Optimum two-stage designs in case–control association studies using false discovery rate

Aya Kuchiba · Noriko Y. Tanaka · Yasuo Ohashi

Received: 19 May 2006 / Accepted: 18 August 2006 / Published online: 27 September 2006
© The Japan Society of Human Genetics and Springer 2006

Abstract Genetic association studies using case–control designs are often done to identify loci associated with disease susceptibility. These studies are often expensive to perform, due to a large number of genetic markers. Several types of two-stage designs are proposed and used from the point of cost effectiveness. We proposed to control the false discovery rate for multiple-testing correction in two-stage designs, using optimal sample sizes and criteria for selecting markers associated with a disease in each stage to minimize the cost of genotyping. The expected power and cost of two-stage designs were compared with those of one-stage designs, under the assumptions that the genetic markers are independent and total sample size is fixed. The results showed that the proposed two-stage procedure usually reduced the cost of genotyping by 40–60%, with a power similar to that of the one-stage designs. In addition, the sample size and selection criteria, which are optimized parameters, are defined as a function of a prior probability that marker–disease association is true. So, the effects of mis-specification of a prior probability on efficiency were also considered.

Keywords Population-based association studies · Common diseases · Power · Genotyping cost · A prior probability

Introduction

Common diseases are considered to be caused by numerous genetic and environmental factors, and many studies have been done on the association between disease and genetic markers. Association studies are expected to be more powerful than linkage studies to detect alleles that confer modest risk to common diseases (Risch and Merikangas 1996), and common and modest-risk alleles are thought to be more important in public health (Carlson et al. 2004). Population-based association studies using case–control designs have become widespread in an attempt to identify genes related to common diseases. Previously, associations between disease and genetic factors targeted several markers within a few candidate genes. Recent advances in high-volume genotyping technology have made it possible to scan whole genomes using hundreds of thousands of single nucleotide polymorphisms (SNPs). Whereas these large-scale studies have many advantages to detect causal genes, the critical issues for efficiency remain, and such studies require stepwise screening processes to increase efficiency in both single and multiple studies (van den Oord and Sullivan 2003; Hirschhorn and Daly 2005; Thomas et al. 2005). In this study we focused on the second screening studies with several hundreds to thousands of sparse markers after the first screening of the whole genome or the selection of tag SNPs to eliminate redundancy in the information provided by the SNPs.

Aya Kuchiba and Noriko Y. Tanaka have contributed equally to this study.

A. Kuchiba (✉) · Y. Ohashi
Department of Biostatistics/Epidemiology and Preventive
Health Sciences, Graduate School of Medicine,
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, Japan
e-mail: kuchiba@epistat.m.u-tokyo.ac.jp

N. Y. Tanaka
Department of Clinical Bioinformatics,
Graduate School of Medicine, University of Tokyo,
Tokyo, Japan

Generally, the disease-associated markers can be identified by hypothesis testing, and the efficiency is considered by power (true positives), sample size, false positives, and cost. Because variants that contribute to common disease have modest effect, large sample sizes are needed. Additionally, significance criteria are rigorous because of multiple-testing correction, requiring further sample sizes. Consequently, these studies are expensive (Botstein and Risch 2003).

In genetic epidemiology studies many multi-stage designs have been suggested and applied to reduce time and cost on average (Bödder and Ziegler 2001). Most of these designs are two-stage. One of these approaches is to increase sample size in subsequent stages for more promising markers. This approach in two-stage design evaluates all markers from a subset of individuals in the first stage, while the most promising markers selected from the first stage are evaluated in the second stage, using additional individuals (Satagopan et al. 2002, 2004; Satagopan and Elston 2003). These studies show that use of the two-stage approach as a sampling strategy for marker selections is more cost-effective than the one-stage approach. However, in addition to the issue of cost, other problems arise from multiple testing, because a large number of markers is tested simultaneously. The commonly controlled family-wise error rate (FWER) is the probability of yielding one or more false positives when all hypotheses are null. The most familiar example of this is Bonferroni's method. Though controlling FWER is a familiar method for multiple comparison problems, what happens in the situation when no marker is associated with disease? Often, studies aim to find a number of interesting features in numerous markers rather than to detect any one marker. The investigators would assume that more than one marker is associated with disease. In this case, FWER confers such strict criteria that many features may be missed. Furthermore, these criteria depend on the number of markers used in the study, making the interpretation of results difficult (Colhoun et al. 2003; Wacholder et al. 2004). Benjamini and Hochberg (1995) introduced a new multiple-testing error measure called the false discovery rate (FDR), which is the expected proportion of falsely rejected null hypotheses among all rejected hypotheses. Although a prior probability, which is the proportion of truly associated markers, must be set to use the FDR criteria, it is easy to interpret the outcome of tests. Another advantage is that the efficiency of the subsequent study can be evaluated by the FDR definition. In brief, it is possible to avoid additional costs incurred by examining waste markers that are not associated with a disease. In recent years the FDR has

