

# Family-Based Association Tests for Different Family Structures Using Pooled DNA

Guohua Zou<sup>1,2</sup> and Hongyu Zhao<sup>1,\*</sup>

<sup>1</sup>*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, USA*

<sup>2</sup>*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China*

---

## Summary

DNA pooling is a cost-effective strategy for genomewide association studies to identify disease genes. In the context of family-based association studies, Risch & Teng (1998) mainly considered families of identical structures to detect associations between genetic markers and disease, and suggested possible approaches to incorporating different family types without a thorough study of their properties. However, families collected in real genetic studies often have different structures and, more importantly, the informativeness of each family structure depends on the disease model which is generally unknown. So there is a need to develop and investigate statistical methods to combine information from diverse family types. In this article, we propose a general strategy to incorporate different family types by assigning each family an “optimal” weight in association tests. In addition, we consider measurement errors in our analysis. When we evaluate our approach under different disease models and measurement errors, we find that our weighting scheme may lead to a substantial reduction in sample size required over the approach suggested by Risch & Teng (1998), and measurement errors may have significant impact on the required sample size when the error rates are not negligible.

---

Keywords: family-based association study; family structure; DNA pooling; measurement error; sample size.

## Introduction

The genome-wide association study is a promising approach to identifying disease genes. However, it is still extremely expensive to genotype hundreds or thousands of individuals at hundreds of thousands marker loci with current technologies. As a result, DNA pooling has received much attention recently due to its potential in saving genotyping cost (Michelmore *et al.* 1991; Lipkin *et al.* 1998; Risch & Teng, 1998; Xu *et al.* 1999; Bader *et al.* 2001; Jawaid *et al.* 2002; Ito *et al.* 2003; Wang *et al.* 2003, among others). Recent developments in quantitative assays and in the design and analysis of pooling studies were reviewed by Sham *et al.* (2002).

For quantitative phenotypes, Bader & Sham (2002) proposed statistical methods to use DNA pooling in family-based association designs. For qualitative phenotypes, Risch & Teng (1998) derived formulae for calculating power to detect associations for identical family structures when DNA pooling is used. Compared to the results of population-based case-control tests under both individual genotyping and DNA pooling (*c.f.*, Zou & Zhao, 2004), their research shows that when families with parents are used pooling leads to higher power, especially as the number of affected children increases, although when case-parent triads are used the power between them is similar. Note that family-based designs are robust to population stratification, and family-based association tests are promising when pooled DNA is used. However, families often have different structures in practical genetic studies, and it is more flexible if a study does not constrain family types. For example, it may not be easy to collect only families with three affected children. More importantly, Risch & Teng's (1998) research

\*Corresponding author: Hongyu Zhao, Ph.D. Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034; Phone: (203) 785-6271; Fax: (203) 785-6912. E-mail: hongyu.zhao@yale.edu.

and our own suggest that the sample sizes required for various family structures depend on the disease model, which is often unknown to researchers. Therefore, there is a need to develop statistical methods to combine families of different types both for practical and design considerations. As pointed out by Risch & Teng (1998), simple pooling of families of different structures, e.g., all affecteds are pooled together and all parents and unaffecteds are pooled together, is not a robust procedure. They proposed two ways to combine families of different structures. The first is to form pools using only families with identical structures, the second is to duplicate individuals for different family types, so that the ratio of the number of affecteds and the number of unaffecteds remains constant. However, they did not investigate the properties nor the power of their proposed methods. In this article, we consider the first method of forming pools using families of the same structures. Through marker score distributions, we first derive formulae for the mean and variance of the test statistic using DNA pooling data from families of identical structures. Our general results cover those of Risch & Teng (1998) as special cases. Based on these results, we propose a weighting scheme to combine data of different family types. When our approach is applied to different disease models we find that it may lead to significant reduction in sample size requirements compared to those through Risch & Teng's approach under certain disease models. We also consider errors in measuring allele frequencies, which are unavoidable for DNA pooling technology. Recent research suggests that for a given DNA pooling sample, the standard deviation of the estimated allele frequency is between 1% and 4% (cf., Buetow *et al.* 2001; Grupe *et al.* 2001; Le Hellard *et al.* 2002, and Sham *et al.* 2002). For example, Le Hellard *et al.* (2002) reported that using the SNaPshot™ Method, which is based on allele-specific extension or minisequencing from a primer adjacent to the site of the SNP, the standard deviations for estimating allele frequency are from 1% to 4% depending on the specific markers being tested. Therefore, we also incorporate measurement errors into our approach. Our numerical results show that the sample size required to attain the desired significance level and power using pooled DNA samples may be seriously affected when the error rates are not negligible.

## Genetic Models and Measurement Error Models

Consider a disease locus with two alleles  $D$  and  $d$ , and a marker locus with two alleles  $A$  and  $a$ . Assume that the penetrance is  $f_2$  for genotype  $DD$ ,  $f_1$  for genotype  $Dd$ , and  $f_0$  for genotype  $dd$ . Denote the frequency of allele  $A$  in the families of structure  $(r, s, a)$  by  $p_{(r,s,a)}$ , where  $r (= 0, 1, 2)$  is the number of available parents in a family,  $s (= 1, 2, \dots)$  is the number of siblings, and  $a (= 1, \dots, s)$  is the number of affected siblings. Let  $q_{(r,s,a)} = 1 - p_{(r,s,a)}$ . In this paper, we assume Hardy-Weinberg equilibrium for parents. If  $r = 1$ , that is, for the family type with only one parent, we further assume random mating between two parents, and the parents are missing at random. This is a reasonable assumption if the parental missingness is not related to the phenotype. For the  $i$ th family of structure  $(r, s, a)$ , where  $i = 1, \dots, n_{(r,s,a)}$  and  $n_{(r,s,a)}$  is the number of families of type  $(r, s, a)$  in the sample, let  $X_{(r,s,a)i}^{(j)}$  be the true (unobservable) number of allele  $A$  at the marker locus for the  $j$ th ( $j = 1, \dots, a$ ) affected sibling,  $Y_{(r,s,a)i}^{(j)}$  be the true (unobservable) number of allele  $A$  for the  $j$ th ( $j = 1, \dots, s - a$ ) unaffected sibling, and  $Z_{(r,s,a)i}^{(j)}$  be the true (unobservable) number of allele  $A$  for the  $j$ th ( $j = 1, \dots, r$ ) parent, who may be the father or mother. Here, to simplify our analysis, we treat the parental phenotypes as unknown. Let  $n$  be the total number of families drawn at random from the ascertainment subpopulation, which consists of all families with at least one affected sibling, then the  $n_{(r,s,a)}$  ( $r = 0, 1, 2; s = 1, 2, \dots; a = 1, \dots, s$ ) are random variables that satisfy  $\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s n_{(r,s,a)} = n$ . For each family type we form two pools, one consisting of affected siblings, and the other consisting of unaffected siblings and available parents. Here, we discard the families in which all children are affected and no parents are available, because the control group cannot be formed for such families.

For a genetic model, denote the mean and variance of  $X_{(r,s,a)i}^{(j)}$  by  $\mu_{(r,s,a)X}$  and  $\sigma_{(r,s,a)X}^2$ , and the covariance and mixed second moment of  $X_{(r,s,a)i}^{(j)}$  and  $X_{(r,s,a)i}^{(k)}$  ( $j \neq k$ ) by  $\gamma_{(r,s,a)XX}$  and  $\Delta_{(r,s,a)XX}$ , respectively. Other notations are defined similarly. The formulae for calculating these means, variances, covariances and mixed second moments are provided in Appendix A. To consider measurement errors, we assume the following measurement

error models

$$\begin{cases} \widehat{p}_{(r,s,a)A} = \frac{\sum_{i=1}^{n_{(r,s,a)}} \sum_{j=1}^a X_{(r,s,a)i}^{(j)}}{2n_{(r,s,a)}a} + \xi_{(r,s,a)}, \\ \widehat{p}_{(r,s,a)U} = \frac{\sum_{i=1}^{n_{(r,s,a)}} \left[ \sum_{j=1}^{s-a} Y_{(r,s,a)i}^{(j)} + \sum_{j=1}^r Z_{(r,s,a)i}^{(j)} \right]}{2n_{(r,s,a)}(s-a+r)} \\ \quad + \eta_{(r,s,a)}, \end{cases} \quad (1)$$

where  $\widehat{p}_{(r,s,a)A}$  is the sample frequency of allele  $A$  among the affected siblings of family type  $(r, s, a)$ , and  $\widehat{p}_{(r,s,a)U}$  is the sample frequency of allele  $A$  among the unaffected siblings and available parents of family type  $(r, s, a)$ . Given  $X_{(r,s,a)i}^{(j)}$ ,  $Y_{(r,s,a)i}^{(j)}$  and  $Z_{(r,s,a)i}^{(j)}$ ,  $\xi_{(r,s,a)}$  and  $\eta_{(r,s,a)}$  are independent normal random variables with mean 0 and variance  $\varepsilon^2$ . Here we assume that for the DNA pooling technology, standard deviation  $\varepsilon$  is not affected by family structures or true allele proportions. However,  $\varepsilon$  may be related to these factors in practice. In this case, we need to replace  $\varepsilon$  by  $\varepsilon_{(r,s,a)}$  for the family type  $(r, s, a)$  in the formulae below.

