Optimal DNA Pooling-Based Two-Stage Designs in Case-Control Association Studies

Yihong Zhao Shuang Wang

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, N.Y., USA

Key Words

Two-stage design • DNA-pooling • Genome-wide association study • Measurement errors • Optimal design

Abstract

Study cost remains the major limiting factor for genome-wide association studies due to the necessity of genotyping a large number of SNPs for a large number of subjects. Both DNA pooling strategies and two-stage designs have been proposed to reduce genotyping costs. In this study, we propose a cost-effective, two-stage approach with a DNA pooling strategy. During stage I, all markers are evaluated on a subset of individuals using DNA pooling. The most promising set of markers is then evaluated with individual genotyping for all individuals during stage II. The goal is to determine the optimal parameters (π_{sample}^p , the proportion of samples used during stage I with DNA pooling; and π_{marker}^p , the proportion of markers evaluated during stage II with individual genotyping) that minimize the cost of a two-stage DNA pooling design while maintaining a desired overall significance level and achieving a level of power similar to that of a one-stage individual genotyping design. We considered the effects of three factors on optimal two-stage DNA pooling designs. Our results suggest that, under most scenarios considered, the optimal two-stage DNA pooling design may be much more cost-effective than the optimal two-stage individual genotyping design, which use individual genotyping during both stages.

Copyright © 2008 S. Karger AG, Basel

Introduction

A genome-wide association (GWA) study is a powerful approach to detect susceptibility genes with small effects associated with complex diseases [1–4]. Recent advances in genotyping technologies have made GWA studies possible; however, even though genotyping costs have dropped significantly, they remain prohibitively high for many GWA studies due to necessity of genotyping a large number of SNPs from a large number of case and control subjects.

DNA pooling [5] and two-stage designs [6–11] have been proposed to reduce genotyping costs. Using a DNA pooling strategy, pooled DNA samples from a group of case subjects and pooled DNA samples from a group of control subjects are formed. The allele frequency of each marker is then estimated by genotyping case pools and control pools instead of individual subjects. This reduces the number of genotyping assays substantially. In a recent publication, Pearson et al. [12] suggested that pooling-based GWA studies might provide a viable alternative for screening disorders with common variations of large effect. Another recent study has demonstrated the applicability of using DNA pools and 500 K Affymetrix GeneChip mapping arrays as a cost-effective, reliable and valid initial screening tool [13]. DNA pooling has also been utilized in a whole-genome association study of neuroticism [14] and a genome-wide scan of progressive supranuclear palsy [15], using Affymetrix 100 and 500 K mapping ar-

rays. In well-developed two-stage designs using individual genotyping, all M markers are evaluated on a subset of N total subjects ($N \times \pi_{\text{sample}}$) during stage I, and the most promising subset of M markers ($M \times \pi_{\text{marker}}$) is then evaluated in the remaining $N \times (1 - \pi_{\text{sample}})$ subjects during stage II. This greatly reduces the total amount of genotyping and therefore leads to a more cost-effective design. Zuo et al. [16] combined DNA pooling and a two-stage design. In their design, all M markers and all N subjects were genotyped with DNA pooling during stage I. A subset of the most promising markers was selected and individually genotyped on a set of newly collected subjects during stage II. However, how to define an optimal cost-effective two-stage design with DNA pooling and compare the cost-effectiveness with the widely studied optimal two-stage individual genotyping design under a fixed study power has not been investigated.

In this study, we propose a cost-effective two-stage approach with a DNA pooling strategy. During stage I, all M markers are evaluated on a subset of subjects using pooled DNA samples. The most promising set of markers is then evaluated with individual genotyping in all subjects during stage II. The goal is to determine the optimal parameters (π_{sample}^p , the proportion of samples used during stage I with DNA pooling; and π_{marker}^p , the proportion of markers evaluated during stage II) that minimize the cost of a two-stage DNA pooling design while maintaining a desired overall significance level and achieving a level of power similar to that of a one-stage individual genotyping design. We compare the costs of the optimal two-stage DNA pooling designs to those of the optimal two-stage individual genotyping designs.

We investigate the effects of three factors on the optimal two-stage DNA pooling designs. The first factor involves the DNA pooling-related measurement errors including, but not limited to, the following: accuracy of pool construction by pipetting, integrity of the pooled genomic DNA and possible unequal amplification of one allele over another [5, 12, 17–19]. For array-based DNA pooling, additional variation may be introduced by the array itself [20, 21]. In this study, we have grouped the above-mentioned DNA pooling-related measurement errors into two types. The first type is the pool-specific error, σ_p , which includes the pool-construction errors and the array errors known for their confounding effects [20]. The second type of error is due to unequal allele amplification, which has been shown to cause a systematic bias in testing allele frequency differences between pools [20]. If it were possible for there to be no pooling-related er-

rors, studies using DNA pooling would be able to achieve the same level of power as studies using individual genotyping. However, in reality, pooling-related errors always exist.

The second factor, denoted as R , is the ratio of the cost per genotyping assay performed in stage II to that of stage I. A recent study [10] of the impact of R on the optimal two-stage individual genotyping design suggests that, despite the markedly higher per-genotyping cost during stage II (R ranges from 15 to 20), two-stage individual genotyping designs are more cost-effective than one-stage designs. In an optimal two-stage individual genotyping design, it has been shown that the optimal proportion of samples (π_{sample}) used during stage I increases as R increases, and the optimal proportion of markers (π_{marker}) evaluated during stage II decreases as R increases [10].

The third factor is the DNA pool size, denoted as s . The pool size is the number of individuals used in constructing a pool that will be genotyped on a single array. Given pool size s , we can obtain the number of pools with a fixed sample size. Pearson et al. [12] suggests a limiting point beyond which increasing the pool size becomes less efficient based on simulation studies.