been thought to be more appropriate than the FWER for exploratory studies (Colhoun et al. 2003; Sabatti et al. 2003; Wacholder et al. 2004).

Benjamini and Yekutieli (2005) have reported that optimal multi-stage designs can be constructed with the goal of decreasing the total cost, while the study still controls the FDR at a nominal level. Also, Zehetmayer et al. (2005) have optimized the significance level at the first stage and the fraction of total sample size used in the first stage by controlling the FDR and maximizing the expected power under the fixed total costs; this showed that these two-stage designs have higher power than one-stage designs. Such results are seen to have value for conducting association studies. However, the sample size and effect size must also be considered and may not be very realistic for an association study.

In practice, the total number of subjects will often be limited. The cohort study in which biological samples have been collected at baseline is an example. In two-stage designs with a total sample size constraint, the cost of the study is affected by the sample sizes and selection criteria used at each stage. Thus, we propose two-stage designs in case-control association studies by controlling the FDR and by optimizing the sample sizes and marker selection criteria in each stage and cost ratios to those of a one-stage design, under the assumption that the markers are independent. By this method we can utilize investigators' knowledge as a prior probability (assumed to be determined subjectively by the investigators). Consequently, the design for each study allows one the flexibility to choose the optimal sample size and selection criteria that fit the actual situation. At the same time, the subjective prior probability would affect the efficiency of the study. Therefore, the robustness of the proposed procedure was evaluated for mis-specification of a prior probability.

Materials and methods

Consider evaluating m markers, using a case-control design with a fixed number of subjects (n) available for each group. It is assumed that the m markers are not in linkage disequilibrium (LD) with each other and can be considered to be independent. Suppose there are D markers associated with disease; it would be assumed that these markers are in complete LD with D disease loci or that the markers themselves would be the disease-susceptible polymorphisms. Also, these D markers have the same effect size $\mu = \mu_d$ on disease, which is the target effect size that we want to detect. π_1 is a prior probability that denotes the proportion of truly associated markers to all markers; $\pi_1 = D/m$. Any

typical test statistic, X , for the association of a marker derived from n subjects will follow an asymptotic normal distribution $N(n\mu, n\sigma^2)$, in which $n\mu$ and $n\sigma^2$ are mean and variance, respectively, of a distribution; μ would be zero when there is no association between the disease and the marker. Without loss of generality, it can be assumed that $\sigma^2=1$ and can be considered one-sided testing. In a case-control design the allele (or genotype or haplotype) frequencies in cases and controls are compared at every marker locus. Generally, the odds ratio (OR) is used to measure the association between a disease and a marker. If we assume that we focus on the effect of certain genotype(s), the relationship between the OR (θ), the effect size (μ_d), and the genotype, the frequency of interest in the controls (p_0) would be:

$$\log(\theta) = \mu_d \times \sqrt{\frac{1}{p_0} + \frac{1}{1-p_0} + \frac{(1-p_0) + \theta p_0}{\theta p_0} + \frac{(1-p_0) + \theta p_0}{1-p_0}}. \quad (1)$$