### Statistical Tests Combining Families of Different Structures

To test the null hypothesis of  $H_0$ : no association between the marker and disease, we form our pools using families with identical structures. To simplify our presentation, we consider the case of perfect linkage disequilibrium between the marker locus and disease locus, i.e., the two loci are identical. More general situations are discussed in the Discussion Section. As in Risch & Teng (1998), a one-sided test will be used. Consider the following general weighting scheme combining information from various family structures:

$$T = \sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s w_{(r,s,a)} (\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}),$$

where the  $w_{(r,s,a)}$  are weights to be discussed below. It can be seen that under the null hypothesis  $H_0$ ,  $\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}$  has mean 0, and variance

$$V(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}) = \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot p_{(r,s,a)} q_{(r,s,a)}$$

$$\begin{aligned} & \times \left[ \frac{1}{a} + \frac{s-a+2r-r^2}{(s-a+r)^2} \right] + 2\varepsilon^2 \\ & \equiv \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \Delta_{(r,s,a)}^0 + 2\varepsilon^2 \\ & \equiv \sigma_{(r,s,a)0}^2, \end{aligned}$$

which can be estimated by

$$\widehat{\sigma}_{(r,s,a)0}^2 = \frac{\widehat{p}_{(r,s,a)}(1 - \widehat{p}_{(r,s,a)})}{4n_{(r,s,a)}} \left[ \frac{1}{a} + \frac{s-a+2r-r^2}{(s-a+r)^2} \right] + 2\varepsilon^2,$$

where

$$\widehat{p}_{(r,s,a)} = \frac{a\widehat{p}_{(r,s,a)A} + (s-a+r)\widehat{p}_{(r,s,a)U}}{s+r}$$

is the sample frequency of allele  $A$  for family type  $(r, s, a)$ , and  $E_1$  denotes the expectation over all possible values of  $n_{(r,s,a)}$ . Other estimation methods of the frequency of allele  $A$  for family type  $(r, s, a)$  under  $H_0$  are possible (c.f., Risch & Teng, 1998). So under  $H_0$ ,

$$V(T) = \sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s w_{(r,s,a)}^2 \sigma_{(r,s,a)0}^2.$$

The optimal value of  $w_{(r,s,a)}$  in the sense of minimizing  $V(T)$  is given by

$$w_{(r,s,a)} = \frac{1/\sigma_{(r,s,a)0}^2}{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s 1/\sigma_{(r,s,a)0}^2}.$$

That is, the weights should be inversely proportional to the corresponding variances. The minimal variance is given by

$$V(T) = \frac{1}{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s 1/\sigma_{(r,s,a)0}^2}.$$

Therefore, we propose the following test statistic for  $H_0$ :

$$t = \frac{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \frac{1}{\widehat{\sigma}_{(r,s,a)0}^2} (\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U})}{\sqrt{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \frac{1}{\widehat{\sigma}_{(r,s,a)0}^2}}}.$$

Using a one-sided test and assuming asymptotic normality (a proof on the asymptotic normality of the test statistic  $t$  under  $H_0$  is provided in Appendix B), the power to reject the null hypothesis with significance level  $\alpha$  can be approximated by

$$\Phi \left( \frac{-z_{\alpha} \sqrt{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s 1/\widehat{\sigma}_{(r,s,a)0}^2} + \sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \mu_{(r,s,a)}/\widehat{\sigma}_{(r,s,a)0}^2}{\sqrt{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \sigma_{(r,s,a)}^2/\widehat{\sigma}_{(r,s,a)0}^4}} \right),$$

where  $\mu_{(r,s,a)}$  and  $\sigma_{(r,s,a)}^2$  are the mean and variance of the difference between the two allele frequency estimates

$$n = \left[ \frac{z_\alpha \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{\lambda_{(r,s,a)}}{\Delta_{(r,s,a)}}} - z_{1-\beta} \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{\lambda_{(r,s,a)} \Delta_{(r,s,a)}^*}{\Delta_{(r,s,a)}^2}}}{2 \sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{\lambda_{(r,s,a)} \mu_{(r,s,a)}}{\Delta_{(r,s,a)}}} \right]^2,$$

$\widehat{p}_{(r,s,a)A}$  and  $\widehat{p}_{(r,s,a)U}$  under the alternative hypothesis  $H_1$ , respectively, whose expressions are given by (A.22) and (A.24) in Appendix A, and

$$\begin{aligned} \widetilde{\sigma}_{(r,s,a)0}^2 &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \frac{\widetilde{p}_{(r,s,a)}(1 - \widetilde{p}_{(r,s,a)})}{4} \\ &\quad \times \left[ \frac{1}{a} + \frac{s - a + 2r - r^2}{(s - a + r)^2} \right] + 2\varepsilon^2 \\ &\equiv \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \widetilde{\Delta}_{(r,s,a)} + 2\varepsilon^2, \end{aligned}$$

with

$$\widetilde{p}_{(r,s,a)} = \frac{a}{s+r} \mu_{(r,s,a)} + \frac{(s-a)\mu_{(r,s,a)Y} + r\mu_{(r,s,a)Z}}{2(s-a+r)}$$

being the expected frequency of allele  $A$  in family type  $(r, s, a)$  under  $H_1$ ,  $\Phi$  is the cumulative standard normal distribution function, and  $z_\alpha$  the upper  $100\alpha$  percentile of the standard normal distribution. If the penetrances are low, then  $\widetilde{p}_{(r,s,a)}$  is simplified to

$$\Phi \left( \frac{-z_\alpha \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)}^2 \widetilde{\sigma}_{(r,s,a)0}^2} + \sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a \lambda_{(r,s,a)} \mu_{(r,s,a)}}{\sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)}^2 \sigma_{(r,s,a)}^2}} \right).$$

$$\widetilde{p}_{(r,s,a)} = \frac{a}{s+r} \mu_{(r,s,a)} + \sum_{u,v} \frac{u+v}{4} m_{uv}^{(r,s,a)},$$

where  $m_{uv}^{(r,s,a)}$  is the conditional probability of the mating type of parents,  $G = (u, v)$ , given the family type being  $(r, s, a)$ . The sample size necessary to obtain a power of  $1 - \beta$  with a significance level of  $\alpha$  satisfies

$$\begin{aligned} z_\alpha \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{1}{\widetilde{\sigma}_{(r,s,a)0}^2}} - \sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{\mu_{(r,s,a)}}{\widetilde{\sigma}_{(r,s,a)0}^2} \\ = z_{1-\beta} \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s \frac{\sigma_{(r,s,a)}^2}{\widetilde{\sigma}_{(r,s,a)0}^4}}. \end{aligned} \quad (2)$$

In particular, when there is no measurement error, i.e.,  $\varepsilon = 0$ , we have

where  $\Delta_{(r,s,a)}^*$  is given in (A.24),  $\lambda_{(r,s,a)}$  is the proportion of families with type  $(r, s, a)$  in the ascertainment subpopulation, and we have used the first order approximation of  $E_1[\frac{1}{n_{(r,s,a)}}]$  (see (A.25) and (A.26) in Appendix A). Note that the resulting sample size  $n$  has included uninformative families, i.e. those families with type  $(0, a, a)$ .

If we use the weights suggested by Risch & Teng (1998), i.e.  $w_{(r,s,a)}$  is proportional to  $an_{(r,s,a)}$ , then the test statistic for  $H_0$  is

$$t^{RT} = \frac{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s an_{(r,s,a)} (\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U})}{\sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 n_{(r,s,a)}^2 \widehat{\sigma}_{(r,s,a)0}^2}}.$$

The corresponding power to reject the null hypothesis with significance level  $\alpha$  is given by

The sample size necessary to obtain a power of  $1 - \beta$  with a significance level of  $\alpha$  satisfies

$$\begin{aligned} z_\alpha \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)}^2 \widetilde{\sigma}_{(r,s,a)0}^2} \\ - \sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a \lambda_{(r,s,a)} \mu_{(r,s,a)} \\ = z_{1-\beta} \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)}^2 \sigma_{(r,s,a)}^2}. \end{aligned} \quad (3)$$

For the case of  $\varepsilon = 0$ , the sample size required is

$$n = \left[ \frac{z_\alpha \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)} \widetilde{\Delta}_{(r,s,a)}} - z_{1-\beta} \sqrt{\sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a^2 \lambda_{(r,s,a)} \Delta_{(r,s,a)}^*}}{2 \sum_{r=0}^2 \sum_{s=1}^\infty \sum_{a=1}^s a \lambda_{(r,s,a)} \mu_{(r,s,a)}} \right]^2.$$

### Numerical Results and Simulation Study

Now we consider an example given by Risch & Teng (1998) to (i) compare the sample sizes required to detect association under our weighting design and the weighting scheme suggested by Risch & Teng (1998); (ii) illustrate the impact of measurement errors on sample size; and (iii) compare the sample sizes required to detect association only using families of the same structures and combining different family structures. In this regard, we should note that because combining different family types needs more pools, our method is slightly more expensive.

