Genetics: Published Articles Ahead of Print, published on April 19, 2006 as 10.1534/genetics.105.042648

Two-stage designs in case-control association analysis

Yijun Zuo[1], Guohua Zou[2], and Hongyu Zhao[3,*]

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P. R. China

[3] Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

*_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

Email: hongyu.zhao@yale.edu.

Abstract

DNA pooling is a cost effective approach for collecting information on marker allele frequency in genetic studies. It is often suggested as a screening tool to identify a subset of candidate markers from a very large number of markers to be followed up by more accurate and informative individual genotyping. In this paper, we investigate several statistical properties and design issues related to this two-stage design, including the selection of the candidate markers for second stage analysis, statistical power of this design, and the probability that truly disease-associated markers are ranked among the top after second stage analysis. We have derived analytical results on the proportion of markers to be selected for second stage analysis. For example, to detect disease-associated markers with an allele frequency difference of 0.05 between the cases and controls through an initial sample of 1000 cases and 1000 controls, our results suggest that when the measurement errors are small (0.005), about 3% of the markers should be selected. For the statistical power to identify disease-associated markers, we find that the measurement errors associated with DNA pooling have little effect on its power. This is in contrast to the one-stage pooling scheme where measurement errors may have large effect on statistical power. As for the probability that the disease-associated markers are ranked among the top in the second stage, we show that there is a high probability that at least one disease-associated marker is ranked among the top when the allele frequency differences between the cases and controls are not smaller than 0.05 for reasonably large sample sizes, even though the errors associated with DNA pooling in the first stage is not small.

Therefore, the two-stage design with DNA pooling as a screening tool offers an efficient strategy in genome-wide association studies, even when the measurement errors associated with DNA pooling are non-negligible. For any disease model, we find that all the statistical results essentially depend on the population allele frequency and the allele frequency differences between the cases and controls at the disease-associated markers. The general conclusions hold whether the second stage uses an entirely independent sample or includes both the samples used in the first stage as well as an independent set of samples.

Key Words: DNA pooling; individual genotyping; measurement errors; power; two-stage design.

Running Title: Two-stage designs for association studies

Introduction

Genome-wide case-control association study is a promising approach to identifying disease genes (Risch 2000). For a specific marker, allele frequency difference between cases and controls may indicate potential association between this marker and disease, although other factors (e.g. population stratification) may account for the observed difference. Allele frequencies among the cases and controls can be obtained either through individual genotyping or DNA pooling. Although individual genotyping provides more accurate estimates of allele frequencies and allows for the inference of haplotypes and the study of genetic interactions, DNA pooling can be more cost effective in genome-wide association studies as individual genotyping needs to collect data from hundreds of thousands markers for each person.

In the absence of measurement errors associated with DNA pooling, there would be no difference between using DNA pooling or individual genotyping for the estimation of allele frequency. However, one major limitation of the current DNA pooling technologies is indeed the errors associated with measuring allele frequencies in the pooled samples. Recent research suggests that for a given pooled DNA 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). LeHellard et al. (2002) reported that using the SNaPshot ^{TM} Method, which is based on allele-specific extension or minisequencing from a primer adjacent to the site of the SNP, the standard deviation ranged

from 1% to 4% depending on the specific markers being tested. Our recent studies have found that the errors of this magnitude may have a large effect on the power of case-control association studies using DNA pooling as the sole source for genotyping (see Zou and Zhao 2004 for unrelated population samples and Zou and Zhao 2005 for family samples). Therefore, a two-stage design where DNA pooling is used as a screening tool followed by individual genotyping for validation in an expanded or independent sample may offer an attractive strategy to balance power and cost (Barcellos et al. 1997, Bansal et al. 2002, Barratt et al. 2002, Sham et al. 2002). In such a design, the first stage evaluates a very large number (e.g. one million) of markers using DNA pooling, and only the most promising ones are selected and studied in the second stage through individual genotyping. Similar two-stage designs have been considered by Elston (1994) and Elston et al. (1996) in the context of linkage analysis, and by Satagopan et al. (2002, 2003, 2004) in the context of association studies. However, these studies primarily assumed that individual genotyping is used in both stages, which may not be as cost-effective as using DNA pooling in the first stage. Moreover, errors associated with genotyping have never been considered in the literature.

When DNA pooling is used as a screening tool in the first stage, the following issues need to be addressed:

(i) How many markers should be chosen after the first stage so that there is a high probability that all or some of the disease-associated markers are included in the individual genotyping (second) stage?

(ii) What is the statistical power that a disease-associated marker is identified when the overall false positive rate is appropriately controlled for?

(iii) When the primary goal is to ensure that some of the disease-associated markers are ranked among the top L markers after the two-stage analysis, what is the probability that at least one of the disease-associated markers is ranked among the top?

The objective of this paper is to provide answers to these practical questions to facilitate the most efficient use of the two-stage design strategy where DNA pooling is used. In genetic studies, the sample in the first stage can be expanded with a set of new samples in the second stage analysis, or the second stage may only involve a new set of samples for individual genotyping, so both these strategies will be considered in our article. We hope that the principles thus learned will provide an effective and practical guide to genetic association studies.

This paper is organized as follows. We will first present our analytical results to treat the above three problems, and then conduct numerical calculations under various scenarios to gain an overview and insights on these design issues. Finally, some future research problems are discussed.

Methods

Genetic models

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

$q = 1 - p$, respectively. For simplicity, 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. When the candidate marker is associated with disease, we assume that the penetrance is 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. Under this genetic model, the probabilities of having *k* copies of *A* among the cases, $m_k = P(X_i = k)$, and those among the controls, $m_k' = P(Y_i = k)$, are

$$m_0 = \frac{q^2 f_0}{p^2 f_2 + 2pqf_1 + q^2 f_0},$$

$$m_1 = \frac{2pqf_1}{p^2 f_2 + 2pqf_1 + q^2 f_0},$$

$$m_2 = \frac{p^2 f_2}{p^2 f_2 + 2pqf_1 + q^2 f_0},$$

$$m_0' = \frac{q^2(1-f_0)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)},$$

$$m_1' = \frac{2pq(1-f_1)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)},$$

$$m_2' = \frac{p^2(1-f_2)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)}.$$

One-stage designs

For useful reference, we first formulate the test statistics and derive statistical power based on a one-stage design using either individual genotyping or DNA pooling. These can be considered as special cases or direct extensions of the results in Zou and Zhao (2004).

(a) Individual genotyping

For individual genotyping, let n_A and n_U denote the observed numbers of allele A in the case group and control group, respectively, p_A and p_U denote the population allele frequencies of allele A in these two groups, and $\hat{p}_A$ and $\hat{p}_U$ denote their maximum likelihood estimates, where $\hat{p}_A = n_A/(2n)$ and $\hat{p}_U = n_U/(2n)$.

Under the null hypothesis of no association between the candidate marker and disease status, $E(\hat{p}_A - \hat{p}_U) = 0$, and $V(\hat{p}_A - \hat{p}_U) = pq/n$. On the other hand, under the genetic model introduced above,

$$E(\hat{p}_A - \hat{p}_U) = m_2 + \frac{1}{2}m_1 - m_2' - \frac{1}{2}m_1' \equiv \mu,$$

and

$$V(\hat{p}_A - \hat{p}_U) = \frac{1}{4n}\left[4m_2 + m_1 - (2m_2 + m_1)^2 + 4m_2' + m_1' - (2m_2' + m_1')^2\right]$$
$$\equiv \frac{\sigma^2}{n}.$$

The statistic to test genetic association between the candidate marker and disease is

$$t_{ind} = \frac{\hat{p}_A - \hat{p}_U}{\sqrt{\hat{p}(1-\hat{p})/n}},$$

where $\hat{p} = (n_A + n_U)/(4n)$.

Consider a one-sided test and use a significance level of α, the power of the test statistic

t_{ind} is

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

where $\tilde{p}=\mu/2+m_2'+m_1'/2$ is the expected frequency of allele *A* under the genetic model, Φ is the cumulative standard normal distribution function, and z_α is the upper 100α th percentile of the standard normal distribution.

