\left(\frac{C - \mu_{ij} - \beta_x x_{ij}}{\sigma}\right)}}_{*} \right) \right] \quad (2.5)$$

term compensating for the truncation; the (*) term in (2.5) reflects the derivative of $\ln$

$$\frac{1}{P(Y > C)}.$$

Calculation of the score statistic based on this score function requires estimation of μ_z , $E(X_{ij} | Z_{ij}, S_i)$, μ_{ij} , β_x and σ . We estimate μ_z , the mean of the distribution of Z , as in the QBAT-I [1] and estimate $E(X_{ij} | Z_{ij}, S_i)$ using a spline function to account for non-linearity induced by the ascertainment sampling (see Figure 2 and Appendix B). In the QBAT-I, μ_{ij} is a parameter that accounts for the contribution of Z and S to Y . Under random ascertainment, the choice of the estimate of μ_{ij} affects only the efficiency of the test and does not affect the validity [1]. To estimate μ_{ij} , we maximize the likelihood of the truncated distribution

$$\frac{\frac{1}{\sigma} \phi\left(\frac{y - \beta_0 - \beta_z Z - \beta_s S - \beta_x X}{\sigma}\right)}{1 - \Phi\left(\frac{C - \beta_0 - \beta_z Z - \beta_s S - \beta_x X}{\sigma}\right)} \quad (2.6)$$

to estimate β_z and β_s , and let $\hat{\mu}_{ij} = \hat{\beta}_0 + \hat{\beta}_z Z_{ij} + \hat{\beta}_s S_i$. An alternative of the E-estimator (2.2) of β_x in model 2.1 can be obtained by regressing Y versus $X - \hat{E}(X | S)$ using OLS with no intercept, and this estimator has the same asymptotic distribution as the E-estimator (2.2) [10]. Therefore, we maximize the likelihood

$$\frac{\frac{1}{\sigma} \phi \left(\frac{y - \hat{\mu} - \beta_x (X - \hat{E}(X|Z, S))}{\sigma} \right)}{1 - \Phi \left(\frac{C - \hat{\mu} - \beta_x (X - \hat{E}(X|Z, S))}{\sigma} \right)} \quad (2.7)$$

to estimate β_x and σ . We use the sample variance of the scores multiplied by the number of families to calculate the test statistic

$$\frac{U^2}{n \hat{V}(U_i)} \sim \chi_1^2 \quad (2.8)$$

2.3 Extended pedigrees and an ad-hoc ascertainment correction

The score test described above is valid for trios and can be extended to more general pedigrees. Since many studies have additional family members either in nuclear families or extended pedigrees, we examined a simple ad-hoc method for an ascertainment correction when at least one sibling is available in addition to the proband. Specifically, when a sibling is available, we exclude the proband's phenotype (and perhaps genotype) from analysis, which is essentially an extreme conditioning on the phenotype of the proband. When parental genotypes are complete, the genotype data from the proband can also be ignored. However, when parental genotype information is incomplete, the minimum sufficient statistic for the parental genotypes, S , is derived using all available offspring genotypes and we retain the proband genotype for this purpose.

When there is correlation between the trait of the proband and the trait of the sibling, the sibling trait distribution is also skewed in ascertained samples. Therefore, when there is also an environmental main effect, Z and X are not independent conditional on parental genotype, violating the assumptions of the QBAT-I. However, this dependency is much weaker among siblings than among probands. In addition, the degree of unequal variance across genotypes is expected to be weaker among siblings. When parental genotypes are missing, S is a function of parental genotypes and offspring genotypes such that offspring genotypes are dependent after conditioning on S which may increase the dependence between Z and X for the sibling. Thus, when including the proband genotype, especially when the genetic main effect is strong, we expect that the reduction in the type I error rate is less than that when parental genotypes are available.

2.4 Type I error rate and power under random selection and ascertainment

In order to evaluate the type I error and power of the QBAT-I and our two proposed methods under both random sampling and ascertainment, we studied a biallelic genetic marker with minor allele frequencies (MAF) 0.05, 0.15, 0.25, 0.35, or 0.45. We focused on two classic genetic study designs: a trio study (two parents and an offspring) and a sibling pair study (sibling pairs recruited with or without parents). 10,000 replicate populations were created. Each population consisted of 3,000 nuclear families, each with two parents and two children, from which we sampled 500 families for study, representing the sampling process. We determined parental genotypes based on the MAF at the SNP of interest and

assuming Hardy-Weinberg Equilibrium; we determined offspring genotypes by gene dropping assuming Mendelian inheritance. We used the same 10,000 populations for each scenario; we ignored data on the other sibling or parents if only trios were considered or if parental genotypes were assumed to be unobserved, respectively. Note that since we determined phenotype values after gene-dropping, generating sibships but ignoring one sibling for the trio analyses did not distort the proband distribution of Y .