Our results suggest that, under most scenarios considered, the optimal two-stage DNA pooling designs are much more cost-effective than the optimal two-stage individual genotyping designs. When the pool-specific error is small, and the pool size is as small as 20, the cost of the two-stage DNA pooling design is only 4–16% that of the two-stage individual genotyping design. When the pool-specific error is large, the two-stage DNA pooling design may not be as cost-effective as the two-stage individual genotyping design, especially when the genetic relative risk (GRR) is small and the allele responsible for the risk is rare. The optimal design parameters and the optimal cost fractions of the two-stage DNA pooling designs and the two-stage individual genotyping designs compared to the cost of the one-stage individual genotyping designs are also provided. We believe that with the prohibitive costs of genome-wide association scans and the great demand for conducting hypothesis-free genome-wide scans, the optimal two-stage DNA pooling design can be a cost-effective high-throughput screening tool for evidence of association. (A program written in R is available upon request from the authors.)

Methods

Notations and Assumptions

We consider M markers and $2N$ (a balanced design with N cases and N controls) samples in a case-control study. Among the M markers, we assume D (≥ 1) markers are the true disease markers.

Assume bi-allelic markers with alleles A and a , which have frequencies p and $1 - p$, respectively. Let X_i and Y_i be the number of copies of the risk-allele A that the i -th case and the i -th control carry. Under the Hardy-Weinberg equilibrium assumption, X_i or $Y_i = 2$ has a probability of p^2 ; X_i or $Y_i = 1$ has a probability of $2p(1 - p)$; and X_i or $Y_i = 0$ has a probability of $(1 - p)^2$. Let $f_0 = \Pr(\text{affected} | aa)$, and ψ_1, ψ_2 be the genetic relative risk (GRR) for genotypes Aa and AA , where

$$\psi_1 = \frac{\Pr(\text{affected} | Aa)}{\Pr(\text{affected} | aa)}, \quad \text{and} \quad \psi_2 = \frac{\Pr(\text{affected} | AA)}{\Pr(\text{affected} | aa)}.$$

The risk allele frequencies within case group (P_A) and control group (P_U) are functions of p, ψ_1, ψ_2 , and f_0 ; that is,

$$P_A = \frac{p^2\psi_2 + p(1-p)\psi_1}{p^2\psi_2 + 2p(1-p)\psi_1 + (1-p)^2},$$

$$P_U = \frac{p^2(1-\psi_2f_0) + p(1-p)(1-\psi_1f_0)}{p^2(1-\psi_2f_0) + 2p(1-p)(1-\psi_1f_0) + (1-p)^2(1-f_0)}.$$

Study Costs

To investigate the cost-effectiveness of two-stage DNA pooling designs, we compared the optimal cost of the two-stage DNA pooling design and the optimal cost of the two-stage individual genotyping design relative to the cost of the one-stage individual genotyping design.

Let C_{ind} and C_{pool} be the per-genotyping-array costs of individual genotyping and DNA pooling during stage I, respectively. Without considering the cost of constructing pools, we assume $C_{pool} = C_{ind}$. With M markers and $2N$ samples, the total cost of the one-stage individual genotyping design, which serves as the reference, is:

$$Cost_1 = 2 \times N \times M \times C_{ind}.$$

Given the ratio of the per-genotyping cost during stage II to that during stage I, R , the per-genotyping cost during stage II, will be $R \times C_{ind}$ for both the two-stage DNA pooling design and the two-stage individual genotyping design under the assumption of $C_{pool} = C_{ind}$. Therefore, the total cost of the two-stage individual genotyping design is:

$$Cost_2 = 2 \times N \times \pi_{sample} \times M \times C_{ind} + 2 \times N \times (1 - \pi_{sample}) \times M \times \pi_{marker} \times R \times C_{ind}.$$

The total cost of the two-stage DNA pooling design is:

$$Cost_3 = 2 \times (N \times \pi_{sample}^p / s) \times M \times C_{pool} + 2 \times N \times M \times \pi_{marker}^p \times R \times C_{ind}.$$

CFI is the ratio of the cost of the two-stage individual genotyping design to that of the one-stage design, and CFP is the ratio of

the cost of the two-stage DNA pooling design to that of the one-stage design. That is,

$$CFI = Cost_2 / Cost_1,$$

$$CFP = Cost_3 / Cost_1.$$

Test Statistics

One-Stage Individual Genotyping Designs

The one-stage designs with individual genotyping serve as reference designs. Let $\hat{P}_A$ and $\hat{P}_U$ be the estimated risk-allele frequencies in case and control groups. The test statistic to test the association between a marker and the disease is:

$$T = \frac{\hat{P}_A - \hat{P}_U}{\sqrt{[\hat{P}_A(1 - \hat{P}_A) + \hat{P}_U(1 - \hat{P}_U)] / (2N)}}.$$

Under the null hypothesis, T follows approximately a normal distribution with mean 0 and variance 1. Under the alternative hypothesis, T follows approximately a normal distribution with mean

$$\mu = \frac{P_A - P_U}{\sqrt{[P_A(1 - P_A) + P_U(1 - P_U)] / (2N)}}$$

and the variance is 1. When the total sample size is fixed, the power of the one-stage design for a one-sided test is $1 - \beta = 1 - \Phi(z_{1-\alpha/M} - \mu)$, where we control the marker-wise false positive rate at α/M using a Bonferroni correction.

