

Improved Association Analyses of Disease Subtypes in Case-Parent Triads

Michael P. Epstein,^{1*} Irwin D. Waldman,² and Glen A. Satten³

¹Department of Human Genetics, Emory University, Atlanta, Georgia

²Department of Psychology, Emory University, Atlanta, Georgia

³Centers for Disease Control and Prevention, Atlanta, Georgia

The sampling of case-parent triads is an appealing strategy for conducting association analyses of complex diseases. In certain situations, one may have interest in using the triads to identify genetic variants that are associated with a specific subtype of disease, perhaps related to a characteristic cluster of symptoms. A straightforward strategy for conducting such a subtype analysis would be to analyze only those triads with the subtype of interest. While such a strategy is valid, we show that triads *without* the subtype of interest can provide additional genetic information that increases power to detect association with the subtype of interest. We incorporate this additional information using a likelihood-based framework that permits flexible modeling and estimation of allelic effects on disease subtypes and also allows for missing parental data. Using simulated data under a variety of genetic models, we show that our proposed association test consistently outperforms association tests that only analyze triads with the subtype of interest. We also apply our method to a triad study of attention-deficit hyperactivity disorder and identify a genetic variant in the dopamine transporter gene that is associated with a subtype characterized by extreme levels of both inattentive and hyperactive-impulsive symptoms. *Genet. Epidemiol.* 30:209–219, 2006. © 2005 Wiley-Liss, Inc.

Key words: triad; subtype; TDT; CPG; likelihood

Contract grant sponsor: University Research Committee of Emory University.

*Correspondence to: Michael P. Epstein, Ph.D., Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322. E-mail: mepstein@genetics.emory.edu

Received 8 August 2005; Accepted 7 November 2005

Published online 22 February 2006 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.20138

INTRODUCTION

A variety of study designs use data from the detailed catalogue of single-nucleotide polymorphisms (SNPs) [International SNP Working Group, 2001; Sherry et al., 2001] to identify genetic variants that are associated with complex disease. One popular association design collects case-parent triads, which are units consisting of an affected subject and the subject's two parents. Association analyses of triads consider the possible pairings of SNP alleles from each parent, with the alleles transmitted to the affected offspring forming the "case" genotype and the other possible offspring genotypes forming the "control" genotypes. Testing for association between a SNP and disease corresponds to testing whether a specific SNP allele is preferentially transmitted from parent to affected offspring. One can assess the significance

of such preferential transmission using the transmission-disequilibrium test (TDT) [Spielman et al., 1993], which is a McNemar statistic for matched-pair data. Alternatively, one can test for association using a likelihood-based version of the TDT called the conditional-on-parental genotypes (CPG) approach, which was developed by Schaid and Sommer [1993, 1994]. The CPG approach models the probability of an affected offspring's genotype conditional on parental genotypes. This probability is a function of genotype relative risk parameters, which can be estimated using standard maximum-likelihood procedures. The CPG and TDT approaches are appealing for association analysis because the matched nature of the genotype data inherently conditions on the genetic ancestry of the parents. Because of this conditioning, these procedures are robust to spurious association arising from population stratification [Schaid, 2004].

Association studies of complex diseases occasionally focus inference on specific disease subtypes, especially if the overall disease phenotype is based on the amalgamation of different symptoms and criteria. Disease subtypes can correspond to clinically observed clusters of symptoms that occur together, or can equate to groupings based on similar age-of-onset, similar severity of symptoms, or similar response to medication (which is of interest in pharmacogenetics studies). Such subtypes are potentially valuable to examine because they may have similar genetic origins, which should facilitate identification of relevant disease-influencing variants. Waldman et al. [1998] implemented such a subtype strategy to test whether a putative high-risk allele in the dopamine transporter gene *DAT1* was associated with attention-deficit hyperactivity disorder (ADHD) in a sample of case-parent triads. The ADHD diagnosis actually comprises three specific and mutually-exclusive subtypes: inattentive, hyperactive-impulsive, and combined. Each subtype is based on surpassing diagnostic thresholds for inattentive and hyperactive-impulsive symptom dimensions. The inattentive subtype is diagnosed in subjects who surpass the inattentive dimension but not the hyperactive-impulsive dimension, whereas the hyperactive-impulsive subtype is diagnosed in subjects who surpass the hyperactive-impulsive dimension but not the inattentive dimension. The combined subtype is diagnosed in subjects who surpass the thresholds for both the inattentive and hyperactive-impulsive symptom dimensions. When Waldman et al. [1998] analyzed each subtype separately in the triads, they found that the high-risk *DAT1* allele was strongly associated only with the combined subtype, which suggests different subtypes might have different genetic origins.

To conduct subtype association analyses in triads, Waldman et al. [1998] applied either the TDT or CPG test to those triads with the subtype of interest and ignored the remaining triads without the subtype. While such an approach is valid, the power of such an approach might be weak in general if the number of eligible triads for the subtype analysis is small. To increase power, we propose an idea based on the work of Nagelkerke et al. [2004], Epstein et al. [2005], and Weinberg and Umbach [2005], who showed that one can use unrelated controls (or another external information source of population allele frequencies) to increase the power of an association analysis of triads. Translating this idea to the

subtype setting, we propose using the parents from triads *without* the subtype of interest as unrelated controls to increase power of the association analysis of triads with the subtype of interest. While such a strategy is appealing, we must be cautious in implementation since we still sample these triad parents without the subtype of interest on the basis of having an affected offspring with some form of the disease. If the tested SNP is associated with a different subtype, then these sampled parents do not constitute a random sample from the population and cannot serve as unrelated controls. Such a situation could easily arise if the SNP allele associated with the subtype of interest is also associated with a different subtype. Additionally, such an occurrence could result from a phenomenon like allelic heterogeneity, where one SNP allele is associated with one subtype and the other SNP allele is associated with a different subtype.