To allow more direct comparison between our results and those reported by Vansteelandt et al., we used a similar model for the relationships among the phenotypic, genetic and environmental variables for our simulation studies. Recall that for the j^{th} subject in the i^{th} family, we assume $E(Y_{ij}) = \beta_z Z_{ij} + \beta_x X_{ij} + \beta_{xz} X_{ij} Z_{ij}$; here X is a dichotomous variable indicating presence of the putative at-risk allele and Z is a continuous or binary (environmental) covariate. We let the marginal distribution of Y be $Normal(E[Y], 1)$ and the marginal distribution of Z be either $Normal(20, 16)$, $Exponential(0.25)$ or $Bernoulli(0.40)$. We fixed $\beta_z = 1$, which corresponds to a correlation between Y and Z of 0.90 when Z is $Normal(20, 16)$. We drew the sibling trait and the environmental covariates from bivariate normal distributions, varying the correlation coefficient between the proband and sibling trait ($\rho_y = 0.20$ or 0.80) and covariate ($\rho_z = 0.20$ or 0.80) distributions.

If families were ascertained on a trait value, we ascertained based on the trait value of one randomly selected child who was then designated as the proband. We used the sample mean and standard deviation of Y to determine a threshold, C , then selected the first 500 families whose proband had a trait value larger than C and used these same families for all 3 methods to eliminate simulation sampling variability between methods. We considered values of C equal to a) the sample mean $\bar{Y}$, $\bar{Y}$ minus 0.5 standard deviations ($\bar{Y} - 0.5\sigma$) and $\bar{Y}$ plus 0.5 standard deviations ($\bar{Y} + 0.5\sigma$), representing a wide range of ascertainment stringency.

We evaluated the empirical type I error rate for the test of a gene-environment interaction under both random selection and ascertainment for both the trio and sibling pair designs assuming no gene-environment interaction ($\beta_{xz} = 0$). Again following Vansteelandt *et al.*, we explored a range of values for the main genetic effect (β_x), corresponding to correlation coefficients of 0.023, 0.047, 0.071, 0.095, and 0.12 between Y and $\beta_x X$. We examined the nominal type I error rate at $\alpha = 0.01$ and 0.05 .

The basic simulation scheme for power analyses was similar to the scheme used for the type I error rate study. However, we assumed that there was no genetic main effect ($\beta_x = 0$) and examined a set of nonzero gene-environment interaction effects (β_{xz}) corresponding to correlation coefficients of 0.016, 0.035, 0.085, 0.135, and 0.183 between Y and $\beta_{xz} XZ$. Note that these interaction effects would result in estimated main effects ($\beta \neq 0$) in a model that did not include an interaction term.

2.5 Data analysis example

Data analysis in trios—We applied our ascertainment-adjusted score statistic to trio data from the Autism Genetic Resource Exchange (AGRE; [11]). AGRE is an ongoing DNA repository and family registry established in 1997 by Cure Autism Now (CAN). Support for AGRE is provided in large part by Cure Autism Now and federally-funded grants. Families

are recruited through a variety of methods (e.g., physician referral, Web site contact, and family meetings and seminars). Family recruitment and phenotypic assessment have been previously described in detail [11]. Briefly, families are ascertained on the basis of at least two family members meeting criteria for a diagnosis of an Autism Spectrum Disorder (ASD) (autism, Aspergers, or pervasive developmental disorders). Diagnosis is established by the Autism Diagnostic Interview-Revised (ADI-R; [12]), which is currently the gold standard for research diagnosis. To be scored as affected, individuals must meet criteria in all three content areas of the ADI-R: (1) quality of social interaction, communication, and language; (2) repetitive, restricted, and stereotyped interests and behavior; and (3) age at onset <3 years. We used the behaviors component score as the quantitative trait, gender as the covariate to demonstrate the application of our score test when the covariate is dichotomous, and used SNP rs12198932 on chromosome 6 as the genetic marker of interest given its marginal association with the behaviors component score. We first randomly selected 694 independent trios from the data set as our population and then performed trait-based ascertainment by selecting those trios for whom the behaviors component score was > 5.63 (the sample mean) for testing the interaction between gender and rs12198932. This ascertainment resulted in 354 trios. To obtain tests of association in the 694 randomly ascertained trios, we used the offspring from all trios to estimate the Pearson correlation between gender and the behaviors component score, computed the FBAT [17] to test for association between rs12198932 and the behaviors component score using all the trios, and computed the QBAT-I to test for an interaction between rs12198932 and gender. To test for an interaction among the 354 ascertained trios, we computed both the original QBAT-I and our proposed ascertainment-adjusted score test

Data analysis in extended pedigrees—We applied our proband exclusion method to the Insulin Resistance and Atherosclerosis Study (IRAS) Family Study data. The IRAS Family Study is a multi-center study designed to investigate the genetic determinants of glucose homeostasis and adiposity. Families were recruited based on large size and structure without regard to phenotype, including glucose homeostasis parameters, diabetes or obesity. Details of the study design and recruitment have been published elsewhere [13]. We included 90 Hispanic families from San Antonio, Texas (649 individuals in 60 families) and the San Luis Valley, Colorado (619 individuals in 30 families) in our analyses.

