Genetic Epidemiology 32 : 520–527 (2008)

Simple Association Analysis Combining Data From Trios/Sibships and Unrelated Controls

Yi-Hau Chen[1]* and Hui-Wen Lin[2]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, Republic of China*
[2]*Department of Statistics, National Chengchi University, Taipei, Taiwan, Republic of China*

We consider genetic association analysis that combines data from case-parent trios/sibships and unrelated controls. A general and simple methodology is proposed, using a weighted least-squares approach to combine separate information from the case-parent/case-sibling analysis and the case-unrelated control analysis. The new proposal improves over the existing methods in that it requires no assumptions and estimation on the mating-type distribution. Simulation results show that the new method competes well with the likelihood-based method when the latter is applicable, and achieves substantial power gains over separate analyses in general. Therefore, the proposed combined association analysis can enjoy wide applications, including the multiallele/locus, haplotype, and genome-wide association studies. *Genet. Epidemiol.* 32:520–527, 2008. © 2008 Wiley-Liss, Inc.

Key words: case-control studies; combined association analysis; family-based association analysis; population-based association analysis

Contract grant sponsor: Genomic Research Center, Academia Sinica; Contract grant number: 94B001-2.
*Correspondence to: Yi-Hau Chen, Ph.D., Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China.
E-mail: yhchen@stat.sinica.edu.tw
Received 7 May 2007; Revised 7 January 2008; Accepted 8 February 2008
Published online 17 March 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20325

INTRODUCTION

The genetic association analysis which detects gene-disease association, either the direct association caused by a disease susceptibility gene itself, or the indirect association caused by the linkage disequilibrium of a disease susceptibility gene with adjacent markers, has now been a very popular and promising tool to locate genes that underlie complex human diseases [Risch, 2000; Cardon and Bell, 2001]. A commonly used design in genetic association studies is the population-based case-control design, in which unrelated cases and controls are collected and compared with respect to the frequencies of some genetic variants. This study design has an advantage that the implementation is very convenient, since recruiting population controls are both time- and cost-effective.

One potential drawback for the population-based study is its lack of protection against confounding due to unmeasured race/ethnicity factors, so that an excess of false-positive results may arise when population stratification exists. To avoid a spurious association by this confounding, the family-based designs using relatives of the cases as controls have been proposed. The simplest family design is the case-parent design, where both parents of the affected subjects are chosen as the family controls [Falk and Rubinstein, 1987; Spielman et al., 1993]. The transmission/disequilibrium test (TDT) proposed by Spielman et al. [1993] in case-parent designs has been popular for testing the gene-disease association, whose validity relies only on the assumption of Mendelian transmissions, and hence is immune to the population stratification bias. When genotype data for parents are not available, such as in the study of late onset diseases, unaffected siblings of affected subjects can be chosen as the family controls [Curtis, 1997; Schaid and Rowland, 1998; Spielman and Ewens, 1998]. The disadvantage of the case-parent/case-sibling design is that recruiting family controls usually requires more resources in terms of time and money [Laird and Lange, 2006].

Recently, some authors have drawn attention to the association study using both family-based and population-based controls [Martin and Kaplan, 2000; Mitchell, 2000; Nagelkerke et al., 2004; Kazeem and Farrall, 2005; Epstein et al., 2005]. Motivations for this type of studies include: (1) using the TDT as a follow-up and confirmatory test for gene-disease association detected in a case-unrelated control study, since the significance of the TDT ensures a true association; or (2) supplementing the case-parent trios with additional unrelated controls, if available, to ensure a sufficient power to detect association, since parental controls may be hard to recruit, especially for late-onset diseases.

In this work, using a weighted least-squares (WLS) approach, a combined association analysis is proposed for combining separate association information from the case-parent/case-sibling analysis and the case-unrelated control analysis. We also propose a procedure for testing whether the family data (trios/sibships) could be combined with the unrelated control data. The new approach improves over the existing methods in that it involves no assumptions and estimation on the mating-type distribution. Power comparisons made in simulations show that the new approach competes well with the likelihood-based method of Epstein et al. [2005] in the setting where the latter is applicable, and achieves substantial power gains

over the separate analyses in general settings for combining data from case-parent trios/sibships and unrelated controls. Our proposal can thus have wide applications, such as the multiallele/locus, haplotype, and genome-wide association analyses.

METHODS

COMBINING DATA FROM CASE-PARENT TRIOS AND UNRELATED CONTROLS

Let D denote disease status with $D = 1$ denoting affected and $D = 0$ denoting unaffected, G the genotype, and $h(G)$ a vector of genotype covariates, which can be coded according to some suitable genetic model [e.g., multiplicative, dominant, or recessive model; Schaid and Sommer, 1993]. Let β be the vector of log genotype relative risk (RR) parameters so that

$$\log\left\{\frac{\mathrm{pr}(D=1|G)}{\mathrm{pr}(D=1|G=g_0)}\right\} = \beta' h(G), \tag{1}$$

where g_0 represents some reference genotype so that $h(g_0)$ is a zero vector.

Suppose that we have a case-parent trio sample, which consists of genotype data $\{G_i\}_{i=1}^{N_1}$ for N_1 affected subjects (cases) and genotype data $\{G_i^P\}_{i=1}^{N_1}$ for both parents of the cases. Following Self et al. [1991] and Schaid and Sommer [1993], we can estimate β by maximizing the conditional on parental genotypes (CPG) likelihood $\prod_{i=1}^{N_1} \mathrm{pr}(G_i|G_i^P, D_i = 1)$ with

$$\mathrm{pr}(G_i|G_i^P, D_i = 1) = \frac{\exp\{\beta' h(G_i)\}\,\mathrm{pr}(G_i|G_i^P)}{\sum_{\tilde{g}_i}\exp\{\beta' h(\tilde{g}_i)\}\,\mathrm{pr}(G=\tilde{g}_i|G_i^P)},$$

where $\mathrm{pr}(G_i|G_i^P)$ is given by the Mendelian proportions, and $\tilde{g}_i$ is over all the possible offspring genotypes for the given parental genotype G_i^P. Denote the resulting estimator by $\hat{\beta}_{\mathrm{TRIO}}$, which is an association measure summarizing from the case-parent data.