One-stage procedure

In this situation m markers are evaluated using all $2n$ subjects and controlling the FDR at a nominal level. Let FDR_1 be the desired FDR level. α is the significance level for each marker, and $1-\beta$ denotes the power for effect size μ_d . FDR_1 is defined as follows:

$$FDR_1 = \frac{\pi_0 \alpha}{\pi_0 \alpha + \pi_1 (1 - \beta)}, \quad (2)$$

in which $\pi_0=1-\pi_1$. Hence, α can be obtained, given n , μ_d , π_1 , and FDR_1 . The false non-discovery rate (FNR), the expected proportion of the falsely accepted null hypotheses among all accepted null hypotheses, can be thought of as another error in this study. FNR_1 , which is the FNR in this design, is calculated by the following:

$$FNR_1 = \frac{\pi_1 \beta}{\pi_0 (1 - \alpha) + \pi_1 \beta}. \quad (3)$$

In addition, the cost of genotyping, T_1 , can be written with the arbitrary constant k as:

$$T_1 = k(n \times m). \quad (4)$$

Proposed two-stage procedure

In the first stage, all m markers are evaluated using n_1 subjects for each group, and ($m' < m$) markers are selected by testing with a significance level of α_1 . The proportion of this selected marker is denoted by s' ($=m'/m$). In the second stage n_2 subjects for each group are added, thereby giving a total of n ($=n_1+n_2$) subjects per group. m' markers are tested at a significance level of α_2 using all $2n$ subjects. The markers that are significant in the second stage are identified as disease-associated markers. In proposed two-stage designs, there are three design parameters: (1) n_1 , sample size for each group in the first stage; (2) α_1 , significance level (i.e., the criterion for selecting markers) in the first stage; and (3) α_2 , significance level in the second stage. These parameters (n_1 , α_1 , α_2) are decided to control the FDR at a predetermined nominal level.

(X_1, X_2) denote the test statistics of a marker in the first and second stages, as $X_1 \sim N(n_1\mu, n_1)$, $X_2 \sim N(n\mu, n)$. Because cases and controls are unrelated individuals, the test statistics follow the Markov property, i.e., $X_2 - X_1$ is independent of X_1 . Therefore, the statistics pair (X_1, X_2) has bivariate normal distribution with a mean ($n_1\mu, n\mu$) and covariance matrix Σ , in which

$$\Sigma = \begin{bmatrix} n_1 & n_1 \\ n_1 & n \end{bmatrix}.$$

In the first stage, the FDR is given by the following:

$$FDR_{2'} = \frac{\pi_0 \alpha_1}{\pi_0 \alpha_1 + \pi_1 (1 - \beta_1)}, \quad (5)$$

in which $1-\beta_1$ denotes the power for the effect size μ_d . The expected proportion of the significant markers in the first stage, $E[s']$, is given by the following:

$$E[s'] = \pi_0 \alpha_1 + \pi_1 (1 - \beta_1). \quad (6)$$

Now consider the second stage. Let T_i and Γ_i denote the statistics and the rejection regions at the i th stage, respectively. The FDR for the two-stage approach is represented below:

$$\begin{aligned} FDR_{\text{for two-stage approach}} &= \frac{\pi_0 \Pr_{H=0}[T_1 \in \Gamma_1] \Pr_{H=0}[T_2 \in \Gamma_2 | T_1 \in \Gamma_1]}{\pi_0 \Pr_{\mu=0}[T_1 \in \Gamma_1] \Pr_{H=0}[T_2 \in \Gamma_2 | T_1 \in \Gamma_1] + \pi_1 \Pr_{H=1}[T_1 \in \Gamma_1] \Pr_{H=1}[T_2 \in \Gamma_2 | T_1 \in \Gamma_1]} \\ &= \frac{\pi_0 \Pr_{H=0}[T_1 \in \Gamma_1, T_2 \in \Gamma_2]}{\pi_0 \Pr_{H=0}[T_1 \in \Gamma_1, T_2 \in \Gamma_2] + \pi_1 \Pr_{H=1}[T_1 \in \Gamma_1, T_2 \in \Gamma_2]}, \end{aligned} \quad (7)$$

in which $\Pr_{H=0}$ and $\Pr_{H=1}$ denote the probabilities under the null hypothesis and the alternative. Hence, this can be obtained by two probabilities: one is the overall α error, or the probability that a true null marker is selected in the first and second stage, and the other is the overall power, or the probability that a true associated marker is selected in the first and second stage. In this case, $P_{\mu=\mu_d}$ denotes the probability that a marker with effect size μ is rejected in the first stage and also in

$$\begin{aligned} E[T_2] &= k(n_1 \times m + n_2 \times m \times E[s_1]) \\ &= k[(Z_{1-\alpha_1} + Z_{1-\beta_1})^2 \times m \\ &\quad + \{n\mu_d^2 - (Z_{1-\alpha_1} + Z_{1-\beta_1})^2\} \\ &\quad \times m \times \{\pi_0\alpha_1 + \pi_1(1 - \beta_1)\}] / \mu_d^2. \end{aligned} \quad (11)$$