For illustrative purposes, we used body mass index (BMI) as the phenotype of interest and SNP rs2606319 from the gene for Toll-like receptor 2 TLR2. Obesity is associated with low-grade inflammation [14, 15] and TLR signaling activates pro-inflammatory processes [16]. We considered blood insulin level at 100 min (INS100) from a frequently sampled intravenous glucose tolerance test as the covariate of interest based on the interrelationships between obesity, inflammation and insulin resistance. We first tested for a main effect of rs2606319 (using FBAT, [17]) and for an interaction between INS100 and rs2606319 using all families (i.e. in an unascertained sample). In order to perform the test for an interaction in an ascertained sample, we ascertained families based on the BMI of a single individual from each family. We randomly chose one person from each family as the proband and selected families whose proband had a BMI > 27.57, the sample mean. We had 46 families in this ascertained subsample. We then tested for an interaction in the ascertained subsample.

Finally, we tested for an interaction while excluding the proband's phenotype and covariate, but including his/her genotype.

3. Results

3.1 Type I error rate of QBAT-I

We first evaluated the empirical type I error rate of the QBAT-I under random selection. The type I error rates reported by Vansteelandt and colleagues [1] for randomly selected trios were conservative. In contrast, we found type I error rates within the 95% CI for $\alpha = 0.05$ based on 10,000 replicates (Table I) when trios were randomly ascertained. This is likely attributable to the fact that Vansteelandt et al. included phase uncertainty in their simulations, whereas we did not. The type I error rates for the QBAT-I were also at nominal levels when the environmental factor followed an exponential distribution and under the sibling pair design (data not shown). These results confirmed that the QBAT-I maintains the nominal type I error rate for samples of randomly selected families. In contrast, the type I error rate of the QBAT-I is inflated under ascertainment (Table II [continuous exposure Z] and Supplemental Table I [dichotomous exposure Z]). As the main genetic effect increases, inflation of the type I error rate also increases. This is consistent with the statements regarding the expected impact of ascertainment on the QBAT-I by Vansteelandt et al and more recent work showing inflation of the type I error rate for the QBAT-I under non-random ascertainment [18]. In addition, we found the type I error rate of the QBAT-I to be slightly inflated when testing for an interaction with a trait that is correlated with the ascertainment trait (Supplemental Table II)

3.2 Type I error rate using proposed methods

In contrast to the QBAT-I, our non-random ascertainment-adjusted score test largely maintains the appropriate type I error rate even under extreme trait-based ascertainment. Most of the observed type I error rates fall within the 95% confidence interval for $\alpha = 0.05$ (Table II [continuous exposure Z] and Supplemental Table I [dichotomous exposure Z]) or $\alpha = 0.01$ (Supplemental Table III) based on 10,000 replicates

In general, the observed type I error rates for the proband exclusion method also fall within the 95% CI for $\alpha = 0.05$ based on 10,000 replicates (Table II) when parental genotypes are available. However, the type I error rate for the highest correlation we initially tested and the weakest ascertainment condition (third column from right, 5th row of Table II) is just above the upper limit of the 95% CI and indicates the potential for an inflated type I error rate that may be present with a more powerful study (larger sample size and/or much bigger genetic main effect). We examined the potential for this inflation by evaluating the type I error rate when the main genetic effect induced a correlation of 91% between $\beta_x X$ and Y . As most of those type I error rates were within the 95% confidence intervals for a nominal $\alpha = 0.05$ (Supplemental Table IV), our method appears to maintain the appropriate type I error rate even when the genetic main effect is very large.

When parental genotypes are missing, the type I error rate for the proband exclusion method is less inflated than when using the uncorrected QBAT-I. However, the type I error rate is

still inflated when the genetic main effect is large or the ascertainment is extreme (Supplemental Table V), as we expected, due to the genotypic correlation among siblings in the absence of complete parental data. This inflation was largely unchanged after varying the magnitude of the correlation between the proband and sibling phenotypes (ρ_y) (data not shown), although the inflation was reduced when the correlation coefficient between the sibling *environmental* covariates was reduced ($\rho_z = 0.20$ vs. 0.80) (Supplemental Table VI).

3.3 Power

Random selection—We studied the power of the QBAT-I using randomly selected trios (Table III). Power increases as the allele frequency increases to a point, then as expected, decreases for further allele frequency increases. Our observed power pattern is different from what Vansteelandt et al. observed due to the lack of haplotype uncertainty in our simulations. When the environmental exposure variable Z followed an exponential distribution, the power was generally higher than the power when Z followed a normal distribution. We believe this is due to the larger difference between the means of the interaction groups due to the heavy tail of the exponential distribution compared to the normal distribution.