Two-Stage Individual Genotyping Designs

In two-stage individual genotyping designs, all M markers are evaluated during stage I using $N \times \pi_{sample}$ cases and $N \times \pi_{sample}$ controls. The most promising set of markers π_{marker} are selected and genotyped in the remaining cases and controls during stage II. We focus on the joint analysis, and follow the test statistics defined by Skol et al. [22]. Let $\hat{P}_{A1}$ and $\hat{P}_{U1}$ be the estimated risk-allele frequencies in case and control groups using $N \times \pi_{sample}$ cases and $N \times \pi_{sample}$ controls, and $\hat{P}_{A2}$ and $\hat{P}_{U2}$ be the estimated risk-allele frequencies using the remaining $N \times (1 - \pi_{sample})$ cases and $N \times (1 - \pi_{sample})$ controls. The test statistics T_1 during stage I and T_{all} during stage II are defined as:

$$T_1 = \frac{\hat{P}_{A1} - \hat{P}_{U1}}{\sqrt{[\hat{P}_{A1}(1 - \hat{P}_{A1}) + \hat{P}_{U1}(1 - \hat{P}_{U1})] / (2N\pi_{sample})}},$$

$$T_2 = \frac{\hat{P}_{A2} - \hat{P}_{U2}}{\sqrt{[\hat{P}_{A2}(1 - \hat{P}_{A2}) + \hat{P}_{U2}(1 - \hat{P}_{U2})] / [2N(1 - \pi_{sample})]}},$$

$$T_{all} = \sqrt{\pi_{sample}} T_1 + \sqrt{1 - \pi_{sample}} T_2.$$

Under the null hypothesis of no association, T_1 and T_{all} follow an approximate bivariate normal distribution $N(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \sqrt{\pi_{sample}} \\ \sqrt{\pi_{sample}} & 1 \end{bmatrix}.$$

Under the alternative hypothesis, T_1 and T_{all} follow an approximate bivariate normal distribution $N(\mu_{ind}, \Sigma)$, where

$$\mu_{ind} = \begin{pmatrix} \mu_{ind1} \\ \mu_{indall} \end{pmatrix} = \begin{pmatrix} \frac{P_A - P_U}{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N\pi_{sample})}} \\ \frac{P_A - P_U}{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N)}} \end{pmatrix}.$$

Given the sample size $2N$, π_{sample} and μ_{marker} must be such that the following two equations are satisfied:

$$P(T_1 > K_1, T_{all} > K_{all} | H_0) = \alpha/M,$$

and

$$P(T_1 > K_1, T_{all} > K_{all} | H_1) = 1 - \beta.$$

This is to achieve an overall study power of $1 - \beta$ (which will be less than, but close to, the power of the one-stage individual genotyping design), while controlling the marker-wise false positive rate at α/M . Here the critical values K_1 and K_{all} of stage I and stage II are functions of π_{sample} , π_{marker} , GRR, and risk-allele frequency. Note that we are able to find multiple pairs $(\pi_{sample}, \pi_{marker})$ of estimates using a grid search method such that the above conditions are met. Among those pairs of parameter estimates, the optimal set $(\pi_{sample}, \pi_{marker})$ is the one that minimizes CFI for a given R.

Two-Stage DNA Pooling Designs

In the two-stage DNA pooling designs, all M markers are genotyped using pooled DNA samples during stage I using $N \times \pi_{sample}^p$ cases and $N \times \pi_{sample}^p$ controls. The most promising set of markers $M \times \pi_{marker}^p$ are individually genotyped during stage II using all cases and controls. Let $\hat{P}_{Ap}$ and $\hat{P}_{Up}$ be the estimated risk-allele frequencies using DNA pooling of case and control groups using $N \times \pi_{sample}^p$ cases and $N \times \pi_{sample}^p$ controls during stage I. Consider m case pools, where each pool contains s cases, and m control pools, where each pool contains s controls, and $m \times s = N \times \pi_{sample}^p$. Assume a simple linear model for the estimated risk allele frequencies with DNA pooling, we have:

$$\hat{P}_{Ap} = \frac{1}{m} \sum_{i=1}^m \hat{P}_{Ap_i} = \frac{1}{m} \sum_{i=1}^m \left(\frac{X_{i1} + X_{i2} + \dots + X_{is}}{2s} + \varepsilon_{A_i} + \varepsilon_{AK_i} \right),$$

and

$$\hat{P}_{Up} = \frac{1}{m} \sum_{i=1}^m \hat{P}_{Up_i} = \frac{1}{m} \sum_{i=1}^m \left(\frac{Y_{i1} + Y_{i2} + \dots + Y_{is}}{2s} + \varepsilon_{U_i} + \varepsilon_{UK_i} \right).$$

Where X_{ij} and Y_{ij} denote the number of copies risk-allele A carried by the j -th case in the i -th case pool and the j -th control in the i -th control pool; ε_{A_i} and ε_{U_i} are the independent pool-specific errors associated with the i -th case pool and the i -th control pool, where they both follow a normal distribution with mean 0 and variance σ_p^2 ; ε_{AK_i} and ε_{UK_i} are the errors due to unequal allele amplification in case and control groups, which are introduced in the allele frequency estimates through error in estimating the correction factor k [17], which is the parameter to adjust for unequal allele amplification. Two steps are involved. First, the coefficient of preferential amplification k , also called k -correction factor, is estimated. For each SNP marker, $\hat{k}$ is estimated by the ratio of A to B from a

number of independent heterozygotes [14], where A and B are the fluorescent signal intensities of alleles A and B , respectively. Therefore, for a particular marker locus, $\hat{k}$ is approximately normally distributed with mean k and variance σ_k^2 . Following the methods of Visscher and Le Hellard [17], we define the coefficient of variation of $\hat{k}$, as $CVK = \sigma_k/k = s.e.(\hat{k})/k$. In the second step, the allele frequency estimates in case and control groups are corrected with the estimated k -correction factor applied,

