

The Impacts of Errors in Individual Genotyping and DNA Pooling on Association Studies

Guohua Zou¹ and Hongyu Zhao^{1,2*}

¹Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut

²Department of Genetics, Yale University School of Medicine, New Haven, Connecticut

Case-control association studies using unrelated individuals may offer an effective approach for identifying genetic variants that have small to moderate disease risks. In general, two different strategies may be employed to establish associations between genotypes and phenotypes: (1) collecting individual genotypes or (2) quantifying allele frequencies in DNA pools. These two technologies have their respective advantages. Individual genotyping gathers more information, whereas DNA pooling may be more cost effective. Recent technological advances in DNA pooling have generated great interest in using DNA pooling in association studies. In this article, we investigate the impacts of errors in genotyping or measuring allele frequencies on the identification of genetic associations with these two strategies. We find that, with current technologies, compared to individual genotyping, a larger sample is generally required to achieve the same power using DNA pooling. We further consider the use of DNA pooling as a screening tool to identify candidate regions for follow-up studies. We find that the majority of the positive regions identified from DNA pooling results may represent false positives if measurement errors are not appropriately considered in the design of the study. *Genet Epidemiol* 26:1–10, 2004. © 2003 Wiley-Liss, Inc.

Key words: case-control study; individual genotyping; DNA pooling; measurement error; sample size; false discovery rate

Grant sponsor: National Institute of Health; Grant number: GM59507.

*Correspondence to: Hongyu Zhao, Ph.D., Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

Received 13 May 2003; Accepted 21 July 2003

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

DOI: 10.1002/gepi.10277

INTRODUCTION

Case-control association studies using unrelated individuals may be an effective approach to identifying genetic variants underlying complex traits. Individual genotyping is routinely used as the genotyping strategy in many studies to assess statistical associations between genetic variants and disease outcomes. Although this approach allows the collection of each individual's genetic composition, the genotyping cost can be considerable when hundreds or thousands of individuals are studied at many genetic markers, even with the ever-decreasing cost in genotyping. Therefore, the methods using DNA pools have been actively developed as a viable alternative approach for collecting allele frequency information from a group of individuals [see, for example, Shaw et al., 1998; Breen et al., 2000; Germer et al., 2000; Hoogendoorn et al., 2000; Norton et al., 2002]. Recent developments in quantitative assays and in

the design and analysis of pooling data are summarized in a thorough review by Sham et al. [2002].

In the ideal world where the technologies are flawless, the only loss of information in DNA pooling is the inability to assign genotypes to each individual. This may result in some loss of information for association studies on quantitative traits [Bader et al., 2001], and difficulty in controlling for population stratification. Despite such information loss, DNA pooling may still be more preferable considering the potential cost reduction with this approach. However, the errors in genotyping or measuring allele frequencies are unavoidable for both technologies. For example, for a given DNA pooling sample, the standard deviation of the estimated allele frequency is between 2 and 4% [Sham et al., 2002]. The effects of measurement errors on statistical inference have been extensively studied in statistical literature. Such errors can lead to increased bias,

reduced precision in estimation, and inflated type-I error and decreased power in test [Cochran, 1968; Fuller, 1987; Carroll et al., 1995]. For association studies on quantitative traits and susceptibility loci, measurement errors can substantially reduce the information retained for DNA pooling [Jawaid et al., 2002]. Barratt et al. [2002] considered the sources of error in each experimental stage for the estimation of allele frequency when DNA pooling sample is used. The authors found that a design based on the formation of numerous small pools of approximately 50 individuals may attain an effective trade-off between accuracy and cost. Bansal et al. [2002] examined 15 single-nucleotide polymorphisms (SNPs) in the cholesteryl ester transfer protein gene using two pools forming from individuals with extremely high and extremely low serum high-density lipoprotein cholesterol levels, respectively, and found the association described previously by other research groups. Therefore, DNA pooling has been advocated as a screening tool to identify regions of interest followed with individual genotyping [Barcellos et al., 1997; Bansal et al. 2002; Barratt et al., 2002; Sham et al., 2002]. In this article, we first consider the impacts of genotyping errors in individual genotyping and measurement errors in DNA pooling on the required sample sizes to attain a desired statistical power to identify associations between genetic variants and diseases at a given significance level. We show that, although genotyping errors have relatively small effects on the required sample sizes for individual genotyping, the required sample sizes are substantially larger with increasing levels of measurement errors using DNA pooling. As a result, for the current levels of measurement errors, the sample size required to detect association can be considerably larger based on the DNA pooling strategy unless many pools are formed to reduce the overall error rate. We also show that, unless the measurement errors are controlled at very low levels, the majority of the positive results obtained through DNA pooling may represent false positives if the measurement errors are not appropriately taken into account in study designs, making DNA pooling a less efficient screening tool for association studies.