(b) DNA pooling

For DNA pooling, we consider *m* pools of cases and *m* pools of controls each having size *s* such that *n=ms.* We assume the following model relating the observed allele frequencies estimated from the pooled samples to the true frequencies of allele *A* in the samples:

$$\hat{p}_{Ai}^{pool}=\frac{X_{i1}+\cdots+X_{is}}{2s}+u_i,$$

$$\hat{p}_{Ui}^{pool}=\frac{Y_{i1}+\cdots+Y_{is}}{2s}+v_i,$$

where X_{ij} denotes the number of allele *A* carried by the *j*th individual in the *i*th case group, and Y_{ij} is defined similarly (*i=1,…,m*; *j=1,…,s*), u_i and v_i are disturbances with mean 0 and variance ε^2 and are assumed to be independent and normally distributed. Define

$$\hat{p}_A^{pool}=\frac{1}{m}\sum_{i=1}^{m}\hat{p}_{Ai}^{pool},$$

and

$$\hat{p}_U^{pool}=\frac{1}{m}\sum_{i=1}^{m}\hat{p}_{Ui}^{pool}.$$

Under the null hypothesis of no association, $E(\hat{p}_A^{pool}-\hat{p}_U^{pool})=0$, and $V(\hat{p}_A^{pool}-\hat{p}_U^{pool})=pq/n+2\varepsilon^2/m$. On the other hand, under the genetic model introduced

above,

$$E(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = \mu,$$

and

$$V(\hat{p}_A^{pool} - \hat{p}_U^{pool}) = \frac{\sigma^2}{n} + \frac{2\varepsilon^2}{m}.$$

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

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

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

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

$$\Phi\left(\frac{-z_\alpha\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \frac{2\varepsilon^2}{m}} + \mu}{\sqrt{\frac{\sigma^2}{n} + \frac{2\varepsilon^2}{m}}}\right).$$

Two-stage designs

(a) *How many markers should be selected after the pooling stage?*

In the first stage, i.e., the DNA pooling stage, we consider *m* pools of cases and *m* pools of controls each having size *s* such that *n* = *ms*. The main objective for the first stage is to select the most promising markers based on pooled DNA data to follow up in the second stage in order to reduce the overall cost. Therefore, the following problem should be addressed: how many of the *M* markers initially screened should be selected for

second-stage analysis so that the probability that the disease-associated markers are selected is high, e.g. 90%? For simplicity, we assume that the associated markers are independent. Let the desired number of markers be M_1. As in Satagopan et al. (2002, 2004), we choose those markers which have the largest test statistic.

For markers not associated with disease, the test statistic can be approximated by

$$t_{pool} = \frac{\sqrt{\frac{pq}{n}}\xi_0 + w}{\sqrt{\frac{pq}{n} + \frac{2\varepsilon^2}{m}}},$$

where $\xi_0 \sim N(0,1)$, $w = \bar{u} - \bar{v} \sim N\left(0, \frac{2\varepsilon^2}{m}\right)$, $\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i$, $\bar{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$, and ξ_0 and w are mutually independent. Whereas for markers associated with disease through the genetic model introduced above, the test statistic can be approximated by:

$$t_{pool} = \frac{\sqrt{\frac{\sigma^2}{n}}\xi_1 + w}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \frac{2\varepsilon^2}{m}}},$$

where $\xi_1 \sim N(\sqrt{n}\mu/\sigma, 1)$, and ξ_1 and w are mutually independent.

Let $t^{(T)}_{pool,1}, \cdots, t^{(T)}_{pool,K}$ be the test statistics corresponding to the K disease-associated markers, $t^{(N)}_{pool,1}, \cdots, t^{(N)}_{pool,M-K}$ be those corresponding to the $M-K$ null markers, and $t^{(N)}_{pool,(1)} \geq \cdots \geq t^{(N)}_{pool,(M-K)}$ are the corresponding ordered test statistics. Let $P_{i_1,\ldots,i_{K_1}}$ denote the probability that the specified K_1 of the K truly associated markers are among the top M_1 markers. Furthermore, denote

$$Z_0 = \min\left\{t^{(T)}_{pool,i_1}, \cdots, t^{(T)}_{pool,i_{K_1}}\right\},$$

and

$$Z^* = \max\left\{t^{(T)}_{pool,j},\ j \in E \equiv \{1,...,K\} \setminus \{i_1,...,i_{K_1}\}\right\}.$$

Note that $t^{(T)}_{pool,j} \sim N\left(\theta_{pool,j}, \lambda^2_{pool,j}\right),\ j = 1,...,K$, where

$$\theta_{pool,j} = \frac{\mu_j}{\sqrt{\tilde{p}_j(1-\tilde{p}_j)/n + 2\varepsilon^2/m}},$$

$$\lambda^2_{pool,j} = \frac{\sigma_j^2/n + 2\varepsilon^2/m}{\tilde{p}_j(1-\tilde{p}_j)/n + 2\varepsilon^2/m},$$

and $\tilde{p}_j$, μ_j and σ_j^2 are defined as $\tilde{p}$, μ and σ^2 with allele frequency p_j, penetrances $f_{2,j}$, $f_{1,j}$ and $f_{0,j}$ at the truly associated marker j in place of p, f_2, f_1 and f_0, respectively, $j = 1,...,K$. In addition, $t^{(N)}_{pool,j} \sim N(0,1),\ j = 1,...,M-K$. For convenience, we denote the distribution and density functions of $t^{(T)}_{pool,j}$ by $F_j(x)$ and $f_j(x)$, and the distribution and density functions of $t^{(N)}_{pool,j}$ by $\Phi(x)$ and $\phi(x)$, respectively. Then it can be shown that the joint density function of $\left(Z_0, Z^*\right)$ is

$$g(z_0, z^*) = g_{Z_0}(z_0) \cdot g_{Z^*}(z^*),$$

where

$$g_{Z_0}(z_0) = \prod_{j=1}^{K_1}\left[1 - F_{i_j}(z_0)\right] \cdot \sum_{j=1}^{K_1} \frac{f_{i_j}(z_0)}{1 - F_{i_j}(z_0)},$$

and

$$g_{Z^*}(z^*) = \prod_{j \in E} F_j(z^*) \cdot \sum_{j \in E} \frac{f_j(z^*)}{F_j(z^*)}.$$

Moreover, the joint density of $\left(t^{(N)}_{pool,(M_1-K_1+1)}, t^{(N)}_{pool,(M_1-K_1)}\right)$ is

$$g_0(u,v) = \frac{(M-K)!}{\left[(M-K)-(M_1-K_1)-1\right]!(M_1-K_1-1)!}\Phi^{(M-K)-(M_1-K_1)-1}(u)\left[1-\Phi(v)\right]^{M_1-K_1-1}\phi(u)\phi(v),$$

$$u < v.$$

Hence,

$$P_{i_1,\ldots,i_{K_1}} = P\left(Z_0 > t^{(N)}_{pool,(M_1-K_1+1)}, Z^* < Z_0, Z^* < t^{(N)}_{pool,(M_1-K_1)}\right)$$
$$= \iint_{u<v} P\left(Z_0 > u, Z^* < Z_0, Z^* < v\right) \cdot g_0(u,v)dudv, \quad (1)$$

where

$$P\left(Z_0 > u, Z^* < Z_0, Z^* < v\right) = P\left(u < Z_0 < v, Z^* < Z_0\right) + P\left(Z_0 \geq v, Z^* < v\right)$$
$$= \int_u^v dz_0 \int_{-\infty}^{z_0} g(z_0, z^*)dz^* + \int_v^{\infty} g_{Z_0}(z_0)dz_0 \cdot \int_{-\infty}^{v} g_{Z^*}(z^*)dz^* .$$

Therefore, the probability that K_1 of the K disease-associated markers are among the top M_1 markers is given by

$$P_1(K_1) = \sum_{i_1<\cdots<i_{K_1}} P_{i_1,\ldots,i_{K_1}} . \quad (2)$$

From this expression, we can determine the value of M_1 such that $P_1(K_1)$ is higher or equal to a given level, e.g. 90%.

For a given M_1, let ζ denote the number of disease-associated markers included in the top M_1 markers, then its expectation is $E(\zeta) = \sum_{l=0}^{K} l \cdot P(\zeta = l) = \sum_{l=0}^{K} l \cdot P_1(l)$. Therefore, we can determine the value of M_1 through this formula such that the average number of disease-associated markers included in the top M_1 markers is K_1, i.e. K_1 disease-associated markers are selected on average.

The above formulas (1) and (2) are exact but somewhat complicated. In the following, we derive their asymptotic expressions so that we can obtain simpler analytical results. It is easy to see that we need only to consider formula (1).

For a fixed proportion p_0, let λ_0 denote the normal distribution quantile corresponding to p_0, that is, $\int_{-\infty}^{\lambda_0}\phi(x)dx = p_0$. Then from the asymptotic property of order statistics, we have

$$t^{(N)}_{pool,((M-K)-[(M-K)p_0]+1)} \xrightarrow{a.s.} \lambda_0, \tag{3}$$

and

$$t^{(N)}_{pool,((M-K)-[(M-K)p_0])} \xrightarrow{a.s.} \lambda_0, \tag{4}$$

when $M-K$ tends to infinity, where $[t]$ denotes the integer part of t, and $\xrightarrow{a.s.}$ denotes convergence almost sure.