Suppose that we have further genotype data $\{G_i\}_{i=N_1+1}^{N_1+N_0}$ for N_0 unrelated controls. Then, based on the case-unrelated control data $\{D_i, G_i\}_{i=1}^{N_1+N_0}$ (where by definition $D_i = 1$ for $i = 1, \ldots, N_1$ and $D_i = 0$ for $i = N_1 + 1, \ldots, N_1 + N_0$), we can obtain an estimate for the genotype odds ratio (OR) parameter β_* defined by

$$\log\left\{\frac{\mathrm{pr}(D=1|G)}{\mathrm{pr}(D=0|G)}\right\} = \alpha + \beta_*' h(G) \tag{2}$$

with α a nuisance intercept parameter, using traditional logistic regression analysis that treats the case-control data as prospectively collected [Prentice and Pyke, 1979]. Denote by $\hat{\beta}_{*,\mathrm{CC}}$ the resulting estimator for β_*.

Therefore, with genotype data from case-parent trios as well as unrelated controls, association information can be acquired, respectively, through $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\beta}_{*,\mathrm{CC}}$. In situations where the two pieces of information are essentially equivalent, such as when the disease is rare [Breslow and Day, 1980, pp 70–71], and the case-parent trios and unrelated controls are sampled from the same population where no population stratification exists [Epstein et al., 2005], it would be advantageous to integrate them to enhance the statistical power. Note that the equivalence of the two sources of information can be empirically checked via a statistical test for $\beta = \beta_*$. A likelihood-ratio statistic for testing $\beta = \beta_*$ with trio and unrelated control data has been suggested by Epstein et al. [2005], and we will propose an alternative Wald-type test later.

Here we assume that the suitable conditions hold such that $\beta \approx \beta_*$ and the two parameters are equivalent measures for gene-disease association. In this case, let Σ_{11} be the variance matrix of $\hat{\beta}_{\mathrm{TRIO}}$, Σ_{22} be the variance matrix of $\hat{\beta}_{*,\mathrm{CC}}$, and Σ_{12} be the covariance matrix between $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\beta}_{*,\mathrm{CC}}$. By the linear model theory [Seber, 1997, pp 61–62], the optimal (most efficient) estimator for β based on the linear combination of $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\beta}_{*,\mathrm{CC}}$ can be obtained by the WLS estimator, which is given by

$$\hat{\beta} = W_1\hat{\beta}_{\mathrm{TRIO}} + W_2\hat{\beta}_{*,\mathrm{CC}}, \tag{3}$$

where $W_1 = (\Sigma_{22} - \Sigma_{12}')Q^{-1}$, $W_2 = (\Sigma_{11} - \Sigma_{12})Q^{-1}$, and $Q = \Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{12}'$ (see the Appendix). In practical implementation, we need to substitute suitable estimates for Σ_{jk}s. Note that estimates for Σ_{11} and Σ_{22} can be, respectively, obtained from the information (negative Hessian) matrices of the CPG and logistic regression analyses. The estimate for Σ_{12} can be obtained by using the scores and information matrices for the CPG and logistic regression analyses. Explicit expressions for estimators of Σ_{jk}s are given in the Appendix.

The variance of $\hat{\beta}$ can be estimated by

$$\mathrm{var}(\hat{\beta}) = \Sigma_{11} - W_1 Q W_1' = \Sigma_{22} - W_2 Q W_2' \tag{4}$$

with Σ_{jk}s substituted with their estimates. It can thus be explicitly seen that the combined estimator $\hat{\beta}$ is more efficient than the two separate estimators $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\beta}_{*,\mathrm{CC}}$. The null hypothesis of no linkage or no association between a locus and disease, i.e., $\beta = 0$, can be tested with the Wald-test statistic $\hat{\beta}'\mathrm{var}(\hat{\beta})^{-1}\hat{\beta}$, which is an association test integrating information from case-parent trios and unrelated controls.

Multiple affected offsprings in a family can also be included in the proposed combined analysis. Suppose that $\hat{\beta}_{\mathrm{TRIO}}$ denotes the CPG estimator for β using all possible case-parent trios formed by the affected offsprings and the parents and treating them as independent, and $\hat{\beta}_{\mathrm{CC}}$ denotes the logistic regression estimator using all affected offsprings and unrelated controls as the case-control sample and treating them as independent. A combined estimator $\hat{\beta}$ and its variance can still be obtained using expressions (3) and (4), except that the involved variance-covariance matrices Σ_{jk}s should be substituted with their "robust" estimates (see the Appendix), which still provide valid variance estimation when trios/subjects are in fact dependent.

It is seen that the proposed combined association analysis does not require any assumptions and estimation on the mating-type distribution. In contrast, the methods by Nagelkerke et al. [2004] and Epstein et al. [2005] require estimation of the mating-type parameters, which may be high dimensional in the multiallele/locus setting and hence may complicate the analysis. The Nagelkerke et al. method further assumes the Hardy-Weinberg equilibrium (HWE).