METHODS

We consider two alleles, A and a , at a candidate marker, whose allele frequencies are p and

$q = 1 - p$, respectively. We assume that the penetrances are f_2 for genotype AA , f_1 for genotype Aa , and f_0 for genotype aa . Note that these two alleles may be true functional alleles or may be in linkage disequilibrium with true functional alleles. For simplification, we consider a case-control study with n cases and n controls. Let X_i denote the number of allele A carried by the i th individual in the case group, and Y_i is similarly defined for the i th individual in the control group. Assuming Hardy-Weinberg equilibrium, each X_i or Y_i has a value of 2, 1, 0 with respective probabilities p^2 , $2pq$, and q^2 under the null hypothesis of no association between the candidate marker and disease. Under the above genetic model, the probabilities for having k copies of A among the cases, $m_k = P(X_i = k)$, and among the controls, $m'_k = P(Y_i = k)$, are

$$\begin{aligned} m_0 &= \frac{q^2 f_0}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m_1 &= \frac{2pq f_1}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m_2 &= \frac{p^2 f_2}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m'_0 &= \frac{q^2(1 - f_0)}{p^2(1 - f_2) + 2pq(1 - f_1) + q^2(1 - f_0)}, \\ m'_1 &= \frac{2pq(1 - f_1)}{p^2(1 - f_2) + 2pq(1 - f_1) + q^2(1 - f_0)}, \\ m'_2 &= \frac{p^2(1 - f_2)}{p^2(1 - f_2) + 2pq(1 - f_1) + q^2(1 - f_0)}. \end{aligned} \tag{1}$$

POWER AND SAMPLE SIZE

Individual genotyping. For individual genotyping, we assume that genotyping errors are introduced independently to each allele, and the error rate from true allele A to erroneous allele a is e_1 and from true allele a to erroneous allele A is e_2 . Note that a more complete model for genotyping errors may be based on genotypes instead of alleles. Let n_A^* and n_U^* denote the observed numbers of allele A in the case group and the control group, respectively, p_A and p_U denote the true frequencies of allele A in these two groups, and \hat{p}_A and \hat{p}_U denote their maximum likelihood estimates assuming known genotyping error rates. It can be shown that

$$\hat{p}_A = \frac{n_A^*/(2n) - e_2}{1 - e_1 - e_2}$$

and

$$\hat{p}_U = \frac{n_U^*/(2n) - e_2}{1 - e_1 - e_2}.$$

A similar formula is obtained by Gastwirth [1987] for estimating the prevalence of a trait from diagnostic/screening tests, where $e_1 = 1 -$ sensitivity of test, and $e_2 = 1 -$ specificity of test, with the sensitivity of test being the probability that a person with the disease is correctly diagnosed, and the specificity of test being the probability that a disease-free individual is correctly diagnosed.

Under the null hypothesis of no association between the candidate marker and disease status, $E(\hat{p}_A - \hat{p}_U) = 0$, and

$$V(\hat{p}_A - \hat{p}_U) = \frac{p^*(1 - p^*)}{n(1 - e_1 - e_2)^2},$$

where $p^* = (1 - e_1)p + e_2q$. On the other hand, under the genetic model we introduced above,

$$E(\hat{p}_A - \hat{p}_U) = \frac{p_A^* - p_U^*}{1 - e_1 - e_2} \equiv \mu,$$

and

$$\begin{aligned} & V(\hat{p}_A - \hat{p}_U) \\ &= \frac{[p_A^*(1 - p_A^*) - p_{A12}^*/4] + [p_U^*(1 - p_U^*) - p_{U12}^*/4]}{n(1 - e_1 - e_2)^2} \\ &\equiv \frac{\sigma^2}{n}, \end{aligned}$$

where

$$\begin{aligned} p_A^* &= p_{A11}^* + p_{A12}^*/2, \\ p_U^* &= p_{U11}^* + p_{U12}^*/2, \end{aligned}$$

with

$$\begin{aligned} p_{A11}^* &= (1 - e_1)^2 m_2 + (1 - e_1)e_2 m_1 + e_2^2 m_0, \\ p_{A12}^* &= 2(1 - e_1)e_1 m_2 + [(1 - e_1)(1 - e_2) + e_1 e_2] m_1 \\ &\quad + 2(1 - e_2)e_2 m_0. \end{aligned}$$

p_{U11}^* and p_{U12}^* are similarly defined for the control group and their values can be calculated by replacing m_i in the above formulas for p_{A11}^* and p_{A12}^* with $m'_i (i = 0, 1, 2)$ which are given in (1). Note that p_{A11}^* or p_{A12}^* is the probability that given an individual is affected, his or her observed genotype is AA or Aa , respectively. p_{U11}^* or p_{U12}^* has similar meaning for unaffected individuals.

Based on the above results, we can construct the test statistic as

$$\begin{aligned} t_{ind} &= \frac{\hat{p}_A - \hat{p}_U}{\sqrt{\hat{p}^*(1 - \hat{p}^*)/[n(1 - e_1 - e_2)^2]}} \\ &= \frac{(n_A^* - n_U^*)/(2n)}{\sqrt{\hat{p}^*(1 - \hat{p}^*)/n}}, \end{aligned}$$

where $\hat{p}^* = (n_A^* + n_U^*)/(4n)$. Note that t_{ind} is independent of genotyping error rates.

As in Risch and Teng [1998], we consider a one-sided test, and the power of the test statistic t_{ind} with a significance level of α is

$$\Phi\left(\frac{-z_\alpha \sqrt{\tilde{p}^*(1 - \tilde{p}^*)}/(1 - e_1 - e_2) + \sqrt{n}\mu}{\sigma}\right),$$

where $\tilde{p}^* = (p_A^* + p_U^*)/2$ is the expected frequency of allele A under the above genetic model, Φ is the cumulative standard normal distribution function, and z_α is the upper 100α th percentile of the standard normal distribution. The sample size necessary to obtain a power of $1 - \beta$ at the significance level α is

$$n = \left(\frac{z_\alpha \sqrt{\tilde{p}^*(1 - \tilde{p}^*)}/(1 - e_1 - e_2) - z_{1-\beta}\sigma}{\mu}\right)^2. \quad (2)$$

DNA pooling. For DNA pooling, we assume the following models relating the observed allele frequencies estimated from the sample to the true frequencies of allele A in the sample:

$$\hat{p}_A^{pool} = \frac{X_1 + \cdots + X_n}{2n} + u, \quad (3)$$

$$\hat{p}_U^{pool} = \frac{Y_1 + \cdots + Y_n}{2n} + v, \quad (4)$$

where X_i and Y_i were defined before, u and v are independent normal random variables with mean 0 and variance e^2 .