If we write $M_1 = K_1 + (M-K) - [(M-K)p_0]$, then we have

$$\begin{aligned} P_{i_1,\ldots,i_{K_1}} &= P\left(Z_0 > t^{(N)}_{pool,(M_1-K_1+1)}, Z^* < Z_0, Z^* < t^{(N)}_{pool,(M_1-K_1)}\right) \\ &\to P\left(Z_0 > \lambda_0, Z^* < Z_0, Z^* < \lambda_0\right) \\ &= P\left(Z_0 > \lambda_0\right) \cdot P(Z^* < \lambda_0), \end{aligned} \tag{5}$$

where

$$P(Z_0 > z_0) = \prod_{j=1}^{K_1}\left[1 - F_{i_j}(z_0)\right], \tag{6}$$

and

$$P\left(Z^* < z^*\right) = \prod_{j\in E} F_j(z^*).$$

Note that the total number of markers M is usually extremely large, the number of disease-associated markers K is extremely small compared to *M*, and

$$(1-p_0) + (K_1 - K(1-p_0))/M \le M_1/M < (1-p_0) + (K_1 - K(1-p_0)+1)/M.$$

Therefore, taking M_1 top markers is equivalent to taking the top markers in the proportion of $q_0 = 1-p_0$.

In particular, when the number of disease-associated markers is $K=1$, we can obtain an analytical expression for the selected proportion q_0 necessary to attain the desired probability that the disease-associated marker is selected. In fact, when $K=1$, from formulas (5) and (6), we have

$$P_1 = P(Z_0 > \lambda_0) = 1 - F_1(\lambda_0)$$

$$= 1 - \Phi\left(\frac{\lambda_0 \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{2\varepsilon^2}{m}} - \mu_1}{\sqrt{\frac{\sigma_1^2}{n} + \frac{2\varepsilon^2}{m}}} \right).$$

Therefore, if we require the probability that the truly associated marker is included in the selected subset from the first stage is at least p_0^*, i.e., $P_1 \geq p_0^*$, then

$$\Phi\left(\frac{\lambda_0 \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{2\varepsilon^2}{m}} - \mu_1}{\sqrt{\frac{\sigma_1^2}{n} + \frac{2\varepsilon^2}{m}}} \right) \leq 1 - p_0^* = \Phi(\lambda_0^*),$$

where λ_0^* is the normal distribution quantile corresponding to $1-p_0^*$. Clearly, the above formula is equivalent to

$$\lambda_0 \leq \frac{\lambda_0^* \sqrt{\frac{\sigma_1^2}{n} + \frac{2\varepsilon^2}{m}} + \mu_1}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{2\varepsilon^2}{m}}} \equiv U_0^*.$$

So the proportion q_0 should satisfy $q_0 \geq \Phi(-U_0^*)$. Therefore, a conservative selection of the proportion q_0 is the maximum of $\Phi(-U_0^*)$ over various genetic models and allele frequencies.

It should be noted that the above selection approach for markers is through comparing the

values of the test statistics at all the markers and no statistical inference is conducted. If statistical tests are performed to select the promising markers, then one would keep those markers showing stronger statistical significance in the first stage. However, the two methods are actually asymptotically equivalent. This is because, if we take $\lambda_0 = z_{\alpha_1}$ (where z_{α_1} is the upper $100\alpha_1$th percentile of the standard normal distribution corresponding to the significance level α_1 for each marker tested in the first stage), that is, $q_0 = \alpha_1$, which means that the selected proportion of markers is the same as the significance level for testing each marker in the first stage, then the asymptotic probability of the specified K_1 of K truly associated markers being selected given in formula (5) is in fact the statistical power of detecting the specified K_1 of K truly associated markers. So for the case of independent markers, selecting the markers through comparing the values of their test statistics is asymptotically equivalent to selecting the markers through statistical tests, a conclusion similar to that of Satagopan et al. (2004) who considered individual genotyping in the first stage. In other words, the selection approach based on statistical tests is the limiting case of that based on comparing the values of test statistics at the markers when the number of total markers is very large.

(b) *The statistical power of the two-stage design*

After a set of promising markers are identified through DNA pooling, these markers will be individually genotyped in the second stage. In this subsection, we first derive the statistical power of the two-stage design to detect the disease-associated markers. In the next subsection, we will investigate the possibility of at least one disease-associated marker

being ranked among the top after the second stage. In addition to the $2n$ individuals used in the pooling stage, we will also consider an additional sample of size $2n_a$. Under the null hypothesis H_0, i.e. the marker is not associated with disease, the test statistic for markers tested in the second stage can be written approximately as

$$t_{ind} = \sqrt{\frac{n}{n+n_a}} \cdot \xi_0 + \sqrt{\frac{n_a}{n+n_a}} \cdot \eta_0 ,$$

where $\eta_0 \sim N(0,1)$ and η_0 is independent of ξ_0 and *w*, which were defined above in the discussion of pooled DNA analysis.

Similarly, for markers associated with disease under the genetic model introduced above, the test statistic for markers tested in the second stage can be written approximately as

$$t_{ind} = \frac{\sqrt{\frac{n}{n+n_a}}\sigma \cdot \xi_1 + \sqrt{\frac{n_a}{n+n_a}}\sigma \cdot \eta_1}{\sqrt{\tilde{p}(1-\tilde{p})}},$$

where $\eta_1 \sim N(\sqrt{n_a}\mu/\sigma, 1)$, and η_1 is independent of ξ_1 and *w*, which were defined above in the discussion of pooled DNA analysis.

Under the null hypothesis of no association, (t_{pool}, t_{ind}) has a joint bivariate normal distribution $N(0, \sum_0)$, where

$$\sum_0 = \begin{pmatrix} 1 & \frac{\sqrt{pq/(n+n_a)}}{\sqrt{pq/n + 2\varepsilon^2/m}} \\ \frac{\sqrt{pq/(n+n_a)}}{\sqrt{pq/n + 2\varepsilon^2/m}} & 1 \end{pmatrix}.$$

Under the alternative hypothesis H_1, (t_{pool}, t_{ind}) has a joint bivariate normal distribution $N(\tilde{\mu}, \sum_1)$, where

$$\tilde{\mu} = \left(\frac{\mu}{\sqrt{\tilde{p}(1-\tilde{p})/n + 2\varepsilon^2/m}} \quad \frac{\mu}{\sqrt{\tilde{p}(1-\tilde{p})/(n+n_a)}} \right),$$

and

$$\sum\nolimits_1 = \begin{pmatrix} \dfrac{\sigma^2/n + 2\varepsilon^2/m}{\tilde{p}(1-\tilde{p})/n + 2\varepsilon^2/m} & \dfrac{\sigma^2}{\sqrt{(n+n_a)\tilde{p}(1-\tilde{p})} \cdot \sqrt{\tilde{p}(1-\tilde{p})/n + 2\varepsilon^2/m}} \\ \dfrac{\sigma^2}{\sqrt{(n+n_a)\tilde{p}(1-\tilde{p})} \cdot \sqrt{\tilde{p}(1-\tilde{p})/n + 2\varepsilon^2/m}} & \dfrac{\sigma^2}{\tilde{p}(1-\tilde{p})} \end{pmatrix}.$$

For a given sample size n and significance level α_1 or power $1-\beta_1$ in the first stage (or a given proportion of markers to be selected for second-stage analysis), we can determine a critical value k_1 by solving $\alpha_1 = P(t_{pool} > k_1 \mid H_0)$ or $1-\beta_1 = P(t_{pool} > k_1 \mid H_1)$. Then for the overall significance level α for testing M markers and an additional sample of size n_a, we can determine the critical value k_2 in the second stage by solving

$$\alpha / M = P\left((t_{pool} > k_1) \bigcap (t_{ind} > k_2) \mid H_0\right) = \int_{k_1}^{\infty}\int_{k_2}^{\infty} h_0(x,y)\,dxdy,$$

where $h_0(x,y)$ is the density function of (t_{pool}, t_{ind}) under H_0, which is given by

$$h_0(x,y) = \frac{1}{2\pi\sqrt{|\sum_0|}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \sum\nolimits_0^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right\},$$

where $|\sum_0|$ is the determinant of the matrix $\sum_0$, and $\sum_0^{-1}$ is the inverse of $\sum_0$.