Trait-based Ascertainment—Power for both the ascertainment-adjusted score test and the proband exclusion method (Table IV) was lower than that of the QBAT-I for an equal number of randomly selected trios due to the reduced variability of the truncated distribution of Y . For both of our proposed methods, power decreases as the ascertainment becomes more extreme. These results are likely due to reduced variability of the environmental covariate Z in the more extreme samples. Decreased variability in Y and Z among probands compared to siblings is also the likely explanation for the observed higher power for the proband exclusion method compared to the ascertainment-adjusted score test. Note that larger simulated interaction effects did not necessarily translate into higher power. This is due to the fact that the trait-based ascertainment changes the allele frequency in the ascertained samples such that the proportion of individuals with $X = 0$ is smaller as the interaction effect increases. This imbalance in the dichotomous variable X can decrease power.

3.4 AGRE data analysis

The mean of the behaviors component score in the randomly ascertained sample of 694 trios was 5.63 with standard deviation 2.52. The correlation between gender and the behaviors component score among the probands was -0.1430 ($p=0.0002$). The SNP rs12198932 was associated with the behaviors component score (FBAT $p=0.008$), but there was no interaction between rs12198932 and gender (QBAT-I $p=0.647$) in the randomly-selected trios. Of the 354 trios ascertained on a score > 5.63 , 219 were informative for testing. After ascertainment, the minor allele frequency of rs12198932 was essentially unchanged compared to the whole sample (23% vs. 24%, respectively); the ascertained sample had 81% males compared to 77% in the whole sample, reflecting the stronger association between gender and the behaviors component score compared to that between rs12198932 and the score. The change in gender distribution also demonstrates that when there is a main effect of the environmental covariate on the trait, trait-based ascertainment can result in a

distortion of the distribution of the environmental covariate. When testing for the interaction between rs12198932 and gender, the p-value for our ascertainment-adjusted score test was 0.407 while the p-value of the QBAT-I was 0.175. Thus, ascertainment based on the trait did decrease the p-value for the gene-environment interaction test based on QBAT-I, and our ascertainment-adjusted score test appropriately increased the p-value.

3.5 IRAS Family Study data analysis

In the IRAS Family Study data, rs2606319 was significantly associated with BMI ($P=0.018$), and without considering the within-family correlation, the sample correlation between $\log_e(\text{INSULIN}_{100})$ and BMI was 0.63. When we included all families (corresponding to random selection), the p-value for the interaction between INSULIN_{100} and rs2606319 was 0.700. After ascertainment on $\text{BMI} > 22$, the p-value for the interaction based on the QBAT-I was 0.098. When we applied our correction by excluding the proband phenotype (46 of 560 individuals), the p-value for the interaction increased to 0.166. Again, ascertainment based on the trait did decrease the p-value for the gene-environment interaction test based on QBAT-I, and excluding the proband genotypes from the analysis appropriately increased the p-value.

4. Discussion

Virtually all human diseases likely result from complicated interactions between an individual's genetic makeup and environmental factors. Therefore, identifying such interactions is of great importance to disease prevention and disease treatment. Until recently, there have been no procedures to estimate and test gene-environment interactions for general pedigree structures that are robust to confounding by ancestry via a family-based test. The QBAT-I provides researchers a much needed tool in this area. However, it is inappropriate to use the QBAT-I when the data are ascertained based on a trait that is associated with a genetic marker that will be tested as part of an interaction. In reality, most family-based studies have ascertained samples and often test for an interaction with genetic markers that are associated with the trait used in ascertainment. Thus, it is important to study the QBAT-I using ascertained family data and extend the QBAT-I to non-randomly selected family data.

In this article, we confirmed that the QBAT-I maintains the appropriate type I error rate when families are randomly selected. We found that the type I error rate is inflated, however, when families are ascertained on a phenotype that is associated with the genetic predictor of interest. As the genetic main effect becomes stronger or the sampling more extreme on the phenotype of interest, the inflation increases. Both phenomena are caused by the association between the genotype and the environmental variables conditional on parental genotypes, the unequal variance of the trait across genotypes and the truncated distribution of the trait in ascertained samples.

We have proposed two methods for correcting the type I error of the QBAT-I under ascertainment. First, when complete trios are sampled (with no other siblings), our ascertainment-adjusted score test restores the type I error rate to nominal levels for all of the simulation conditions that we considered. Second, if at least one sibling is available, we

propose to completely exclude the proband from the analysis when parental genotypes are observed or only the phenotype of the proband when parental genotypes are incomplete. Our simulation results show that our approach restores the appropriate type I error rate of the QBAT-I when parental genotypes are observed, even when the genetic main effect is very large. When parental genotypes are incomplete, the proband genotypes are required to infer the missing parental genotypes. We found that this approach may not maintain the appropriate type I error rate when the main genetic effect is large, although type I error rates were at nominal levels for most of the main genetic effects we considered. Our simulation results were corroborated using real data from the IRAS Family Study. The IRAS Family Study families are large, with an average of 9 genotyped and phenotyped individuals per family, indicating that the correction we suggest for sibling pair data is likely applicable to pedigrees of arbitrary structure. One potential limitation of our approach is lack of information on the ascertainment threshold, C . We found that using the minimum phenotype value (not adjusted for any covariates) among the sample probands gave nearly identical results as using the true threshold, providing both an operational solution and evidence that minor misspecification of the threshold does not result in increased type I error.