We consider two types of family structures: (a) (0, 3, 2), i.e., two affected and one unaffected children, and no parents; and (b) (0, 3, 1), i.e., one affected and two unaffected children, and no parents. From formulae (2) and (3) we calculate the sample size necessary to attain the significance level of  $\alpha = 5 \times 10^{-8}$  and power of  $1 - \beta = 80\%$ , the levels suggested by Risch & Merikangas (1996) for a genome scan, under various genetic models and measurement errors. The results are presented in Table 2 for low penetrances and Table 3 for high penetrances. Note that the sample sizes provided in the tables are the number of families required. It should be mentioned that the sample sizes obtained in our calculations do not include uninformative families with the structure (0, a, a) because there are no such families in the population we considered. Based on the results in these tables, we can see that (i) For low penetrances, the sample sizes required under our weighting

**Table 1** Conditional probability  $m_{uv}^{(s,a)}$  of mating type given a affected and s - a unaffected children

Mating type	population frequency	$m_{uv}^{(s,a)}$
(2,2)	$g_{22}$	$f_2^a (1 - f_2)^{s-a} g_{22} / K_{s,a}$
(2,1)	$g_{21}$	$\left(\frac{f_2+f_1}{2}\right)^a \left(1 - \frac{f_2+f_1}{2}\right)^{s-a} g_{21} / K_{s,a}$
(2,0)	$g_{20}$	$f_1^a (1 - f_1)^{s-a} g_{20} / K_{s,a}$
(1,2)	$g_{12}$	$\left(\frac{f_2+f_1}{2}\right)^a \left(1 - \frac{f_2+f_1}{2}\right)^{s-a} g_{12} / K_{s,a}$
(1,1)	$g_{11}$	$\left(\frac{f_2+2f_1+f_0}{4}\right)^a \left(1 - \frac{f_2+2f_1+f_0}{4}\right)^{s-a} g_{11} / K_{s,a}$
(1,0)	$g_{10}$	$\left(\frac{f_1+f_0}{2}\right)^a \left(1 - \frac{f_1+f_0}{2}\right)^{s-a} g_{10} / K_{s,a}$
(0,2)	$g_{02}$	$f_1^a (1 - f_1)^{s-a} g_{02} / K_{s,a}$
(0,1)	$g_{01}$	$\left(\frac{f_1+f_0}{2}\right)^a \left(1 - \frac{f_1+f_0}{2}\right)^{s-a} g_{01} / K_{s,a}$
(0,0)	$g_{00}$	$f_0^a (1 - f_0)^{s-a} g_{00} / K_{s,a}$

\*  $K_{s,a}$  is the sum of all numerators in the third column.

**Table 2** Sample size required to detect genetic associations combining different family structures for low penetrances\*

	$\epsilon = 0$	$\epsilon = 0.005$	$\epsilon = 0.01$
Dominant			
$p = 0.05$	532(533)	1702(1701)	$\infty^{**}(\infty)$
$p = 0.20$	355(357)	456(457)	2751(2969)
$p = 0.70$	4286(4306)	$\infty(\infty)$	$\infty(\infty)$
Recessive			
$p = 0.05$	59117(59092)	$\infty(\infty)$	$\infty(\infty)$
$p = 0.20$	1494(1495)	45486( $\infty$ )	$\infty(\infty)$
$p = 0.70$	270(270)	353(353)	3312(4511)
Multiplic.			
$p = 0.05$	2026(2030)	$\infty(\infty)$	$\infty(\infty)$
$p = 0.20$	653(654)	1135(1135)	$\infty(\infty)$
$p = 0.70$	639(640)	1256(1255)	$\infty(\infty)$
Additive			
$p = 0.05$	1209(1212)	$\infty(\infty)$	$\infty(\infty)$
$p = 0.20$	524(525)	788(789)	97207( $\infty$ )
$p = 0.70$	984(986)	3547(3571)	$\infty(\infty)$

\*The values in brackets are based on the weighting scheme suggested by Risch & Teng (1998); \*\* $\infty$  means that 80% power cannot be attained or the sample size required is unrealistically large (greater than 100 000); Significance level  $\alpha = 5 \times 10^{-8}$ ; power  $1 - \beta = 0.80$ ; Dominant model:  $f_2 = f_1 = 0.004, f_0 = 0.001$ ; Recessive model:  $f_2 = 0.004, f_1 = f_0 = 0.001$ ; Multiplicative model:  $f_2 = 0.004, f_1 = 0.002, f_0 = 0.001$ ; Additive model:  $f_2 = 0.004, f_1 = 0.0025, f_0 = 0.001$ .

scheme and that suggested by Risch & Teng (1998) are almost the same; both are close to the case of using only families of type (0, 3, 1). This can be easily understood by noting that for the low penetrances, there are much more families with structure (0, 3, 1) than those with structure (0, 3, 2). For high penetrances, the sample sizes required under our weighting scheme are generally smaller than those under that of Risch & Teng (1998). The difference is largest for dominant models, and smallest for recessive models. It can also be observed that for the recessive model and not large allele frequencies, the weighting method of Risch & Teng is even slightly better, although the difference is small (the largest relative difference is about 5%). This is not surprising because our weighting scheme will not necessarily result in a uniformly optimal power. (ii) The impact of measurement errors on sample size is generally large. Relative to the case of low penetrances, for high penetrances the impact is not very large when the error rates are small and the allele frequencies are not small. But the impact can be substantial for moderate error rates ( $\epsilon = 0.01$ ) or

**Table 3** Sample size required to detect genetic associations combining different family structures for high penetrances\*

	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$
<b>Dominant</b>			
$p = 0.05$	338(373)	537(500)	2124( $\infty^{**}$ )
$p = 0.20$	185(217)	202(232)	269(290)
$p = 0.70$	1746(2104)	4663(5890)	$\infty$ ( $\infty$ )
<b>Recessive</b>			
$p = 0.05$	48005(45598)	$\infty$ ( $\infty$ )	$\infty$ ( $\infty$ )
$p = 0.20$	1139(1126)	2238(2121)	$\infty$ ( $\infty$ )
$p = 0.70$	155(159)	167(170)	215(216)
<b>Multiplic.</b>			
$p = 0.05$	1513(1669)	$\infty$ ( $\infty$ )	$\infty$ ( $\infty$ )
$p = 0.20$	443(486)	564(587)	1458(1568)
$p = 0.70$	333(358)	385(409)	704(724)
<b>Additive</b>			
$p = 0.05$	869(967)	4448(9965)	$\infty$ ( $\infty$ )
$p = 0.20$	335(376)	398(430)	762(752)
$p = 0.70$	485(534)	598(650)	1804(1867)

\*The values in brackets are based on the weighting scheme suggested by Risch & Teng (1998); \*\* $\infty$  means that 80% power cannot be attained or the sample size required is unrealistically large (greater than 100 000); Significance level  $\alpha = 5 \times 10^{-8}$ ; power  $1 - \beta = 0.80$ ; Dominant model:  $f_2 = f_1 = 0.4, f_0 = 0.1$ ; Recessive model:  $f_2 = 0.4, f_1 = f_0 = 0.1$ ; Multiplicative model:  $f_2 = 0.4, f_1 = 0.2, f_0 = 0.1$ ; Additive model:  $f_2 = 0.4, f_1 = 0.25, f_0 = 0.1$ .

small allele frequencies, especially for low penetrances. In these cases, there is a dramatic increase in sample sizes.

To compare the sample sizes required by the design combining different family structures, and by the design only using families of the same structures, we further calculate the sample sizes by using the families with types (0, 3, 2) and (0, 3, 1) for the case of no measurement errors, respectively. The results for high penetrances are provided in Table 4. It is clear from this table that the sample sizes required for incorporating different family structures are between those separately using families with type (0, 3, 2) or (0, 3, 1). One family structure is not always preferable over the other, and the relative information for disease association depends on specific disease models. Similar conclusions can be drawn for low penetrances. Therefore, there is added benefit in incorporating various family structures when the mode of inheritance is unknown.

We conduct some simulation studies to confirm our large sample results. We first generate the genotypes of parents assuming Hardy-Weinberg Equilibrium, and the genotypes of three children assuming Mendelian trans-

**Table 4** Sample sizes required to detect genetic associations combining different family structures and using only the same family structures for high penetrances\* and no measurement errors

	Using both (0, 3, 2) and (0, 3, 1)	Using only (0, 3, 2)	Using only (0, 3, 1)
<b>Dominant</b>			
$p = 0.05$	338(373)	208	355
$p = 0.20$	185(217)	187	184
$p = 0.70$	1746(2104)	2539	1417
<b>Recessive</b>			
$p = 0.05$	48005(45598)	17237	55431
$p = 0.20$	1139(1126)	534	1275
$p = 0.70$	155(159)	150	152
<b>Multiplic.</b>			
$p = 0.05$	1513(1669)	1116	1556
$p = 0.20$	443(486)	359	459
$p = 0.70$	333(358)	348	322
<b>Additive</b>			
$p = 0.05$	869(967)	615	898
$p = 0.20$	335(376)	299	342
$p = 0.70$	485(534)	534	459

\*The values in brackets are based on the weighting scheme suggested by Risch & Teng (1998); Significance level  $\alpha = 5 \times 10^{-8}$ ; power  $1 - \beta = 0.80$ ; Dominant model:  $f_2 = f_1 = 0.4, f_0 = 0.1$ ; Recessive model:  $f_2 = 0.4, f_1 = f_0 = 0.1$ ; Multiplicative model:  $f_2 = 0.4, f_1 = 0.2, f_0 = 0.1$ ; Additive model:  $f_2 = 0.4, f_1 = 0.25, f_0 = 0.1$ .