The expected cost ratio in the one-stage design, F , is written as follows:

$$F = \frac{E[T_2]}{T_1} = \frac{(Z_{1-\alpha_1} + Z_{1-\beta_1})^2 + \{n\mu_d^2 - (Z_{1-\alpha_1} + Z_{1-\beta_1})^2\} \{\pi_0\alpha_1 + \pi_1(1 - \beta_1)\}}{n\mu_d^2}. \quad (12)$$

the second stage, and the overall α error and the overall power for effect size μ_d can be represented by $P_{\mu=0}$ and $P_{\mu=\mu_d}$, respectively. Note that this overall power, $P_{\mu=\mu_d}$, is thought of as the expected proportion of rejected markers among all truly associated markers, because it is assumed that truly associated markers have the same effect size, μ_d . The expected power is called $P_{\mu=\mu_d}$. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and the probability density function of a standard normal distribution. Z_a denotes the 100 $\times a^{\text{th}}$ percentile of a standard normal distribution. $P_{\mu=\mu}$ is given by the following:

$$\begin{aligned} P_{\mu=\mu} &= \text{Prob}[X_2 > Z_{1-\alpha_2}\sqrt{n}, X_1 > Z_{1-\alpha_1}\sqrt{n_1}] \\ &= \int_{Z_{1-\alpha_1}}^{\infty} [1 - \Phi(\{Z_{1-\alpha_2}\sqrt{n} - z\sqrt{n_1} - n_2\mu\}/\sqrt{n_2})] \\ &\quad \times \phi(z - \mu\sqrt{n_1})dz. \end{aligned} \quad (8)$$

The FDR after the second stage can be written, using $P_{\mu=0}$ and $P_{\mu=\mu_d}$, as follows:

$$FDR_2 = \frac{\pi_0 P_{\mu=0}}{\pi_0 P_{\mu=0} + \pi_1 P_{\mu=\mu_d}}. \quad (9)$$

FDR_2 means the proportion of non-associated markers among the identified markers by this two-stage procedure. The FNR can also be written as follows:

$$\begin{aligned} FNR_2 &= \{\pi_1\beta_1 + \pi_1(1 - P_{\mu=\mu_d})\} / \{\pi_0(1 - \alpha_1) + \pi_1\beta_1 \\ &\quad + \pi_0(1 - P_{\mu=0}) + \pi_1(1 - P_{\mu=\mu_d})\}. \end{aligned} \quad (10)$$

The expected cost of genotyping, T_2 , is given by the following:

Estimating optimal parameters

The purpose of this study was to minimize the expected cost by optimizing the three unknown design parameters n_1 , α_1 , and α_2 . Given parameters of n , π_1 , μ_d , and FDR_2 , the optimal parameters are defined as follows. The FDR is controlled by FDR_2 , the difference between the power of the two-stage design and the power of the one-stage design, which is $\leq 1\%$, and the expected cost is the minimum among the parameters that satisfy the former two conditions.

Let the values of n , π_1 , μ_d , and FDR_2 be given. The n_1 value is fixed, and α_1 and β_1 are calculated under the given FDR_2' . Using these values, the expected cost ratio can be obtained using Eq. 12, and $P_{\mu=\mu}$ is expressed as a function of α_2 . The value of α_2 can be estimated from Eq. 9 by a predetermined FDR_2 . The expected power, $P_{\mu=\mu_d}$, can also be obtained at once. Every point is searched on a grid with values that correspond to the values of n_1 and FDR_2' . The design parameters (n_1 , α_1 , α_2), expected power, and expected cost are estimated at every grid point; then, the optimal parameters that satisfy the above conditions can be identified.