In recognition of the limitations of the QBAT-I in non-randomly ascertained samples, the QBAT-I has been extended to non-randomly ascertained families when the ascertainment trait is dichotomous [19] using a score test statistic based on a conditional likelihood in a manner similar to our score test for quantitative traits. In addition, an ascertainment-adjusted QBAT-I applicable to quantitative traits has very recently been developed which estimates and tests the interaction only in the subset of families who would have been selected regardless of genotype, thereby removing the conditional dependence of X and Z given S [18]. The authors chose this approach to avoid making assumptions regarding the model for $E(X_{ij} | Z_{ij}, S_i)$ in ascertained samples and instead rely on the assumption of conditional independence of X and Z given S in the population that gave rise to the trait-based ascertained sample. This untestable conditional independence assumption is common for tests of gene-environment interactions [20] and is necessary for application of our proband exclusion method. Our score test does not require this assumption since we directly model $E(X_{ij} | Z_{ij}, S_i)$, which allows the use of traditional model fit diagnostics [21] to evaluate the appropriateness of the model. In addition, both of our approaches utilize all informative families rather than restricting the analysis to only those families who would have been ascertained regardless of genotype. As noted by Fardo et al, that restriction has the potential to result in limited power for their test in the setting of a haplotype analysis due to the genotype covariate, X , taking on many levels.

We have described tests for an interaction effect with a single genetic marker and single environmental covariate. Given the existence of many genome-wide association studies with information on hundreds of thousands or millions of markers, several methods have been developed to reduce the dimensionality of interaction testing with many genetic markers [22]. Several of these methods are based on the Multifactor-Dimensionality Reduction (MDR) method [23] which was designed to detect interactions between genetic and/or discrete environmental factors when empty cells are possible in high-order interaction modeling. Of particular relevance to this paper, the Pedigree-based Generalized Multifactor

Dimensionality Reduction (PGMDR; [24]) and FAM-MDR [25] are extensions of MDR approaches to pedigree data. The MDR and related approaches provide a framework for detecting interactions between genetic and/or environmental factors that requires a statistic to place cells (e.g. combinations of the genetic and environmental covariate values) into either high-risk or low-risk groups in order to reduce the dimension of the testing problem. As such, implementation of these methods requires using the correct test statistic to counter the effects of population stratification and to adjust for different sample ascertainment schemes. Our study provides valuable information for investigators who will expand the MDR-related framework to allow for testing of gene-environment interactions in families ascertained on the trait of interest genome-wide.

The ability to test for gene-environment interactions in family-based studies is important given the resources that have been devoted to developing such studies and the proposed use of such samples in following-up genetic markers identified in genome-wide association studies. The development of the QBAT-I is a major step forward in this endeavor. In this paper, we have described methods that allow appropriate use of the QBAT-I in ascertained samples which make up the majority of family-based studies. Further study is needed to extend and test our proposed methods for more general pedigrees and traits correlated with the ascertainment trait.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the investigators of the Autism Genetic Resource Exchange (AGRE) for use of data as an example. TEF was supported by supported by an American Diabetes Association (ADA) Junior Faculty Award. TEF and CDL were supported by the IRAS Family Study (NIH Grants HL-60944-02, HL-61210-02, HL-61019-02, HL-60894, and HL-60931-02) and the GUARDIAN study (DK-085175-11A1).

Appendix A

$\frac{f(Y|X, Z, S)I_A(x)}{P(A)}$ is a member of the exponential family of distributions.

The QBAT-I is a score test derived from the likelihood $f(Y|X, Z, S)$, where S is the set of parental genotypes when they are observed and the sufficient statistics for those genotypes when they are not observed [17]. Under ascertainment, let the phenotypic sample space be a set A that is a fixed subset of the original sample space. The likelihood for ascertained

samples is then $\frac{f(Y|X, Z, S)}{P(A)}$ and the score test should be derived from this truncation-adjusted likelihood.

The theoretical requirements for a valid score test have been detailed previously [26]. Here we consider the exponential family of distributions for simplicity. The k -dimensional exponential family is the set of distributions that have densities of the form

$$f_{\theta} = \exp \left[\sum_{i=1}^k \eta_i(\theta) T_i(t) - B(\theta) \right] h(t),$$

where θ is a vector of parameters.