$$\hat{p} = \frac{A}{A + \hat{k} \times B},$$

where A and B are the observed peak heights of alleles A and B from a DNA pool. Note that the same k -correction factor is applied to the correction of the allele frequency estimates for both case and control groups at a given marker locus. Thus, the errors introduced to the allele frequency estimates in the case and control groups due to unequal allele amplification are not independent. ε_{AK_i} and ε_{UK_i} follow an approximate bivariate normal distribution $N(0, \Sigma_k)$ [17], where

$$\Sigma_k = \begin{pmatrix} p_A^2(1-p_A)^2 \frac{\sigma_k^2}{k^2} & p_A(1-p_A)p_U(1-p_U) \frac{\sigma_k^2}{k^2} \\ p_A(1-p_A)p_U(1-p_U) \frac{\sigma_k^2}{k^2} & p_U^2(1-p_U)^2 \frac{\sigma_k^2}{k^2} \end{pmatrix}.$$

When the test statistic during stage I is T_{pool} , we have:

$$T_{pool} = \frac{\hat{P}_{Ap} - \hat{P}_{Up}}{\sqrt{[\hat{P}_{Ap}(1-\hat{P}_{Ap}) + \hat{P}_{Up}(1-\hat{P}_{Up})]/(2N\pi_{sample}^p) + 2\sigma_p^2/m}}.$$

The test statistic T during stage II is defined similarly as in the one-stage design. Under the null hypothesis of no association, T_{pool} and T follow an approximate bivariate normal distribution $N(0, \Sigma_0)$, where

$$\Sigma_0 = \begin{pmatrix} 1 & a_{12} \\ a_{12} & 1 \end{pmatrix},$$

and

$$a_{12} = \frac{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N)}}{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N\pi_{sample}^p) + 2\sigma_p^2/m}}.$$

Under the alternative hypothesis, T_{pool} and T follow an approximate bivariate normal distribution $N(\mu_2, \Sigma_1)$, where

$$\mu_2 = \begin{pmatrix} \mu_{pool} \\ \mu_{ind} \end{pmatrix} = \begin{pmatrix} \frac{P_A - P_U}{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N\pi_{sample}^p) + 2\sigma_p^2/m}} \\ \frac{P_A - P_U}{\sqrt{[P_A(1-P_A) + P_U(1-P_U)]/(2N)}} \end{pmatrix},$$

$$\Sigma_1 = \begin{pmatrix} b_{11} & a_{12} \\ a_{12} & 1 \end{pmatrix},$$

and

$$b_{11} = \frac{[P_A(1-P_A) + P_U(1-P_U)]/(2N\pi_{sample}^p) + 2\sigma_p^2/m + \sigma_k^2[P_A(1-P_A) - P_U(1-P_U)]^2/(mk^2)}{[P_A(1-P_A) + P_U(1-P_U)]/(2N\pi_{sample}^p) + 2\sigma_p^2/m}.$$

Given the sample size $2N$, π_{sample}^p and π_{marker}^p must be such that the following two equations are satisfied:

$$P(T_{pool} > K_{pool}, T > K | H_0) = \alpha/M,$$

and

$$P(T_{pool} > K_{pool}, T > K | H_1) = 1 - \beta.$$

This is to achieve an overall study power of $1 - \beta$ (which will be less than, but close to, the power of the one-stage individual genotyping design), while controlling the marker-wise false positive rate at α/M . Here the critical values K_{pool} and K of stage I and stage II are functions of π_{sample}^p , π_{marker}^p , GRR, pool size, and risk-allele frequency. Note that we are able to find multiple pairs of $(\pi_{sample}^p, \pi_{marker}^p)$ estimates using a grid search method such that the above conditions are met. Among those pairs of parameter estimates, the optimal set $(\pi_{marker}^p, \pi_{sample}^p)$ is the one that minimizes CFP for a given R .

Fixed Power

The above scenarios can be readily adapted to a situation where a constant study power (e.g., 80% power), rather than a sample size, is fixed. Let α/M be the marker-wise false positive rate and $1 - \beta$ be the desired study power. Assume a balanced case-control design with N samples in each group. The sample size needed to have $1 - \beta$ power in a one-stage design is given by:

$$N = \frac{(Z_{1-\alpha/M} - Z_\beta)^2 \times [P_A(1 - P_A) + P_U(1 - P_U)]}{2(P_A - P_U)^2}.$$

With the calculated N , the optimal CFP and CFI can be obtained following the same procedure as described above with a fixed sample size.

Results

Parameter Settings

We considered two sets of GRRs and four genetic models. The first GRR set is: $\psi_1 = \psi_2 = 1.5$ for the dominant models; $\psi_1 = 1, \psi_2 = 1.5$ for the recessive models; $\psi_1 = 1.5, \psi_2 = 2.25$ for the multiplicative models and $\psi_1 = 1.5, \psi_2 = 3$ for the additive models. For illustration simplicity, we refer to this set as GRR = 1.5. The second set is: $\psi_1 = \psi_2 = 4$ for the dominant models; $\psi_1 = 1, \psi_2 = 4$ for the recessive models; $\psi_1 = 4, \psi_2 = 16$ for the multiplicative models and $\psi_1 = 4, \psi_2 = 8$ for the additive model. We refer to this set as GRR = 4. We set f_0 to 0.01 for all models considered. We also considered risk-allele frequencies of $p = 0.05, 0.1, 0.2, 0.3$ and 0.4 ; the ratio of the per-genotyping cost during stage II to that during stage I being $R = 1, 10$ and 15 ; the pool-specific errors of $\sigma_p = 0, 0.005, 0.01, 0.03$ and 0.05 and the pool sizes of $s = 20, 40, 60$, and 100 . We considered three CVK levels, $0.05, 0.2$ and 0.5 , which were chosen based on suggestions from a simulation study by Yang et

