J Hum Genet (2007) 52:650–658
DOI 10.1007/s10038-007-0159-9

ORIGINAL ARTICLE

A grid-search algorithm for optimal allocation of sample size in two-stage association studies

S. H. Wen · C. K. Hsiao

Received: 29 December 2006 / Accepted: 10 May 2007 / Published online: 30 June 2007

Abstract Multiple testing occurs commonly in genome-wide association studies with dense SNPs map. With numerous SNPs, not only the genotyping cost and time increase dramatically, many family wise error rate (FWER) controlling methods may fail for being too conservative and of less power when detecting SNPs associated with disease is of interest. Recently, several powerful two-stage strategies for multiple testing have received great attention. In this paper, we propose a grid-search algorithm for an optimal design of sample size allocation for these two-stage procedures. Two types of constraints are considered, one is the fixed overall cost and the other is the limited sample size. With the proposed optimal allocation of sample size, bearable false-positive results and larger power can be achieved to meet the limitations. The simulations indicate, as a general rule, allocating at least 80% of the total cost in stage one provides maximum power, as opposed to other methods. If per-genotyping cost in stage two differs from that in stage one, downward proportion of the total cost in earlier stage maintains good power. For limited total sample size, evaluating all the markers on 55% of the subjects in the first stage provides the maximum power while the cost reduction is approximately 43%.

Keywords Association studies · Optimal design · Grid-search algorithm · Cost-efficiency · Truepositive rate

S. H. Wen (✉)
Department of Public Health, College of Medicine,
Tzu-Chi University, Hualien 97004, Taiwan
e-mail: shwen@mail.tcu.edu.tw

C. K. Hsiao
Department of Public Health and Institute of Epidemiology,
College of Public Health, National Taiwan University,
Taipei 100, Taiwan

Introduction

With the recent advances in high-throughput genotyping technology, many genome-wide association studies are conducted to unravel the relation between disease and genes. However, the advancement in biotechnology often confronts with a statistical issue when dealing with large-scale data (Hirschhorn and Daly 2005; Thomas et al. 2005). It encounters the multiple testing dilemmas that most traditional statistical tests fail in reducing both chances of making true-positives and false-positives (Botstein and Risch 2003; Cardon 2001; Long and Langley 1999; Risch and Merikangas 1996). In addition, as the number of markers escalates, the amount of genotyping cost increases dramatically (Thomas 2006). For the first difficulty, it is common to adopt Bonferroni correction to solve the multiplicity effect when analyzing large-scale association studies. For example, Klein et al. (2005) analyzed the relationship between age-related macular degeneration and numerous single nucleotide polymorphisms (SNPs). They used Bonferroni correction to adjust the significance level as the ratio of original nominal level to the total number of SNPs. Although it controls the family wise error rate (FWER), probability of claiming more than one false alarm, the downward level results in loss of power in detecting the relevant SNPs. Furthermore, such a single-stage strategy is not cost-efficient under limited resources, especially when testing a large number of markers. Hence, a procedure that saves cost and maintains a satisfactory power simultaneously is in urgent need. The great majority of unassociated markers can be eliminated via multi-stage procedures, in particular the two-stage methods have been proposed to optimize the power and conserve the cost of such studies (Hirschhorn and Daly 2005; Skol et al. 2006).

From the design viewpoint, there are two types of two-stage procedures. One uses independent subjects at different stages (Miller et al. 2001; Saito and Kamatani 2002), where a large significance level is adopted to select promising markers first, and then a stringent level is applied at the next stage to control the FWER. Ohashi and Clark (2005) took cost-efficiency further into account and conducted a stage-wise approach under limited total cost. Instead of FWER, other studies control the false discovery rate (FDR), false-positive proportion of significant markers (Benjamin and Hochberg 1995), that attains larger power to detect associated SNPs. van den Oord et al.(2003) suggested to use independent samples at different stages and to choose a suitable threshold for controlling FDR at an arbitrary bound. However, one limitation of the approach is that the whole information contained in the data is not fully utilized. For instance, the data of subjects recruited in the first stage are usually discarded, thus it does not satisfy the purpose of preserving the cost as much as possible. Second, when the primary concern is to reduce the number of false-positive markers, usually less attention is placed on the proportion of true-positive markers, which seems conflict with the scientific interest of identifying the markers with association. Third, the FWER-controlling method becomes stringent when the total number of markers is huge.