The density of the truncated distribution is $\frac{f_{\theta}(t)I_A(t)}{P_{\theta}(t \in A)}$ where $P_{\theta}(t \in A) = \int_A f_{\theta}(t)dt$, when A is a fixed subset of the sample space. Therefore, $P_{\theta}(t \in A)$ is a function of the parameter θ . $I_A(t)$ is an indicator function for membership in A and is a function of T since A is a fixed subset of the sample space. The density of the truncated distribution can be rewritten as

$$\begin{aligned} & f_{\theta}(t)I_A(t)/P_{\theta}(t \in A) \\ &= \exp \left[\sum_{i=1}^k \eta_i(\theta) T_i(t) - B(\theta) \right] \\ & \quad \times \exp(-\ln(P_{\theta}(t \in A)))h(t)I_A(t) \\ &= \exp \left[\sum_{i=1}^k \eta_i(\theta) T_i(t) - (B(\theta) + \ln(P_{\theta}(t \in A))) \right] (h(t)I_A(t)) \\ &= \exp \left[\sum_{i=1}^k \eta_i(\theta) T_i(t) - B^*(\theta) \right] h^*(t). \end{aligned}$$

Thus, the distribution with density $\frac{f_{\theta}(t)I_A(t)}{P_{\theta}(t \in A)}$ constitutes a member of the exponential family. Therefore, when $f(Y|X, Z, S)$ is a member of the exponential family and A is a fixed

set, $\frac{f(Y|X, Z, S)I_A(x)}{P(A)}$ is also a member of the exponential family. As such, the truncation-adjusted likelihood satisfies the theoretical requirement for a valid score test.

Appendix B

Estimation of $E[X | Z, S]$

The assumption of Mendelian inheritance cannot be used to calculate the conditional expectation of X given Z and S due to the non-random sampling scheme. Under non-random ascertainment, Z is associated with both the trait variable Y and the genetic variable X . Therefore, in the E-estimation procedure, we need to model $E[X | Z, S]$ to additionally adjust for the new confounder Z when we estimate the main genetic effect and test for the gene-environment interaction.

We found that using a spline function for Z when modeling $E[X | Z, S]$ provided the best fit to the both simulated and real data and provided the necessary flexibility to be broadly applicable. A B-spline with degree 1 and degrees of freedom 4 was adopted, resulting in a

piecewise linear regression. The model we considered was

$\text{logit}(X_{ij}) = \beta_0 + \sum_{n=1}^{u-1} \beta_n S_{n+1} + \sum_{m=1}^k \gamma_m Z_m$, where S_{n+1} , $n = 1 \dots u - 1$ are the indicator variables for the u parental genotype groups, and Z_m , $m = 1 \dots k$ are the spline basis functions of covariate Z . For the genetic variable X based on a dominant model of a biallelic SNP marker, there are two informative parental genotype configurations. Clearly (e.g. in Figure 2), the spline function fit the observed data better than the logistic regression. When Z is dichotomous, a logistic regression model with Z as the covariate is appropriate.

References

1. 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. [PubMed: 17970814]
2. Rahman A, Isenberg DA. Systemic lupus erythematosus. *N Engl J Med*. 2008; 358:929–939. [PubMed: 18305268]
3. 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; 52:506–516. [PubMed: 8447318]
4. Boehnke M, Langefeld CD. Genetic Association Mapping Based on Discordant Sib Pairs: The Discordant-Alleles Test. *Am J Hum Genet*. 1998; 62:950–961. [PubMed: 9529345]
5. Spielman RS, Ewens WJ. A Sibship Test for Linkage in the Presence of Association: The Sib Transmission/Disequilibrium Test. *Am J Hum Genet*. 1998; 62:450–458. [PubMed: 9463321]
6. Lunetta KL, Faraone SV, Biederman J, Laird NM. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet*. 2000; 66:605–614. [PubMed: 10677320]
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; 67:146–154. [PubMed: 10825280]
8. Whittemore AS, Halpern J, Ahsan H. Covariate adjustment in family-based association studies. *Genet Epidemiol*. 2005; 28:244–255. [PubMed: 15593089]
9. Lu AT, Cantor RM. Weighted variance FBAT: a powerful method for including covariates in FBAT analyses. *Genet Epidemiol*. 2007; 31:327–337. [PubMed: 17323371]
10. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics*. 1992; 48:479–495. [PubMed: 1637973]
11. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet*. 2001; 69:463–466. [PubMed: 11452364]
12. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*. 1994; 24:659–685. [PubMed: 7814313]
13. Henkin L, Bergman RN, Bowden DW, Ellsworth DL, Haffner SM, Langefeld CD, Mitchell BD, Norris JM, Rewers M, Saad MF, Stamm E, Wagenknecht LE, Rich SS. Genetic Epidemiology of Insulin Resistance and Visceral Adiposity: The IRAS Family Study Design and Methods. *Annals of Epidemiology*. 2003; 13:211–217. [PubMed: 12684185]
14. Hotamisligil G. Inflammation, TNF-alpha, and insulin resistance. In: LeRoith, D.; Olefsky, J., editors. *Inflammation, TNF-alpha, and insulin resistance*. New York: Lippincott, Williams, and Wilkins; 2003.
15. Tataranni PA, Ortega E. A Burning Question: Does an Adipokine-Induced Activation of the Immune System Mediate the Effect of Overnutrition on Type 2 Diabetes? *Diabetes*. 2005; 54:917–927. [PubMed: 15793228]