Evaluating the performance for mis-specification of a prior probability

To perform the study using optimal parameters, investigators have to determine a prior probability, π_1 . The current situation is that the optimal parameters of expected power and expected cost depend on π_1 , so π_1 must be subjectively determined by investigators. Hence, π_1^* denotes the subjective prior probability, while π_1 is the true one. Controlling the FDR at a nominal level to achieve the expected power and

expected cost requires π_1^* to equal π_1 . Therefore, we evaluated the performance related to mis-specification of a prior probability by estimating the FDR, the expected power, and the expected cost ratio. These estimates were obtained from Eqs. 8, 9, and 12, with the optimal parameters corresponding to π_1^* and π_1 .

Results

Optimal parameters

The optimal parameters were estimated for combinations of $n=200, 500$, and $1,000$ for each group, and $\pi_1=0.001, 0.002, 0.005, 0.01, 0.02, 0.025, 0.05$, and 0.1 . In addition, the target effect sizes, μ_d , were evaluated from 0.1 to 0.3 at intervals of 0.02 . Table 1 shows the ORs that correspond to $\mu_d (=0.1-0.3)$ and $p_0 (=0.1-0.4)$. The FDR levels are evaluated for $FDR_2=0.05$ and $FDR_2=0.1$. The optimal parameters, expected power, expected cost ratio, and FNR_2 were calculated. Note that the proposed optimal parameters and these performances depend on π_1 rather than on the number of markers, because of the use of FDR criteria against the multiple-comparison problems.

The examples of the optimal parameters when $n=1,000$ are shown in Table 2. Suppose that the genotype frequency of interest in the controls is 0.3 and the target OR is 1.5 ; then, 1% of all markers used in the study are likely to be associated with the disease. This OR corresponds to $\mu_d=0.14$ (Table 1). In Table 2, in the first stage, each marker is tested at a significance level of 0.153 using 421 of the $1,000$ subjects, and in the second stage, the selected markers from the first stage are tested at a significance level of 0.00048 , using all $1,000$ subjects. Table 3 illustrates the expected values of the proportion of selected markers at each stage, the power and the cost ratio by using the optimal param-

Table 1 OR (θ) corresponding to effect size (μ_d) and genotype frequency of interest in the controls (p_0)

μ_d	$p_0=0.1$	$p_0=0.2$	$p_0=0.3$	$p_0=0.4$
0.10	1.55	1.40	1.35	1.33
0.12	1.68	1.50	1.43	1.41
0.14	1.82	1.60	1.52	1.49
0.16	1.97	1.70	1.61	1.58
0.18	2.13	1.81	1.71	1.67
0.20	2.30	1.93	1.81	1.77
0.22	2.48	2.05	1.92	1.88
0.24	2.67	2.19	2.04	1.99
0.26	2.88	2.33	2.16	2.11
0.28	3.10	2.48	2.29	2.24
0.30	3.34	2.63	2.43	2.38

Table 2 Examples of the two-stage designs with the optimal parameters in which $n=1,000$ and $FDR_2=0.05$ (μ_d target effect size, π_1 a prior probability, n_1 sample size in the first stage, α_1 significance level in the first stage, α_2 significance level in the second stage)

μ_d	π_1	Optimal parameters		
		n_1	α_1	α_2
0.1	0.001	305	0.056	0.00001
0.1	0.005	282	0.196	0.00007
0.1	0.01	361	0.156	0.00019
0.1	0.05	420	0.215	0.00168
0.1	0.1	474	0.217	0.00423
0.14	0.001	399	0.092	0.00004
0.14	0.005	398	0.156	0.00022
0.14	0.01	421	0.153	0.00048
0.14	0.05	393	0.235	0.00280
0.14	0.1	423	0.222	0.00615
0.18	0.001	362	0.097	0.00005
0.18	0.005	323	0.160	0.00028
0.18	0.01	353	0.133	0.00059
0.18	0.05	338	0.165	0.00331
0.18	0.1	336	0.172	0.00738

eters, and also the power of one-stage design. In the above situation the expected power was then approximately 86% , and the expected cost was 51% of the one-stage design; the proportion of the identified markers associated with disease was 0.9% . In Table 3 the number of markers finally selected in the study increased with π_1 . In contrast, the number of markers selected in the first stage did not have a monotonic relationship with π_1 , due to restriction on the expected