Under the null hypothesis of no association,

$$\begin{aligned} E\left(\frac{X_1 + \cdots + X_n}{2n} - \frac{Y_1 + \cdots + Y_n}{2n}\right) &= 0, \\ V\left(\frac{X_1 + \cdots + X_n}{2n} - \frac{Y_1 + \cdots + Y_n}{2n}\right) &= \frac{pq}{n}, \end{aligned}$$

and under the genetic model,

$$\begin{aligned} E\left(\frac{X_1 + \cdots + X_n}{2n} - \frac{Y_1 + \cdots + Y_n}{2n}\right) \\ = m_2 + \frac{1}{2}m_1 - m'_2 - \frac{1}{2}m'_1 \equiv \mu, \\ V\left(\frac{X_1 + \cdots + X_n}{2n} - \frac{Y_1 + \cdots + Y_n}{2n}\right) \\ = \frac{1}{4n} \left[4m_2 + m_1 - (2m_2 + m_1)^2 \right. \\ \left. + 4m'_2 + m'_1 - (2m'_2 + m'_1)^2 \right] \\ \equiv \frac{\sigma^2}{n}. \end{aligned}$$

Therefore, when there is no association,

$$E(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = 0, V(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = \frac{pq}{n} + 2\epsilon^2.$$

Under the genetic model introduced above,

$$E(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = \mu, V(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = \frac{\sigma^2}{n} + 2\epsilon^2.$$

We can use the following test statistic to test genetic associations based on DNA pooling data:

$$t_{pool} = \frac{\hat{p}_A^{pool} - \hat{p}_U^{pool}}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n} + 2\epsilon^2}},$$

where

$$\hat{p}_{pool} = \frac{1}{2}(\hat{p}_A^{pool} + \hat{p}_U^{pool}).$$

Consider a one-sided test and use the significance level of α , the power of the test statistic t_{pool} is

$$FDR \approx \frac{\sum_{i=1}^K P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ true}) P(H_0^{(i)} \text{ true})}{\sum_{i=1}^K [P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ true}) P(H_0^{(i)} \text{ true}) + P(\text{reject } H_0^{(i)} | H_1^{(i)} \text{ true}) P(H_1^{(i)} \text{ true})]}, \quad (7)$$

$$\Phi\left(\frac{-z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + 2\epsilon^2} + \mu}{\sqrt{\frac{\sigma^2}{n} + 2\epsilon^2}}\right),$$

where $\hat{p} = \mu/2 + m'_2 + m'_1/2$ is the expected frequency of allele A under the genetic model introduced before.

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

following equation

$$\frac{z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + 2\epsilon^2} - \mu}{\sqrt{\frac{\sigma^2}{n} + 2\epsilon^2}} = z_{1-\beta}. \quad (5)$$

FALSE DISCOVERY RATES

In a genome-wide association study, many candidate markers are examined and there is, in general, no prior information on which markers are more likely to be involved in the disease of interest. Due to stochastic variations, some markers that are not associated with the disease may be found to be significant. If a genetic study is planned under a certain genetic model assuming the absence of genotyping errors or measurement errors, this particular study may be underpowered to detect true associations. As a result, the study may not have adequate power to distinguish true signals from false signals, resulting in a large proportion of false positives among all positive results. In this article, in addition to the effects of genotyping errors in individual genotyping and measurement errors in DNA pooling on sample sizes, we also consider their effects on the proportion of false positives among all positive results. This proportion is called the false discovery rate (FDR) [Benjamini and Hochberg, 1995]. In a test of K hypotheses $H_0^{(i)}, i = 1, \dots, K$, let $\eta_i = 1$ when $H_0^{(i)}$ is true and 0 otherwise, and let $\zeta_i = 1$ if we reject $H_0^{(i)}$ and 0 otherwise, then the FDR is formally defined as

$$FDR = E\left(\frac{\sum_{i=1}^K \eta_i \zeta_i}{\sum_{i=1}^K \zeta_i} \mid \sum_{i=1}^K \zeta_i > 0\right) \quad (6)$$

Using the first-order approximation to FDR, we have

where $H_1^{(i)}$ denotes the i th alternative hypothesis, $i = 1, \dots, K$ (see the Appendix). Therefore, if we have prior knowledge on $H_0^{(i)}$ being true, we can obtain the value of FDR using the above formula.