al. [23]. The total sample size was fixed at $2N = 2,000$, with 1,000 cases and 1,000 controls. For the fixed-power scenario, the power was fixed at 80%. The total number of markers was fixed at $M = 500,000$. For simplicity, among 500,000 markers, one marker was assumed to be the true disease marker ($D = 1$). The overall Type I error rate was controlled at $\alpha = 0.05$ level by controlling the marker-wise false positive rate at $\alpha_{marker} = 1 \times 10^{-7}$. The optimal cost fractions were evaluated when the powers of the two-stage designs reached at least 99% of the powers of the one-stage designs.

Effect of R on the Optimal Cost Fractions

The optimal cost fractions of the two-stage DNA pooling designs (CFP) and the optimal cost fractions of the two-stage individual genotyping designs (CFI) under the multiplicative model are summarized in table 1 with different risk-allele frequencies, p , and different ratios of the per-genotyping cost during stage II to that during stage I, R . For the two-stage DNA pooling designs, the pool size s was fixed at 20, the pool-specific error $\sigma_p = 0.03$, and $CVK = 0.2$. We noticed that under the current parameter setting, the cost of the two-stage DNA pooling design is around 10-fold lower than that of the two-stage individual genotyping design no matter what the R or p values are. Similar trends were also observed in the other genetic models (data not shown).

Although increasing trends were observed in both CFI and CFP as R increases (table 1), CFP increases only slightly on the absolute scale. For example, under the multiplicative model when $GRR = 1.5, p = 0.05$ and $\sigma_p = 0.03$, when increases from 1 to 15, CFP increases from 4.7 to 8.4%, while CFI increases from 36 to 55%. One explanation could be that when the per-genotyping cost during stage II is very high, the DNA pooling design tends to use more samples during stage I, however increasing the number of samples does not increase the number of genotyping assays as dramatically as with the individual genotyping design. A similar effect of R on CFP and CFI was also observed under the dominant model and when the GRR and pool-specific error, σ_p , are large (fig. 1).

Distributions of the Optimal Parameters

Also presented in table 1 are the optimal proportions of samples used during stage I (π_{sample}^p or π_{sample}) and the optimal proportions of markers evaluated during stage II (π_{marker}^p or π_{marker}) for the two-stage DNA pooling designs or the two-stage individual genotyping designs under the multiplicative model. For the two-stage DNA pooling designs, π_{sample}^p ranges from 56 to 100% and π_{marker} ranges

Table 1. Effect of the ratio of the per-genotyping cost during stage II to that during stage I, R , on the optimal cost fractions, CFP and CFI, and the distribution of the optimal parameters in the two-stage DNA pooling designs and the two-stage individual genotyping designs under the multiplicative model with different levels of the risk-allele frequency when $GRR = 1.5$.

R	Risk allele frequency	CFP	CFI	π_{sample}^p	π_{sample}	π_{marker}^p	π_{marker}	One-stage power
1	0.05	0.0472	0.3608	0.8600	0.2802	0.0042	0.1120	0.0173
	0.1	0.0399	0.3663	0.7199	0.2861	0.0039	0.1123	0.1609
	0.2	0.0353	0.3587	0.6400	0.2814	0.0033	0.1076	0.6197
	0.3	0.0330	0.3462	0.5999	0.2720	0.0030	0.1019	0.8377
	0.4	0.0318	0.3383	0.5599	0.2665	0.0038	0.0978	0.9006
10	0.05	0.0605	0.5220	0.9995	0.4600	0.0011	0.0115	0.0173
	0.1	0.0491	0.5254	0.8977	0.4640	0.0004	0.0115	0.1609
	0.2	0.0433	0.5132	0.7978	0.4535	0.0003	0.0109	0.6197
	0.3	0.0410	0.4961	0.7500	0.4360	0.0003	0.0107	0.8377
	0.4	0.0393	0.4854	0.7188	0.4316	0.0003	0.0095	0.9006
15	0.05	0.0835	0.5467	1.0000	0.4869	0.0025	0.0078	0.0173
	0.1	0.0511	0.5498	0.9091	0.4907	0.0003	0.0077	0.1609
	0.2	0.0450	0.5377	0.8556	0.4700	0.0001	0.0085	0.6197
	0.3	0.0419	0.5195	0.7780	0.4626	0.0002	0.0071	0.8377
	0.4	0.0406	0.5084	0.7586	0.4521	0.0002	0.0069	0.9006

For the two-stage DNA pooling designs, the pool-specific error $\sigma_p = 0.03$, the pool size $s = 20$, and $CVK = 0.2$.

π_{sample}^p and π_{marker}^p are the optimal parameters for the two-stage DNA pooling designs. π_{sample} and π_{marker} are the optimal parameters for the two-stage individual genotyping designs.

from 0.01 to 0.42%. For the two-stage individual genotyping designs, a smaller proportion of samples are genotyped during stage I, with π_{sample} ranging from 27 to 50%, and a larger proportion of markers are evaluated during stage II, with π_{marker} ranging from 0.69 to 11%. Similar trends were also observed under the other genetic models (data not shown). This observation is consistent with what we expected because the cost savings of the two-stage DNA pooling designs mainly come from the usage of DNA pooling during stage I. If all samples are pooled for genotyping during stage I (i.e., $\pi_{sample}^p = 100\%$), with a pool size of 20 and the pool-specific error less than 0.05, CFP is always around 5%, regardless of the genetic models, risk-allele frequencies, and R (data not shown).