The second type of two-stage procedures combines all available data, including those of previously selected promising markers. One advantage is the complete utilization of all information. Another is putting more emphasis on the power to detect associated markers than focusing simply on false-positive rate (Wen et al. 2006). Kuchiba et al. (2006) controlled the FDR with optimal sample size and reduced cost. They also emphasized the influence of the true proportion of associated markers on the performance of two-stage designs. Satagopan et al. (2002, 2004) proposed to employ a fraction of resources (either cost or individuals) at an earlier stage, and to use all available individuals in the final stage. Zehetmayer et al. (2005) advocated two-stage designs with controlled FDR and split sample size into two stages for gene-expression studies. Wang et al. (2006) considered various configurations of per-genotype cost ratio and significance levels in both stages to achieve the desired power with minimum cost. Wen et al. (2006) recommended excluding mostly irrelevant markers while adopting a large significance level in the first stage, and controlling the overall false-positives with a downward significance level in the second stage. Different from Satagopan et al. (2002, 2004), they can choose the promising markers at a pre-specified significance level and control the false-positive rate (FPR) adequately. However, the optimal allocation of subjects remains an open issue. In practice, the costs or subjects are limited and it affects the recruitment in both stages with respect to error rates and power. In this paper, we propose optimal designs in this two-stage setting to distribute subjects and select associated markers under two different situations, where one is fixed total genotyping cost (FTGC) and the other is fixed sample sizes (FSS). In the following sections, we introduce the rationale and implementation of the optimal design for both FTGC and FSS. Simulation studies are conducted to evaluate the performances of the proposed approach based on limited cost and sample size. The comparison with other existing alternatives is also discussed.

Methods

In this section, we first brief the notation and then explain the derivation of optimal allocation of sample size under limited cost or total number of subjects. To detect the association between markers and disease phenotype, we consider SNPs as testing markers for illustration. Let δ denote the difference in the mean allele frequency between cases and controls, let N_1 be allele data for each group in the first stage, M the total number of markers in linkage equilibrium, and w the proportion of truly unassociated SNPs. In the earlier stage, if the individual P-value for a marker is less than the uncorrected level α_1 (=0.05), the marker is considered promising and will be verified further with additional N_2 allele data in the second stage. Here we assume a balance population-based case control design, and the total number of subjects in stage one and two are N_1 and $(N_1 + N_2)$, respectively. Suppose a total of R promising markers are considered in the second stage, and a stringent significance level, α_2=0.05/R, is adopted in this stage to reduce the overall inflated type I error due to large α_1. We considered two indices, TPR (true-positive rate) and FPR, to evaluate the performance of the two-stage procedure. According to Wen et al. (2006), both overall FPR and TPR are functions of sample sizes (N_1, N_2), significance levels (α_1, α_2), number of total markers M, and the irrelevant proportion w. In addition, the disease model parameters such as the allele frequency and effect size of tests also affect the FPR and TPR. Therefore, an optimal design must take these into account.

In the following, we introduce a grid-search algorithm for optimal allocation of (N_1, N_2) with a desired power and constrained resources. First, under FTGC, the total genotyping cost is given by $T = MN_1 + RN_2$, where R can be replaced with $E(R)$=$Mw\alpha_1$+$M(1–w)(1–\beta_1)$, and $(1–\beta_1)$ represents the power in the first stage. For simplicity, let N_2=kN_1, we maximize TPR with respect to k and N_1 under the constraint that $T = MN_1 + \mathrm{E}(R)(kN_1)$. Since N_1 is related to other factors, e.g. E(R) and significance level in stage 2, in the overall power, it is more flexible to keep the optimization algorithm in a low-dimension setting than in high-dimension of optimizing all factors simultaneously. Besides, it is not feasible to compute the analytical form, and hence we suggest a grid search for a wide range of (N_1, N_2=kN_1) and the one with maximum TPR would be the optimal allocation of sample size. The second limitation concerns the fixed total sample size N (= N_1 + N_2) used in the genetic study. For simplicity, let N_1 = π N, here π is the proportion of N_1 in N and ranges from 0 to 1. Given N, M, and w, both the TPR and required costs are proportional to π. Over a plausible range of π, one can conduct a grid search to find the optimal π that attains the maximum TPR, as well as a substantial cost reduction to strike a balance between power and cost. We use a program written in S-plus 7.0 to perform the searches (The program is available upon request and more details of optimization are given in Appendix.).