16. Medzhitov R. Toll-like receptors and innate immunity. *Nat. Rev. Immunol.* 2001; 2:135–145. [PubMed: 11905821]
17. Rabinowitz D, Laird N. A Unified Approach to Adjusting Association Tests for Population Admixture with Arbitrary Pedigree Structure and Arbitrary Missing Marker Information. *Human Heredity.* 2000; 504:227–233. [PubMed: 10782014]
18. Fardo DW, Liu J, Demeo DL, Silverman EK, Vansteelandt S. Gene-environment interaction testing in family-based association studies with phenotypically ascertained samples: a causal inference approach. *Biostatistics.* 2011; 13:468–481. [PubMed: 22084302]
19. Moerkerke B, Vansteelandt S, Lange C. A doubly robust test for gene-environment interaction in family-based studies of affected offspring. *Biostatistics.* 2010; 11:213–225. [PubMed: 20154305]
20. Umbach DM, CR W. The Use of Case-Parent Triads to Study Joint Effects of Genotype and Exposure. *Am J Hum Genet.* 2000; 66:251–261. [PubMed: 10631155]
21. Hosmer, DW.; Lemeshow, S. *Applied logistic regression.* 2nd edn. New York: Wiley; 2000.
22. Aschard H, Lutz S, Maus Br, Duell E, Fingerlin T, Chatterjee N, Kraft P, Steen K. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Human Genetics.* 2012; 131:1591–1613. [PubMed: 22760307]
23. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *American journal of human genetics.* 2001; 69:138–147. [PubMed: 11404819]
24. Lou X-Y, Chen G-B, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD. A Combinatorial Approach to Detecting Gene-Gene and Gene-Environment Interactions in Family Studies. *American journal of human genetics.* 2008; 83:457–467. [PubMed: 18834969]
25. Cattaert T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards TL, Van Steen K. FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. *PLoS One.* 2010; 5:e10304. [PubMed: 20421984]
26. Godfrey, LG. *Misspecification Tests in the Econometrics: The Lagrange Multiplier Principle and Other Approaches.* First edn. Cambridge, United Kingdom: Cambridge University Press; 1988.

Random Ascertainment

Non-random Ascertainment

Figure 1.

This figure demonstrates the conditional dependence between the proband genotype variable (X) and environmental covariate (Z) in samples ascertained based on the trait value (Y) of the proband. For illustrative purposes, we used a rather large main genetic effect ($\beta_x = 20$, $\beta_{xz} = 0$) and $C = 22$ as the threshold for ascertainment. We used a dominant effect of the minor allele, labeled “2”. The informative families are those whose parental genotype configuration is either 1/1 and 1/2 or 1/2 and 1/2. The upper panel shows the data from randomly ascertained families. The distribution of Z is the same across the proband

genotypes within each parental genotype configuration. The lower panel shows the data from families ascertained via a proband with phenotype > 22 . The distribution of Z is clearly different across proband genotypes within each parental genotype configuration under this non-random ascertainment, demonstrating the dependence of Z and X even after conditioning on parental genotypes, S .

Deviation of Offspring Genotype Frequencies from Mendelian Expectations Under Non-Random Ascertainment

	Expected Under Mendelian Inheritance		Random Ascertainment		Non-Random Ascertainment	
Parental Genotype Pair						
Offspring Genotype	1/1 and 1/2	1/2 and 1/2	1/1 and 1/2	1/2 and 1/2	1/1 and 1/2	1/2 and 1/2
1/1	0.50	0.25	0.50	0.25	0.24	0.10
1/2	0.50	0.50	0.50	0.50	0.76	0.60
2/2	0.00	0.25	0.00	0.25	0.00	0.30

Figure 2.

This figure demonstrates the deviation of offspring genotype frequencies from those expected under Mendel's Laws in samples non-randomly ascertained based on the trait value (Y) of the proband when there is a main effect of the genotype on Y . For illustrative purposes, we generated trios assuming a large main genetic effect ($\beta_x = 20$, $\beta_{xz} = 0$) under a dominant model for allele "2" and strong ascertainment ($C=22$). Genotype frequencies observed in randomly-ascertained trios were consistent with frequencies expected based on Mendel's Laws given the parental genotypes while those in the trait-based ascertained samples deviate strongly from expectation given only the parental genotypes.

Figure 3.

This figure shows the observed and fitted values of $E[X | Z, S]$. To generate the data in the figure, we assumed a large main effect ($\beta_x = 20$), no interaction effect ($\beta_{xz} = 0$) and $C > 22$ as the threshold for ascertainment. We considered a dominant effect of the rare allele labeled “2”. The informative families were those whose parental genotype configuration was either 1/1 and 1/2 or 1/2 and 1/2. The data were stratified based on parental genotypes. We estimated $E[X | Z, S]$ via both a usual logistic regression with the environmental covariate Z and parental genotypes S as independent variables and a spline regression adjusted for S with a B-spline function for Z .