Effect of DNA Pooling-Related Measurement Errors on the Optimal Cost Fractions

Table 2 summarizes the effect of the pool-specific errors, σ_p , on CFP under the multiplicative model when $GRR = 1.5$, $R = 15$ and the pool size $s = 20$. There is an increasing trend in CFP as σ_p increases. However, the magnitude of increase in CFP decreases as the risk-allele

Table 2. Effect of the pool-specific errors, σ_p , on the optimal cost fractions (CFP) under the multiplicative model with different levels of the risk-allele frequency when $GRR = 1.5$, $R = 15$, and the pool size $s = 20$

Risk allele frequency	CFP				
	$\sigma_p = 0$	$\sigma_p = 0.005$	$\sigma_p = 0.01$	$\sigma_p = 0.03$	$\sigma_p = 0.05$
0.05	0.0386	0.0392	0.0410	0.0835	0.6713
0.1	0.0386	0.0389	0.0399	0.0511	0.1331
0.2	0.0378	0.0380	0.0384	0.0450	0.0575
0.3	0.0364	0.0366	0.0370	0.0419	0.0513
0.4	0.0357	0.0359	0.0363	0.0406	0.0491

frequency increases. For example, when the risk-allele frequency is 0.05, CFP increases from 3.9 to 67% as σ_p increases from 0 to 0.05. When the risk-allele frequency is 0.4, CFP increases only slightly from 3.6 to 4.9%. One possible explanation is, as the true risk-allele frequency increases, the relative influence σ_p of on the allele frequency estimates decreases and the study power also de-

Fig. 1. Effect of the ratio of the per-genotyping cost during stage II to that during stage I, R , on the optimal cost fractions CFI and CFP when $\sigma_p = 0.05$ (and the pool size $s = 20$ for the two-stage DNA pooling designs). Note: Solid lines represent the two-stage individual genotyping designs; dotted lines represent the two-stage DNA pooling designs.

creases. Similar trends were also observed in the other genetic models (data not shown).

Figure 2 shows the effect of CVK on CFP, which suggests that CVK has little impact on CFP when the risk-allele frequency is high. When the risk allele is rare, increasing CVK increases CFP. This is because the errors due to unequal allele amplification (introduced in the allele frequency estimates through errors in estimating the k -correction factor) involve not only CVK but also risk-

allele frequency (see Σ_k). When the risk allele is rare, the relative influence of CVK on the allele frequency estimates increases and the study power increases as well. Similar effects of CVK on CFP were also observed in the other genetic models (data not shown).

Effect of the Pool Sizes on the Optimal Cost Fractions

The total costs of stage I in the two-stage DNA pooling designs mainly depend on the number of DNA pools

Fig. 2. Effect of CVK on the optimal cost fractions, CFP, under the multiplicative model when GRR = 1.5, $R = 10$, the pool size $s = 20$ and $\sigma_p = 0.03$.

formed, which is determined by the pool size when the sample size is fixed. The effects of the pool sizes on CFP under the multiplicative model when $R = 15$, $p = 0.1$, and GRR = 1.5 are summarized in table 3. The effect of the pool sizes on CFP depends heavily on σ_p . When $\sigma_p = 0.01$, there is a consistent decreasing trend in CFP as the pool size is increased from 20 to 100. However, when $\sigma_p = 0.05$, there is a consistent increasing trend in CFP as the pool size is increased from 20 to 100. When $\sigma_p = 0.03$, CFP decreases first, then increases when the pool size is increased from 20 to 100. The different effects of the pool sizes at different levels of σ_p are due to the small number of pools generated during stage I when the pool size is very large. This leads to a relatively larger effect of σ_p on the allele-frequency estimates and the study power when σ_p is large. Similar patterns were also observed in π_{marker}^p , as the pool size was increased at different levels of σ_p , but π_{sample}^p increased consistently as the pool size was increased at all levels of σ_p considered. Similar trends were also observed in the other genetic models (data not shown).

Effects of All Three Factors on the Optimal Cost Fractions when the Study Power Is Fixed

We also investigated the effects of the R , pooling-related measurement errors and pool sizes on the optimal cost fractions when the study power was fixed at 0.8. Similar patterns of the effects of R , the pool-specific errors and the pool sizes on CFP were observed compared

Table 3. Effect of the pool size, s , on the optimal cost fractions (CFP) under the multiplicative model with different levels of σ_p when GRR = 1.5, $R = 15$ and the risk-allele frequency $p = 0.1$

σ_p	Pool size s	CFP	π_{marker}^p	π_{sample}^p
0.01	20	0.0399	0.00019	0.7395
	40	0.0218	0.00012	0.7970
	60	0.0156	0.00004	0.8913
	100	0.0105	0.00003	0.9746
0.03	20	0.0511	0.00034	0.9091
	40	0.0489	0.00159	0.9991
	60	0.1261	0.00727	0.9997
	100	0.6673	0.04382	1.0000
0.05	20	0.1331	0.00554	0.9998
	40	0.8758	0.05672	1.0000
	60	2.1788	0.14412	1.0000
	100	5.1556	0.34304	1.0000

As a comparison, for two-stage individual genotyping designs, when GRR = 1.5, $R = 15$, and the risk-allele frequency $p = 0.1$, CFI = 0.5498, $\pi_{marker}^p = 0.00773$ and $\pi_{sample}^p = 0.4907$.

to when the sample size was fixed. A set of tables and figures are provided in the supplementary materials (www.karger.com/doi/10.1159/000164398).