Table 3 The expected proportion of selected markers, expected power, expected cost ratio, and the power of one-stage design in which $n=1,000$ and $FDR_2=0.05$ (μ_d target effect size, π_1 a prior probability, s' proportion of selected markers in the first stage, s proportion of selected markers in the second stage)

μ_d	π_1	Expected values in proposed two-stage design				One-stage design
		s'	s	Power	Cost ratio	Power
0.1	0.001	0.056	0.0001	0.102	0.34	0.111
0.1	0.005	0.199	0.0013	0.247	0.42	0.257
0.1	0.01	0.163	0.0035	0.335	0.46	0.344
0.1	0.05	0.249	0.0304	0.577	0.56	0.587
0.1	0.1	0.287	0.0724	0.688	0.62	0.697
0.14	0.001	0.093	0.0007	0.665	0.45	0.675
0.14	0.005	0.160	0.0043	0.810	0.49	0.818
0.14	0.01	0.161	0.0090	0.859	0.51	0.867
0.14	0.05	0.272	0.0497	0.944	0.56	0.949
0.14	0.1	0.298	0.1012	0.961	0.59	0.971
0.18	0.001	0.098	0.0010	0.955	0.42	0.964
0.18	0.005	0.165	0.0051	0.977	0.43	0.987
0.18	0.01	0.141	0.0103	0.983	0.44	0.992
0.18	0.05	0.206	0.0521	0.989	0.47	0.998
0.18	0.1	0.254	0.1042	0.990	0.5	0.999

cost. These results show that the proposed two-stage design has power similar to that of the one-stage design and that the expected costs are reduced by 40–60%.

When n and FNR_2 are given, the values of the optimal parameters, expected power, and expected cost ratio, vary with μ_d and π_1 . Figure 1 illustrates the relationships between a prior probability and the optimal parameters, expected power, expected cost ratio, and FNR_2 . n_1/n was about 0.4 and α_1 was from 0.1 to 0.2 in any prior probability (Fig. 1a). These two parameters influence one another to minimize the expected cost. In contrast, as α_2 is decided by a predetermined FDR_2 , not the expected cost, α_2 increased monotonically with a prior probability (Fig. 1b). In Figure 1a the expected power naturally increased with a prior probability, and the expected cost ratio also increased monotonically with a prior probability but was convergent to approximately 0.55. This is because the expected cost ratios depend on n_1 and α_1 as well as

on a prior probability. FNR_2 also increased with a prior probability (Fig. 1b), but even so, FNR_2 remained $<5\%$, which is considered low enough and an acceptable range. These tendencies in Fig. 1 were the same, regardless of the effect size and sample size that we set.

Mis-specification of a prior probability

As mentioned earlier, the results are assured only when the subjective prior probability, π_1^* , equals the true prior probability, π_1 . Therefore, we evaluated the performance related to the mis-specification of the prior probability. Figure 2 illustrates the results when the optimal parameters for $\pi_1^*=0.001, 0.005, 0.01, 0.05$ and 0.1 were used and when π_1 is 0.01 (Fig. 2a) and 0.05 (Fig. 2b). These results are for $n=1,000$, $\mu_d=0.14$, and $FDR_2=0.05$. The vertical and horizontal reference lines in each figure show that $\pi_1^*=\pi_1$ and $FDR=0.05$, respectively. The optimal parameters for lower π_1^*

Fig. 1 Characteristics of the optimal two-stage design in which $n=1,000$, $FNR_2=0.05$, and $\mu_d=0.14$. **a** The relationships between a prior probability and significance level in the first stage (α_1), proportion of sample size in the first stage to all sample size (n_1/n), expected power ($power$) and expected cost ratio ($cost$). **b** The relationships between a prior probability and significance level in the second stage (α_2) and false negative rate (FNR_2)

n), expected power ($power$) and expected cost ratio ($cost$). **b** The relationships between a prior probability and significance level in the second stage (α_2) and false negative rate (FNR_2)

Fig. 2 Influence of mis-specified prior probability on FDR, expected power, and expected cost ratio, in which $n=1,000$, $FNR_2=0.05$, and $\mu_d=0.14$. **a**, **b** $\pi_1=0.01$ and $\pi_1=0.05$, respectively

lead to loss of power, because the expected power depends on only the optimal parameters. Although the expected cost ratio and the FDR are related to both π_1 and π_1^* , the expected cost largely depended on the optimal parameters with π_1^* , regardless of π_1 . The main difference between these two figures is the FDR. The FDR was also affected by the degree of misspecification. When π_1^* was not more than π_1 , the FDR always provided the conservative bias, i.e., control at a nominal level. Also, the FDR was inflated as π_1^* was overestimated and had a nearly linear relationship with π_1^* .