Table 1

Type I error rate of QBAT-I at $\alpha = .05$ for randomly ascertained trios

MAF	$\rho(Y, \beta_X)$						
	0	0.023	0.047	0.071	0.095	0.12	
0.05	5.03%	4.99%	5.03%	5.06%	5.02%	4.91%	
0.15	4.56%	4.54%	4.70%	4.71%	4.76%	4.84%	
0.25	4.54%	4.67%	4.69%	4.69%	4.77%	4.76%	
0.35	5.08%	5.10%	5.04%	4.97%	5.02%	4.94%	
0.45	4.98%	4.75%	4.85%	4.86%	4.88%	5.07%	

MAF: minor allele frequency. The 95% confidence interval for nominal type I error rate of 0.05 is (0.0457, 0.0543) with 10,000 replicates. $Z \sim N(20,16)$

Table II

Type I error rate under random and non-random ascertainment.

$\rho(Y, \beta, X)$	Random Ascertainment	Trait-based Ascertainment Trios QBAT-I			Trait-based Ascertainment Trios Ascertainment Corrected Score Test			Trait-based Ascertainment Sibling QBAT-I Proband Exclusion		
		$\hat{Y}^- - 0.5\hat{\sigma}$	$\hat{Y}^-$	$\hat{Y}^- + 0.5\hat{\sigma}$	$\hat{Y}^- - 0.5\hat{\sigma}$	$\hat{Y}^-$	$\hat{Y}^- + 0.5\hat{\sigma}$	$\hat{Y}^- - 0.5\hat{\sigma}$	$\hat{Y}^-$	$\hat{Y}^- + 0.5\hat{\sigma}$
0.023	4.67%	5.08%	5.09%	4.99%	5.00%	4.93%	5.00%	5.06%	4.80%	5.22%
0.047	4.69%	5.35%	6.35%	6.62%	4.63%	4.91%	4.65%	4.91%	4.70%	4.69%
0.071	4.69%	7.01%	8.45%	10.08%	4.74%	4.48%	4.49%	4.68%	5.00%	5.02%
0.095	4.77%	10.02%	14.25%	18.27%	4.92%	4.90%	4.85%	4.53%	4.58%	5.12%
0.12	4.76%	15.07%	23.7%	33.53%	5.20%	5.00%	4.89%	5.44%	5.04%	5.11%

MAF=0.25. The 95% confidence interval for nominal type I error rate of 0.05 is (0.0457, 0.0543) with 10,000 replicates. Parental genotypes known. $Z \sim N(20,16)$

Table III

Power of QBAT-I for randomly ascertained trios

Distribution	MAF	$\rho(Y, \beta_{xz}, XZ)$				
		0.016	0.035	0.085	0.135	0.183
$Z \sim N(20,16)$	0.05	17.62%	27.11%	38.49%	50.68%	61.39%
	0.15	36.38%	54.08%	69.23%	80.17%	87.18%
	0.25	39.85%	54.97%	67.02%	74.87%	80.22%
	0.35	35.77%	47.85%	56.22%	61.97%	66.53%
	0.45	28.76%	36.5%	42.17%	46.28%	48.78%
$Z \sim EXP(0.25)$	0.05	16.16%	24.81%	33.96%	43.31%	52.15%
	0.15	36.65%	56.40%	74.48%	86.94%	93.95%
	0.25	45.68%	68.01%	84.84%	94.29%	98.31%
	0.35	45.17%	66.72%	82.84%	92.87%	97.08%
	0.45	37.95%	57.12%	73.41%	84.61%	91.60%

Table IV
Power of QBAT-I for random ascertainment and power of proposed methods for trait-based ascertainment

$\rho(Y, \beta_{xz} Z)$	Random Ascertainment*	Trait-based Ascertainment Trios Ascertainment Corrected Score Test			Trait-based Ascertainment QBAT-I Proband Exclusion ⁺		
		$Y^- - \hat{\sigma}^-$	$Y^- - 0.5\hat{\sigma}^-$	Y^-	$Y^- - \hat{\sigma}^-$	$Y^- - 0.5\hat{\sigma}^-$	Y^-
0.016	39.85%	36.59%	28.78%	21.89%	30.86%	26.05%	23.84%
0.035	54.97%	51.80%	37.07%	27.34%	43.42%	36.86%	33.61%
0.085	67.02%	60.78%	40.37%	29.49%	53.02%	45.73%	41.92%
0.135	74.87%	63.07%	38.71%	25.64%	60.62%	52.67%	49.13%
0.183	80.22%	62.88%	33.92%	21.08%	66.75%	57.83%	57.57%

500 nuclear families, MAF=0.25, $Z \sim N(20,16)$

* QBAT-I test

⁺ Parental genotypes known