The probability that a disease-associated marker is identified by the two-stage design is then given by

$$1-\beta = P\left((t_{pool} > k_1) \bigcap (t_{ind} > k_2) \mid H_1\right) = \int_{k_1}^{\infty}\int_{k_2}^{\infty} h_1(x,y)\,dxdy,$$

where $h_1(x,y)$ is the density function of (t_{pool}, t_{ind}) under H_1, which is given by

$$h_1(x,y) = \frac{1}{2\pi\sqrt{|\sum_1|}} \exp\left\{ -\frac{1}{2} \left((x,y) - \tilde{\mu} \right) \sum\nolimits_1^{-1} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \tilde{\mu}' \right) \right\}.$$

In the above two-stage design, the sample in the first stage is re-used in the second stage, and this introduces correlation between the two test statistics, t_{pool} and t_{ind}. Therefore, we will call this two-stage scheme *the two-stage dependent design* in the following discussion. On the other hand, we may use two separate samples in the two stages with one sample used for screening and another independent sample used for individual genotyping. In this scenario, the two test statistics, t_{pool} and t_{ind}, are independent. Hereafter we call such a two-stage scheme *the two-stage independent design*. For the two-stage independent design, the type-I error rate and power are simply the products of those in both stages. That is,

$$P\left((t_{pool} > k_1)\bigcap(t_{ind} > k_2) \mid H_0\right) = P\left(t_{pool} > k_1 \mid H_0\right)\cdot P\left(t_{ind} > k_2 \mid H_0\right),$$

and

$$P\left((t_{pool} > k_1)\bigcap(t_{ind} > k_2) \mid H_1\right) = P\left(t_{pool} > k_1 \mid H_1\right)\cdot P\left(t_{ind} > k_2 \mid H_1\right).$$

(c) The chance of at least one marker associated with disease being ranked among the top L markers after individual genotyping

We suppose that, among the M_1 markers selected from the first stage, there are K_1 markers associated with disease and $M_1 - K_1$ null markers. Without loss of generality, we assume that they are $t^{(T)}_{pool,1},\cdots,t^{(T)}_{pool,K_1}$ and $t^{(N)}_{pool,1},\cdots,t^{(N)}_{pool,M_1-K_1}$, respectively. In this case, let Z_0 and Z^* denote $\min\left\{t^{(T)}_{pool,1},\cdots,t^{(T)}_{pool,K_1}\right\}$ and $\max\left\{t^{(T)}_{pool,K_1+1},\cdots,t^{(T)}_{pool,K}\right\}$, respectively. Let $t^{(T)}_{ind,j}$ $(j=1,...,K_1)$ be the test statistic for the *j*th truly associated marker, $t^{(N)}_{ind,j}$ $(j=1,...,M_1-K_1)$ be the test statistic for the *j*th null marker in the second stage, and $t^{(T)}_{ind,(1)} \geq \cdots \geq t^{(T)}_{ind,(K_1)}$ and $t^{(N)}_{ind,(1)} \geq \cdots \geq t^{(N)}_{ind,(M_1-K_1)}$ be their order statistics. Then in the second stage, the probability that none of the truly associated markers are ranked among the top L

markers is

$$P_2' = P\left(X < Y \middle| Z_0 > U, Z^* < Z_0, Z^* < V, V > U\right), \tag{7}$$

where

$$X = \max\left\{t_{ind,1}^{(T)}, \cdots, t_{ind,K_1}^{(T)}\right\},$$
$$Y = t_{ind,(L)}^{(N)},$$
$$U = \max\left\{t_{pool,M_1-K_1+1}^{(N)}, \cdots, t_{pool,M-K}^{(N)}\right\},$$

and

$$V = \min\left\{t_{pool,1}^{(N)}, \cdots, t_{pool,M_1-K_1}^{(N)}\right\}.$$

Like formula (1), an exact expression for calculating the probability P_2' can be derived (Appendix). Therefore, the probability that at least one truly associated marker is ranked among the top L markers is obtained by $P_2 = 1 - P_2'$. Because the exact formula is quite complicated, we provide an approximate one below to simplify the calculation of this probability. First note that $t_{ind,j}^{(T)} \sim N\left(\theta_{ind,j}, \lambda_{ind,j}^2\right)$, where

$$\theta_{ind,j} = \frac{\mu_j}{\sqrt{\tilde{p}_j(1-\tilde{p}_j)/(n+n_a)}},$$

and

$$\lambda_{ind,j}^2 = \frac{\sigma_j^2}{\tilde{p}_j(1-\tilde{p}_j)},$$

$j = 1,..., K_1$. We denote the distribution function of $t_{ind,j}^{(T)}$ by $G_j(x)$. Also, let $H_j^{(T)}(x, y)$ denote the joint distribution function of $\left(t_{pool,j}^{(T)}, t_{ind,j}^{(T)}\right)$, $j = 1,..., K_1$.

Now for a fixed proportion p_0', we have

$$t_{ind,((M_1-K_1)-[(M_1-K_1)p_0'])}^{(N)} \approx \lambda_0',$$

when $M_1 - K_1$ is large, where λ_0' is a normal distribution quantile corresponding to p_0', that is, $\int_{-\infty}^{\lambda_0'} \phi(x)dx = p_0'$, and $[t]$ denotes the integer part of t as before. Denote $L = M_1 - K_1 - [(M_1 - K_1)p_0']$, then $t_{ind,(L)}^{(N)} \approx \lambda_0'$. Therefore, we substitute λ_0' for $Y = t_{ind,(L)}^{(N)}$ in formula (7). This means that as long as $X < \lambda_0'$, we think no truly associated markers are ranked among the top L markers, regardless of the null markers chosen from the first stage. On the other hand, we have demonstrated that in the first stage, selecting a proportion q_0 of the markers through comparing the values of the test statistics is asymptotically equivalent to selecting the significant markers through statistical tests with significance level $\alpha_1 (= q_0)$, that is, the critical value can be taken as λ_0. Therefore, we obtain

$$P_2' \approx P\left(X < \lambda_0' \middle| Z_0 > \lambda_0, Z^* < \lambda_0, V > \lambda_0, U < \lambda_0\right)$$

$$= \frac{P(Z_0 > \lambda_0, X < \lambda_0')}{P(Z_0 > \lambda_0)}, \tag{8}$$

where $P(Z_0 > z_0)$ is given in formula (6), and

$$P(Z_0 > z_0, X < x) = \prod_{j=1}^{K_1} \left[G_j(x) - H_j^{(T)}(z_0, x)\right].$$

For the two-stage independent design, the probability of at least one truly associated marker being ranked among the top L markers after the second stage can be easily obtained as:

$$P_2^* = 1 - P(X < Y) = 1 - \int_{-\infty}^{\infty} P(X < y) \cdot g^*(y)dy ,$$

where

$$P(X < y) = \prod_{j=1}^{K_1} G_j(y),$$

and

$$g^*(y) = \frac{(M_1 - K_1)!}{(M_1 - K_1 - L)!(L-1)!} \Phi^{M_1 - K_1 - L}(y)[1 - \Phi(y)]^{L-1} \phi(y).$$

An approximation to P_2^* is

$$P_2^* \approx 1 - P(X < \lambda_0') = 1 - \prod_{j=1}^{K_1} G_j(\lambda_0'). \tag{9}$$

Results

To see how many markers should be chosen from the pooling stage, we conduct some calculations using formula (5) first under various genetic models and allele frequencies. The following four genetic models are considered: a dominant model with $f_2 = f_1 = 0.04$, $f_0 = 0.01$; a recessive model with $f_2 = 0.04$, $f_1 = f_0 = 0.01$; a multiplicative model with $f_2 = 0.04$, $f_1 = 0.02$, $f_0 = 0.01$; and an additive model with $f_2 = 0.04$, $f_1 = 0.025$ and $f_0 = 0.01$ (Risch and Teng 1998, Zou and Zhao 2004). The population frequency of allele *A* is varied from 0.05, 0.2, to 0.7. We take the sample size to be $n = 1000$ and assume that the number of the disease-associated markers is $K = 5$.

Table 1 provides the probabilities of i ($i = 1, \cdots, 5$) truly associated markers being among the top 1/1000 markers when we assume the same genetic model and allele frequency at each disease-associated marker and no measurement errors. It is clear from the table that for most cases, the probability that all truly associated markers are among the top 1/1000 markers is high. The probability that these top markers include only some of the truly associated markers is often very low. An explanation is that when there is a signal that the

marker is associated with disease, the corresponding test statistic should often be large when the sample size is reasonably large. So the chance for such a marker to be ranked low is rather small. The exceptional cases are the recessive models with small allele frequencies or dominant models with large allele frequencies. This is because the allele frequency difference between the cases and controls is often small in these scenarios and the sample sizes are not large enough to distinguish the signals from noises. However, we can observe from the table that the probability of at least one truly associated marker being among the top 1/1000 markers is uniformly very large except for the recessive models with small allele frequencies. The conclusion still holds for the case in which genetic models and allele frequencies are different at each truly associated marker or the case of different sample sizes (data not shown). So in the following analysis, we consider the chance that at least one truly associated marker is among the top $100q_0\%$ of the markers.