Discussion

We have proposed an optimal two-stage DNA pooling design as a screening tool to identify genetic susceptibility loci in GWA studies, where we focused on the joint analysis in the two-stage design given its demonstrated greater power than the replication analysis [22]. The effects of the following three factors on the cost of the optimal two-stage DNA pooling designs were studied: (1) DNA pooling-related measurement errors; (2) the ratio of the per-genotyping assay cost during stage II to that during stage I, and (3) the DNA pool sizes. We considered a wide range of parameter settings. Our results suggest that the optimal two-stage DNA pooling designs are much more cost-effective than the optimal two-stage individual genotyping designs under most scenarios considered.

The effects of the three factors on the optimal study cost of the two-stage DNA pooling design can be summarized as follows. First, increasing the pool-specific errors σ_p increases the optimal cost fraction CFP, the proportion of samples evaluated during stage I with DNA pooling (π_{sample}^p) and the proportion of markers evaluated

during stage II with individual genotyping (π_{marker}^p). In most scenarios, CFP is smaller than CFI, providing a more cost-effective design. In some scenarios, when σ_p is large and the risk allele is rare, or when σ_p is moderate and the pool size is large, the relative influence of σ_p on the allele-frequency estimates increases and the study power increases. In these scenarios, the two-stage DNA pooling designs might not be as cost-effective as the two-stage individual genotyping designs. The coefficient of variation of $\hat{k}$, CVK, hardly affects CFP when the risk allele frequency is high. When the risk allele is rare, increasing CVK increases CFP. This is because the errors introduced by estimating the k correction factor involve not only CVK but also the true risk-allele frequency. When the risk allele is rare, the relative influence of CVK on the allele-frequency estimates increases and the study power increases as well.

Second, increasing the ratio of the per-genotyping cost during stage II to that during stage I increases both CFP and CFI. Note that when R increases from 1 to 15, the absolute increase of 17.5% for CFI might have a huge impact on the actual study cost, while the absolute increase of 1.4% for CFP will have much less impact on the actual study cost. The reason that R has a smaller effect on CFP than CFI is because when R increases, more samples could be evaluated during stage I with DNA pooling and increasing the number of samples does not increase the number of genotyping assays as dramatically as with individual genotyping. For example, as shown in table 1, when the risk-allele frequency is 0.1, about 20% more samples will be evaluated during stage I in both the two-stage DNA pooling designs and the two-stage individual genotyping designs when R increases from 1 to 15. However, the increase in the number of genotyping assays during stage I with 20% more samples will be much smaller with DNA pooling than with individual genotyping.

Third, increasing the pool size does not always decrease CFP. The effect of pool size on CFP is heavily dependent on σ_p . When σ_p is smaller than 0.05, CFP is always lower than CFI if the pool size is between 20 and 60. When σ_p is 0.05, CFP is lower than CFI only when the pool size is 20. Therefore, in the DNA pooling-based GWA studies when σ_p is unknown, a pool size of 20 may be the best choice.

Although our results clearly suggest the cost-effectiveness of the two-stage DNA pooling designs, there are disadvantages to using DNA pooling as well. Note that since no genotypic data is available with DNA pooling-based studies, we could not perform tests of genotypic associa-

tion based on the assumptions of different genetic models such as dominant or recessive models. We could only perform tests of allelic association. In addition, it is common for large-scale genotyping studies to have some missing genetic markers in some subjects. Existing imputation methods to infer missing genotypes such as impute [24] and TUNA [25], which utilize haplotype information, were developed for individual genotyping and provide benefits to individual genotyping but not to DNA pooling. We expect more effort from researchers to develop new imputation methods for DNA pooling. Moreover, pooled genotype data pose even greater challenges in haplotype-based analysis than individual genotype data. Several studies have been done on the efficiency of haplotype estimation and on haplotype-disease associations with pooled DNA [26–28]. Although software has been developed for haplotype estimation with pooled DNA [27] and for directly analyzing SNP array intensity data [12], the lack of software for analyzing pooled data with more sophisticated methods makes the amount of time spent analyzing pooled data much greater than that for individual genotyping data. More research effort is needed in developing software to further promote DNA pooling designs.

We did not take linkage disequilibrium (LD) into account. We assumed all SNP markers to be independent. In analysis of disease association with real data, the standard significance tests assuming independence may not be valid. A permutation procedure that maintains the autocorrelation between test statistics of the adjacent SNP markers could be applied to obtain the empirical Type I error. We did not consider technical replicates of DNA pools but assume only one replicate per pool. When pooling error is quite large, running two or more arrays per pool might be helpful. This point was demonstrated nicely in the work of Ji et al. [29]. However, in array-based DNA pooling, a recent study [20] suggests that the most pooling variation is attributable to array errors rather than pool-construction errors. We only considered a rather stringent Type I error rate with the Bonferroni correction. We believe a similar relationship between CFP and CFI would be observed when applying other types of multiple comparison adjustments, such as the False Discovery Rate (FDR). We assumed genetic models with both DNA pooling strategy and individual genotyping in the design stage. If a certain genetic model is assumed, the genetic relative risks of genotypes AA and Aa will follow some relationship (e.g., under the dominant model, GRR of genotype AA will be equal to GRR of genotype Aa). If no genetic model is assumed, the genetic relative

risks of genotypes AA and Aa will not necessarily follow any relationship. We believe a similar relationship between CFP and CFI would be observed when no genetic model is assumed.