Rather than using the parental genotypes from triads without the subtype of interest, we show that we can still extract useful information on population allele frequencies using the genotypes of the entire triad (parents and offspring) without the subtype of interest. This additional information will increase the power to detect association with the disease subtype of interest. To incorporate this information from triads without the subtype of interest, we develop a likelihood-based procedure similar in form to that proposed by Epstein et al. [2005]. In remaining sections, we describe this likelihood and develop estimation procedures and statistical tests for assessing association between the SNP and the disease subtype of interest. We evaluate the performance of our proposed approach using both simulated data and real data from the triad study of ADHD described in Waldman et al. [1998].

MATERIALS AND METHODS

ASSUMPTIONS AND NOTATION

We assume a sample of case-parent triads that were collected for the study of a complex disease. We assume that triads are genotyped at a SNP of interest with alleles S and s . We further assume these triads can be classified without error into two mutually-exclusive disease subtypes, denoted by A and B . Without loss of generality, we assume subtype A is the primary subtype of interest. In this situation, subtype B can be either a unique subtype or a composite of all remaining subtypes.

Interest focuses on assessing whether the SNP of interest is associated with disease subtype *A*.

Each genotype is coded to equal the number of copies of the *S* allele within the genotype. For a given triad, let G_{p_1} and G_{p_2} denote the two parental genotypes with the convention that $G_{p_1} \geq G_{p_2}$, and let $G_p = (G_{p_1}, G_{p_2})$. Let G_o denote the offspring genotype. Finally, let D_o denote the offspring phenotype, with $D_o = A$ indicating disease subtype *A* and $D_o = B$ indicating disease subtype *B*.

We focus inference on the SNP-based genotype relative risks (RRs) of subtype *A*, which we model using

$$\psi_g = \frac{P(D_o = A | G_o = g)}{P(D_o = A | G_o = 0)} = \exp(\beta^T \cdot X_g), \quad g = 1, 2$$

where β is a vector of risk parameters and X_1 and X_2 are design vectors for offspring genotypes $g = 1$ and $g = 2$, respectively. This RR model permits flexible modeling of allelic effects on disease subtype *A*. We can implement a general RR model by assuming the β and X_g vectors each have two components with $X_1 = (1, 0)$ and $X_2 = (0, 1)$. We can also let β and X_g be scalar and specify non-general RR models, such as multiplicative ($X_1 = 1, X_2 = 2$), dominant ($X_1 = X_2 = 1$), and recessive ($X_1 = 0, X_2 = 1$). Additional effects such as imprinting or parent-of-origin effects can also be included in this framework [Weinberg et al., 1998] but are not considered further in this report. Our goal in the association analysis of subtype *A* is to conduct inference on β or, equivalently, ψ_g .

LIKELIHOOD DERIVATION

Following Nagelkerke et al. [2004] and Epstein et al. [2005], we conduct association analysis of subtype *A* by constructing the joint likelihood of the genotype data from triads with subtype *A* and the genotype data from triads with subtype *B*. To model the genotype data from subtype *A* triads, we use the full likelihood

$$L_A = \prod_{i \in I_A} P[G_{oi} | G_{pi}, D_{oi} = A] \cdot P[G_{pi} | D_{oi} = A], \quad (1)$$

where the subscript *i* indexes individual triads and I_A denotes the set of indices for triads with subtype *A*. Here, $P[G_{oi} | G_{pi}, D_{oi} = A]$ denotes the probability for the standard CPG approach [Schaid and Sommer, 1993] and has the form

$$\begin{aligned} &P[G_{oi} = g | G_{pi} = g_p, D_{oi} = A] \\ &= \frac{\psi_g P(G_o = g | G_p = g_p)}{\sum_{g^*} \psi_{g^*} P(G_o = g^* | G_p = g_p)} \end{aligned}$$

where $P(G_o = g | G_p = g_p)$ are known combinatorial coefficients determined by Mendelian transmission.

The term $P[G_{pi} | D_{oi} = A]$ in L_A in (1) takes the form [Schaid, 1999]

$$\begin{aligned} &P[G_{pi} = g_p | D_{oi} = A] \\ &= \frac{\sum_g \psi_g P(G_o = g | G_p = g_p) \mu_{g_p}}{\sum_{g^*} \sum_{g_p^*} \psi_{g^*} P(G_o = g^* | G_p = g_p^*) \mu_{g_p^*}} \quad (2) \end{aligned}$$

where $\mu_{g_p} = P[G_p = g_p]$ denotes the distribution of parental genotypes g_p in the target population. Following Weinberg et al. [1998], we specify μ_{g_p} by assuming mating symmetry in the target population, such that we can describe the distribution of parental mating types using the vector $\mu = (\mu_{(2,2)}, \mu_{(2,1)}, \mu_{(2,0)}, \mu_{(1,1)}, \mu_{(1,0)}, \mu_{(0,0)})$. The elements of μ may take any positive values that sum to 1. Note this model does not assume Hardy-Weinberg Equilibrium (HWE) or random mating in the population.