Discussion

We proposed optimal sample sizes and criteria to select markers for each stage using the FDR in a two-stage design and then evaluated those performances. Differing from methods in which FWER has been controlled, the FDR approach has an advantage because it can reflect prior information. Therefore, the proposed method allowed us to optimize the designs by utilizing π_1 , which is the proportion of truly associated markers to all markers used in the study. This flexible strategy allows each study to be done effectively.

The cost was defined as proportionate to the number of genotypings, that is, sample size $\times$ the number of markers. Regardless of choice of the design parameters, which are μ_d and π_1 , at least a 40% reduction in cost was demonstrated. These results show that costs, in both time and personnel associated with genotyping, can be reduced. In this study we assumed the total sample size was fixed. Future improvements in genotyping technology will reduce the cost of genotyping, and the substantial study costs will then depend on only the sample size. To reduce sample size, some investigators (Sobell et al. 1993; König et al. 2001, 2003) have proposed designs that are similar to group sequential methods with application to clinical trials. Although these procedures are considered to be complicated for practical applications, it would be a challenge to improve efficiency by considering the required sample size.

We examined the problem of multiple testing in association studies. Previous proposed two-stage designs, for example, Satagopan and Elston (2003), have proposed controlling the overall α error of the two-stage procedure. Also, Wen et al. (2006) showed the overall α error and FDR when certain fixed design parameters were used, of which the only second-stage significance level was corrected by Bonferroni's method. Recently, controlling the FDR has been considered

to be more relevant for eliminating false positives, especially in the exploratory studies. A lot of measures, estimation methods, and properties have been examined (Benjamini and Hochberg 2000; Efron et al. 2001; Storey and Tibshirani 2001, Storey and Tibshirani 2003; Efron and Tibshirani 2002; Genovese and Wasserman 2002; Storey 2002, 2003; Tsai et al. 2003; Black 2004; Fernando et al. 2004; Pounds and Cheng 2004; Storey et al. 2004). In addition to the strategies to control either FWER or FDR, another type of two-stage strategy using both FWER and FDR criteria has been proposed. Rosenberg et al. (2005) assumes the association studies with several SNPs in a number of candidate genes, and the association between each gene and disease is summarized by a single gene-wise P value that controls the FWER in each gene in the first stage. In the second stage, these gene-wise P values are adjusted to control the FDR to detect particular genes associated with the disease. Accordingly, the application range of the FDR would be considered to become broader.

To control the FDR in our proposed designs, however, π_1 must be decided before the study is performed. Because π_1 affects the performance of the design, it is important to examine its impact on efficiency. The use of an underestimated π_1^* , while leading to loss of the expected power, ensured the FDR was controlled at a nominal level. On the other hand, an overestimated π_1^* led to an inflated FDR. However, the expected power increased with π_1^* , and the FDR was kept at less than double the nominal level if $\pi_1^* \leq 2\pi_1$. This tendency would be acceptable for exploratory association studies using a number of markers, because the purpose of such studies is generally to select as many true positives as possible, accepting a few false positives. It will also be useful to examine the degree of the loss of power or the inflation of FDR or cost for some range of prior probabilities (e.g., Fig. 2). Particularly, when π_1 was expected to be very low, the FDR increased sharply with π_1^* . From this point, we considered that the proposed procedure is more appropriate for π_1 at least ≥ 0.01 . Here, π_1 was determined subjectively by investigators; other proposed methods to estimate π_1 have used available data, and Hsuesh et al. (2003) compared these latter methods. The optimal parameters and performances of the designs will be affected by the method used to estimate π_1 . Therefore, the optimization to include the stage to estimate π_1 might be an issue in the future.

The FNR, which is a measure of the truly associated markers left behind, should also be considered. While the FNR was always kept low because of low prior probability, in this study, it is possible to determine the

error measure using the FNR together with the FDR and to optimize the design parameters, as both are controlled.