mission. Using the penetrances  $f_2, f_1$  and  $f_0$  we simulate the disease status of each child. For a given sample size we confine ourselves to the families with one or two affected children. The test statistic  $t$  is used to calculate the empirical type I error rate and power. Note that to see whether our method leads to a correct type I error rate, a very large number of simulations is needed to consider  $\alpha = 5 \times 10^{-8}$ . So we consider the nominal significance level of  $\alpha = 0.05$  instead. The empirical type I error rate is the proportion of significant replicates out of the total number of replicates under  $H_0$ . By making use of the sample sizes suggested by the asymptotic power approximations (when the sample size suggested is  $\infty$ , we do not report power), we can calculate the empirical power and hence check whether a power of 80% can be attained. The empirical power is the proportion of significant replicates out of the total number of replicates under  $H_1$ . Based on 500 replicates (100 replicates for the case of recessive model and very low allele frequency,  $p = 0.05$ ) our results are summarized in Table 5 for empirical type I error rate and in Table 6 for empirical power. It can be seen that the empirical

**Table 5** Empirical type I error rate for prevalence of 0.1\*

	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$
$p = 0.05$	0.064	0.049	0.057
$p = 0.20$	0.048	0.042	0.044
$p = 0.70$	0.060	0.070	0.068

\*The critical value is 1.6449 (which corresponds to the significance level of 0.05 under normality), and the sample size is 200.

**Table 6** Empirical power using sample sizes obtained through asymptotic approximation for high penetrances\*

	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$
Dominant			
$p = 0.05$	0.840	0.842	0.866
$p = 0.20$	0.840	0.846	0.840
$p = 0.70$	0.896	0.896	
Recessive			
$p = 0.05$	0.780		
$p = 0.20$	0.808	0.702	
$p = 0.70$	0.796	0.836	0.804
Multiplic.			
$p = 0.05$	0.872		
$p = 0.20$	0.796	0.830	0.852
$p = 0.70$	0.790	0.818	0.814
Additive			
$p = 0.05$	0.774	0.796	
$p = 0.20$	0.800	0.764	0.890
$p = 0.70$	0.780	0.730	0.816

\*The critical value is 5.3267 (which corresponds to the significance level of  $5 \times 10^{-8}$  under normality); Dominant model:  $f_2 = f_1 = 0.4, f_0 = 0.1$ ; Recessive model:  $f_2 = 0.4, f_1 = f_0 = 0.1$ ; Multiplicative model:  $f_2 = 0.4, f_1 = 0.2, f_0 = 0.1$ ; Additive model:  $f_2 = 0.4, f_1 = 0.25, f_0 = 0.1$ .

type I error rates and empirical powers are generally close to the significance level of 0.05 and power of 80%, respectively.

## Discussion

In this article, we have developed a general weighting scheme to combine families of different structures in the detection of genetic associations using DNA pooling through family-based association designs. In addition, we explicitly modelled the measurement errors in our approach. It is observed that our weighting scheme is usually better than that suggested by Risch & Teng (1998). In the example we considered, where the families have two different types of structures, the efficiency of the design combining families of different structures

is always between those of the designs only using families with one of the two structures. However, because it is generally much easier to collect families of different structures in practice and, more importantly, the informativeness of each family structure depends on the disease model, which is often unknown, we advocate the use of a study design that maximizes the usage of available family data. We also studied the impact of measurement errors on the sample size required. Our numerical results showed that, similar to the case of pooled population data (Zou & Zhao, 2004), the sample size required to attain a desired significance level and power using pooled DNA may significantly increase as the measurement errors increase for family-based association tests. However, such impact can be reduced if multiple replicates of each pooled sample are measured. For example, if four replicate measurements are used and accordingly,  $\hat{p}_{(r,s,a)A}$  and  $\hat{p}_{(r,s,a)U}$  are replaced by the averages of these four measurements, then the standard deviation  $\varepsilon$  will be reduced by half,  $\varepsilon/2$ , and all formulae in the paper are still true. Thus, if  $\varepsilon = 0.01$ , then the standard deviation after four replicate measurements will be 0.005. From Tables 2 and 3 we see that the sample sizes required are greatly reduced. As a result of measurement errors it is possible that a specified power, e.g. 80%, may never be achieved under certain disease models (see results in Tables 2 and 3). Therefore, our analysis emphasizes the importance of reducing measurement errors in DNA pooling studies. Note that in our discussion the standard deviation  $\varepsilon$  is assumed to be known. If  $\varepsilon$  is unknown then we can infer it from laboratory experiments or from the distributions of the test statistics (Jawaid *et al.* 2002). Although a precise value of  $\varepsilon$  is impossible, our findings based on asymptotic results and simulation studies in Section 4 suggest that for high penetrances the effect of a minor misspecification of  $\varepsilon$  (for example, the estimate of  $\varepsilon$  is 0.005 but  $\varepsilon = 0.0075$  in reality) on the association test is not very large. Relatively speaking the effect is slightly larger for low allele frequencies. However, for low penetrances such an effect can be large (data not shown).

It should be pointed out that we have assumed that the parental phenotypes are unknown in order to simplify our analysis. If the parental disease prevalence is low, then our results are close to the case of unaffected parents. Generally we can use separate pools for

families with affected parents and for families with unaffected parents. More precisely we can consider the following family types separately: two affected parents, two unaffected parents, and one affected parent and one unaffected parent for the families with two parents; one affected parent, and one unaffected parent for the families with only one parent; and the families with no parents. Such consideration should provide additional information. The analytical details can be given along the line devised here. However, this will be more complicated and remains to be studied in our future work.

In this discussion we have assumed perfect linkage disequilibrium between the disease locus and marker locus. However, it is more likely that the marker being examined is in incomplete linkage disequilibrium with the genetic variant of interest. In this case it is necessary to derive the penetrances of the genotypes at a marker locus for each family structure  $(r, s, a)$ . Then all the formulae obtained previously can be used. Risch & Teng's (1998) results can serve this purpose, although the problem will be more difficult if several markers are considered together. In fact, let  $p_{(r,s,a)}$  and  $q_{(r,s,a)}$  be the frequencies of alleles  $D$  and  $d$  at the disease locus, and  $f_2, f_1$  and  $f_0$  still be the penetrances of genotypes  $DD, Dd$  and  $dd$ , respectively. Further, let  $p'_{(r,s,a)}$  and  $q'_{(r,s,a)}$  be the frequencies of alleles  $A$  and  $a$  at the marker locus, and  $f'_2, f'_1$  and  $f'_0$  be the penetrances of genotypes  $AA, Aa$  and  $aa$ , respectively. If we use Bengtsson & Thomson's (1981) definition of the linkage disequilibrium parameter  $\delta$ :

$$\delta = \frac{P(A|D) - P(A)}{1 - P(A)},$$

then from Risch & Teng (1998) we have

$$P(D|A) = p_{(r,s,a)} + \frac{p_{(r,s,a)}q'_{(r,s,a)}}{p'_{(r,s,a)}}\delta_{(r,s,a)},$$

$$P(d|A) = q_{(r,s,a)} - \frac{p_{(r,s,a)}q'_{(r,s,a)}}{p'_{(r,s,a)}}\delta_{(r,s,a)},$$

$$P(D|a) = p_{(r,s,a)} - p_{(r,s,a)}\delta_{(r,s,a)},$$

and

$$P(d|a) = q_{(r,s,a)} + p_{(r,s,a)}\delta_{(r,s,a)},$$

where  $\delta_{(r,s,a)}$  is the linkage disequilibrium measure in the family structure  $(r, s, a)$ . Therefore,

$$f'_2 = f_2 P^2(D|A) + 2f_1 P(D|A)P(d|A) + f_0 P^2(d|A),$$

$$f'_1 = f_2 P(D|A)P(D|a) + f_1 [P(D|A)P(d|a) + P(d|A)P(D|a)] + f_0 P(d|A)P(d|a),$$

and

$$f'_0 = f_2 P^2(D|a) + 2f_1 P(D|a)P(d|a) + f_0 P^2(d|a).$$

Note that  $f'_2, f'_1$  and  $f'_0$  may be dependent on the family structures. But the formulae in this paper can still be used for this case as long as we substitute them for  $f_2, f_1$  and  $f_0$ , respectively.

In this paper we have assumed the random missingness of parental genotypes so that the available and missing parents have the same marker score distributions. This is plausible if the parental missingness is not related to the phenotype. For example, the random missingness assumption holds if we are unable to locate the parents because of death from some other disease or accident. However, in the situation where the missingness of a parent is related to the phenotype, this assumption may not be reasonable. For instance, in a study of genetic factors in an aggressive form of cancer, it is more likely that parents carrying the disease-predisposing allele are missing. A detailed discussion can be found in Allen *et al.* (2003). The construction of appropriate test statistics under this scenario warrants further research.

### Acknowledgments

This work was supported in part by grant GM59507 from the National Institutes of Health and by grant No. 70221001 from the National Natural Science Foundation of China. The authors thank two reviewers for their helpful comments.