Figure 1 presents the probability of at least one truly associated marker being included among the top $100q_0\%$ of the markers for a fixed population allele frequency, p and allele frequency difference between the case and control groups, $p_A - p_U$ (where f_0 is taken as 0.01. When f_0 is taken to be other values, the results are similar (data not shown)). It can be observed from the figure that for given p and $p_A - p_U$, the probabilities are almost the same under different genetic models. This shows that the probability that at least one truly associated marker is included among the top markers depends on the genetic model and allele frequency mostly through the population allele frequency and allele frequency difference between the case and control groups. Because the exact genetic model is often

unavailable to researchers, this fact makes it possible to select the proportion q_0 based on the assumed population allele frequency and allele frequency difference between the cases and controls at the candidate marker. Note that the effect of the number of truly disease-associated markers on the probability that at least one such marker is included is not very small (data not shown). So we require that the value of q_0 is chosen so that the probability is greater than 80% for the case of having only one truly associated marker and not smaller than 99% for the case of five truly associated markers. For the case of five truly associated markers, the allele frequency differences at four markers are assumed to be at least 0.03. Note that when the number of truly associated markers K is greater than five, the probability that at least one truly associated marker is included is larger.

Figure 2 gives the probability that the disease-associated marker is included among the top q_0 = 6.7% of the markers for various population allele frequencies and allele frequency differences between the cases and controls when there is only one truly associated marker. The figure shows that when the error rate is 0.01, choosing $q_0 = 6.7\%$ can detect the truly associated marker with an allele frequency difference of 0.05 with more than 80% chance. Furthermore, when there are five disease-associated markers, to detect at least one such marker with more than 99% chance, the selection proportion should be 7% (data not shown). Therefore, to detect the disease markers with an allele frequency difference of 0.05 at one marker, the selection proportion of 7% is recommended when the error rate is 0.01 and the sample consists of 1000 cases and 1000 controls. To select the truly associated markers with an allele frequency difference of 0.03 at one marker, the

proportion q_0 should be about 29% (data not shown). If the error rate is reduced to 0.005, the proportion q_0 can be reduced to 3% or 19% to select the truly associated markers with an allele frequency difference of 0.05 or 0.03 at one marker, respectively. The required proportions for including at least one truly associated marker with an allele frequency difference of $p_A - p_U = 0.03$, 0.05, 0.07 or 0.10 are summarized in Table 2 when the sample size is $n = 1000$. Generally, the effect of sample size on selecting the disease-associated markers is not very large, especially for the extreme allele frequencies (data not shown). However, it can be seen from Table 2 that reducing the measurement errors can greatly reduce the required proportion q_0. Therefore, it is important to reduce the measurement errors in the first stage.

To investigate the statistical power of the two-stage design, we set the sample size in the first stage to be $n = 500$, and the supplemental sample size in the second stage to be $n_a = 500$. Note that the main purpose in the first stage is to screen for those truly associated markers. Therefore, we hope that the probability of the truly associated markers being included is large. Thus, we set the power to be 95% in the pooling stage. The significance level of the two-stage design for a single marker test is taken to be $\alpha = 5 \times 10^{-8}$, a level suggested by Risch and Merikangas (1996) for genome-wide association studies. The results for the two-stage dependent design under the previous four genetic models are presented in Table 3. Clearly, the power depends on the genetic model and allele frequency. In general, the power is very high for the sample sizes we consider here. The exceptions are the recessive models with a small allele frequency or dominant

models with a large allele frequency. From this table, we can see that the measurement errors in DNA pooling have little impact on the statistical power of the two-stage design. Our previous studies showed that such effect can be large for a one-stage design, especially when the error rates are not small (Zou and Zhao 2004). Our finding shows that the impact of measurement errors on the case-control association studies can almost be neglected by using the two-step design, although a larger measurement error will lead to more markers to be selected in the first stage. Compared to the one-stage design, the two-stage strategy has slightly smaller power due to the selection in the first stage (data not shown). When the two-stage independent design is used, the power is higher than that of the two-stage dependent design (Table 4). In our calculation, we assume that the same number of the cases and controls are typed at the second stage for both designs, which implies that more efforts are needed for the two-stage independent design to collect additional cases and controls compared to the two-stage dependent design. Our calculation shows that if we ignore the correlation between the two stages for a two-stage dependent design, then we will slightly overestimate the power. On the other hand, from Table 4, the two-stage independent design is more affected by the measurement errors than the two-stage dependent design but less affected than the one-stage pooling scheme.

Table 5 gives the statistical power of the two-stage dependent design for the fixed allele frequency and allele frequency difference between the cases and controls (where f_0 is still taken as 0.01). It can be observed from the table that for given p and $p_A - p_U$, the power is almost the same under different genetic models. This shows that the power of the

two-stage design depends on the genetic model and allele frequency almost only through the population allele frequency and allele frequency difference between the case and control groups. As before, this observation is useful in practice because that, although the genetic models are often unknown to us, we can estimate the sample size to attain the desired significance level and power under different genetic models as long as the allele frequencies in the general population and the allele frequency differences between the cases and controls can be assumed.

We use the approximate formula (8) to calculate the probability of at least one truly associated marker being ranked among the top L markers after the second stage for the two-stage dependent design. Likewise, the probabilities are almost the same under different genetic models for the same population allele frequency and allele frequency difference between the case and control groups (data not shown). As an example, we consider a recessive model with a population allele frequency of 0.2 and allele frequency difference of 0.05. The results are presented in Figure 3. It can be seen that there is a high probability for the top 50 markers to include at least one truly associated marker when 1% of the markers are selected from the first stage, even though the measurement errors are not small. However, this probability may not be high for detecting disease-associated markers with small allele frequency differences, e.g. 0.03 (data not shown). Essentially, the chance that at least one truly associated marker is ranked among the top L markers after the second stage is higher for markers with larger allele frequency differences. The conclusion is similar for the two-stage independent design (data not shown). In general, the

probabilities are not larger for the two-stage independent design than those for the two-stage dependent design. This can be understood by noting the positive correlation between the two stages for the two-stage dependent design which leads to the smaller value of the right-hand side of formula (8) than $P(X < \lambda_0')$.

Discussion

In this paper, we have investigated the two-stage design with DNA pooling used in the first stage screening. Three related problems have been considered: (i) How many markers should be chosen from the first stage? (ii) What is the overall statistical power when the two-stage design is used? and (iii) What is the probability that at least one of the disease-associated markers is ranked among the top after the second stage? Our analyses show that the answers to these questions are dependent on the genetic models and allele frequencies essentially through the population allele frequencies and allele frequency differences between the case and control groups at the candidate markers. For the first problem, we have derived the proportion of markers that needs to be selected to include the truly associated markers. For instance, when the measurement errors are small (0.005), 3% of the markers need to be selected to include a disease-associated marker with an allele frequency difference of 0.05 between the case and control groups for a sample consisting of 1000 cases and 1000 controls. When the measurement errors are not small, multiple pools can be formed to reduce measurement errors. For the second problem, we have derived the formula for calculating the statistical power of a two-stage strategy. We find that the measurement errors in pooled DNA have little effect on the power when the

two-stage design, especially the two-stage dependent design, is used, contrary to the single stage pooling scheme. Recall our conclusion that reducing measurement errors can greatly reduce the selection proportion of markers in the pooling stage, we see that for a two-stage design, measurement errors have a large impact only on the first stage. Once the markers are selected, the effect of measurement errors can be very small. Three strategies, the two-stage dependent design, the two-stage independent design, and the one-stage design, have been compared. Overall, the two-stage independent design has the highest power, the one-stage design with individual genotyping has slightly higher power than the two-stage dependent design. However, their difference in power is not large. On the other hand, the one-stage design will be either too expensive (for individual genotyping) in genome-wide search or seriously affected by measurement errors (for DNA pooling). Furthermore, for the two-stage independent design, extra sample collection is needed, although the genotyping cost is the same as in the two-stage dependent design. In fact, if in our calculations, we use exactly the same number of individuals as that in the two-stage dependent design with 500 used to screen and the other 500 for follow-up analyses, the statistical power for such a two-stage independent design can be much lower than that of the two-stage dependent design. For example, the power under the multiplicative model with a population allele frequency of 0.05 and a measurement error rate of 0.005 is 0.209 for the above two-stage independent design but 0.599 for the two-stage dependent design. For the third problem, our studies show that the chance that at least one truly associated marker selected from the first stage is ranked among the top markers after the second stage is high when the allele frequency differences are not smaller than 0.05 for samples of

reasonable sizes, even though the measurement errors are not small.