Now we consider the impact of errors on FDR through formula (7) for both individual genotyping and DNA pooling. Suppose that for testing hypothesis $H_0^{(i)}$, a sample size of $n_0^{(i)}$ is needed to obtain a power of $1 - \beta_0^{(i)}$ at the significance level

of $\alpha_0^{(i)}$ in the absence of genotyping or measurement errors. Then the corresponding FDR can be approximated by

$$FDR \approx \frac{\sum_{i=1}^K \alpha_0^{(i)} P(H_0^{(i)} \text{ true})}{\sum_{i=1}^K [\alpha_0^{(i)} P(H_0^{(i)} \text{ true}) + (1 - \beta_0^{(i)}) P(H_1^{(i)} \text{ true})]}.$$

Individual genotyping. When there are genotyping errors in individual genotyping, the actual type-I error and power for a study with sample size $n_0^{(i)}$ derived assuming the absence of genotyping errors are

$$\begin{aligned} & P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ true}) \\ &= P\left(\frac{n_A^{*(i)}/(2n_0^{(i)}) - n_U^{*(i)}/(2n_0^{(i)})}{\sqrt{\hat{p}_i^*(1 - \hat{p}_i^*)/n_0^{(i)}}} > z_{\alpha_0^{(i)}} | H_0^{(i)} \text{ true}\right) \\ &= P\left(\frac{\hat{p}_A^{(i)} - \hat{p}_U^{(i)}}{\sqrt{p_i^*(1 - p_i^*)/[n_0^{(i)}(1 - e_1^{(i)} - e_2^{(i)})^2]}} > \frac{z_{\alpha_0^{(i)}} \sqrt{\hat{p}_i^*(1 - \hat{p}_i^*)}}{\sqrt{p_i^*(1 - p_i^*)}} | H_0^{(i)} \text{ true}\right) \\ &= \Phi(-z_{\alpha_0^{(i)}}) = \alpha_0^{(i)} \equiv \alpha^{(i)}, \end{aligned} \tag{8}$$

and

$$\begin{aligned} & P(\text{reject } H_0^{(i)} | H_1^{(i)} \text{ true}) \\ &= \Phi\left(\frac{-z_{\alpha_0^{(i)}} \sqrt{\tilde{p}_i^*(1 - \tilde{p}_i^*)}/(1 - e_1^{(i)} - e_2^{(i)}) + \sqrt{n_0^{(i)} \mu^{(i)}}}{\sigma^{(i)}}\right) \\ &\equiv 1 - \beta^{(i)}, \end{aligned}$$

respectively, where $n_A^{*(i)}$, p_i^* , $\mu^{(i)}$, and $\sigma^{(i)2}$ etc. are defined for the i th marker in the same way as n_A^* , p^* , μ , and σ^2 etc. were defined above for individual genotyping. Then the actual FDR in the presence of genotyping errors for neglecting the genotyping errors and using the sample size $n_0^{(i)}$ is

$$FDR = \frac{\sum_{i=1}^K \alpha^{(i)} P(H_0^{(i)} \text{ true})}{\sum_{i=1}^K [\alpha^{(i)} P(H_0^{(i)} \text{ true}) + (1 - \beta^{(i)}) P(H_1^{(i)} \text{ true})]} \tag{9}$$

DNA pooling. Assuming the measurement error models given by (3) and (4) for DNA pooling, the

actual type-I error and power for testing hypothesis $H_0^{(i)}$ when we neglect the measurement errors and use the sample size $n_0^{(i)}$ in the absence of errors are

$$\begin{aligned} & P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ true}) \\ &= \Phi\left(-\frac{z_{\alpha_0^{(i)}} \sqrt{p_i(1 - p_i)/n_0^{(i)}}}{\sqrt{p_i(1 - p_i)/n_0^{(i)} + 2\epsilon_i^2}}\right) \\ &\equiv \alpha^{(i)}, \end{aligned}$$

and

$$\begin{aligned} & P(\text{reject } H_0^{(i)} | H_1^{(i)} \text{ true}) \\ &= \Phi\left(\frac{-z_{\alpha_0^{(i)}} \sqrt{\tilde{p}_i(1 - \tilde{p}_i)/n_0^{(i)} + \mu^{(i)}}}{\sqrt{\sigma^{(i)2}/n_0^{(i)} + 2\epsilon_i^2}}\right) \\ &\equiv 1 - \beta^{(i)}, \end{aligned}$$

respectively, where p_i , \tilde{p}_i , $\mu^{(i)}$, $\sigma^{(i)2}$ and ϵ_i are defined for the i th marker in the same way as p , \tilde{p} , μ , σ^2 and ϵ were defined for DNA pooling. Therefore, when the measurement errors are neglected and the sample size $n_0^{(i)}$ is used, the actual FDR in the presence of measurement errors can be approximated as in (9).

RESULTS

As in Risch and Teng [1998], we consider four genetic models: dominant model with $f_2 = f_1 = 0.04$, $f_0 = 0.01$, recessive model with $f_2 = 0.04$, $f_1 = f_0 = 0.01$, multiplicative model with $f_2 = 0.04$, $f_1 = 0.02$, $f_0 = 0.01$, and additive model with $f_2 = 0.04$,

$f_1 = 0.025$ and $f_0 = 0.01$. The population frequencies of allele A are taken as 0.05, 0.2, and 0.7, respectively. Using formulas (2) and (5), we calculate the sample size necessary to attain the power of $1 - \beta = 80\%$ at a statistical significance level of $\alpha = 5 \times 10^{-8}$. The same power and significance level were used by Risch and Merikangas [1996] in the discussion of genome-wide association studies. The results are summarized in Tables I and II for individual genotyping when the error rates e_1 and e_2 are equal and unequal, respectively, and Table III for DNA pooling. Tables I and II indicate that although genotyping errors lead to increased sample size for individual genotyping studies, the impacts are very moderate for the genotyping error rates normally encountered. The impacts also depend on the mode of inheritance. This is in contrast to DNA pooling as shown in Table III, where measurement errors have a strong impact on sample size requirements. For example, without exception, when the standard deviation ϵ is 3%, it is impossible to attain the desired power of 80%. This can be understood from considering formula (5), from which we can calculate the sample size n to obtain 80% power through DNA pooling as:

$$5.33\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} + 2\epsilon^2 + 0.84\sqrt{\frac{\sigma^2}{n}} + 2\epsilon^2 = \mu.$$