In summary, the optimal two-stage DNA pooling designs are more cost-effective than the optimal two-stage individual genotyping designs under most scenarios considered. Because costs remain the major limiting factor for GWA studies, applying the optimal two-stage DNA pooling designs will make hypothesis-free GWA studies much more feasible, especially in the beginning stages of disease mapping when interests are focused on detecting the differences in allele frequencies between the case and control groups. Our results suggest that the optimal two-stage DNA pooling designs have only 1–5% of the costs of the one-stage individual genotyping designs, leading to a 4- to 20-fold decrease in costs compared to the optimal two-stage individual genotyping designs, which have 20–50% of the costs of the one-stage individual genotyping

designs. When the pool-specific error σ_p is large, relatively small pool size should be considered. However, a pool size of 20 or smaller will lead to a cost-effective design even when pooling error is large. In general, two-stage DNA pooling designs use a relatively large proportion (at least 70% when the per-genotyping cost is higher during stage II) of samples during stage I, and evaluate a relatively small proportion (less than 0.5%) of markers during stage II. To get the optimal two-stage DNA pooling designs in specific scenarios, we recommend the readers to use our R program.

Acknowledgments

This work was supported in part by grant 5-T32-CA09529 from the National Cancer Institute. We thank Dr. Susan E. Hodge for helpful discussions.

References

- 1 Risch N, Merikangas K: The future of genetic studies of complex diseases. *Science* 1996; 273:1516–1517.
- 2 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* 2003;suppl 33:228–237.
- 3 Carlson C, Eberle M, Kruglyak L, Nickerson D: Mapping complex disease loci in whole-genome association studies. *Nature* 2004; 429:446–452.
- 4 Thomas DC: Are we ready for genome-wide association studies? *Cancer Epidemiol Biomarkers Prev* 2006;15:595–598.
- 5 Sham P, Bader JS, Craig I, O'Donovan M, Owen M: DNA Pooling: A tool for large-scale association studies. *Nat Rev Genet* 2002;3:862–871.
- 6 Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB: Two-stage designs for gene-disease association studies. *Biometrics* 2002;58:163–170.
- 7 Satagopan JM, Elston RC: Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 2003;25: 149–157.
- 8 Satagopan JM, Venkatraman ES, Begg CB: Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 2004;60:589–597.
- 9 Zehetmayer S, Bauer P, Posch M: Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* 2005;21:3771–3777.
- 10 Wang H, Thomas DC, Pe'er I, Stram DO: Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol* 2006;30:356–368.
- 11 Kraft P: Efficient two-stage genome-wide association designs based on false positive report probabilities. *Pac Symp Biocomput* 2006;11:523–534.
- 12 Pearson JV, Huettelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL, Webster JA, Beach T, Sando SB, Aasly JO, Heun R, Jensen F, Kolsch H, Tsolaki M, Daniilidou M, Reiman EM, Papassotiropoulos A, Hutton ML, Stephan DA, Craig DW: Identification of the genetic basis for complex disorders by use of pooling-based genome-wide single-nucleotide-polymorphism association studies. *Am J Hum Genet* 2007;80:126–139.
- 13 Docherty AJ, Butcher LM, Schalkwyk LC, Polmin R: Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genome-wide association studies. *BMC Genomics* 2007;8:214.
- 14 Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, Hetttema JM, An SS, Neale MC, van den Oord EJCG, Kendler KS, Chen X, Boomsma DI, Middeldorp CM, Hottenga JJ, Slagboom PE, Flint J: A Whole genome association study of neuroticism using DNA pooling. *Mol Psychiatr* 2007;1359: 4184.
- 15 Melquist S, Craig DW, Huettelman MJ, Crook R, Pearson JV, Baker M, Zismann VL, Gass J, Adamson J, Szelinger S, Corneveaux J, Cannon A, Coon KD, Lincoln S, Adler C, Tuite P, Calne DB, Bigio EH, Uitti RJ, Wszolek ZK, Golbe LI, Caselli RJ, Graff-Radford N, Litvan I, Farrer MJ, Dickson DW, Hutton M, Stephan DA: Identification of a Novel Risk Locus for Progressive Supranuclear Palsy by a Pooled Genomewide Scan of 500,288 Single-Nucleotide Polymorphisms. *Am J Hum Genet* 2007;80:769–778.
- 16 Zuo Y, Zou G, Zhao H: Two-stage designs in case-control association analysis. *Genetics* 2006;173:1747–1760.
- 17 Visscher PM, Hellard SL: Simple method to analyze SNP-based association studies using DNA pools. *Genet Epidemiol* 2003;24:291–296.
- 18 Le Hellard SL, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple C, Muir WJ, Blackwood D, Porteous D, Evans K: SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res* 2002; 30:e74.
- 19 Prentice R, Qi L: Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics* 2006;7:339–354.
- 20 Macgregor S: Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur J Hum Genet* 2007;15:501–504.

- 21 Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'donovan MC: Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* 2006;7:27.
- 22 Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–213.
- 23 Yang HC, Pan CC, Lu RC, Fann CS: New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification. *Genetics* 2005;169:399–410.
- 24 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 2007;39:906–913.
- 25 Nicolae DL: Testing Untyped Alleles (TUNA) – applications to genome-wide association studies. *Genet Epidemiol* 2006;30:718–727.
- 26 Wang S, Kidd KK, Zhao H: On the use of DNA pooling to estimate haplotype frequencies. *Genet Epidemiol* 2003;24:74–82.
- 27 Yang Y, Zhang J, Hoh J, Matsuda F, Xu P, Lathrop M, Ott J: Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc Natl Acad Sci USA* 2003;100:7225–7230.
- 28 Zeng D, Lin DY: Estimating haplotype-disease associations with pooled genotype data. *Genet Epidemiol* 2005;28:70–82.
- 29 Ji F, Finch SJ, Haynes C, Mendell NR, Gordon D: Incorporation of genetic model parameters for cost-effective designs of genetic association studies using DNA pooling. *BMC Genomics* 2007;8:238.