We can also use equations of the same form as (1) and (2) to describe data of subtype *B* triads, with ψ replaced by

$$\phi_g \equiv \frac{P(D_o = B | G_o = g)}{P(D_o = B | G_o = 0)},$$

and μ_g replaced by μ_g^B . This result holds even if subtype *B* represents a composite of several distinct subtypes, although it may be difficult to interpret ϕ_g in this situation. If the mating frequencies μ_g and μ_g^B are not constrained in any way other than each describes a probability distribution, then the joint analysis of triads with subtype *A* and *B* will not yield increased power with respect to a separate analysis of subtype *A* triads alone. However, if we constrain $\mu_g = \mu_g^B$, then joint analysis of triads with subtypes *A* and *B* using $L = L_A \cdot L_B$, where

$$L_B = \prod_{i \in I_B} P[G_{oi} | G_{pi}, D_{oi} = B] \cdot P[G_{pi} | D_{oi} = B]$$

can increase the efficiency of estimates of β and, hence, can lead to a more powerful test of subtype *A*. A similar effect was seen by Nagelkerke et al. [2004], Epstein et al. [2005] and Weinberg and Umbach [2005].

Finally, although we maximize $L = L_A \cdot L_B$ with respect to β, μ, ϕ_1 and ϕ_2 , we consider ϕ_1 and ϕ_2 to be nuisance parameters. For this reason, we assume a general (saturated) model for ϕ_1 and ϕ_2 to avoid any potential bias arising from model misspecification of these parameters. However, in

some situations, it may be worth attempting to model ϕ_g to increase the efficiency of estimation of β , especially if it can be established that $\phi_g \equiv 1$.

ASSOCIATION TESTING AND INFERENCE BETWEEN SNP AND DISEASE SUBTYPE A

Using the likelihood $L = L_A \cdot L_B$, we can estimate β , μ , ϕ_1 and ϕ_2 using standard maximum-likelihood procedures. We can also develop likelihood-ratio (LR) and generalized score statistics [Boos, 1992] for testing the null hypothesis of no association between a SNP of interest and disease subtype A. Testing for no association corresponds to testing $H_0: \beta_1 = \beta_2 = 0$, assuming a general RR model, or testing $H_0: \beta = 0$ assuming a non-general RR model. Under the null hypothesis, the LR and score statistics each follow a χ^2_2 distribution under a general RR model and a χ^2_1 distribution otherwise.

An interesting question arises whether it is more advantageous to apply the proposed approach under a general RR model or a non-general RR model for β . Analyses based on the general RR model make no modeling assumptions and, therefore, are more robust to model misspecification than those based on a non-general RR model. However, analyses under a general RR model subsequently will be less efficient than those of a non-general RR model when the latter model is correctly specified. Rather than choosing to go exclusively with a general or non-general model, we instead suggest using the Akaike Information Criterion (AIC) [Akaike, 1985] to help guide appropriate inference. For a given model, we calculate the AIC as $-2\log(L^*) + 2p$, where L^* denotes L evaluated using the maximum-likelihood parameter estimates and p denotes the number of model parameters. Using AIC values, we can conduct inference either using the "best" model (which has the smallest AIC value) or by using a weighted model average (where the weights are functions of the AIC values) as described by Buckland et al. [1997].

COMBINING DATA FROM TRIADS WITH AND WITHOUT SUBTYPE OF INTEREST

For subtype analysis, the efficiency gains in β using our proposed approach only arise when the mating-type frequencies from subtype A triads are assumed to be the same as the mating-type frequencies from subtype B triads (that is, $\mu_g = \mu_g^B$). If this assumption does not hold, then the (inappropriate) μ information from subtype

B triads will bias the estimates of μ in subtype A triads, leading to bias in estimates of β . Unfortunately, situations may arise where the mating-type frequencies do indeed differ between these two sets of triads. For example, a form of population stratification could exist where subtype A and B triads are sampled in different proportions from two discrete populations, each with its own distinct mating-type frequencies. As a result, the mating-type frequencies in subtype A triads will not match the frequencies of subtype B triads, which will lead to bias in estimates of β .

To avoid potential bias in β , we must conduct a test prior to analysis to ensure that we can safely combine the two types of triads together. We recommend testing $H_0^{(B)}: \mu_g = \mu_g^B$, which is the null hypothesis that the mating-type frequencies of the two subtypes are the same. We can test $H_0^{(B)}$ using LR or score statistics based on $L = L_A \cdot L_B$. Each statistic follows a χ^2_5 distribution under $H_0^{(B)}$. If we fail to reject the null hypothesis, then we constrain the mating-type frequencies from each subtype to be the same and use $L = L_A \cdot L_B$ to conduct association analysis of subtype A triads. If we reject the null hypothesis, we discard the information from subtype B triads and base analyses on subtype A triads only using the CPG approach.

ALLOWING FOR MISSING PARENTAL DATA IN TRIADS

Case-parent triad analyses of disease must often deal with the issue of missing parental data, which can arise from factors such as death or refusal to participate. As a result of the missing parental data, studies often contain triads with 1 missing parent (defined as dyads) or 2 missing parents (defined as monads). In the context of this report, missing parental genotype data lead to partially or completely missing values of G_p , which complicates inference.