When the standard deviation of allele frequency estimates is 3%, the left-hand side of the above

TABLE I. Sample size required to detect associations based on individual genotyping for equal error rates $e_1 = e_2 = e^a$

	$e=0$	$e=0.005$	$e=0.01$	$e=0.03$
Dominant				
$p=0.05$	304	321	339	415
$p=0.20$	214	219	225	249
$p=0.70$	2,707	2,777	2,849	3,160
Recessive				
$p=0.05$	38,101	41,888	45,791	62,669
$p=0.20$	949	975	1,001	1,116
$p=0.70$	191	196	202	228
Multiplicative				
$p=0.05$	1,220	1,312	1,408	1,819
$p=0.20$	404	415	426	473
$p=0.70$	428	440	452	505
Additive				
$p=0.05$	714	764	814	1,034
$p=0.20$	322	330	339	376
$p=0.70$	647	664	682	761

^aSignificance level $\alpha=5 \times 10^{-8}$; power $1 - \beta=0.80$; Dominant model: $f_2=f_1=0.04$, $f_0=0.01$; Recessive model: $f_2=0.04$, $f_1=f_0=0.01$; Multiplicative model: $f_2=0.04$, $f_1=0.02$, $f_0=0.01$; Additive model: $f_2=0.04$, $f_1=0.025$, $f_0=0.01$.

TABLE II. Sample size required to detect associations based on individual genotyping for unequal error rates where $e_1 = 0.02$ and e_2^a

	$e_2=0$	$e_2=0.005$	$e_2=0.01$	$e_2=0.03$
Dominant				
$p=0.05$	311	327	343	409
$p=0.20$	220	224	228	245
$p=0.70$	2,911	2,933	2,955	3,046
Recessive				
$p=0.05$	38,916	42,619	46,362	61,734
$p=0.20$	974	995	1,015	1,100
$p=0.70$	209	210	212	218
Multiplicative				
$p=0.05$	1,247	1,336	1,425	1,793
$p=0.20$	416	424	432	466
$p=0.70$	464	468	471	485
Additive				
$p=0.05$	730	777	824	1,019
$p=0.20$	331	337	344	370
$p=0.70$	700	705	710	731

^aSignificance level $\alpha=5 \times 10^{-8}$; power $1 - \beta=0.80$; Dominant model: $f_2=f_1=0.04$, $f_0=0.01$; Recessive model: $f_2=0.04$, $f_1=f_0=0.01$; Multiplicative model: $f_2=0.04$, $f_1=0.02$, $f_0=0.01$; Additive model: $f_2=0.04$, $f_1=0.025$, $f_0=0.01$.

TABLE III. Sample size required to detect associations based on DNA pooling^a

	$\epsilon=0$	$\epsilon=0.005$	$\epsilon=0.01$	$\epsilon=0.03$
Dominant				
$p=0.05$	304	366	938	∞
$p=0.20$	214	226	272	∞
$p=0.70$	2,707	8,654	∞	∞
Recessive				
$p=0.05$	38,101	∞	∞	∞
$p=0.20$	949	1,271	∞	∞
$p=0.70$	191	202	247	∞
Multiplicative				
$p=0.05$	1,220	13,512	∞	∞
$p=0.20$	404	451	693	∞
$p=0.70$	428	485	808	∞
Additive				
$p=0.05$	714	1,386	∞	∞
$p=0.20$	322	351	479	∞
$p=0.70$	647	782	2,103	∞

^aSignificance level $\alpha=5 \times 10^{-8}$; power $1 - \beta=0.80$; Dominant model: $f_2=f_1=0.04$, $f_0=0.01$; Recessive model: $f_2=0.04$, $f_1=f_0=0.01$; Multiplicative model: $f_2=0.04$, $f_1=0.02$, $f_0=0.01$; Additive model: $f_2=0.04$, $f_1=0.025$, $f_0=0.01$.

formula is greater than $6.17 \times \sqrt{2}\epsilon = 6.17\sqrt{2} \times 0.03 \approx 0.2617$. However, it can be shown that μ is smaller than 0.22 for any p under all four genetic models considered in this article. Therefore, it is impossible to achieve the desired power to identify genetic associations.

The effects of genotyping and measurement errors on FDR are summarized in Tables IV and V for individual genotyping when the error rates e_1

TABLE IV. FDR for individual genotyping in the presence of genotyping errors and equal error rates $e_1=e_2=e$ when an association study is conducted using the sample size calculated assuming no errors ($\times 10^4$)^a

	$e=0.005$	$e=0.01$	$e=0.03$
Dominant			
$p=0.05$	6.69	7.23	10.74
$p=0.20$	6.45	6.67	7.87
$p=0.70$	6.44	6.65	7.78
Recessive			
$p=0.05$	7.01	8.03	16.10
$p=0.20$	6.43	6.63	7.69
$p=0.70$	6.46	6.70	8.02
Multiplicative			
$p=0.05$	6.83	7.57	12.80
$p=0.20$	6.43	6.64	7.75
$p=0.70$	6.45	6.67	7.88
Additive			
$p=0.05$	6.78	7.44	12.00
$p=0.20$	6.44	6.65	7.78
$p=0.70$	6.44	6.66	7.84

^aFDR in the absence of errors is 6.25×10^{-4} .