Finally, we assumed that the markers are not in LD with each other. This is considered reasonable for exploratory studies using spaced markers, in which it would be appropriate to use the FDR. With several recent advances, however, whole-genome association studies using a fixed array of high density SNPs are now feasible. Because there would be LD between these SNPs, the proposed procedure could not guarantee that the FDR would be controlled; in this case, making an informed SNP selection could be important for efficiency. Some work has been done on tag SNP selection, which could also improve efficiency by utilizing the information of LD (Zhang et al. 2002, 2004; Stram et al. 2003; Stram 2004; Thomas et al. 2004). Incorporating these approaches into the optimization of designs is one of the challenges of the future.

Acknowledgments We thank Dr. Yutaka Matsuyama, Dr. Takuhiro Yamaguchi, and Ms. Yoichi Ito for their helpful discussions.

References

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300
- Benjamini Y, Hochberg Y (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat* 25:60–83
- Benjamini Y, Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. *Genetics* 171:783–790
- Black MA (2004) A note on the adaptive control of false discovery rates. *J R Stat Soc B* 66:297–304
- Böddiker IR, Ziegler A (2001) Sequential designs for genetic epidemiological linkage or association studies. A review of the literature. *Biom J* 43:501–525
- Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* 33:228–237
- Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature* 429:446–452
- Colhoun HM, McKeigue PM, Smith GD (2003) Problems of reporting genetic associations with complex outcomes. *Lancet* 361:865–872
- Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 23:70–86
- Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96:1151–1160
- Fernando RL, Nettleton D, Southey BR, Dekkers JCM, Rothschild MF, Soller M (2004) Controlling the proportion of false positives in multiple dependent tests. *Genetics* 166:611–619
- Genovese C, Wasserman L (2002) Operating characteristics and extensions of the false discovery rate procedure. *J R Stat Soc B* 64:499–517
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6:95–108
- Hsuesh HM, Chen JJ, Kodell RL (2003) Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J Biopharm Stat* 13:675–689
- König IR, Schäfer H, Müller H-H, Ziegler A (2001) Optimized group sequential study designs for tests of genetic linkage and association in complex diseases. *Am J Hum Genet* 69:590–600
- König IR, Schäfer H, Ziegler A, Müller H-H (2003) Reducing sample sizes in genome scans: group sequential study designs with futility stops. *Genet Epidemiol* 25:339–349
- van den Oord EJ, Sullivan PF (2003) False discoveries and models for gene discovery. *Trends Genet* 19:537–542
- Pounds S, Cheng C (2004) Improving false discovery rate estimation. *Bioinformatics* 20:1737–1745
- Risch N, Merikangas K (1996) The future of genetics studies of complex human disease. *Science* 273:1516–1517
- Rosenberg PS, Che A, Chen BE (2005) Multiple hypothesis testing strategies for genetic case-control association studies. *Stat Med* 25:3134–3149 [Epub ahead of print]
- Sabatti C, Service S, Freimer N (2003) False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 164:829–833
- Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 25:149–156
- Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. *Biometrics* 58:163–170
- Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60:589–597
- Sobell JL, Heston LL, Sommer SS (1993) Novel association approach for determining the genetic predisposition to schizophrenia: case-control resource and testing of a candidate gene. *Am J Med Genet* 48:28–35
- Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc B* 64:479–498
- Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat* 31:2013–35
- Storey JD, Tibshirani R (2001) Estimating false discovery rate under dependence, with applications to DNA microarrays. Technical report 2001–28, Department of Statistics, Stanford University, Stanford, CA, USA
- Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci U S A* 100:9440–9445
- Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc B* 66:187–205
- Stram DO (2004) Tag SNP selection for association studies. *Genet Epidemiol* 27:365–74
- Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered* 55:27–36
- Thomas D, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. *Genet Epidemiol* 27:401–414

- Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet* 77:337–345
- Tsai CA, Hsueh HM, Chen JJ (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* 59:1071–1081
- Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 96:434–442
- Wen SH, Tzeng JY, Kao JT, Hsiao CK (2006) A two-stage design for multiple testing in large-scale association studies. *J Hum Genet* 51:523–532
- Zehetmayer S, Bauer P, Posch M (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* 21:3771–3777
- Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci U S A* 99:7335–7339
- Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res* 14:908–916