To incorporate dyads and monads with subtypes A and B into a valid association analysis, we use missing-data procedures for triads similar to those proposed by Weinberg [1999] and Epstein et al. [2005]. For dyads and monads with subtype A, we replace $P[G_{oi}|G_{pi}, D_{oi} = A]P[G_{pi}|D_{oi} = A]$ in L_A in equation (1) with

$$\sum_{G_{pi} \in S_{pi}} P[G_{oi}|G_{pi}, D_{oi} = A]P[G_{pi}|D_{oi} = A] \quad (3)$$

where S_{pi} denotes the set of parental genotypes consistent with the observed genotype

data. For incomplete triads with subtype B , we similarly replace $P[G_{oi}|G_{pi}, D_{oi} = B] \cdot P[G_{pi}|D_{oi} = B]$ in L_B with

$$\sum_{G_{pi} \in S_{pi}} P[G_{oi}|G_{pi}, D_{oi} = B]P[G_{pi}|D_{oi} = B] \quad (4)$$

Using the probabilities in (3) and (4) for incomplete triads with subtypes A and B , respectively, we can maximize the resulting likelihood using on Expectation–Maximization (EM) algorithm [Dempster et al., 1977] that is similar to the algorithm described in the appendix of Epstein et al. [2005].

APPLICATION TO ADHD DATASET

We applied our association test to data based on the study of Waldman et al. [1998], which investigated the role of the dopamine transporter gene DAT1 as a risk factor for ADHD. DAT1 marker data consisted of genotype data from a 40-bp VNTR found within the 3' UTR of the gene. For simplicity, we represent the marker in biallelic form by focusing inference on the common 10-repeat DAT1 allele and pooling all remaining alleles into a second category. Prior evidence [Cook et al., 1995; Gill et al., 1997] suggests this common allele is a high-risk allele for ADHD.

The sample we consider consist of 85 triads with the combined ADHD subtype, 44 triads with the inattentive ADHD subtype, 11 triads with hyperactive-impulsive ADHD subtype, and 92 triads with no ADHD diagnosis. We note that these sample sizes are larger than those used in Waldman et al. [1998], due to additional sampling of triads since that report's publication. The majority of the 232 triads were Caucasian, although a small proportion were African-American and Hispanic [see Waldman et al., 1998, for an ethnic breakdown of the sample]. The frequency of the 10-repeat DAT1 allele is approximately the same (≈ 0.72) in these three ethnic groups [Doucette-Stamm et al., 1995], which suggests bias in β estimates due to population stratification is unlikely.

We first investigated whether the common DAT1 allele was associated with the overall ADHD phenotype. Of the 140 ADHD triads in the dataset, 82 were complete, 42 were dyads, and 16 were monads. We analyzed the triad data under multiplicative, dominant, recessive, and general mechanisms using the EM-based CPG approach of Weinberg [1999]. After completing these analyses, we next examined whether the

allele was associated with the combined ADHD subtype. Our sample then consists of 85 triads with the subtype of interest and 147 triads without the subtype of interest. Of the 85 triads with the subtype of interest, 49 were complete, 28 were dyads, and 8 were monads. Of the 147 triads without the subtype of interest, 105 were complete, 33 were dyads, and 9 were monads.

We tested whether the common DAT1 allele was associated with the subtype of interest under multiplicative, dominant, recessive, and general mechanisms using two separate procedures. First, we applied the CPG approach to the 85 triads with the subtype of interest and used the EM algorithm of Weinberg [1999] to allow for missing parental data. Next, we considered our proposed approach that uses both the 85 triads with and 147 triads without the subtype of interest for analysis. Using our proposed EM algorithm to account for missing parental data, we first tested whether it was safe to combine the two sets of triads by testing the null hypothesis that their mating-type frequencies were the same. If we failed to reject the null hypothesis, we then applied our proposed association test to the combined dataset. To determine the most likely mechanism of genetic action, we calculated the Akaike Information Criterion (AIC) under each RR model and chose the best model as the one with the smallest AIC value [Akaike, 1985].

SIMULATIONS

Using the ADHD subtype analyses as a basis, we conducted additional simulations to assess the type I error and power of our test for detecting association between a SNP and the subtype of interest. We assumed the high-risk allele of the SNP corresponded to the common DAT1 allele used in the ADHD analyses. We then simulated SNP data for triads using estimates of μ from the best model from the ADHD subtype analyses. Using the definitions for the design-vectors X_1 and X_2 described earlier, we varied the true RR model $\psi_g = \exp(\beta \cdot X_g)$ ($g = 1, 2$) among multiplicative, dominant, and recessive mechanisms. We generated subtype RR values ($\exp(\beta)$) that ranged between 1.0 (null) and 2.0. To assess the effect of missing data on results, we varied the percentage of missing parental data between 0 and 40%. We analyzed each dataset twice; once applying the CPG method to triads with the subtype of interest only and once applying our proposed method that uses triads with and

without the subtype of interest. For each method, we analyzed the data both under the true model as well as the general model (which does not require knowledge of the true model). For a particular simulation design, we based power and type I error estimates on 10,000 replicates of the data.

We also conducted simulations to assess the impact of population stratification on the results of our proposed approach. To induce stratification, we sampled triads with and without the subtype of interest in different proportions from two discrete strata, each with its own unique set of mating-type frequencies for the simulated SNP. For the first stratum, we assumed the high-risk allele frequency was 0.72, which corresponded to the frequency of the common DAT1 allele in the ADHD analyses, and used mating-type frequencies corresponded to those from the best ADHD model. For the second stratum, we assumed the high-risk allele frequency was 0.44, which corresponded to the frequency of the common DAT1 allele in a Yemenite Jewish population [Kang et al., 1999] as reported in the ALFRED database [Cheung et al., 2000]. For simplicity, we simulated mating-type frequencies in the second stratum assuming random mating and HWE. Given the difference in allele frequencies between the two strata, we induced stratification by sampling triads with the subtype of interest in equal proportions from the two strata, while sampling triads without the subtype of interest in unequal proportions. For each dataset, we first assessed whether it was appropriate to combine triads with and without the subtype of interest by testing the null hypothesis that the mating-type frequencies of the two sets were the same. If we failed to reject the null, then we combined the two sets of triads together in the analysis. If we rejected the null, then

we based inference only on those triads with the subtype of interest using the CPG approach. Based on this strategy, we recorded the RR estimates from the best model (CPG approach if the null hypothesis is rejected, our proposed approach otherwise). For a particular stratification design, we base results on 10,000 replicates of the data.