A TEST FOR THE APPROPRIATENESS OF COMBINING THE DATA

We further propose a test statistic for testing whether it is appropriate to combine the association information from

the case-parent and case-control analyses. Since the equality of β and β_* would encourage a combined analysis while a discrepancy between them would not, the proposed test statistic, T_C, is based on a direct comparison of $\hat{\beta}_{\text{TRIO}}$ and $\hat{\beta}_{*,\text{CC}}$:

$$T_C = (\hat{\beta}_{\text{TRIO}} - \hat{\beta}_{*,\text{CC}})' \hat{Q}^{-1} (\hat{\beta}_{\text{TRIO}} - \hat{\beta}_{*,\text{CC}}), \quad (5)$$

where $\hat{Q} = \hat{\Sigma}_{11} + \hat{\Sigma}_{22} - \hat{\Sigma}_{12} - \hat{\Sigma}_{12}'$ is an estimate for the variance of $\hat{\beta}_{\text{TRIO}} - \hat{\beta}_{\text{CC}}$, with $\hat{\Sigma}_{jk}$ denoting the estimate for Σ_{jk}. Under the null hypothesis of $\beta = \beta_*$ T_C is distributed as a p-df χ^2 random variable, where p is the dimension of β.

COMBINING DATA FROM SIBSHIPS AND UNRELATED CONTROLS

When parental genotype data are lacking, such as when the disease under study is late onset, it is a common practice to use siblings as family controls. Here we extend our proposal to the situation where genotype data are available for a set of sibships and a set of unrelated controls, and it is suitable to combine information from them. Note that the likelihood-based method by Epstein et al. [2005] cannot be applied to this type of data.

With each sibship served as a matched set, the conditional logistic regression analysis, originally developed for matched case-control study, can be applied to obtain the estimate $\hat{\beta}_{*,\text{SIB}}$ for the parameter β_* in model (2) [Siegmund et al., 2000], where the intercept α is allowed to be sibship-specific to account for shared but unmeasured genetic/environmental factors in the sibship. On the other hand, suppose that genotype data for a set of unrelated controls are also available, and the unrelated controls and sibships are from the same population where no population stratification exists. We can then perform an unconditional logistic regression analysis, using the case-unrelated control sample formed by the affected sibs and unrelated controls, to obtain the "unconditional" estimate $\hat{\beta}_{*,\text{CC}}$. Note that the unconditional estimate $\hat{\beta}_{*,\text{CC}}$ is approximately unbiased for β_* in (2), even in the presence of a sibship-specific random intercept α if the disease is rare so that $\text{pr}(D = 1|G) \approx \exp\{\alpha + \beta_* h(G)\}$ [Zeger et al., 1988]. Similarly for combining $\hat{\beta}_{\text{TRIO}}$ and $\hat{\beta}_{*,\text{CC}}$, we can use the WLS approach to obtain the estimator β_* combining $\hat{\beta}_{*,\text{SIB}}$ and $\hat{\beta}_{*,\text{CC}}$, where

$$\hat{\beta}_* = W_1 \hat{\beta}_{*,\text{SIB}} + W_2 \hat{\beta}_{*,\text{CC}}$$

with $W_1 = (\Sigma_{22} - \Sigma_{12}')(\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{12}')^{-1}$ and $W_2 = (\Sigma_{11} - \Sigma_{12})(\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{12}')^{-1}$. Here, Σ_{11} is the variance matrix of $\hat{\beta}_{*,\text{SIB}}$, Σ_{22} is the variance matrix of $\hat{\beta}_{*,\text{CC}}$, and Σ_{12} is the covariance matrix of $\hat{\beta}_{*,\text{SIB}}$ and $\hat{\beta}_{*,\text{CC}}$. The variance-covariance matrices Σ_{jk}s can be estimated by the scores and information matrices from the conditional and logistic regression likelihoods. When there are multiple affected sibs in a sibship, to ensure valid variance estimation with the resulting correlated data, robust variance estimates should be used for Σ_{11} [Siegmund et al., 2000; Fay et al., 1998] and for Σ_{22} [Liang and Zeger, 1986], and also a robust covariance estimate should be used for Σ_{12}. See the Appendix for expressions for robust estimators of Σ_{jk}s. The variance of $\hat{\beta}_*$ is again of form (4), and a combined association test can be performed by the Wald test based on $\hat{\beta}_*$ and its variance estimate. Further, a test statistic analogous to (5) that compares $\hat{\beta}_{*,\text{SIB}}$ and $\hat{\beta}_{*,\text{CC}}$ can be used for checking the appropriateness of combining data from sibships and unrelated controls.

When genotype data on the affected subjects, their parents and unaffected siblings, and a set of unrelated controls are all available, the proposed WLS approach can still be applied to yield integrated information. In this case we can obtain three separate estimates for the association parameters: $\hat{\beta}_{\text{TRIO}}$ from the case-parent analysis, $\hat{\beta}_{*,\text{SIB}}$ from the case-sibling analysis, and $\hat{\beta}_{*,\text{CC}}$ from the logistic regression analysis with the case-control sample formed by affected subjects and unrelated controls. Let $Y = (\hat{\beta}_{\text{TRIO}}, \hat{\beta}_{*,\text{SIB}}, \hat{\beta}_{*,\text{CC}})'$ be the vector obtained by stacking the three set of parameter estimates and $X = [I_p|I_p|I_p]'$ the matrix formed by stacking three p-dimensional identity matrices I_p (p is the dimension of β). The proposed combined estimator is the optimal linear combination of the three separate estimators given as

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y,$$

where Σ is the variance-covariance matrix for $Y = (\hat{\beta}_{\text{TRIO}}, \hat{\beta}_{*,\text{SIB}}, \hat{\beta}_{*,\text{CC}})'$, whose component submatrices can be estimated in the way described previously. The variance of $\hat{\beta}$ is obtained as $\text{var}(\hat{\beta}) = (X'\Sigma^{-1}X)^{-1}$. We proposed the Wald test based on $\hat{\beta}$ as an association test employing joint information from cases, parents, unaffected siblings, and unrelated controls.