It is of practical interest how to allocate the sample sizes in the two stages to maximize the power (or minimize the total cost) for a given cost (or given power), as Satagopan et al. (2002), Satagopan and Elston (2003), and Satagopan et al. (2004) have done. For example, let C be the total cost, C_1 be the cost of recruiting an individual, C_{pool} be the cost of measuring allele frequency at a single marker for a DNA pool, C_{ind} be the cost of genotyping a single marker for an individual, and C_0 be the other cost such as administration. Then we have

$$C = C_0 + C_1 \cdot 2(n + n_a) + C_{pool} \cdot 2mM + C_{ind} \cdot 2(n + n_a)M_1$$

for the two-stage dependent design, and

$$C = C_0 + C_1 \cdot 2(n + n_a) + C_{pool} \cdot 2mM + C_{ind} \cdot 2n_a M_1$$

for the two-stage independent design. In particular, we take the number of total markers to be $M = 10^6$, the number of the truly disease-associated markers to be $K = 1$, and the number of pool pairs to be $m = 1$. Further, we take $C = 5 \times 10^5$ (Unit: USD), $C_1 = 200$, $C_{ind} = 0.02$, $C_{pool} = 0.02$, $C_0 = 0$, and $\varepsilon = 0.01$. Then our preliminary calculation results showed that for the given cost, the optimal design that lead to highest power is to allocate exactly (nearly) the same sample size to each stage for the two-stage dependent (independent) design (data not shown). For the two-stage dependent design, this means that all individuals should be used at both stages and no additional sample is needed at the second stage. This is similar to the two-stage individual genotyping design with sample size constraint (Satagopan et al. 2004) but is different from the design with individual genotyping at both stages in which the optimal design maximizing power is to allocate approximately 25% of the individuals to the first stage and the remaining individuals to the

second stage (Satagopan et al. 2002, Satagopan and Elston 2003). Clearly, an overall investigation is needed in this regard. This warrants our further research.

To simplify our analyses, we have assumed independence among the markers. This would be reasonable when the marker density is low. However, for a genome-wide association study, the marker density is high and adjacent markers may be highly correlated. But it is not evident how to model the correlation among markers. One way to avoid this difficulty is to study many subsets of the whole marker set such that they cover the entire genome yet the markers are independent. However, this is clearly less than satisfactory due to the loss of information in the data. On the other hand, this question can be examined empirically to assess the effect of correlations among markers on our results. For example, we have investigated the effect of correlation on the selection of markers in the first stage through the HapMap data. We considered the SNPs on the 500K SNP Array and used the HapMap data approximate the level of correlations among SNPs. The HapMap data consist of 270 individuals from four populations, and the information for the 500K data can be downloaded from http://www.affymetrix.com/support/downloads/data/500K_HapMap270.zip (For the missing alleles, we imputed them by the corresponding frequencies of the existing alleles). For simplicity, we have only considered the first 300 markers and let the 140^{th} marker be disease-associated to illustrate the impact of marker dependence and a more through investigation will be reported in future reports. Assuming a dominant model with $f_2 = f_1 = 0.4, f_0 = 0.1$, the allele frequency difference between the case and control groups is $p_A - p_U = 0.088$. We considered the sample sizes of the two pools to

be $n = 100$. Using the results established before under the independence assumption, we found that if we took the top $q_0 = 18.3\%, 18.7\%, 20.0\%$ and 30.9% of the markers when $\varepsilon = 0,\ 0.005,\ 0.01$ and 0.03, respectively, then we would have the chance of 80% to select the disease-associated marker (i.e., 140^{th} marker) in the first stage. When we applied these q_0 s obtained under the independence assumption to the HapMap data, we observed that in 10,000 simulations, we had the chances of 72%, $72\%,\ 71\%$, and 65% to include the disease-associated marker when $\varepsilon = 0,$ $0.005,\ 0.01$ and 0.03, respectively. This shows that the correlation among markers can reduce the chance that the truly disease-associated marker is selected but such reduction is not large. Further, the impact of correlation is larger (smaller) for less (more) stringent requirement on the chance of including the disease-associated marker under the independence assumption (data not shown). Clearly, to eliminate the effect of correlation, the best way is to develop similar methods to those given in this paper incorporating the correlations among markers, and this will be addressed in our future work.

Throughout the paper, we have assumed that there exist measurement errors in the DNA pooling stage but no errors in the individual genotyping stage. How genotyping errors at both stages can affect the efficiency of the two-stage scheme also warrants future research.

Note that family-based data are often used in genetic epidemiological studies in addition to population-based data. Association studies using pooled DNA family data have been

considered for the one stage scheme (e.g. Risch and Teng 1998, Zou and Zhao 2005). The research on the two-stage designs using family data is no doubt an interesting topic for future research.

Acknowledgments

We thank the Associate Editor and two anonymous reviewers for their insightful and constructive comments. This work was supported in part by grants DMS0234078 from NSF (to Y. Zuo), Nos. 70221001 and 10471043 from NNSFC (to G. Zou), and GM59507 from NIH (to H. Zhao).

References

Bansal A., van den Boom D., Kammerer S., Honisch C., Adam G., Cantor C. R., Kleyn P., and Braun A. (2002). Association testing by DNA pooling: an effective initial screen. *Proc Natl Acad Sci* USA. 99: 16871-16874.

Barcellos L. F., Klitz W., Field L. L., Tobias R., Bowcock A. M., Wilson R., Nelson M. P. , Nagatomi J., and Thomson G. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 61: 734-747.

Barratt B. J., Payne F., Rance H. E., Nutland S., Todd J. A., and Clayton D. G. (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.

Buetow, K. H., Edmonson, M., MacDonald, R., Clifford, P., Yip, P., Kelley, J., Little, D. P., Strausberg, R., Koester, H., Cantor, C. R., and 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.

Elston, R. C. (1994). P values, power, and pitfalls in the linkage analysis of psychiatric

disorders. In: Genetic approaches to mental disorders, Gershon ES, Clonings CR Ed., 3-21. Proceedings of the Annual Meeting of the American Psychopathological Association. Washington, DC: American Psychiatric Press.

Elston R. C., Guo X., and Williams L. V. (1996). Two-stage global search designs for linkage analysis using pairs of affected relatives. *Genet Epidemiol* 13: 535-558.

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

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., and 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.

Risch NJ. (2000) Searching for genetic determinants in the new millennium. *Nature.* 405: 847-856.

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

Risch, N. and 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.

Satagopan J. M. and Elston R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 25: 149-157.

Satagopan J. M., Verbel D. A., Venkatraman E.S., Offit K. E., and Begg C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* 58: 163-170.

Satagopan J. M., Venkatraman E. S., and Begg C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60: 589-597.

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

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

Zou G. and Zhao H. (2005). Family-based association tests for different family structures using pooled DNA. *Ann Hum Genet* 69: 429-442.

Appendix

The calculation of the probability that none of the truly associated markers are ranked among the top L markers

Clearly, P_2' can be written as

$$P_2' = \frac{P\left(X < Y, Z_0 > U, Z^* < Z_0, Z^* < V, V > U\right)}{P\left(Z_0 > U, Z^* < Z_0, Z^* < V, V > U\right)}$$

$$= \frac{P\left(X < Y, V > Z_0 > U, Z^* < Z_0, V > U\right) + P\left(X < Y, Z_0 \geq V, Z^* < V, V > U\right)}{P\left(V > Z_0 > U, Z^* < Z_0, V > U\right) + P\left(Z_0 \geq V, Z^* < V, V > U\right)}. \qquad \text{(A.1)}$$

We have known $t_{ind,j}^{(T)} \sim N\left(\theta_{ind,j}, \lambda_{ind,j}^2\right)$, $j = 1,...,K_1$; and $t_{ind,j}^{(N)} \sim N(0,1)$, $j = 1,...,M_1 - K_1$. We denote the distribution and density functions of $t_{ind,j}^{(T)}$ by $G_j(x)$ and $g_j(x)$, respectively. The distribution and density functions of $t_{ind,j}^{(N)}$ are still denoted as $\Phi(x)$ and $\phi(x)$, respectively. Further, let $H_j^{(T)}(x,y)$ denote the joint distribution of $\left(t_{pool,j}^{(T)}, t_{ind,j}^{(T)}\right)$, $j = 1,...,K_1$; and $H_j^{(N)}(x,y)$ denote the joint distribution of $\left(t_{pool,j}^{(N)}, t_{ind,j}^{(N)}\right)$, $j = 1,...,M_1 - K_1$. Moreover, $h_j^{(T)}(x,y)$ and $h_j^{(N)}(x,y)$ denote the corresponding density functions. Then it can be shown that