RESULTS

ANALYSIS OF ADHD DATASET

We first conducted CPG analyses that assessed association between the common DAT1 allele and the overall ADHD phenotype in the 140 triads with an ADHD diagnosis. Results suggested there was no association between DAT1 and ADHD, with P values ranging between 0.253 (for a dominant model) and 0.652 (for a recessive model). These results suggest the common DAT1 allele does not increase risk for the general ADHD phenotype.

While the common DAT1 allele is not associated with the overall ADHD phenotype, it is still possible that the allele could be associated with a subtype group that is more genetically homogeneous. To explore this possibility, we next conducted DAT1 association analyses of the ADHD combined subtype and report the results in Table I. We first applied the CPG approach to the 85 triads with the subtype of interest and found no significant association between the common DAT1 allele and the subtype under each RR model considered, with P values ranging between 0.400 and 0.589. While these results contrast the results of Waldman et al. [1998], we note that we are using a larger dataset than the one used by the authors for subtype analysis.

TABLE I. Results from DAT1 subtype analysis

	Model			
	Multiplicative	Dominant	Recessive	General
Parameter estimates				
Combined subtype RR	1.54	2.73	1.50	2.41, 3.31
$\mu_{(2,2)}$	0.25	0.27	0.25	0.25
$\mu_{(2,1)}$	0.45	0.44	0.46	0.45
$\mu_{(2,0)}$	0.03	0.03	0.03	0.03
$\mu_{(1,1)}$	0.17	0.17	0.17	0.18
$\mu_{(1,0)}$	0.09	0.08	0.08	0.09
$\mu_{(0,0)}$	0.01	0.01	0.01	0.01
AIC	746.76	747.64	748.44	748.11
P value: combined test	0.04	0.07	0.12	0.09
P value: CPG test	0.40	0.46	0.52	0.59

We next applied our proposed approach that uses both the 85 triads with and the 147 triads without the subtype of interest for analysis. Prior to analysis, we first tested whether it was safe to combine the 85 triads with and 147 triads without the subtype of interest together for association analysis. P values for these tests were 0.395, 0.265, 0.260, and 0.410 under multiplicative, dominant, recessive, and general models, respectively, which suggest it is acceptable to combine the two sets of triads together.

Application of our proposed method to those triads with and without the subtype provided more evidence of an association between the common DAT1 allele and the subtype of interest compared to the CPG approach. On the basis of the AIC values, our approach best fits the DAT1 marker data under a multiplicative model. Under this model, we found significant evidence of an association between the common DAT1 allele and the combined subtype ($p = 0.041$). As the RR estimate of the common variant is >1 , this result suggests there is moderate evidence that the common DAT1 variant increases susceptibility to the ADHD combined subtype.

Taken collectively, these ADHD results illustrate two important points. First, in general, the analysis of meaningful disease subtypes is valuable, since it may identify unique genetic signals that would otherwise be lost in the analysis of a more global (and more genetically heterogeneous) phenotype. Second, while subtype analyses may have limited power due to the reduced number of

eligible triads, the incorporation of triads without the subtype of interest into such analyses provides valuable information for inference and can substantially improve the ability to identify relevant subtype-influencing variants.

SIMULATION RESULTS

Table II reports type I error and power results at nominal $\alpha = 0.05$ for both our proposed test and the CPG test assuming simulated data from a sample of 150 complete triads with and 150 complete triads without a subtype of interest. For a given test, we report analysis results both assuming the true (simulation) model, as well as the general model that makes no assumptions on the genetic mechanisms of the alleles on disease. As expected, for a given test, we found the power under the true model was greater than that under the general model.

Under the null model assuming a RR of 1.0, both the CPG and our proposed tests had appropriate type-I error. For RR values >1.0 , results indicate our novel approach consistently outperforms the CPG approach across a broad range of RR values and genetic mechanisms. Power increases were most noticeable for multiplicative models, followed by recessive and dominant models. We expect these power increases since our proposed approach incorporates additional genetic information (from triads without the subtype of interest) that the CPG approach ignores.

TABLE II. Power results assuming nominal $\alpha = 0.05^a$

True model	Analysis model	Approach	RR value				
			1.00	1.25	1.50	1.75	2.00
Multiplicative	True	Combined test	0.049	0.300	0.723	0.936	0.988
		CPG test	0.050	0.241	0.606	0.852	0.956
	General	Combined test	0.055	0.236	0.633	0.885	0.970
		CPG test	0.057	0.185	0.511	0.764	0.917
Dominant	True	Combined test	0.054	0.109	0.212	0.333	0.455
		CPG test	0.055	0.092	0.182	0.284	0.377
	General	Combined test	0.053	0.088	0.162	0.253	0.348
		CPG test	0.056	0.075	0.132	0.227	0.315
Recessive	True	Combined test	0.050	0.213	0.557	0.827	0.947
		CPG test	0.050	0.187	0.484	0.735	0.882
	General	Combined test	0.055	0.170	0.463	0.743	0.900
		CPG test	0.056	0.151	0.370	0.642	0.796

^aSimulations assumed 150 complete triads with and 150 complete triads without subtype of interest. Power results are based on 10,000 replicates using estimated mating types from the ADHD study given in Table I.