SIMULATION RESULTS

We examine the performance of the proposed combined association test through simulation studies. These simulations are conducted under settings where the combining of the data is suitable. We suppose that both the disease and marker loci are diallelic with a common minor allele frequency (MAF) equal to 0.1 (rare variant) or 0.4 (common variant), and the standardized linkage disequilibrium coefficient [Lewontin, 1988] between them is fixed at 0.8. The recombination rate is set to 0. Given the haplotype frequencies so determined, the diplotype (haplotype pair) for a subject is generated assuming HWE and random mating. The disease outcome is generated by model (1) or (2) with G given by the genotype at the disease locus and $h(G)$ specified according to a multiplicative, dominant, or recessive model. The disease prevalence is fixed at 5%. The size of the tests considered is evaluated with the genotype RR or the genotype OR at the disease locus set to 1, and the power are evaluated at two values of RR or OR that are greater than 1. The size and power are evaluated under a significance level of 0.05. When performing the analysis, we use only the marker genotypes as the genetic data, and code $h(G)$ as the number of copies of the minor allele; that is, we treat the true genetic model as unknown and use a multiplicative working model. All results are based on 1,000 simulation replications.

We first examine the size and power of the proposed association test combining data from 100 trios and 100 unrelated controls, and compare the performance with that of the likelihood-ratio test proposed by Epstein et al. [2005]. In Figure 1 we show the size and power for the combined association tests from our proposal and Epstein et al. Also shown are the size and power from the CPG analysis using only the trio data, and the logistic

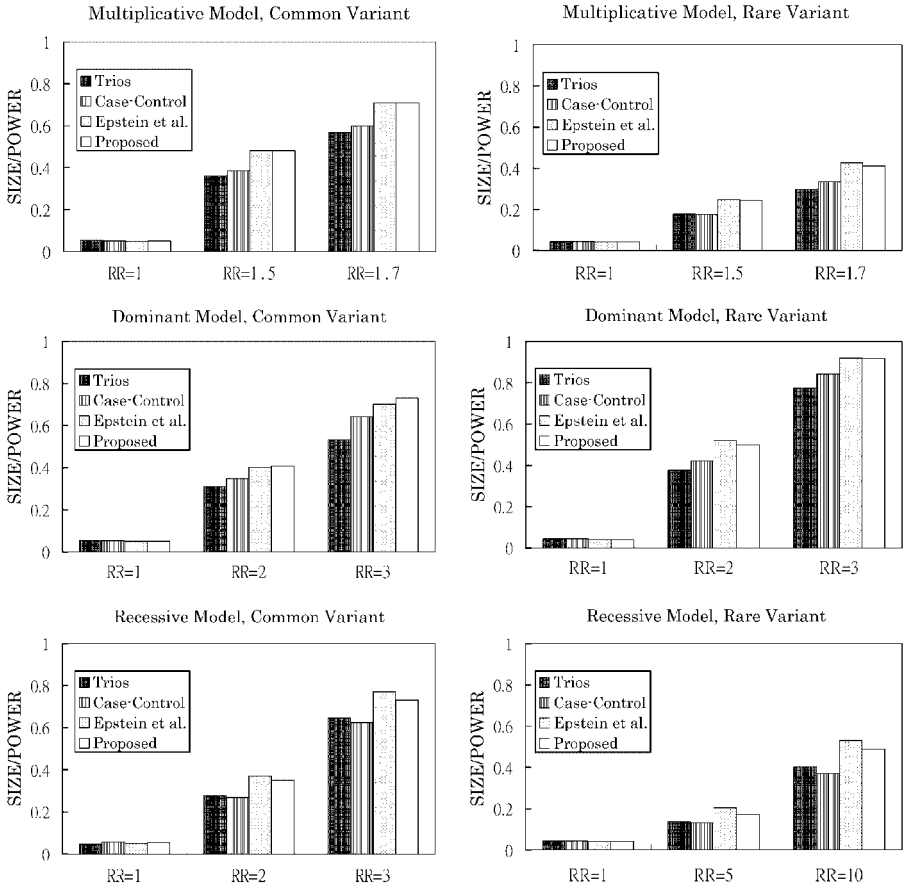

Fig. 1. Size and power at 5% significance level of the tests based on case-parent analysis (Trios), case-unrelated control analysis (Case-Control), and the combined association analyses of the Epstein et al. and the proposed methods. RR denotes the genotype relative risk at the disease locus. The MAF (for both the disease and marker loci) is 0.1 (rare variant) or 0.4 (common variant). Results are based on 100 trios and 100 unrelated controls, and 1,000 simulation replications. MAF, minor allele frequency. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

regression analysis using only data from affected offsprings and unrelated controls. It is seen that all the association tests considered have correct size (type-I error rate) under the null hypothesis (RR = 1). Compared to the likelihood-based method of Epstein et al., the proposed Wald test based on $\hat{\beta}$ has comparable power for detecting gene-disease association; it is equally or slightly more powerful when the disease/marker allele is common and the genetic effect is multiplicative or dominant, while is slightly less powerful when the allele is rare or the genetic effect is recessive. The combined analyses (the proposed and the Epstein et al. methods) do achieve higher power than the separate analyses (the CPG and the logistic regression analyses). Since in this study it is suitable to combine the trio and unrelated control data, we also examine whether the proposed test statistic T_C for checking the appropriateness of combining the data can reveal the appropriateness with sufficient probability. The empirical size given in Table I for various genetic models and various values of RR shows that the proposed test does have correct type-I error rates.

Based on the above setting for the multiplicative model with RR = 1.7, we further vary the number of unrelated trios from 100 to 250 while fix the number of unrelated

TABLE I. Size at 5% significance level of the proposed test for checking the appropriateness of combining the trio and unrelated control data[a]

	Rare variant[b]	Common variant[c]
Multiplicative model		
RR = 1	0.051	0.056
RR = 1.5	0.040	0.042
RR = 1.7	0.047	0.057
Dominant model		
RR = 1	0.051	0.056
RR = 2	0.057	0.050
RR = 3	0.057	0.060
Recessive model		
RR = 1	0.051	0.051
RR = 2	0.045	0.060
RR = 3	0.046	0.050

RR, relative risk; MAF, minor allele frequency.
[a]Based on 1,000 simulations (100 trios and 100 unrelated controls in each), under the settings same as those in Figure 1.
[b]MAF (for both the disease and marker loci) = 0.1.
[c]MAF (for both the disease and marker loci) = 0.4.