(i) $P\left(X < Y, V > Z_0 > U, Z^* < Z_0, V > U\right)$

$$= \int_{-\infty}^{\infty} dy \iint_{u<v} \left[\int_{-\infty}^{y} dx \int_{u}^{v} P\left(Z^* < z_0\right) \cdot p(x, z_0) dz_0 \right] \cdot p(y,v)\, p(u) du dv,$$

(ii) $P\left(X < Y, Z_0 \geq V, Z^* < V, V > U\right)$

$$= \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} P(Z_0 > v, X < y) P\left(Z^* < v\right) P(U < v) \cdot p(y,v) dv,$$

(iii) $P\left(V > Z_0 > U, Z^* < Z_0, V > U\right)$

$$= \iint\limits_{u<v} \left[\int_u^v P\left(Z^* < z_0\right) p(z_0) dz_0 \right] \cdot p(u)\, p(v) du dv,$$

and

(iv) $P\left(Z_0 \geq V, Z^* < V, V > U\right)$

$$= \int_{-\infty}^{\infty} P(Z_0 > v) P\left(Z^* < v\right) P(U < v) \cdot p(v) dv,$$

where

$$P(U < u) = [\Phi(u)]^{(M-K)-(M_1-K_1)},$$

$$p(u) = [(M-K)-(M_1-K_1)][\Phi(u)]^{(M-K)-(M_1-K_1)-1} \phi(u),$$

$$p(v) = (M_1-K_1)[1-\Phi(v)]^{M_1-K_1-1} \phi(v),$$

$$p(z_0, x) = \prod_{j=1}^{K_1} \left[G_j(x) - H_j^{(T)}(z_0, x)\right] \left\{ \sum_{j=1}^{K_1} \frac{g_j(x) - c_j}{G_j(x) - H_j^{(T)}(z_0, x)} \cdot \sum_{j=1}^{K_1} \frac{b_j}{G_j(x) - H_j^{(T)}(z_0, x)} \right.$$

$$\left. + \sum_{j=1}^{K_1} \frac{h_j^{(T)}(z_0, x)\left[G_j(x) - H_j^{(T)}(z_0, x)\right] - b_j\left[g_j(x) - c_j\right]}{\left[G_j(x) - H_j^{(T)}(z_0, x)\right]^2} \right\}$$

with $b_j = \int_{-\infty}^{x} h_j^{(T)}(z_0, t) dt$, and $c_j = \int_{-\infty}^{z_0} h_j^{(T)}(s, x) ds,$ and

$$p(y, v) = \sum_{l=M_1-K_1-L+1}^{M_1-K_1} \sum_{i_1 < \cdots < i_l} \varphi(v, y)\left(D_1 D_2 - D_{12}\right),$$

with $i_1, ..., i_l$ being some l numbers of 1,..., $M_1 - K_1$, and

$$\varphi(v, y) = \prod_{j=1}^{l} \left[\Phi(y) - H_{i_j}^{(N)}(v, y)\right] \cdot \prod_{j=l+1}^{M_1-K} \left[1 - \Phi(y) - \Phi(v) + H_{i_j}^{(N)}(v, y)\right],$$

$$D_1 = \sum_{j=1}^{l} \frac{d_{i_j}}{\Phi(y) - H_{i_j}^{(N)}(v, y)} + \sum_{j=l+1}^{M_1-K_1} \frac{\phi(v) - d_{i_j}}{1 - \Phi(y) - \Phi(v) + H_{i_j}^{(N)}(v, y)},$$

$$D_2 = \sum_{j=1}^{l} \frac{\phi(y) - e_{i_j}}{\Phi(y) - H_{i_j}^{(N)}(v, y)} + \sum_{j=l+1}^{M_1-K_1} \frac{-\phi(y) + e_{i_j}}{1 - \Phi(y) - \Phi(v) + H_{i_j}^{(N)}(v, y)},$$

$$D_{12} = \sum_{j=1}^{l} \frac{-h_{i_j}^{(N)}(v, y)\left[\Phi(y) - H_{i_j}^{(N)}(v, y)\right] + d_{i_j}\left[\phi(y) - e_{i_j}\right]}{\left[\Phi(y) - H_{i_j}^{(N)}(v, y)\right]^2}$$

$$+\sum_{j=l+1}^{M_1-K_1}\frac{h_{i_j}^{(N)}(v,y)\left[1-\Phi(y)-\Phi(v)+H_{i_j}^{(N)}(v,y)\right]-\left[\phi(v)-d_{i_j}\right]\left[\phi(y)-e_{i_j}\right]}{\left[1-\Phi(y)-\Phi(v)+H_{i_j}^{(N)}(v,y)\right]^2},$$

and $d_{i_j}=\int_{-\infty}^{y}h_{i_j}^{(N)}(v,t)dt$ and $e_{i_j}=\int_{-\infty}^{v}h_{i_j}^{(N)}(s,y)ds$.

Combining (A.1) and (i)-(iv), we can obtain P_2'. Thus, the probability that at least one truly associated marker is ranked among the top L markers can be calculated by $P_2=1-P_2'$.

Table 1. The probability of i $(i = 1, \cdots, 5)$ disease-associated markers ranked among the top 1/1000 markers for the case of the same genetic model and allele frequency at each truly associated marker*

	$i = 5$	$i = 4$	$i = 3$	$i = 2$	$i = 1$	$i \geq 1$
Dominant						
p = 0.05	1.000	0.000	0.000	0.000	0.000	1.000
p = 0.20	1.000	0.000	0.000	0.000	0.000	1.000
p = 0.70	0.234	0.394	0.266	0.090	0.015	0.999
Recessive						
p = 0.05	0.000	0.000	0.000	0.004	0.099	0.103
p = 0.20	0.995	0.005	0.000	0.000	0.000	1.000
p = 0.70	1.000	0.000	0.000	0.000	0.000	1.000
Multiplic.						
p = 0.05	0.970	0.030	0.000	0.000	0.000	1.000
p = 0.20	1.000	0.000	0.000	0.000	0.000	1.000
p = 0.70	0.999	0.001	0.000	0.000	0.000	1.000
Additive						
p = 0.05	1.000	0.000	0.000	0.000	0.000	1.000
p = 0.20	1.000	0.000	0.000	0.000	0.000	1.000
p = 0.70	1.000	0.000	0.000	0.000	0.000	1.000

* 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: $f_2 = 0.04$, $f_1 = 0.025$, $f_0 = 0.01$.

** The sample size is $n = 1000$, and no measurement errors are assumed with the number of disease-associated markers being $K = 5$.

Table 2. The recommended proportion q_0 of markers selected from the first stage for including at least one truly associated marker with an allele frequency difference of $p_A - p_U$ at one marker*

$p_A - p_U$	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$	$\varepsilon = 0.03$
0.03	15%	19%	29%	58%
0.05	2%	3%	7%	40%
0.07	0.4%	0.9%	3%	25%
0.10	5×10^{-5}	0.02%	0.4%	18%

* The sample size in the first stage is $n = 1000$, and the number of pools formed for either the cases or the controls is $m = 1$.

Table 3. The power of the two-stage dependent design for the sample sizes of $n = 500$ and $n_a = 500$ *

	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$	$\varepsilon = 0.03$
Dominant				
$p = 0.05$	0.950	0.950	0.950	0.950
$p = 0.20$	0.950	0.950	0.950	0.950
$p = 0.70$	0.046	0.046	0.046	0.046
Recessive				
$p = 0.05$	0.000	0.000	0.000	0.000
$p = 0.20$	0.829	0.827	0.824	0.817
$p = 0.70$	0.950	0.950	0.950	0.950
Multiplic.				
$p = 0.05$	0.600	0.599	0.595	0.584
$p = 0.20$	0.950	0.950	0.950	0.950
$p = 0.70$	0.950	0.950	0.950	0.950
Additive				
$p = 0.05$	0.941	0.939	0.936	0.931
$p = 0.20$	0.950	0.950	0.950	0.950
$p = 0.70$	0.948	0.947	0.946	0.943

* The significance level for the two-stage design is $\alpha = 5 \times 10^{-8}$, and the power in the pooling stage is $1 - \beta_1 = 95\%$. 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: $f_2 = 0.04$, $f_1 = 0.025$, $f_0 = 0.01$.