### References

- Allen, A., Rathouz, P. & Satten, G. (2003) Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet* **72**, 671–680.
- Bader, J., Bansal, A. & Sham, P. (2001) Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *GeneScreen* **1**, 143–150.
- Bader, J. & Sham, P. (2002) Family-based association tests for quantitative traits using pooled DNA. *Eur J Hum Genet* **10**, 870–878.
- Bengtsson, B. O. & Thomson, G. (1981) Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* **18**, 356–363.

Buetow, K. H., Edmonson, M., MacDonald, R., Clifford, P., Yip, P., Kelley, J., Little, D. P., Strausberg, R., Koester, H., Cantor, C. R. & Braun, A. (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci USA* **98**, 581–584.

Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J. K., Klein, R. F., Ahluwalia, M. K., Higuchi, R. & Peltz, G. (2001) In silico mapping of complex disease-related traits in mice. *Science* **292**, 1915–1918.

Ito, T., Chiku, S., Inoue, E., Tomita, M., Morisaki, T., Morisaki, H. & Kamatani, N. (2003) Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am J Hum Genet* **72**, 384–398.

Jawaid, A., Bader, J., Purcell, S., Cherny, S. & Sham, P. (2002) Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Genet* **10**, 125–132.

Le, Hellard S., Ballereau, S. J., Visscher, P. M., Torrance, H. S., Pinson, J., Morris, S. W., Thomson, M. L., Semple, C. A., Muir, W. J., Blackwood, D. H., Porteous, D. J. & Evans, K. L. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res* **30** e74.

Lipkin, E., Mosig, M. O., Darvasi, A., Ezra, E., Shalom, A., Friedmann, A. & Soller, M. (1998) Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics* **149**, 1557–1567.

Michelmore, R., Paran, I. & Kesseli, R. (1991) Identification of marker linked to disease resistance gene by bulk segregant analysis: a rapid method to detect markers in specific genomic regions using segregating populations. *Proc Natl Acad Sci USA* **88**, 9828–9832.

Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Risch, N. & Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. *Genome Res* **8**, 1273–1288.

Sham, P., Bader, J., Craig, I., O'Donovan, M. & Owen, M. (2002) DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics* **3**, 862–871.

Stephan, F. F. (1945) The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate. *Ann Math Statist* **16**, 50–61.

Xu, C., Donnelly, C., Montgomery, D., Allan, C. & Purvis, I. (1999) Determination of SNP allele frequency by a DNA pooling method. *Am J Hum Genet*, **65**, 2577.

Wang, S., Kidd, K. K. & Zhao, H. (2003) On the use of DNA pooling to estimate haplotype frequencies. *Genet Epidemiol* **24**, 74–82.

Zou, G. & Zhao, H. (2004) The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet Epidemiol* **26**, 1–10.

## Appendix A

In this appendix, we first derive the marginal distributions and joint distributions of the marker scores  $X_{(r,s,a)i}^{(j)}$  for the affected siblings,  $Y_{(r,s,a)i}^{(j)}$  for the unaffected siblings, and  $Z_{(r,s,a)i}^{(j)}$  for the parents whose phenotypes are assumed to be unknown, then their means, variances and covariances, and finally the mean and variance of the difference between the two allele frequency estimates  $\hat{p}_{(r,s,a)A}$  and  $\hat{p}_{(r,s,a)U}$  for the families with identical structures under the null hypothesis  $H_0$  and alternative hypothesis  $H_1$ . We give only the results under  $H_1$  as the distributions under  $H_0$  can be obtained by replacing the penetrances  $f_2, f_1$  and  $f_0$  by the disease prevalence for each family type  $(r, s, a)$ .

### Marginal marker score distributions

Let  $G = (u, v)$  be the mating type of parents and  $m_{uv}^{(r,s,a)}$  be the conditional probability of  $G = (u, v)$  given the family type being  $(r, s, a)$ . When the parents are missing at random, the values of  $m_{uv}^{(r,s,a)}$  are given in Table 1 and this reduces to Table 1 of Risch & Teng (1998) if the penetrances are low so that the unaffected individuals can be regarded as having unknown phenotypes. Denote  $m_{(uv)}^{(r,s,a)} = m_{uv}^{(r,s,a)} + m_{vu}^{(r,s,a)}$  when  $u \neq v$ . Then the distribution of  $X_{(r,s,a)i}^{(j)}$  is

$$\left\{ \begin{array}{l} P\left(X_{(r,s,a)i}^{(j)} = 2\right) = m_{22}^{(r,s,a)} + \frac{f_2}{f_2 + f_1} m_{(21)}^{(r,s,a)} + \frac{f_2}{f_2 + 2f_1 + f_0} m_{11}^{(r,s,a)} \equiv \alpha_{(r,s,a)X(2)}, \\ P\left(X_{(r,s,a)i}^{(j)} = 1\right) = \frac{f_1}{f_2 + f_1} m_{(21)}^{(r,s,a)} + m_{(20)}^{(r,s,a)} + \frac{2f_1}{f_2 + 2f_1 + f_0} m_{11}^{(r,s,a)} + \frac{f_1}{f_1 + f_0} m_{(10)}^{(r,s,a)} \\ \equiv \alpha_{(r,s,a)X(1)}, \\ P\left(X_{(r,s,a)i}^{(j)} = 0\right) = \frac{f_0}{f_2 + 2f_1 + f_0} m_{11}^{(r,s,a)} + \frac{f_0}{f_1 + f_0} m_{(10)}^{(r,s,a)} + m_{00}^{(r,s,a)} \equiv \alpha_{(r,s,a)X(0)}. \end{array} \right. \quad (A.1)$$

The distribution of  $Y_{(r,s,a)i}^{(j)}$  can be obtained by replacing  $f_w$  by  $1 - f_w$  in formula (A.1), where  $w = 0, 1$  and  $2$  and the probabilities are denoted by  $\alpha_{(r,s,a)Y}(w')$ , where  $w' = 0, 1$  and  $2$ . In the following discussion,  $u, v, w$ , and  $w'$  always take a value of  $0, 1$ , or  $2$ .

Denote the numbers of allele  $A$  of the father and mother in the  $i$ th family of structure  $(r, s, a)$  by  $Z_{(r,s,a)i}^{(f)}$  and  $Z_{(r,s,a)i}^{(m)}$ , respectively. Note that the notation  $Z_{(r,s,a)i}^{(1)}$  in the previous sections is not necessarily the same as  $Z_{(r,s,a)i}^{(f)}$  and can be equal to  $Z_{(r,s,a)i}^{(m)}$  depending on the observed results. The distribution of marker scores for the father is given by

$$\left\{ \begin{aligned} P\left(Z_{(r,s,a)i}^{(f)} = 2\right) &= m_{22}^{(r,s,a)} + m_{21}^{(r,s,a)} + m_{20}^{(r,s,a)} \\ &\equiv \alpha_{(r,s,a)Z^{(f)}}(2), \\ P\left(Z_{(r,s,a)i}^{(f)} = 1\right) &= m_{12}^{(r,s,a)} + m_{11}^{(r,s,a)} + m_{10}^{(r,s,a)} \\ &\equiv \alpha_{(r,s,a)Z^{(f)}}(1), \\ P\left(Z_{(r,s,a)i}^{(f)} = 0\right) &= m_{02}^{(r,s,a)} + m_{01}^{(r,s,a)} + m_{00}^{(r,s,a)} \\ &\equiv \alpha_{(r,s,a)Z^{(f)}}(0), \end{aligned} \right. \quad (\text{A.2})$$

and the distribution for the mother can be obtained by replacing  $m_{uv}^{(r,s,a)}$  by  $m_{vu}^{(r,s,a)}$  in formula (A.2), and is denoted as  $\alpha_{(r,s,a)Z^{(m)}}(w')$ .

*Joint marker score distributions*

Now we consider the joint distributions for marker scores. Let  $\alpha_{(r,s,a)XY}(u, v)$  denote the probability that one affected sibling and one unaffected sibling in the family with type  $(r, s, a)$  have  $u$  and  $v$  alleles  $A$ , respectively. Then it can be shown that

$$\begin{aligned} \alpha_{(r,s,a)XY}(2, 2) &= m_{22}^{(r,s,a)} + \frac{f_2(1 - f_2)}{(f_2 + f_1)[(1 - f_2) + (1 - f_1)]} m_{21}^{(r,s,a)} \\ &+ \frac{f_2(1 - f_2)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(2, 1) &= \frac{f_2(1 - f_1)}{(f_2 + f_1)[(1 - f_2) + (1 - f_1)]} m_{21}^{(r,s,a)} \\ &+ \frac{2f_2(1 - f_1)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.4})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(2, 0) &= \frac{f_2(1 - f_0)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.5})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(1, 2) &= \frac{f_1(1 - f_2)}{(f_2 + f_1)[(1 - f_2) + (1 - f_1)]} m_{21}^{(r,s,a)} \\ &+ \frac{2f_1(1 - f_2)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.6})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(1, 1) &= \frac{f_1(1 - f_1)}{(f_2 + f_1)[(1 - f_2) + (1 - f_1)]} m_{21}^{(r,s,a)} \\ &+ m_{20}^{(r,s,a)} \\ &+ \frac{4f_1(1 - f_1)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)} \\ &+ \frac{f_1(1 - f_1)}{(f_1 + f_0)[(1 - f_1) + (1 - f_0)]} m_{10}^{(r,s,a)}, \end{aligned} \quad (\text{A.7})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(1, 0) &= \frac{2f_1(1 - f_0)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)} \\ &+ \frac{f_1(1 - f_0)}{(f_1 + f_0)[(1 - f_1) + (1 - f_0)]} m_{10}^{(r,s,a)}, \end{aligned} \quad (\text{A.8})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(0, 2) &= \frac{f_0(1 - f_2)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.9})$$