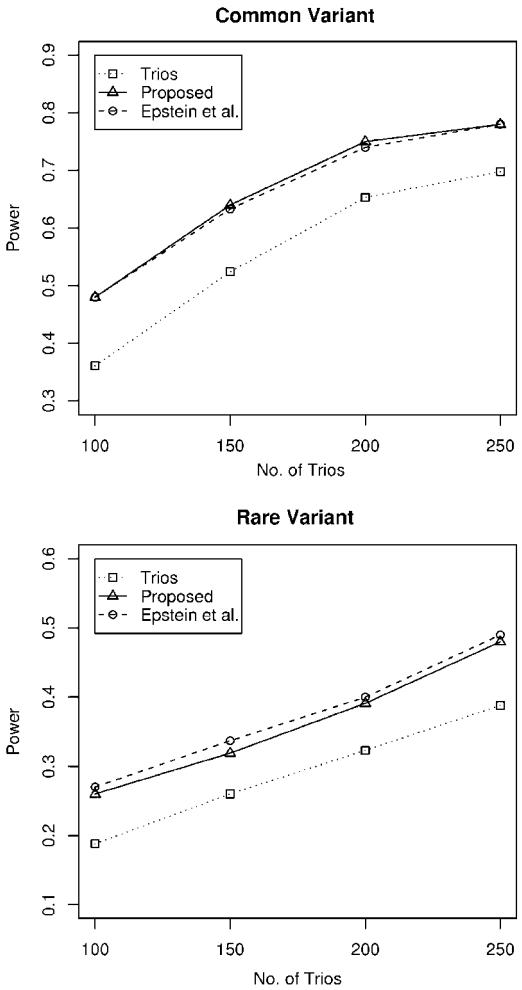

Fig. 2. Power at 5% significance level of the combined association tests from the proposed and Epstein et al. methods, when the number of trios increases from 100 to 250. The number of unrelated controls is fixed at 100. Data are simulated under the multiplicative model with RR = 1.7. The MAF is 0.1 (rare variant) or 0.4 (common variant). Results are based on 1,000 simulation replications. RR, relative risk; MAF, minor allele frequency. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

controls at 100. The power curves in Figure 2 show that the proposed and Epstein et al. association tests gain power as the number of trios increases. The proposed combined association test tends to be slightly more powerful than the Epstein et al. test in the common-variant settings, but is slightly less powerful in the rare-variant settings. We also evaluate the power of the association tests with the number of trios fixed at 100 while varying the number of unrelated controls from 100 to 300. Both the tests from our proposal and Epstein et al. gain power as the number of unrelated controls increases, and the two tests have virtually equivalent performance (data not shown).

Next we assess the power of the proposed association test combining data from sibships and unrelated controls. Genotypes and disease outcomes for sibships of size 3 are generated, with the intercept α in model (2) being a normal random variable which has a variance of 1 and a mean to yield an overall disease prevalence of 5%. A total of 100 sibships with at least one affected and one unaffected subject are selected for analysis. A set of 100 unrelated controls are sampled from the same population. Figure 3 displays the size and power for the association tests based, respectively, on $\hat{\beta}_{*,SIB}$, $\hat{\beta}_{*,CC}$, and the proposed estimate $\hat{\beta}$ combining $\hat{\beta}_{*,SIB}$ and $\hat{\beta}_{*,CC}$. It is seen that, relative to the separate analyses based only on sibship data or data from affected sibs and unrelated controls, the proposed combined test can achieve substantial power gains. Note that in this case there may be multiple affected sibs in a family and there is linkage between the disease and the marker locus; hence, there exists residual familial correlation. All the tests considered are thus based on the robust variance estimates presented in the Appendix, and the correct type-I error rates of these tests shown in Figure 3 imply that the robust variance estimation works well.

To examine the power of the proposed test for checking the appropriateness of combining the family data with unrelated controls, an additional set of simulations is conducted under two scenarios where combining the data is not valid. The first scenario is a setting where the unrelated controls and the trios are sampled from different populations: the MAF (for both the disease and marker loci) is set to 0.2 in the population where the trios are sampled, while that in the population where the unrelated controls are sampled is varied from 0.15 to 0.35. In each of the two populations, parental genotypes are generated under HWE and random mating. There exists no gene-disease association (RR = 1) and the disease prevalence is 5%. The second scenario is a setting with population stratification, where the population at large consists of two strata with constituent proportions (50, 50%). The trios and unrelated controls are randomly sampled from the population at large. There exists no gene-disease association (RR = 1) and the disease prevalence in the two strata is 2 and 10%, respectively. The MAF in the first stratum is 0.1, while that in the second stratum is varied from 0.2 to 0.5. Each of the two strata is under HWE, and the mating is random and restricted within the same stratum.

It is seen from Table II that, at significance level of 5%, the proposed test has satisfactory power to detect the inappropriateness of combining data from trios and unrelated controls. In the scenarios considered, the performance of our proposal is quite comparable to that of the Epstein et al. method: both the two methods have similar amount of bias in their respective estimates for the association parameter, and have similar power for detecting the inappropriateness of combining the data.