Results in Table II assume complete triads with no missing parental data. To assess the impact of missing parental data on our proposed method, we again simulated 150 triads with and 150 triads without the subtype of interest but assumed a percentage of the parental genotype data were missing. We analyzed all datasets using our proposed EM algorithm assuming both the true (simulated) and general models and report power and type-I error results in Table III. Under the null model, we found our method had appropriate type I error for all genetic mechanisms and missing parental rates. Under alternative models, we found that power was hardly affected by missing parental data when the missing rate was <20%. For missing parental rates >20%, we did find moderate decreases in power relative to the ideal condition of no missing data. However, we also found our proposed EM-based approach consistently outperformed the EM-based version of the CPG approach [Weinberg, 1999] under all genetic mechanisms, RR values, and missing-parent rates considered (results not shown).

We conducted additional simulations to assess the impact of population stratification on the results of our proposed method. We again sampled 150 complete triads with and 150 complete triads without the subtype of interest. To induce stratification, we sampled the former set of triads in equal proportions from two discrete strata (described in Materials and Methods), but sampled the latter set of triads in

unequal proportions. For these latter triads, we sampled units from the first stratum with probability q , which we varied between 0.10 and 0.90 (with 0.50 corresponding to no stratification). For each dataset, we tested whether triads with and without the subtype of interest could be combined by testing the null hypothesis that the mating-type frequencies of the two sets were equal. We then recorded RR estimates under the best model (RR from the CPG approach when the null hypothesis is rejected, RR from our proposed approach otherwise). For comparison, we also recorded the RR estimates under the naive model that used our procedure to analyze the combined dataset ignoring the results of the test of mating-type equality.

Figure 1 (top) reports the power of our approach to reject the null hypothesis that the mating-type frequencies of triads with and without the subtype are the same as a function of the sampling probability q , assuming a multiplicative model and a true RR value of 1.0 (results for other models are similar). When $q = 0.50$ (corresponding to no stratification), our proposed approach has appropriate type I error. Under alternative models, the power increases with an increase in the magnitude of the difference between the assumed value of q and the null value of $q = 0.50$.

At least for simulation models based on the DAT1 results in this report, the power to detect stratification is generally modest. That being said, our test for stratification is still valuable in that the resulting inference reduces the bias that can occur

TABLE III. Power results assuming nominal $\alpha = 0.05$: missing parental data^a

True model	RR value	Analysis model	Percentage of missing parental data				
			0	10	20	30	40
Multiplicative	1.0	True	0.049	0.052	0.053	0.056	0.048
		General	0.055	0.050	0.053	0.053	0.052
	1.5	True	0.723	0.709	0.669	0.651	0.628
		General	0.633	0.601	0.580	0.541	0.527
Dominant	1.0	True	0.054	0.053	0.050	0.051	0.055
		General	0.053	0.053	0.052	0.053	0.055
	1.5	True	0.212	0.198	0.196	0.188	0.186
		General	0.162	0.161	0.150	0.144	0.142
Recessive	1.0	True	0.050	0.050	0.048	0.046	0.049
		General	0.055	0.054	0.054	0.056	0.053
	1.5	True	0.557	0.545	0.530	0.515	0.488
		General	0.463	0.439	0.419	0.407	0.397

^aSimulations assumed 150 triads with and 150 triads without subtype of interest. Power results are based on 10,000 replicates using estimated mating types from the ADHD study given in Table I.