$$\begin{aligned} \alpha_{(r,s,a)XY}(0, 1) &= \frac{2f_0(1 - f_1)}{(f_2 + 2f_1 + f_0)[(1 - f_2) + 2(1 - f_1) + (1 - f_0)]} \\ &\times m_{11}^{(r,s,a)} \\ &+ \frac{f_0(1 - f_1)}{(f_1 + f_0)[(1 - f_1) + (1 - f_0)]} m_{10}^{(r,s,a)}, \end{aligned} \quad (\text{A.10})$$

and

$$\begin{aligned} \alpha_{(r,s,a)XY}(0,0) = & \frac{f_0(1-f_0)}{(f_2+2f_1+f_0)[(1-f_2)+2(1-f_1)+(1-f_0)]} \\ & \times m_{11}^{(r,s,a)} \\ & + \frac{f_0(1-f_0)}{(f_1+f_0)[(1-f_1)+(1-f_0)]} m_{10}^{(r,s,a)} + m_{00}^{(r,s,a)}. \end{aligned} \quad (\text{A.11})$$

The joint distribution for two affected (unaffected) siblings can be obtained by replacing  $1-f_w(f_w)$  by  $f_w(1-f_w)$ . Note that at this time,  $1-f_w$  in the formulas remains unchanged in the formulas (A.3)–(A.11), and is denoted as  $\alpha_{(r,s,a)XX}(u, \nu)$  ( $\alpha_{(r,s,a)YY}(u, \nu)$ ).

Likewise, if we let  $\alpha_{(r,s,a)XZ^{(l)}}(u, \nu)$  be the probability that one affected sibling and the father in the family with type  $(r, s, a)$  have  $u$  and  $\nu$  alleles  $A$ , respectively, then we obtain

$$\alpha_{(r,s,a)XZ^{(l)}}(2,2) = m_{22}^{(r,s,a)} + \frac{f_2}{f_2+f_1} m_{21}^{(r,s,a)}, \quad (\text{A.12})$$

$$\begin{aligned} \alpha_{(r,s,a)XZ^{(l)}}(2,1) = & \frac{f_2}{f_2+f_1} m_{12}^{(r,s,a)} \\ & + \frac{f_2}{f_2+2f_1+f_0} m_{11}^{(r,s,a)}, \end{aligned} \quad (\text{A.13})$$

$$\alpha_{(r,s,a)XZ^{(l)}}(2,0) = 0, \quad (\text{A.14})$$

$$\alpha_{(r,s,a)XZ^{(l)}}(1,2) = \frac{f_1}{f_2+f_1} m_{21}^{(r,s,a)} + m_{20}^{(r,s,a)}, \quad (\text{A.15})$$

$$\begin{aligned} \alpha_{(r,s,a)XZ^{(l)}}(1,1) = & \frac{f_1}{f_2+f_1} m_{12}^{(r,s,a)} \\ & + \frac{2f_1}{f_2+2f_1+f_0} m_{11}^{(r,s,a)} \\ & + \frac{f_1}{f_1+f_0} m_{10}^{(r,s,a)}, \end{aligned} \quad (\text{A.16})$$

$$\alpha_{(r,s,a)XZ^{(l)}}(1,0) = m_{02}^{(r,s,a)} + \frac{f_1}{f_1+f_0} m_{01}^{(r,s,a)}, \quad (\text{A.17})$$

$$\alpha_{(r,s,a)XZ^{(l)}}(0,2) = 0, \quad (\text{A.18})$$

$$\begin{aligned} \alpha_{(r,s,a)XZ^{(l)}}(0,1) = & \frac{f_0}{f_2+2f_1+f_0} m_{11}^{(r,s,a)} \\ & + \frac{f_0}{f_1+f_0} m_{10}^{(r,s,a)}, \end{aligned} \quad (\text{A.19})$$

and

$$\alpha_{(r,s,a)XZ^{(l)}}(0,0) = \frac{f_0}{f_1+f_0} m_{01}^{(r,s,a)} + m_{00}^{(r,s,a)}. \quad (\text{A.20})$$

The joint distribution for one affected sibling and the mother can be obtained by replacing  $m_{uv}^{(r,s,a)}$  by  $m_{vu}^{(r,s,a)}$  in the formulae (A.12)–(A.20), and is denoted by  $\alpha_{(r,s,a)XZ^{(m)}}(u, \nu)$ , and the joint distribution for one unaffected sibling and the father can be obtained by replacing  $f_w$  by  $1-f_w$  in the formulae (A.12)–(A.20), and is denoted by  $\alpha_{(r,s,a)YZ^{(l)}}(u, \nu)$ . As for the joint distribution for one unaffected sibling and the mother, this can be obtained by replacing  $m_{uv}^{(r,s,a)}$  by  $m_{vu}^{(r,s,a)}$  and  $f_w$  by  $1-f_w$  in the formulae (A.12)–(A.20) and is denoted by  $\alpha_{(r,s,a)YZ^{(m)}}(u, \nu)$ .

*Mean, variance, and covariance of marker scores*

From the marker score distributions of the affected and unaffected siblings and their parents, and the joint distributions of the family members, we can obtain the corresponding expectations, variances, and covariances:

$$E[X_{(r,s,a)i}^{(j)}] = 2\alpha_{(r,s,a)X}(2) + \alpha_{(r,s,a)X}(1) \equiv \mu_{(r,s,a)X};$$

Similarly, we can obtain the expressions for the expectations of  $Y_{(r,s,a)i}^{(j)}$ ,  $Z_{(r,s,a)i}^{(f)}$  and  $Z_{(r,s,a)i}^{(m)}$ ;

$$\begin{aligned} V[X_{(r,s,a)i}^{(j)}] = & 4\alpha_{(r,s,a)X}(2) + \alpha_{(r,s,a)X}(1) - (\mu_{(r,s,a)X})^2 \\ & \equiv \sigma_{(r,s,a)X}^2; \end{aligned}$$

The variances of  $Y_{(r,s,a)i}^{(j)}$ ,  $Z_{(r,s,a)i}^{(f)}$  and  $Z_{(r,s,a)i}^{(m)}$  have similar forms;

$$\begin{aligned} Cov(X_{(r,s,a)i}^{(j)}, Y_{(r,s,a)i}^{(k)}) = & 4\alpha_{(r,s,a)XY}(2,2) \\ & + 2[\alpha_{(r,s,a)XY}(2,1) \\ & + \alpha_{(r,s,a)XY}(1,2)] \\ & + \alpha_{(r,s,a)XY}(1,1) \\ & - \mu_{(r,s,a)X}\mu_{(r,s,a)Y} \\ & \equiv \Delta_{(r,s,a)XY} - \mu_{(r,s,a)X}\mu_{(r,s,a)Y} \\ & \equiv \gamma_{(r,s,a)XY}; \end{aligned}$$

The expressions of the covariances between  $X_{(r,s,a)i}^{(j)}$  and  $X_{(r,s,a)i}^{(k)}$ , and  $Y_{(r,s,a)i}^{(j)}$  and  $Y_{(r,s,a)i}^{(k)}$  are similar;

$$\begin{aligned} Cov(Z_{(r,s,a)i}^{(f)}, Z_{(r,s,a)i}^{(m)}) = & 4m_{22}^{(r,s,a)} + 2m_{(21)}^{(r,s,a)} + m_{11}^{(r,s,a)} \\ & - \mu_{(r,s,a)Z^{(l)}}\mu_{(r,s,a)Z^{(m)}} \\ & \equiv \Delta_{(r,s,a)ZZ} \\ & - \mu_{(r,s,a)Z^{(l)}}\mu_{(r,s,a)Z^{(m)}} \\ & \equiv \gamma_{(r,s,a)ZZ}; \end{aligned}$$

$$\begin{aligned}
 \text{Cov}\left(X_{(r,s,a)i}^{(j)}, Z_{(r,s,a)i}^{(f)}\right) &= 4m_{22}^{(r,s,a)} + \frac{2f_2 + f_1}{f_2 + f_1} \\
 &\quad \times \left(2m_{21}^{(r,s,a)} + m_{12}^{(r,s,a)}\right) \\
 &\quad + \frac{2(f_2 + f_1)}{f_2 + 2f_1 + f_0} m_{11}^{(r,s,a)} \\
 &\quad + 2m_{20}^{(r,s,a)} + \frac{f_1}{f_1 + f_0} m_{10}^{(r,s,a)} \\
 &\quad - \mu_{(r,s,a)X} \mu_{(r,s,a)Z^{(f)}} \\
 &\equiv \Delta_{(r,s,a)XZ^{(f)}} \\
 &\quad - \mu_{(r,s,a)X} \mu_{(r,s,a)Z^{(f)}} \\
 &\equiv \gamma_{(r,s,a)XZ^{(f)}}; \tag{A.21}
 \end{aligned}$$

The covariance  $\text{Cov}(X_{(r,s,a)i}^{(j)}, Z_{(r,s,a)i}^{(m)})$  has a similar form to  $\text{Cov}(X_{(r,s,a)i}^{(j)}, Z_{(r,s,a)i}^{(f)})$  except  $m_{uv}^{(r,s,a)}$  and  $\mu_{(r,s,a)Z^{(m)}}$  are in place of  $m_{vu}^{(r,s,a)}$  and  $\mu_{(r,s,a)Z^{(f)}}$  in the above formula. Likewise, the expressions for the covariance  $\text{Cov}(Y_{(r,s,a)i}^{(j)}, Z_{(r,s,a)i}^{(f)})$  can be obtained by replacing  $f_w$  by  $1 - f_w$  in formula (A.21), and the covariance  $\text{Cov}(Y_{(r,s,a)i}^{(j)}, Z_{(r,s,a)i}^{(m)})$  can be obtained by replacing  $f_w$  by  $1 - f_w$ ,  $m_{uv}^{(r,s,a)}$  by  $m_{vu}^{(r,s,a)}$  and  $\mu_{(r,s,a)Z^{(f)}}$  by  $\mu_{(r,s,a)Z^{(m)}}$  in formula (A.21).