DISCUSSION

The methodology proposed can be extended to more complicated situations. Since the resulting test statistics involve only outputs from standard analyses (e.g., the CPG analysis and the unconditional or conditional logistic regression), the proposed method can be conveniently and efficiently implemented and hence is very suitable for genome-wide association analyses. Another important extension is to the haplotype-disease association analysis. To be specific, let $\hat{\beta}_{TRIO}$ now denote some estimate for the log haplotype RR parameters obtained by the CPG analysis based on haplotypes. Such an estimate can be obtained by, for example, Cordell and Clayton [2002] or

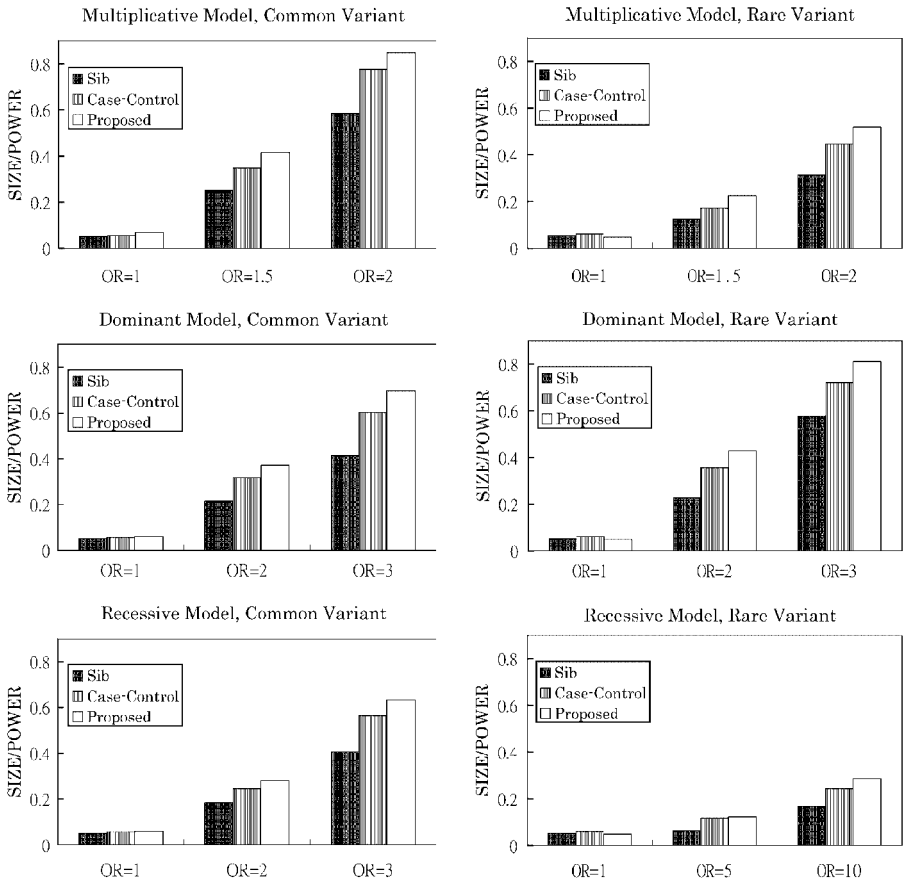

Fig. 3. Size and power at 5% significance level of the tests based on case-sibling analysis (Sib), case-unrelated control analysis (Case-Control), and the proposed combined association analysis. OR denotes the genotype odds ratio at the disease locus. The MAF is 0.1 (rare variant) or 0.4 (common variant). Results are based on 100 trios and 100 unrelated controls, and 1,000 simulation replications. MAF, minor allele frequency. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

Allen and Satten [2007]. Let $\hat{\beta}_{*,CC}$ now denote some estimate for the log haplotype OR parameters from the case-unrelated control analysis based on haplotypes, which can be obtained by, for example, Epstein and Satten [2003], Stram et al. [2003], Zhao et al. [2003], and Allen and Satten [2008]. An estimate $\hat{\beta}$ combining $\hat{\beta}_{TRIO}$ and $\hat{\beta}_{*,CC}$ can then be obtained by expression (3), with Σ_{ij}s being estimated using scores (or estimating functions) and information matrices for these two separate estimates. Note that some existing haplotype-based analyses [e.g., Epstein and Satten, 2003] depend crucially on further assumptions such as HWE [Allen and Satten, 2008].

In some situations, genotype data may also be available for, in addition to case-parent trios/sibships and unrelated controls, a set of unrelated cases. In this case, in addition to the estimate $\hat{\beta}_{TRIO}/\hat{\beta}_{*,SIB}$ from the case-parent/case-sibling analysis, and the estimate $\hat{\beta}_{*,CC}$ from the logistic regression analysis with data from the affected offsprings/sibs and unrelated controls, we can also have the association parameter estimate $\tilde{\beta}_{*,CC}$ from the logistic regression analysis with data from unrelated cases and unrelated controls. When combining the data is suitable, our WLS approach can be easily applied to combine separate information from $\hat{\beta}_{TRIO}/\hat{\beta}_{*,SIB}$, $\hat{\beta}_{*,CC}$, and $\tilde{\beta}_{*,CC}$, and obtain the combined estimator $\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ and its variance $\text{var}(\hat{\beta}) = (X'\Sigma^{-1}X)^{-1}$; here, $Y =$

TABLE II. Power at 5% significance level of the tests for checking the appropriateness of combining the trio and unrelated control data, and the bias of the combined estimates for log genotype relative risk[a]

	Proposed		Epstein et al.	
	Power	Bias	Power	Bias
Scenario I: different populations[b]				
$p_A = 0.15$	0.244	0.132	0.259	0.147
$p_A = 0.25$	0.197	−0.171	0.214	−0.164
$p_A = 0.30$	0.642	−0.310	0.643	−0.300
$p_A = 0.35$	0.916	−0.440	0.912	−0.432
Scenario II: population stratification[c]				
$p_A = (0.1, 0.2)$	0.141	0.105	0.146	0.117
$p_A = (0.1, 0.3)$	0.322	0.169	0.356	0.185
$p_A = (0.1, 0.4)$	0.554	0.219	0.612	0.245
$p_A = (0.1, 0.5)$	0.702	0.285	0.769	0.317

MAF, minor allele frequency.
[a]Based on 1,000 simulations (100 trios and 100 unrelated controls in each).
[b]The populations where trios and unrelated controls are sampled have MAFs 0.2 and p_A, respectively.
[c]Trios and controls are randomly sampled from a population consisting of two strata with disease prevalence (2, 10%) and MAFs given as p_A.