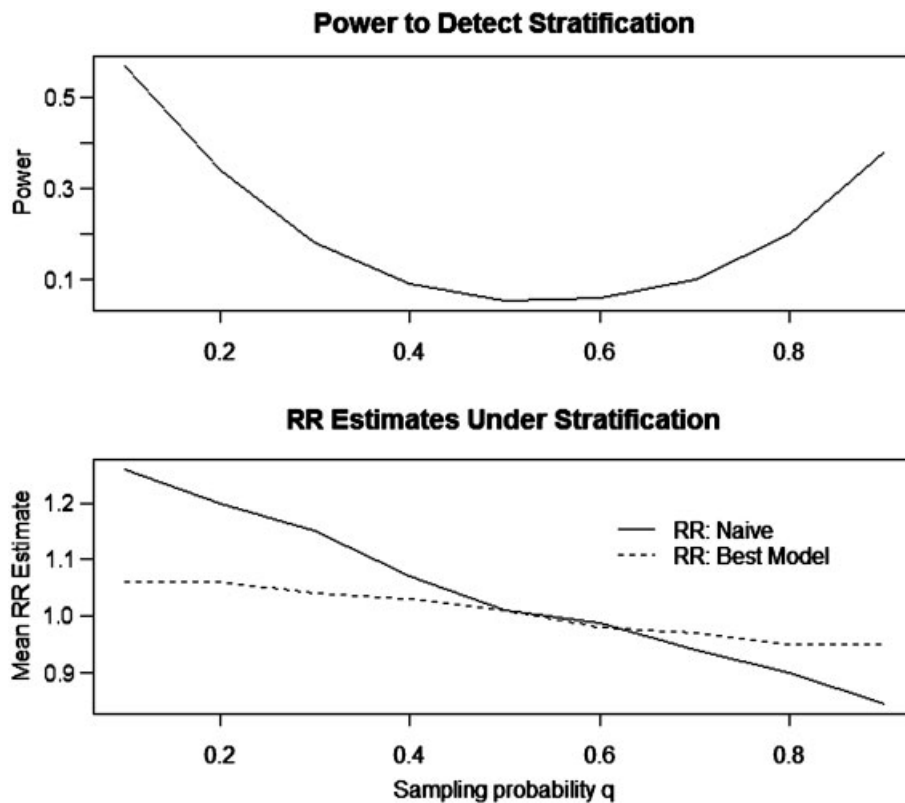


Fig. 1. Suitability of combining SNP data from 150 triads with and 150 triads without subtype of interest under population stratification. Simulations assume a multiplicative model and a RR value of 1.0. Triads with subtype of interest are sampled from two strata in equal proportions. Triads without subtype of interest are sampled from stratum 1 with probability q , with $q=0.50$ corresponding to no stratification. Top: Power of combined test to reject null hypothesis at $\alpha=0.05$ that mating-type frequencies of triads with and without subtype are the same. Bottom: The average naive RR estimate (solid line) using the combined approach, as well as the RR estimate (dashed line) for the best model (RR estimate from the CPG approach if null hypothesis is rejected, RR estimate from the combined approach otherwise). Results are based on 10,000 replicates of data.

in estimates of RR parameters. To show this, we present the mean RR estimates as a function of the sampling proportion q in Figure 1 (bottom). From this, the results clearly show that when stratification exists, naive analysis of triads using our proposed approach can lead to substantial bias in RR estimates, with the bias increasing steadily with an increase in the magnitude of the difference between the true value of q and the null value of $q = 0.50$. However, Figure 1 (bottom) also shows that the mean RR estimates based on the best model (RR estimates from the CPG approach when the null hypothesis of mating-type equality is rejected, RR estimates from our proposed approach otherwise) show little or no bias in these situations. These results show that testing whether triads with and without the subtype of interest can be combined prior to analysis is necessary to ensure valid results using our proposed approach.

DISCUSSION

Case-parent triad study designs are popular for association analyses of disease since results are robust to spurious association arising from population stratification. Rather than analyze all triads with the disease phenotype together, situations may arise where it is advantageous to partition triads into groups with different subtypes based on criteria such as similar symptoms or response to medication. Such subsets potentially are more genetically homogenous than the overall disease phenotype, which may facilitate the mapping of influential genetic variants. However, one potential drawback of triad analysis of disease subtypes is that the resulting analyses may be weakly powered if the number of eligible triads with the subtype of interest is small. To increase power in these situations, we have developed a novel likelihood-based approach for conducting

association analyses of disease subtypes in triads. Using both a real ADHD dataset as well as simulated data, we demonstrated that our proposed

approach is more powerful than the typical approach of analyzing only those triads with the subtype of interest using the CPG approach. We intend to implement these proposed methods for subtype analysis in a software package for public use, which can be downloaded free of charge from our Web site (see Epstein software Web site).

While we developed our subtype method for triads only, our method can easily be extended to accommodate additional unrelated subjects (including cases with the subtype of interest and unaffected controls) within the likelihood using similar approaches as described by Nagelkerke et al. [2004] and Epstein et al. [2005]. Cases provide additional information on RR parameters and mating-type frequencies, while controls provide information on the mating-type frequencies only. Inclusion of this additional information will further increase the power to detect association between a SNP and the subtype of interest.

We have also developed statistical tests to ensure that triads with and without the subtype of interest can be safely combined for analysis. As demonstrated by the simulation results, substantial bias in naive RR parameter estimates can occur when population stratification exists between triads with and without the subtype of interest. This arises from the fact that information on β from $P[G_p|D_o = A]$ in (2) comes from comparing the parental mating-type frequencies in subtype *A* triads with the mating-type frequencies originating from the parental genotypes of subtype *B* triads in L_B . Differences in the distribution of these two sets of frequencies can grossly bias the information on β from (2), resulting in bias in the overall estimate of β using $L = L_A \cdot L_B$. We note that our proposed statistic for assessing the suitability of combining triads with subtypes *A* and *B* differs from a strategy described by Epstein et al. [2005], which translates in this setting to assessing whether the estimates from β in (2) differ from the β estimates from the CPG-based portion of the likelihood. The motivation behind this test is that the former RR estimates are susceptible to bias arising from mating-type frequency differences between triads with subtypes *A* and *B*, whereas the latter RR estimates will still be unbiased in these situations. Therefore, if the two sets of RR estimates are the same, then this suggests the two samples can be

safely combined. We evaluated this other test for assessing the suitability of combining triads with and without the subtype of interest under population stratification and found it to be less powerful relative to our proposed test that is based on direct comparison of the mating-type frequencies of the two samples (results not shown). We, therefore, recommend this latter test to ensure valid association analysis of disease subtypes.

We have not considered the effect of imprinting or parental genotype effects on offspring phenotype. These effects could be incorporated into our approach by adding appropriate parameters to the relative risk model, as shown by Weinberg et al. [1998]. If imprinting or parent-of-origin effects are suspected, then we recommend that they be included in the nuisance model for subtype *B* as well as the model for subtype *A*.