TABLE V. FDR for individual genotyping in the presence of genotyping errors and unequal error rates where $e_1=0.02$ and e_2 when an association study is conducted using the sample size calculated assuming no errors ($\times 10^4$)^a

	$e_2=0$	$e_2=0.005$	$e_2=0.01$	$e_2=0.03$
Dominant				
$p=0.05$	6.42	6.86	7.38	10.39
$p=0.20$	6.48	6.65	6.82	7.65
$p=0.70$	6.85	6.92	7.00	7.32
Recessive				
$p=0.05$	6.39	7.18	8.20	15.47
$p=0.20$	6.42	6.58	6.74	7.52
$p=0.70$	7.00	7.06	7.13	7.44
Multiplicative				
$p=0.05$	6.40	6.99	7.72	12.35
$p=0.20$	6.45	6.60	6.77	7.55
$p=0.70$	6.91	6.98	7.06	7.37
Additive				
$p=0.05$	6.41	6.94	7.59	11.59
$p=0.20$	6.46	6.61	6.78	7.58
$p=0.70$	6.89	6.96	7.03	7.35

^aFDR in the absence of errors is 6.25×10^{-4} .

and e_2 are equal and unequal, respectively, and Table VI for DNA pooling. Tables IV–VI show similar effects of errors on FDR in that although they generally only lead to moderate increases in FDR for individual genotyping, the impacts can be profound for DNA pooling (Table VI). In our calculations, we assumed that the prior probability of the alternative hypothesis being true, i.e.

TABLE VI. FDR for DNA pooling in the presence of measurement errors when an association study is conducted using the sample size calculated assuming no errors^a

	$\epsilon=0.005$	$\epsilon=0.01$	$\epsilon=0.03$
Dominant			
$p=0.05$	0.022	0.738	0.999
$p=0.20$	0.0016	0.014	0.966
$p=0.70$	0.181	0.973	1.000
Recessive			
$p=0.05$	1.000	1.000	1.000
$p=0.20$	0.018	0.683	0.999
$p=0.70$	0.0012	0.0061	0.882
Multiplicative			
$p=0.05$	0.744	0.996	1.000
$p=0.20$	0.0033	0.085	0.995
$p=0.70$	0.0025	0.045	0.991
Additive			
$p=0.05$	0.274	0.983	1.000
$p=0.20$	0.0024	0.043	0.990
$p=0.70$	0.0045	0.158	0.997

^aFDR in the absence of errors is 6.25×10^{-4} .

there is an association, is 0.0001. It can be noted that the impact of genotyping errors on FDR is almost irrespective of the underlying genetic models for individual genotyping (Fig. 1). Also, for unequal error rates, the results are qualitatively the same as those for equal error rates and the sample size required and FDR are between those corresponding to the two equal genotyping error rates (Tables I, II, IV, and V).

DISCUSSION

In this report, we have considered the impacts of genotyping errors in individual genotyping and measurement errors in DNA pooling on the sample size required to detect gene-disease associations and as well as their impacts on FDR for the case-control study. Our results showed that the impact of measurement errors is much greater on DNA pooling than that of genotyping errors on individual genotyping. Therefore, measurement error reduction in association studies is extremely important to make DNA pooling a useful strategy for studying gene-disease associations. One way to reduce measurement errors for DNA pooling is to form multiple pools, either using the same set of samples across pools or dividing all the samples into several pools [Hammick and Gastwirth, 1994; Barratt et al., 2002; Le Hellard et al., 2002; Mohlke et al., 2002; Sham et al., 2002; Visscher and Le Hellard, 2003]. For example, in Barratt et al.'s

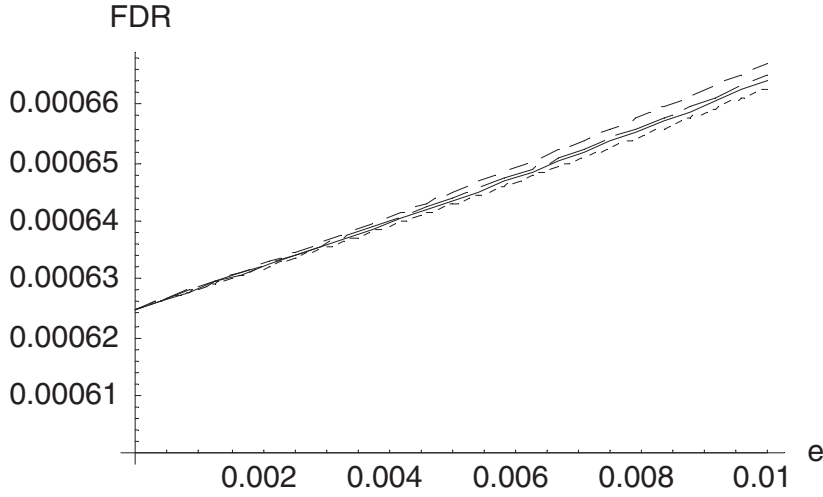


Figure 1. FDR in the presence of genotyping errors and equal error rates $e_1 = e_2 = e$ for individual genotyping under four different genetic models when the allele frequency is $p = 0.2$. From top to bottom, the dashed line represents the dominant model, the long dashed line represents the additive model, the solid line represents the multiplicative model, and the dotted line represents the recessive model.