$(\hat{\beta}'_{TRIO}, \hat{\beta}'_{*,CC}, \tilde{\beta}'_{*,CC})'$ or $(\hat{\beta}'_{*,SIB}, \hat{\beta}'_{*,CC}, \tilde{\beta}'_{*,CC})'$ is the vector obtained by stacking the three sets of parameter estimates, $X = [I_p|I_p|I_p]'$ the matrices formed by stacking three p-dimensional identity matrices I_p ($p =$ dimension of β), and Σ the variance-covariance matrix of Y. The association test combining all the data can be performed by the Wald test based on β. Also, the test for checking the appropriateness of combining the family and unrelated data can be based on the Wald-type test statistic which checks the equality of the limiting values of the three estimates $\hat{\beta}_{TRIO}$, $\hat{\beta}_{*,CC}$, and $\tilde{\beta}_{*,CC}$.

In conclusion, we have proposed a simple method for integrating gene-disease association information from the family-based and population-based analyses. It applies in general association studies without any assumptions and estimation on the mating-type distribution.

ACKNOWLEDGMENTS

This research was supported by a Research Grant from the Genomic Research Center, Academia Sinica (94B001-2).

REFERENCES

Allen AS, Satten GA. 2007. Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. Genet Epidemiol 31:211–223.

Allen AS, Satten GA. 2008. Robust estimation and testing of haplotype effects in case-control studies. Genet Epidemiol 32:29–40.

Breslow NE, Day NE. 1980. Statistical Methods in Cancer Research, Volume I—The Analysis of Case-Control Studies. Lyon: IARC Scientific Publications. p 70–71.

Cardon LR, Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet 2:91–99.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141.

Curtis D. 1997. Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. 2005. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 76:592–608.

Falk CT, Rubinstein P. 1987. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233.

Fay MP, Graubard BI, Freedman LS, Midthune DN. 1998. Conditional logistic regression with sandwich estimators: application to a meta analysis. Biometrics 54:195–208.

Kazeem GR, Farrall M. 2005. Integrating case-control and TDT studies. Ann Hum Genet 69:329–335.

Laird NM, Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394.

Lewontin RC. 1988. On measures of gametic disequilibrium. Genetics 120:849–852.

Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Martin ER, Kaplan NL. 2000. A Monte Carlo procedure for two-stage tests with correlated data. Genet Epidemiol 18:48–62.

Mitchell LE. 2000. Relationship between case-control studies and the transmission/disequilibrium test. Genet Epidemiol 19:193–201.

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.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. Biometrika 66:403–411.

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Schaid DJ, Rowland C. 1998. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506.

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.

Seber GA. 1997. Linear Regression Analysis. New York: Wiley. p 61–62.

Self SG, Longton G, Kopecky KJ, Liang KY. 1991. On estimating HLA-disease association with application to a study of aplastic anemia. Biometrics 47:53–61.

Siegmund KD, Langholtz B, Kraft P, Thomas DC. 2000. Testing linkage disequilibrium in sibships. Am J Hum Genet 67:244–248.

Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458.

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.

Stram DO, Pearce L, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Zeger SL, Liang KY, Albert PS. 1988. Models for longitudinal data: a generalized estimating equation approach. Biometrics 44:1049–1060.

Zhao LP, Li SS, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250.

APPENDIX

DERIVATION OF EXPRESSIONS (3) AND (4)

The following derivation is based on well-known results in the linear model theory. Let $Y = (\hat{\beta}'_{TRIO}, \hat{\beta}'_{*,CC})'$ be the vector obtained by stacking $\hat{\beta}_{TRIO}$ and $\beta_{*,CC}$ and $X = [I_p|I_p]'$ the matrix formed by stacking two p-dimensional identity matrices I_p (p is the dimension of β). Note that, when the case-parent trios and unrelated controls are sampled from the same population and the parameters β in (1) and β_* in (2) are essentially equivalent, the $2p \times 1$ vector Y follows asymptotically a $2p$-variate normal distribution with mean $X\beta$ and variance-covariance matrix Σ. Denote the component submatrices of Σ by Σ_{11}, Σ_{12}, Σ_{21}, and Σ_{22}, where Σ_{11} is the asymptotic variance matrix of $\hat{\beta}_{TRIO}$, Σ_{22} is the asymptotic variance matrix of $\hat{\beta}_{*,CC}$, and $\Sigma_{12} = \Sigma'_{21}$ is the asymptotic covariance matrix between $\hat{\beta}_{TRIO}$ and $\hat{\beta}_{*,CC}$. By the linear model theory [Seber, 1997, pp 61–62], the optimal (most efficient) estimator for β based on the linear combination of $\hat{\beta}_{TRIO}$ and $\hat{\beta}_{*,CC}$ is given by the WLS estimator

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y,$$

which, by some matrix algebra, leads to

$$\hat{\beta} = W_1\hat{\beta}_{\mathrm{TRIO}} + W_2\hat{\beta}_{*,\mathrm{CC}},$$

where $W_1 = (\Sigma_{22} - \Sigma_{12}^{'})Q^{-1}$ and $W_2 = (\Sigma_{11} - \Sigma_{12})Q^{-1}$ with $Q = \Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{12}^{'}$. The variance matrix of $\hat{\beta}$ is given by $(X'\Sigma^{-1}X)^{-1}$, and by matrix algebra we have

$$(X'\Sigma^{-1}X)^{-1} = \Sigma_{11} - W_1QW_1^{'} = \Sigma_{22} - W_2QW_2^{'}.$$