While we developed our proposed approach for the analysis of single SNPs with disease, we could extend our approach to conduct subtype analysis of a multiallelic locus. Let s_1 and s_2 denote the two alleles representing offspring genotype g . We can model the allele effects on disease within the genotype RR using an additive mechanism, such that $\psi_g = \exp(\beta_{s_1} + \beta_{s_2})$, where β_{s_1} and β_{s_2} are the disease-risk parameters for the two alleles. Other allele-effect models are also possible. We can conduct inference on these disease-risk parameters using a variation of the likelihood $L = L_A \cdot L_B$. As described previously, L_A is a function of ψ_g and $P[G_p]$, while L_B is a function of $P[G_p]$ and ϕ_g . For a multiallelic locus, the modeling of $P[G_p]$ is complicated if one uses a symmetric-mating type distribution in analysis. While a biallelic SNP only requires specification of 6 mating-type parameters, a 3-allele locus requires 21 mating-type parameters and a 4-allele locus requires 55 mating-type parameters for analysis. This increased number of mating-type parameters that must be modeled may result in subtype analyses that are numerically unstable and computationally inefficient. Additionally, many of the mating-types will have small (or zero) counts, which may result in the need of permutation-based procedures for proper inference. Because of these unappealing limitations of using mating types, we feel it is likely advantageous to model $P[G_p]$ under the assumption of random mating since it significantly reduces the number of parameters in $P[G_p]$ that have to be estimated (6 parameters for a 3-allele locus, 10 parameters for a 4-allele locus).

We note that we can implement an alternative strategy for analysis of subtype *A* that does not

require modeling of the subtype B RR parameters ϕ_1 and ϕ_2 . Rather than consider the genotypes of an entire subtype B triad for analysis, one instead considers the distribution of the *untransmitted* parental alleles G_u conditional on the offspring genotype G_o in the triad, which we denote by $P[G_{ui} = g_u | G_{oi} = g, D_{oi} = B]$. One can rewrite this probability as

$$P[G_{ui} = g_u | G_{oi} = g, D_{oi} = B] = \frac{P[G_{ui} = g_u, G_{oi} = g]}{\sum_{g_u^*} P[G_{ui} = g_u^*, G_{oi} = g]}.$$

$P[G_{ui} = g_u, G_{oi} = g]$ denotes the joint probability of the untransmitted and transmitted alleles in the population, which is a function of μ only. Therefore, use of the above probability facilitates the extraction of useful information on μ without characterization of ϕ_g . We evaluated the performance of this approach for subtype analysis and found power and type I error to be equivalent to our proposed approach (results not shown). While using untransmitted parental alleles avoids specification of the nuisance parameters ϕ_1 and ϕ_2 , it has the limitation of being harder to generalize to more complicated situations such as imprinting, parental genotype effects, or loci having more than two alleles.

ACKNOWLEDGMENTS

We thank Dr. Michael Boehnke for his helpful comments on an earlier form of this manuscript, and also acknowledge a useful conversation with Dr. Clarice Weinberg. This research was supported by the University Research Committee of Emory University (to M.P.E.).

ELECTRONIC-DATABASE INFORMATION

The URLs for methods presented herein are as follows:

ALFRED, <http://alfred.med.yale.edu/alfred/index.asp>

Epstein software, <http://server2k.genetics.emory.edu/mepstein/software.php>

REFERENCES

Akaike H. 1985. Prediction and entropy. In: Atkinson AC, Fienberg SE, editors. A celebration of statistics. New York: Springer. p 1–24.
Boos DD. 1992. On generalized score tests. *Am Stat* 46:327–333.

Buckland ST, Burnham KP, Augustin NH. 1997. Model selection: an integral part of inference. *Biometrics* 53:603–618.
Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ. 2000. ALFRED: a Web-accessible allele frequency database. *Pac Symp Biocomput* 639–650.
Cook EH Jr, Stein MA, Krasowski MD, Cox NJ, Olkon DM, Kieffer JE, Leventhal BL. 1995. Association of attention-deficit disorder and the dopamine transporter gene. *Am J Hum Genet* 56:993–998.
Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 39:1–22.
Doucette-Stamm LA, Blakely DJ, Tian J, Mockus S, Mao JI. 1995. Population genetic study of the human dopamine transporter gene (DAT1). *Genet Epidemiol* 12:303–308.
Epstein MP, Veal CD, Trembath RC, Barker J NWN, Li C, Satten GA. 2005. Genetic association analyses using data from triads and unrelated subjects. *Am J Hum Genet* 76:592–608.
Gill M, Daly G, Heron S, Hawi Z, Fitzgerald M. 1997. Confirmation of association between attention deficit hyperactive disorder and a dopamine transporter polymorphism. *Mol Psychiatry* 2:311–313.
International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
Kang AM, Palmatier MA, Kidd KK. 1999. Global variation of a 40-bp VNTR in the 3'-untranslated region of the dopamine transporter gene (SLC6A3). *Biol Psychiatry* 46:151–160.
Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. 2004. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet* 12:964–970.
Schaid DJ. 1999. Likelihoods and TDT for the case-parents design. *Genet Epidemiol* 16:250–260.
Schaid DJ. 2004. Transmission disequilibrium methods for family-based studies. Mayo Clinic Technical Report 72. Rochester, MN: Department of Health Sciences Research, Mayo Clinic.
Schaid DJ, Sommer SS. 1993. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 53:1114–1126.
Schaid DJ, Sommer SS. 1994. Comparison of statistics for candidate-gene association studies using case and parents. *Am J Hum Genet* 55:402–409.
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29:308–311.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516.
Waldman ID, Rowe DC, Abramowitz A, Kozel ST, Mohr JH, Sherman SL, Cleveland HH, Sanders ML, Gard JMC, Stever C. 1998. Association and linkage of the dopamine transporter gene and attention-deficit hyperactivity disorder in children: heterogeneity owing to diagnostic subtype and severity. *Am J Hum Genet* 63:1767–1776.
Weinberg CR. 1999. Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 64:1186–1193.
Weinberg CR, Umbach DM. 2005. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet* 77:627–636.
Weinberg CR, Wilcox AJ, Lie RT. 1998. 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* 62:969–978.