[2002] study the authors considered various sources of error and showed that forming numerous small pools will lead to the most effective design. Also, Sham et al. [2002] pointed out that assuming other things are equal, measurement error variance will be reduced by a factor of k if k distinct pool pairs are formed. However, the results in Table III suggest that the number of pools needed to make the results comparable to

the null hypothesis H_0 when there are no genotyping errors, then neglecting errors will not affect type-I error in the presence of genotyping errors. However, if the population frequency p of allele A is known, then genotyping errors will indeed affect the type-I error. In this case, the actual type-I error due to neglecting genotyping errors and using the sample size $n_0^{(i)}$ in the absence of errors is

$$\begin{aligned}
 & P(\text{reject } H_0^{(i)} | H_0^{(i)} \text{ true}) \\
 &= P\left(\frac{n_A^{*(i)}/(2n_0^{(i)}) - n_U^{*(i)}/(2n_0^{(i)})}{\sqrt{p_i(1-p_i)/n_0^{(i)}}} > z_{\alpha_0^{(i)}} | H_0^{(i)} \text{ true}\right) \\
 &= P\left(\frac{\hat{p}_A^{(i)} - \hat{p}_U^{(i)}}{\sqrt{p_i^*(1-p_i^*)/[n_0^{(i)}(1-e_1^{(i)}-e_2^{(i)})^2]}} > \frac{z_{\alpha_0^{(i)}}\sqrt{p_i(1-p_i)}}{\sqrt{p_i^*(1-p_i^*)}} \Big| H_0^{(i)} \text{ true}\right) \\
 &= \Phi\left(-\frac{z_{\alpha_0^{(i)}}\sqrt{p_i(1-p_i)}}{\sqrt{p_i^*(1-p_i^*)}}\right).
 \end{aligned}$$

individual genotyping depends on specific genetic models. It is apparent that the optimal DNA pooling strategy needs further investigation.

It is interesting to note from formula (8) that for the case of individual genotyping, if the population frequency p of allele A is unknown and naturally estimated by $\hat{p}^* = (n_A^* + n_U^*)/(4n)$ since it is a consistent and unbiased estimator of p under

For example, when $p_i = 0.05$ and $e_1^{(i)} = e_2^{(i)} = 0.01$, the actual type-I error rate due to neglecting genotyping errors is 8.35 times of $\alpha_0^{(i)} = 5 \times 10^{-8}$.

Although our analyses focused on case-control data, our approach can also be applied to study family-based data through the methods proposed by Risch and Teng [1998]. We expect that the

effects of errors will be similar to what we have found for case-control studies.

The genotyping error models we assumed in this article may be simplistic. For example, the error patterns may depend on a particular genotype, instead of individual allele. However, we hypothesize that the qualitative nature of the results is similar under more general genotyping error models.

If the error rates for individual genotyping or DNA pooling are known, our methods can be used to adjust for genotyping or measurement errors to study associations between trait of interest and candidate markers. The error rates may be estimated from laboratory experiments or inferred from the distribution of the test statistics [Jawaid et al., 2002]. In addition, genotyping error rates can be estimated by genotyping a subset of samples multiple times (Zou and Zhao, unpublished results). When the error rates are estimated, the variability of the estimate of allele frequency will be increased [see Theorem A.1 of Gastwirth, 1987]. At this time, the sample size required can be obtained by replacing the true error rates by their estimated values in formula (2). From Tables I and II, this can result in oversampling or undersampling. As an example, we consider the case of dominant model with allele frequency $p = 0.05$. Assume that the true genotyping error rates are $e_1 = e_2 = 0.01$. If their estimates are $\hat{e}_1 = \hat{e}_2 = 0.005$, then from Table I, we see that the sample size we calculate is 321; if their estimates are $\hat{e}_1 = \hat{e}_2 = 0.03$, then the sample size we calculate is 415. However, the sample size actually required is 339. Note that even when the error rates are estimated, the test statistic t_{ind} for individual genotyping remains unchanged because it does not depend on the error rates, true or estimated. For DNA pooling, the test statistic t_{pool} will be affected by the estimate of error rate, and the conclusion on sample size required (and hence power) is similar to that for individual genotyping provided that the estimate of error rate is consistent.

This report has considered the effect of errors on association studies when individual genotyping and DNA pooling are used. Another major effect on case-control association studies is population stratification. Although several approaches have been developed to control for population stratification using genomic markers, all these approaches require the availability of individual genotyping data. Therefore, there is a need to develop effective methods to control for population stratification for DNA pooling studies to make this strategy robust against population

stratification. We plan to pursue this research in the future.

ACKNOWLEDGMENTS

The authors thank two reviewers for their valuable comments and suggestions.

REFERENCES

- Bader J, Bansal A, Sham P. 2001. Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *GeneScreen* 1:143–150.
- Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. 2002. Association testing by DNA pooling: an effective initial screen. *Proc Natl Acad Sci USA* 99:16871–16874.
- Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 61:734–747.
- Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. 2002. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 66:393–405.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300.
- Breen G, Harold D, Ralston S, Shaw D, St Clair D. 2000. Determining SNP allele frequencies in DNA pools. *Biotechniques* 28:464–466, 468, 470.
- Carroll RJ, Ruppert D, Stefanski LA. 1995. *Measurement error in nonlinear models*. London: Chapman & Hall.
- Cochran WG. 1968. Errors of measurements in statistics. *Technometrics* 10:637–666.
- Fuller WA. 1987. *Measurement error models*. New York: John Wiley & Sons.
- Gastwirth JL. 1987. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. *Stat Sci* 2:213–222.
- Germer S, Holland MJ, Higuchi R. 2000. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res* 10:258–266.
- Hammick PA, Gastwirth JL. 1994. Group testing for sensitive characteristics: extension to higher prevalence levels. *Int Stat Rev* 62:319–331.
- Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC. 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet* 107:488–493.
- 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 SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, Porteous DJ, Evans KL. 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.

- Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, Hollstein P, Boehnke M, Collins FS. 2002. High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc Natl Acad Sci USA* 99:16928–16933.
- Norton N, Williams NM, Williams HJ, Spurlock G, Kirov G, Morris DW, Hoogendoorn B, Owen MJ, O'Donovan MC. 2002. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet* 110:471–478.
- 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. *Nat Rev Genet* 3:862–871.
- Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 8:111–123.
- Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann Statist.* (in press).
- Visscher PM, Le Hellard S. 2003. Simple method to analyze SNP-based association studies using DNA pools. *Genet Epidemiol* 24:291–296.

APPENDIX: THE PROOF OF FDR BOUND GIVEN IN FORMULA (7)

As in the text, we consider a test of K hypotheses $H_0^{(i)}, i = 1, \dots, K$, and let $\eta_i = 1$ when $H_0^{(i)}$ is true and 0 otherwise, and $\xi_i = 1$ if we reject $H_0^{(i)}$ and 0 otherwise. Then it is easily seen that

$$\begin{aligned} E\left(\xi_i \mid \sum_{i=1}^K \xi_i > 0\right) &= P\left(\xi_i = 1 \mid \sum_{i=1}^K \xi_i > 0\right) \\ &= \frac{P(\xi_i = 1)}{P\left(\sum_{i=1}^K \xi_i > 0\right)} \\ &= \frac{P(\text{reject } H_0^{(i)})}{P\left(\sum_{i=1}^K \xi_i > 0\right)} \end{aligned}$$

and

$$\begin{aligned} \text{FDR} &\approx \frac{\sum_{i=1}^K P\left(\text{reject } H_0^{(i)} \mid H_0^{(i)} \text{ true}\right) P\left(H_0^{(i)} \text{ true}\right)}{\sum_{i=1}^K P\left(\text{reject } H_0^{(i)}\right)} \\ &= \frac{\sum_{i=1}^K P\left(\text{reject } H_0^{(i)} \mid H_0^{(i)} \text{ true}\right) P\left(H_0^{(i)} \text{ true}\right)}{\sum_{i=1}^K \left[P\left(\text{reject } H_0^{(i)} \mid H_0^{(i)} \text{ true}\right) P\left(H_0^{(i)} \text{ true}\right) + P\left(\text{reject } H_0^{(i)} \mid H_1^{(i)} \text{ true}\right) P\left(H_1^{(i)} \text{ true}\right) \right]} \end{aligned}$$

$$\begin{aligned} &E\left(\eta_i \xi_i \mid \sum_{i=1}^K \xi_i > 0\right) \\ &= P\left(\eta_i \xi_i = 1 \mid \sum_{i=1}^K \xi_i > 0\right) \\ &= P\left(\eta_i = 1, \xi_i = 1 \mid \sum_{i=1}^K \xi_i > 0\right) \\ &= \frac{P(\eta_i = 1, \xi_i = 1)}{P\left(\sum_{i=1}^K \xi_i > 0\right)} \\ &= \frac{P(\eta_i = 1 \mid \xi_i = 1) \cdot P(\xi_i = 1)}{P\left(\sum_{i=1}^K \xi_i > 0\right)} \\ &= \frac{P\left(H_0^{(i)} \text{ true} \mid \text{reject } H_0^{(i)}\right) \cdot P\left(\text{reject } H_0^{(i)}\right)}{P\left(\sum_{i=1}^K \xi_i > 0\right)} \end{aligned}$$

Therefore, using the first-order approximation to FDR, we have

$$\begin{aligned} \text{FDR} &\approx \frac{\sum_{i=1}^K E\left(\eta_i \xi_i \mid \sum_{i=1}^K \xi_i > 0\right)}{\sum_{i=1}^K E\left(\xi_i \mid \sum_{i=1}^K \xi_i > 0\right)} \\ &= \frac{\sum_{i=1}^K P\left(H_0^{(i)} \text{ true} \mid \text{reject } H_0^{(i)}\right) P\left(\text{reject } H_0^{(i)}\right)}{\sum_{i=1}^K P\left(\text{reject } H_0^{(i)}\right)}. \end{aligned} \tag{A.1}$$

Note that formula (A.1) for multiple tests applies to independence replicates. For genomewide studies, the tests will be positively correlated and hence the bound for FDR will be conservative.

When $K = 1$, formula (6) leads to

$$\begin{aligned} \text{FDR} &= E(\eta_1 \mid \xi_1 > 0) = P(\eta_1 = 1 \mid \xi_1 = 1) \\ &= P\left(H_0^{(1)} \text{ true} \mid \text{reject } H_0^{(1)}\right). \end{aligned} \tag{A.2}$$

This shows that for a single test, formula (A.1) holds exactly, and FDR is the posterior probability of the null hypothesis being true [see also Storey, 2003]. From formulas (A.1) and (A.2), we see that the overall FDR for all K tests is, in fact, a weighted average of FDR for each single test by its proportionality of rejecting the null hypothesis.

Formula (A.2) can also be written in the form

where $H_1^{(i)}$ denotes the i th alternative hypothesis, $i = 1, \dots, K$. Hence, formula (7) is proved.