*Mean and variance of the difference between the two allele frequency estimates  $\hat{p}_{(r,s,a)A}$  and  $\hat{p}_{(r,s,a)U}$  for families with identical structures.*

In the following, we derive the mean and variance of  $\hat{p}_{(r,s,a)A} - \hat{p}_{(r,s,a)U}$  under the alternative hypothesis. It can be shown that the means of allele frequency estimates in the case group and control group are

$$E(\hat{p}_{(r,s,a)A}) = \mu_{(r,s,a)X}/2,$$

and

$$E(\hat{p}_{(r,s,a)U}) = \frac{(s-a)\mu_{(r,s,a)Y} + \sum_{j=1}^r \mu_{(r,s,a)Z^{(j)}}}{2(s-a+r)},$$

respectively. If we define

$$\mu_{(r,s,a)Z} \equiv \frac{1}{2} [\mu_{(r,s,a)Z^{(1)}} + \mu_{(r,s,a)Z^{(2)}}],$$

and note that

$$\sum_{j=1}^r \mu_{(r,s,a)Z^{(j)}} = r \mu_{(r,s,a)Z}$$

since we have assumed random mating of parents when  $r = 1$ , then we have

$$\begin{aligned}
 E(\hat{p}_{(r,s,a)A} - \hat{p}_{(r,s,a)U}) &= \frac{1}{2} \left[ \mu_{(r,s,a)X} \right. \\
 &\quad \left. - \frac{(s-a)\mu_{(r,s,a)Y} + r \mu_{(r,s,a)Z}}{s-a+r} \right] \\
 &\equiv \mu_{(r,s,a)}. \tag{A.22}
 \end{aligned}$$

Likewise, we can show that

$$\begin{aligned}
 V(\hat{p}_{(r,s,a)A}) &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \\
 &\quad \cdot \frac{\sigma_{(r,s,a)X}^2 + (a-1)\gamma_{(r,s,a)XX}}{4a} + \varepsilon^2,
 \end{aligned}$$

and

$$\begin{aligned}
 V(\hat{p}_{(r,s,a)U}) &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \frac{1}{4(s-a+r)^2} \\
 &\quad \times \left[ (s-a)\sigma_{(r,s,a)Y}^2 \right. \\
 &\quad + (s-a)(s-a-1)\gamma_{(r,s,a)YY} \\
 &\quad + \sum_{j=1}^r \sigma_{(r,s,a)Z^{(j)}}^2 + r(r-1)\gamma_{(r,s,a)ZZ} \\
 &\quad \left. + 2(s-a) \sum_{j=1}^r \gamma_{(r,s,a)YZ^{(j)}} \right] + \varepsilon^2,
 \end{aligned}$$

where  $E_1$  denotes the expectation over all possible values of  $n_{(r,s,a)}$ . Now define

$$\sigma_{(r,s,a)Z}^2 \equiv \frac{1}{2} \left[ \sigma_{(r,s,a)Z^{(1)}}^2 + \sigma_{(r,s,a)Z^{(2)}}^2 \right],$$

and

$$\gamma_{(r,s,a)YZ} \equiv \frac{1}{2} \left[ \gamma_{(r,s,a)YZ^{(1)}} + \gamma_{(r,s,a)YZ^{(2)}} \right].$$

Then from the formulae for  $\sigma_{(r,s,a)Z^{(i)}}^2$  and  $\gamma_{(r,s,a)YZ^{(i)}}$  ( $i = 1, 2$ ) provided above (noting that  $\sigma_{(r,s,a)Z}^2$  etc., defined above, depend only on the sum of the variances etc. for both parents, we can regard the first parent as the father, and the second parent as the mother), we have

$$\begin{aligned}
 \sigma_{(r,s,a)Z}^2 &= 2 \left( 2m_{22}^{(r,s,a)} + m_{(21)}^{(r,s,a)} + m_{(20)}^{(r,s,a)} \right) \\
 &\quad + \frac{1}{2} \left( m_{(21)}^{(r,s,a)} + 2m_{11}^{(r,s,a)} + m_{(10)}^{(r,s,a)} \right) \\
 &\quad - \frac{1}{2} \left[ (\mu_{(r,s,a)Z^{(1)}})^2 + (\mu_{(r,s,a)Z^{(2)}})^2 \right] \\
 &\equiv \Delta_{(r,s,a)Z} - \frac{1}{2} \left[ (\mu_{(r,s,a)Z^{(1)}})^2 + (\mu_{(r,s,a)Z^{(2)}})^2 \right],
 \end{aligned}$$

and

$$\begin{aligned}
 \mathcal{Y}_{(r,s,a)YZ} &= 4m_{22}^{(r,s,a)} + \frac{3[2(1-f_2) + (1-f_1)]}{2[(1-f_2) + (1-f_1)]} m_{(21)}^{(r,s,a)} \\
 &+ m_{(20)}^{(r,s,a)} \\
 &+ \frac{2[(1-f_2) + (1-f_1)]}{(1-f_2) + 2(1-f_1) + (1-f_0)} m_{11}^{(r,s,a)} \\
 &+ \frac{1-f_1}{2[(1-f_1) + (1-f_0)]} m_{(10)}^{(r,s,a)} \\
 &- \mu_{(r,s,a)Y} \cdot \frac{1}{2} [\mu_{(r,s,a)Z^{(1)}} + \mu_{(r,s,a)Z^{(2)}}] \\
 &\equiv \Delta_{(r,s,a)YZ} - \mu_{(r,s,a)Y} \mu_{(r,s,a)Z}. \quad (\text{A.23})
 \end{aligned}$$

Note that the mating of the parents is assumed to be random when  $r = 1$ . Hence

$$\sum_{j=1}^r \sigma_{(r,s,a)Z^{(j)}}^2 = r \sigma_{(r,s,a)Z}^2,$$

and

$$\sum_{j=1}^r \mathcal{Y}_{(r,s,a)YZ^{(j)}} = r \mathcal{Y}_{(r,s,a)YZ}.$$

Therefore, the variance of  $\widehat{p}_{(r,s,a)U}$  can be expressed as

$$\begin{aligned}
 V(\widehat{p}_{(r,s,a)U}) &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \frac{1}{4(s-a+r)^2} \\
 &\times [(s-a)\sigma_{(r,s,a)Y}^2 \\
 &+ (s-a)(s-a-1)\mathcal{Y}_{(r,s,a)YY} \\
 &+ r\sigma_{(r,s,a)Z}^2 + r(r-1)\mathcal{Y}_{(r,s,a)ZZ} \\
 &+ 2(s-a)r\mathcal{Y}_{(r,s,a)YZ}] + \varepsilon^2.
 \end{aligned}$$

Further, we can obtain

$$\begin{aligned}
 \text{Cov}(\widehat{p}_{(r,s,a)A}, \widehat{p}_{(r,s,a)U}) &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \\
 &\cdot \frac{(s-a)\mathcal{Y}_{(r,s,a)XY} + r\mathcal{Y}_{(r,s,a)XZ}}{4(s-a+r)},
 \end{aligned}$$

where

$$\begin{aligned}
 \mathcal{Y}_{(r,s,a)XZ} &\equiv \frac{1}{2} \sum_{j=1}^2 \mathcal{Y}_{(r,s,a)XZ^{(j)}} \\
 &\equiv \Delta_{(r,s,a)XZ} - \mu_{(r,s,a)X} \mu_{(r,s,a)Z}
 \end{aligned}$$

has similar meaning to  $\mathcal{Y}_{(r,s,a)YZ}$  given in (A.23) except we should replace  $1 - f_w$  by  $f_w$  in the expression.