** The number of pools formed for either the cases or the controls is $m = 1$.

Table 4. The power of the two-stage independent design for the sample sizes of 500 in the first stage and 1000 in the second stage [*]

	$\varepsilon = 0$	$\varepsilon = 0.005$	$\varepsilon = 0.01$	$\varepsilon = 0.03$
Dominant				
p = 0.05	0.950	0.950	0.950	0.950
p = 0.20	0.950	0.950	0.950	0.950
p = 0.70	0.092	0.084	0.071	0.051
Recessive				
p = 0.05	0.000	0.000	0.000	0.000
p = 0.20	0.933	0.925	0.902	0.830
p = 0.70	0.950	0.950	0.950	0.950
Multiplic.				
p = 0.05	0.833	0.767	0.678	0.593
p = 0.20	0.950	0.950	0.950	0.950
p = 0.70	0.950	0.950	0.950	0.950
Additive				
p = 0.05	0.950	0.949	0.946	0.933
p = 0.20	0.950	0.950	0.950	0.950
p = 0.70	0.950	0.950	0.950	0.946

[*] The significance level for the two-stage design is $\alpha = 5\times10^{-8}$, and the power in the pooling stage is $1-\beta_1 = 95\%$. 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: $f_2 = 0.04$, $f_1 = 0.025$, $f_0 = 0.01$.

[**] The number of pools formed for either the cases or the controls is $m = 1$.

Table 5. The power of the two-stage dependent design for the fixed allele frequency and allele frequency difference between the case and control groups*

	$p_A - p_U = 0.03$	$p_A - p_U = 0.05$	$p_A - p_U = 0.07$	$p_A - p_U = 0.10$
$p = 0.05$				
Dominant	0.0685	0.748	0.949	0.950
Recessive	0.0915	0.717	0.944	0.950
Multiplic.	0.0704	0.744	0.948	0.950
Additive	0.0697	0.746	0.949	0.950
$p = 0.20$				
Dominant	0.00115	0.0585	0.457	0.941
Recessive	0.00174	0.0722	0.460	0.931
Multiplic.	0.00127	0.0618	0.458	0.938
Additive	0.00126	0.0612	0.458	0.939
$p = 0.70$				
Dominant	4.58×10^{-4}	0.0301	0.352	0.926
Recessive	6.96×10^{-4}	0.0389	0.376	0.934
Multiplic.	6.24×10^{-4}	0.0366	0.374	0.936
Additive	6.16×10^{-4}	0.0362	0.373	0.937

* The significance level for the two-stage design is $\alpha = 5\times10^{-8}$, and the power in the pooling stage is $1-\beta_1 = 95\%$.

** The sample sizes are $n = 500$ and $n_a = 500$, the error rate is $\varepsilon = 0.01$, and the number of pools formed for either the cases or the controls is $m = 1$.

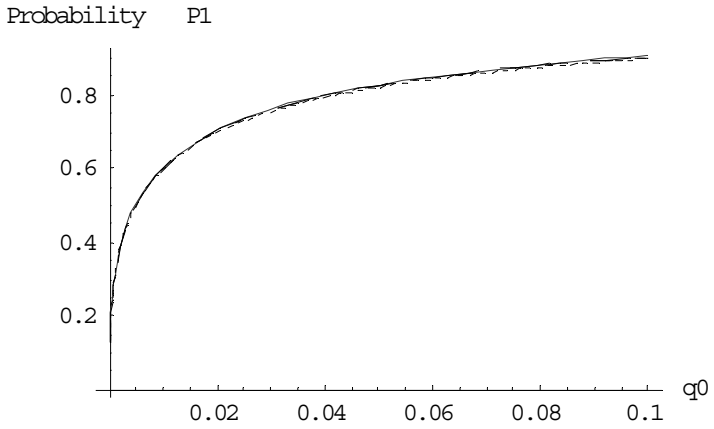

Figure 1. The probability of the truly associated marker being included among the top $100q_0\%$ of the markers under different genetic models for the same population allele frequency (0.20) and allele frequency difference between the case and control groups (0.05). From top to bottom, the curves correspond to the dominant model, additive model, multiplicative model, and recessive model, respectively. The sample size is $n = 1000$, the error rate is $\varepsilon = 0.01$, and the number of pools formed for either the cases or the controls is $m = 1$. We assume that the number of disease-associated markers is $K = 1$.

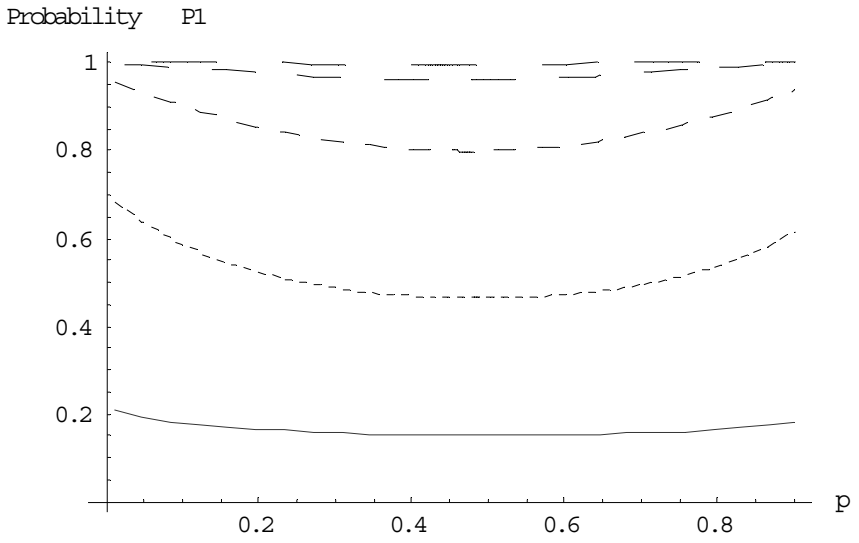

Figure 2. The probability of the truly associated marker being included among the top 6.7% of the markers when the number of disease-associated markers is $K = 1$. The sample size is $n = 1000$, the error rate is $\varepsilon = 0.01$, and the number of pools formed for either the cases or the controls is $m = 1$. From top to bottom, the curves correspond to allele frequency differences of 0.10, 0.07, 0.05, 0.03, and 0.01, respectively.

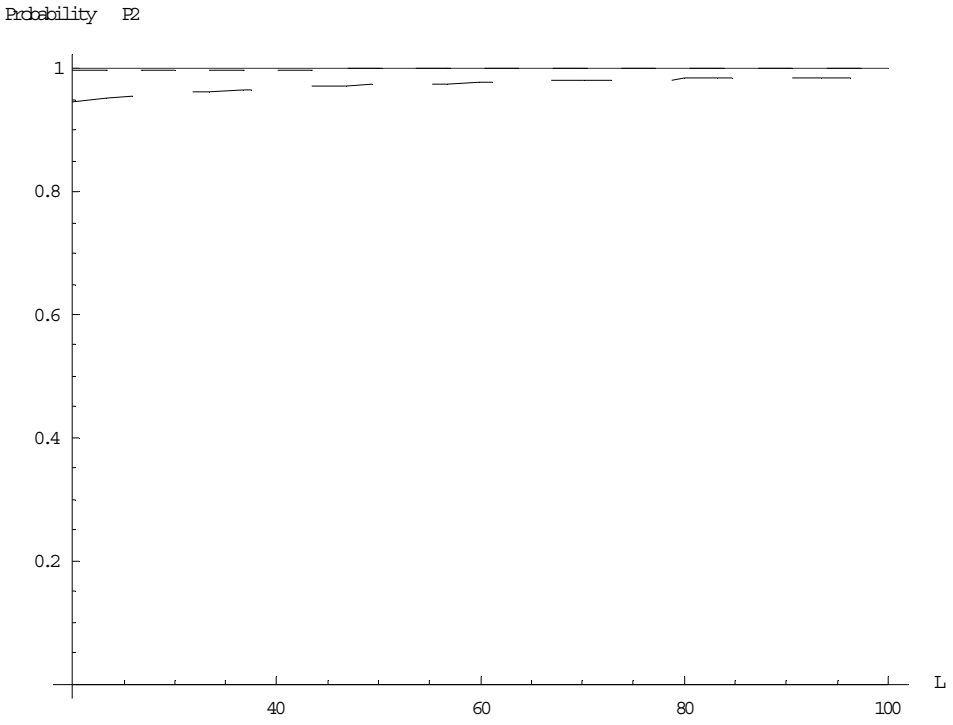

Figure 3. The probability of at least one truly associated marker being ranked among the top L markers after the second stage for the two-stage dependent design where the sample sizes are $n = 500$ and $n_a = 500$, the error rate is $\varepsilon = 0.01$, and the number of pools formed for the cases or the controls is $m = 1$. The allele frequency difference is 0.05, and the population allele frequency is $p = 0.2$. From top to bottom, the curves correspond to the cases of $K_1 = 5$, 2, and 1, respectively (Assume the number of the whole markers is $M = 10^6$ and top 1% markers are chosen from the first stage in which K_1 truly associated markers are included).