ESTIMATORS OF Σ_{jk} WHEN THERE IS ONLY ONE AFFECTED OFFSPRING/SIB IN EACH FAMILY

For $i = 1, \ldots, N_1$, let $S_{\mathrm{CPG},i}$ and $S_{\mathrm{CC},i}$ be the scores of the CPG and logistic regression likelihoods for the ith affected subjects. For $i = N_1 + 1, \ldots, N_1 + N_0$, let $S_{\mathrm{CC},i}$ denote the logistic regression scores for the unrelated subjects. Let I_{CPG} and I_{CC} be the respective information (negative Hessian) matrices from the CPG analysis with trio data and the logistic regression analysis with case-unrelated control data. Note that the score and information matrices for the logistic regression include additional components for the intercept parameter α. Let $\hat{\beta}_{*,\mathrm{CC}} = L\hat{\theta}$, where $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_{*,\mathrm{CC}})'$ and $L = [0|I_p]$, where I_p is the p-dimensional identity matrix. Estimates for $\Sigma_{11} = I_{\mathrm{CPG}}^{-1}$ and $\Sigma_{22} = LI_{\mathrm{CC}}^{-1}L'$ can be obtained from outputs of standard CPG and logistic regression analyses. To estimate Σ_{12} (the covariance between $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\beta}_{*,\mathrm{CC}}$), recall that the covariance between $\hat{\beta}_{\mathrm{TRIO}}$ and $\hat{\theta}$, denoted by $\mathrm{cov}(\hat{\beta}_{\mathrm{TRIO}}, \hat{\theta})$, can be estimated by $I_{\mathrm{CPG}}^{-1}\{\sum_{i=1}^{N_1} S_{\mathrm{CPG},i}S_{\mathrm{CC},i}^{'}\}I_{\mathrm{CC}}^{-1}$, and that $\Sigma_{12} = \mathrm{cov}(\hat{\beta}_{\mathrm{TRIO}}, \hat{\theta})L'$; the estimate of Σ_{12} can thus be obtained as

$$I_{\mathrm{CPG}}^{-1}\left\{\sum_{i=1}^{N_1} S_{\mathrm{CPG},i}S_{\mathrm{CC},i}^{'}\right\}\tilde{I}_{\mathrm{CC}}^{-1}, \qquad \text{(A1)}$$

where $\tilde{I}_{\mathrm{CC}}$ is the information matrix for the logistic regression with the first column (corresponding to the intercept α) dropped

In the case where the case-sibling data are used in place of the case-parent data, Σ_{11} is now obtained from the standard conditional logistic regression analysis with sibship data, and $\Sigma_{12} = \mathrm{cov}(\hat{\beta}_{*,\mathrm{SIB}}, \hat{\beta}_{*,\mathrm{CC}})$ can be obtained as (A1), with the CPG scores and information matrix replaced by those from the conditional logistic regression.

ROBUST ESTIMATORS OF Σ_{jk} WHEN THERE ARE MULTIPLE AFFECTED OFFSPRINGS/SIBS IN A FAMILY

In the case where each family may have multiple affected offsprings/siblings, we use the subscripts $i = 1, \ldots, M$ to index families in the case-parent (case-sibling) sample, and use $i = M + 1, \ldots, M + N_0$ to index unrelated subjects. Suppose that there are n_i affected subjects in family i, $i = 1, \ldots, M$. Recall that in this case the estimator $\hat{\beta}_{\mathrm{TRIO}}$ is a CPG estimator that treats the n_i case-parent trios in the ith family as independent, and the estimator $\hat{\beta}_{*,\mathrm{CC}}$ is a logistic regression parameter estimator that treats the affected offsprings and the unrelated controls as an independent case-control sample. Let $S_{\mathrm{CPG},il}$ and $S_{\mathrm{CC},il}$ be the CPG and logistic regression scores for the lth affected offspring in the ith family, $l = 1, \ldots, n_i$, $i = 1, \ldots, M$, and $S_{\mathrm{CPG},i} = \sum_{l=1}^{n_i} S_{\mathrm{CPG},il}$, $S_{\mathrm{CC},i} = \sum_{l=1}^{n_i} S_{\mathrm{CC},il}$ are the respective CPG and logistic regression scores contributed from the ith family, $i = 1, \ldots, M$. For $i = M + 1, \ldots, M + N_0$, let $S_{\mathrm{CC},i}$ denote the logistic regression scores for the unrelated subjects. Then the robust estimator for Σ_{11} is given as $I_{\mathrm{CPG}}^{-1}\{\sum_{i=1}^{M} S_{\mathrm{CPG},i}S_{\mathrm{CPG},i}^{'}\}I_{\mathrm{CPG}}^{-1}$, and the robust estimator of Σ_{22} is $I_{\mathrm{CC}}^{-1}\{\sum_{i=1}^{M+N_0} S_{\mathrm{CC},i}S_{\mathrm{CC},i}^{'}\}\tilde{I}_{\mathrm{CC}}^{-1}$. The Σ_{12} can be estimated by $I_{\mathrm{CPG}}^{-1}\{\sum_{i=1}^{M} S_{\mathrm{CPG},i}S_{\mathrm{CC},i}^{'}\}\tilde{I}_{\mathrm{CC}}^{-1}$. Here again, the I_{CPG} and I_{CC} are information matrices from the current CPG and logistic regression analyses, respectively, and $\tilde{I}_{\mathrm{CC}}$ is the submatrix of I_{CC} with the first column (corresponding to the intercept parameter α) dropped.

When the case-sibling data are used in place of the case-parent data, the estimates of Σ_{11} and Σ_{12} are obtained as above by replacing I_{CPG} and $S_{\mathrm{CPG},i}$ with the information matrix and the ith family's score from the conditional logistic regression with sibship data [Fay et al., 1998]. An alternative robust estimate for Σ_{11} can be obtained as in Siegmund et al. [2000].