Consequently,

$$\begin{aligned}
 V(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}) &= \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \\
 &\cdot \left\{ \frac{\sigma_{(r,s,a)X}^2 - \mathcal{Y}_{(r,s,a)XX}}{a} + \mathcal{Y}_{(r,s,a)XX} + \frac{1}{(s-a+r)^2} \right. \\
 &[(s-a)(\sigma_{(r,s,a)Y}^2 - \mathcal{Y}_{(r,s,a)YY}) + (s-a)^2 \mathcal{Y}_{(r,s,a)YY} \\
 &+ r\sigma_{(r,s,a)Z}^2 + r(r-1)\mathcal{Y}_{(r,s,a)ZZ} + 2(s-a)r\mathcal{Y}_{(r,s,a)YZ}] \\
 &- 2 \cdot \frac{(s-a)\mathcal{Y}_{(r,s,a)XY} + r\mathcal{Y}_{(r,s,a)XZ}}{s-a+r} \left. \right\} + 2\varepsilon^2 \\
 &= \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \left\{ \frac{\sigma_{(r,s,a)X}^2 - \mathcal{Y}_{(r,s,a)XX}}{a} + \Delta_{(r,s,a)XX} \right. \\
 &+ \frac{1}{(s-a+r)^2} [(s-a)(\sigma_{(r,s,a)Y}^2 - \mathcal{Y}_{(r,s,a)YY}) \\
 &+ (s-a)^2 \Delta_{(r,s,a)YY} + r\Delta_{(r,s,a)Z} + r(r-1)\Delta_{(r,s,a)ZZ} \\
 &+ 2(s-a)r\Delta_{(r,s,a)YZ}] \\
 &- 2 \cdot \frac{(s-a)\Delta_{(r,s,a)XY} + r\Delta_{(r,s,a)XZ}}{s-a+r} \\
 &\left. - 4(\mu_{(r,s,a)})^2 \right\} + 2\varepsilon^2 \\
 &\equiv \frac{1}{4} E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \cdot \Delta_{(r,s,a)}^* + 2\varepsilon^2 \\
 &\equiv \sigma_{(r,s,a)}^2. \quad (\text{A.24})
 \end{aligned}$$

From Stephan (1945), we have

$$E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] = \frac{1}{n\lambda_{(r,s,a)}} + \frac{1 - \lambda_{(r,s,a)}}{n^2 \lambda_{(r,s,a)}^2} + o\left(\frac{1}{n^2}\right), \quad (\text{A.25})$$

where as before,  $\lambda_{(r,s,a)}$  is the proportion of families with type  $(r, s, a)$  in the ascertainment subpopulation, and can be accurately estimated when the sample size is large. If we use the first order approximation of  $E_1\left[\frac{1}{n_{(r,s,a)}}\right]$ , then we get

$$\sigma_{(r,s,a)}^2 \approx \frac{\Delta_{(r,s,a)}^*}{4n\lambda_{(r,s,a)}} + 2\varepsilon^2. \quad (\text{A.26})$$

We use formula (A.26) to estimate the sample size required to detect association in this study, although it is straightforward to approximate power and sample size using the second order approximation of  $E_1\left[\frac{1}{n_{(r,s,a)}}\right]$ . In fact, our numerical calculation for the example given in Section 4 shows that using the first and second order

approximations lead to almost identical power (data not shown). If we, like Risch & Teng (1998), assume that the penetrance is low, then the expectation and variance of the difference between the sample frequencies among the affecteds and controls,  $\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}$ , under the alternative hypothesis reduce to

$$\mu_{(r,s,a)} = \pi_{21}m_{(21)}^{(r,s,a)} + \pi_{11}m_{11}^{(r,s,a)} + \pi_{10}m_{(10)}^{(r,s,a)}, \quad (\text{A.27})$$

and

$$\begin{aligned} V(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}) &= E_1 \left[ \frac{1}{n_{(r,s,a)}} \right] \\ &\cdot \left\{ \pi_{21}^2 m_{(21)}^{(r,s,a)} + \pi_{11}^2 m_{11}^{(r,s,a)} + \pi_{10}^2 m_{(10)}^{(r,s,a)} \right. \\ &+ \frac{1}{a} \left( \psi_{21}^2 m_{(21)}^{(r,s,a)} + \psi_{11}^2 m_{11}^{(r,s,a)} + \psi_{10}^2 m_{(10)}^{(r,s,a)} \right) \\ &+ \frac{1}{s-a+r} \left( \frac{1}{16} m_{(21)}^{(r,s,a)} + \frac{1}{8} m_{11}^{(r,s,a)} + \frac{1}{16} m_{(10)}^{(r,s,a)} \right) \\ &+ \frac{r}{4(s-a+r)^2} \left[ (1-r) \left( \frac{1}{4} m_{(21)}^{(r,s,a)} + m_{(20)}^{(r,s,a)} \right) \right. \\ &+ \left. \frac{1}{4} m_{(10)}^{(r,s,a)} \right] + \left( m_{(20)}^{(r,s,a)} - \frac{1}{2} m_{11}^{(r,s,a)} \right) \\ &\left. - (\mu_{(r,s,a)})^2 \right\} + 2\varepsilon^2, \end{aligned} \quad (\text{A.28})$$

respectively, where

$$\pi_{21} = \frac{f_2 - f_1}{4(f_2 + f_1)}, \quad \pi_{11} = \frac{f_2 - f_0}{2(f_2 + 2f_1 + f_0)},$$

$$\pi_{10} = \frac{f_1 - f_0}{4(f_1 + f_0)},$$

and

$$\psi_{21}^2 = \frac{f_2 f_1}{4(f_2 + f_1)^2}, \quad \psi_{11}^2 = \frac{f_2 f_1 + 2f_2 f_0 + f_1 f_0}{2(f_2 + 2f_1 + f_0)^2},$$

$$\psi_{10}^2 = \frac{f_1 f_0}{4(f_1 + f_0)^2}.$$

These results are the same as those in Risch & Teng (1998). Equations (A.27) and (A.28) give unified formulae for various family structures. By taking different values of  $r$ ,  $s$  and  $a$ , we can obtain the corresponding results in Risch & Teng (1998).

## Appendix B

In this appendix, we prove the asymptotic normality of our proposed test statistic  $t$  under  $H_0$ . For convenience we denote the variance of the measurement error by  $\varepsilon_n^2$  when the sample size is  $n$ . When  $t_n$  is a nonzero

constant, the proof is obvious. Now we assume that  $\sqrt{n}\varepsilon_n \rightarrow \ell < \infty$ , where  $\ell$  is a constant. It can be seen that

$$\begin{aligned} \sqrt{n}(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}) &= \\ \sqrt{n} \left( \frac{1}{n_{(r,s,a)}} \sum_{i=1}^{n_{(r,s,a)}} \zeta_{(r,s,a)i} + \xi_{(r,s,a)} - \eta_{(r,s,a)} \right), \end{aligned}$$

where

$$\zeta_{(r,s,a)i} = \frac{\sum_{j=1}^a X_{(r,s,a)i}^{(j)}}{2a} - \frac{\sum_{j=1}^{s-a} Y_{(r,s,a)i}^{(j)} + \sum_{j=1}^r Z_{(r,s,a)i}^{(j)}}{2(s-a+r)}$$

has mean zero and variance

$$\sigma_{(r,s,a)0}^{*2} = \frac{1}{4} p_{(r,s,a)} q_{(r,s,a)} \left[ \frac{1}{a} + \frac{s-a+2r-r^2}{(s-a+r)^2} \right]$$

under  $H_0$ . Note that when  $n \rightarrow \infty$ ,

$$\frac{n_{(r,s,a)}}{n} \rightarrow^{p\cdot} \lambda_{(r,s,a)},$$

where  $\rightarrow^{p\cdot}$  means convergence in probability. So from the central limit theorem and the assumptions given in Section 2, we have

$$\sqrt{n}(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U}) \rightarrow^d N \left( 0, \frac{\sigma_{(r,s,a)0}^{*2}}{\lambda_{(r,s,a)}} + 2\ell^2 \right),$$

where  $\rightarrow^d$  means convergence in distribution. On the other hand, under  $H_0$ ,  $\widehat{p}_{(r,s,a)} \rightarrow^{p\cdot} p_{(r,s,a)}$ . Hence,

$$\begin{aligned} n\widehat{\sigma}_{(r,s,a)0}^2 &= \frac{n}{n_{(r,s,a)}} \cdot \frac{1}{4} \widehat{p}_{(r,s,a)} (1 - \widehat{p}_{(r,s,a)}) \\ &\times \left[ \frac{1}{a} + \frac{s-a+2r-r^2}{(s-a+r)^2} \right] + 2n\varepsilon_n^2 \\ &\rightarrow^{p\cdot} \frac{1}{\lambda_{(r,s,a)}} \cdot \frac{1}{4} p_{(r,s,a)} (1 - p_{(r,s,a)}) \\ &\times \left[ \frac{1}{a} + \frac{s-a+2r-r^2}{(s-a+r)^2} \right] + 2\ell^2 \\ &= \frac{\sigma_{(r,s,a)0}^{*2}}{\lambda_{(r,s,a)}} + 2\ell^2. \end{aligned}$$

Thus,

$$\begin{aligned} t &= \frac{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \frac{1}{n\widehat{\sigma}_{(r,s,a)0}^2} \cdot \sqrt{n}(\widehat{p}_{(r,s,a)A} - \widehat{p}_{(r,s,a)U})}{\sqrt{\sum_{r=0}^2 \sum_{s=1}^{\infty} \sum_{a=1}^s \frac{1}{n\widehat{\sigma}_{(r,s,a)0}^2}}} \\ &\rightarrow^d N(0, 1). \end{aligned}$$

Received: 2 July 2004

Accepted: 5 November 2004