Results

Our purpose is to compare the TPR of the proposed method with that of other single-stage design where all markers on all samples are genotyped, and other alternative two-stage designs. All strategies were tailored to use pre-specified cost or sample size, and we compared the false-positive results (i.e. FPR or FDR) and power for a broad range of sample sizes. We also investigated the influence of different allele frequencies, effect sizes, and the allelic odds ratios (*OR*). Under limited cost *T*, the sample size of a single-stage design would be *T*/*M*. Bonferroni method for this design is denoted as B(*T*/*M*). We denoted M(S) for a single-stage method with the same significance level α_2 for the proposed two-stage method with fixed FPR. Table 1 lists the simulation results of TPR and FPR for the two-stage method, B(*T*/*M*), and M(S) under FTGC for several values of (w, $\bar{p}$,δ, *OR*) under a fixed array of 5,000 SNPs and equal per-genotyping cost. The fifth column also shows the proportion of cost in stage one, i.e. $c_1 = \frac{MN_1}{T}$. Numbers were close to the analytical results (data not shown). In these examples, the false-positives such as FPR and FDR (in Fig. 1) were bearably small under various combinations of N_1 and N_2. However, the TPR of the proposed method varied greatly with respect to $(N_1, N_2, \bar{p}, \delta, OR)$, and was often larger than that of single-stage methods, irrespective

Table 1 Simulation results for the proportion of total cost in stage one, TPR and FPR of two-stage method and two single-stage methods under FTGC

$(\bar{p}, \delta, OR)$	N_1	k	N	c_1(%)	TPR			FPR*10^3		
					B(*T*/*M*)	M(S)	Two-stage	B(T/*M*)	M(S)	Two-stage
(0.5, 0.1, 1.49)[a]	414	8.85	4,078	69.95	0.172	0.003	0.817	0.011	<0.001	<0.001
	488	4.52	2,692	81.12	0.168	0.127	0.887	0.010	0.004	0.002
	531	2.56	1,888	87.89	0.170	0.403	0.911	0.011	0.190	0.107
	563	1.29	1,291	93.56	0.169	0.387	0.862	0.014	0.196	0.155
	594	0.20	712	98.96	0.180	0.401	0.518	0.078	0.190	0.190
(0.5, 0.1, 1.49)[b]	414	8.34	3,868	69.95	0.172	0.006	0.818	0.010	<0.001	<0.001
	488	4.24	2,557	81.12	0.169	0.154	0.881	0.011	0.008	0.004
	531	2.39	1,803	87.89	0.169	0.390	0.901	0.006	0.173	0.108
	563	1.21	1,244	93.56	0.167	0.390	0.859	0.010	0.163	0.151
	594	0.19	705	98.96	0.171	0.389	0.506	0.009	0.171	0.166
(0.3, 0.08, 1.47)[a]	514	3.30	2,209	85.62	0.077	0.144	0.781	0.078	0.046	0.026
	536	2.35	1,797	89.34	0.079	0.240	0.797	0.012	0.198	0.110
	557	1.52	1,404	92.82	0.086	0.244	0.736	0.011	0.191	0.150
	578	0.75	1,012	96.32	0.072	0.244	0.571	0.010	0.207	0.181
	594	0.20	713	98.99	0.079	0.233	0.329	0.076	0.190	0.193
(0.3, 0.08, 1.47)[b]	514	3.11	2,114	85.62	0.082	0.170	0.781	0.010	0.067	0.039
	536	2.22	1,725	89.34	0.080	0.237	0.779	0.009	0.189	0.111
	557	1.43	1,355	92.82	0.078	0.232	0.722	0.010	0.180	0.155
	578	0.71	986	96.32	0.079	0.239	0.541	0.009	0.180	0.166
	594	0.19	706	98.99	0.078	0.232	0.320	0.011	0.186	0.171

The number of replication is 1,000 in simulation (*T*/*M* = 600, *M* = 5,000, w = 0.999, 0.995, and α_1= 0.05)

c_1: the proportion of total cost in stage one

[a] w = 0.999, [b] w = 0.995

B(*T*/*M*) Bonferroni method with *T*/*M* subjects. M(S): single-stage method with the same FPR with two-stage method

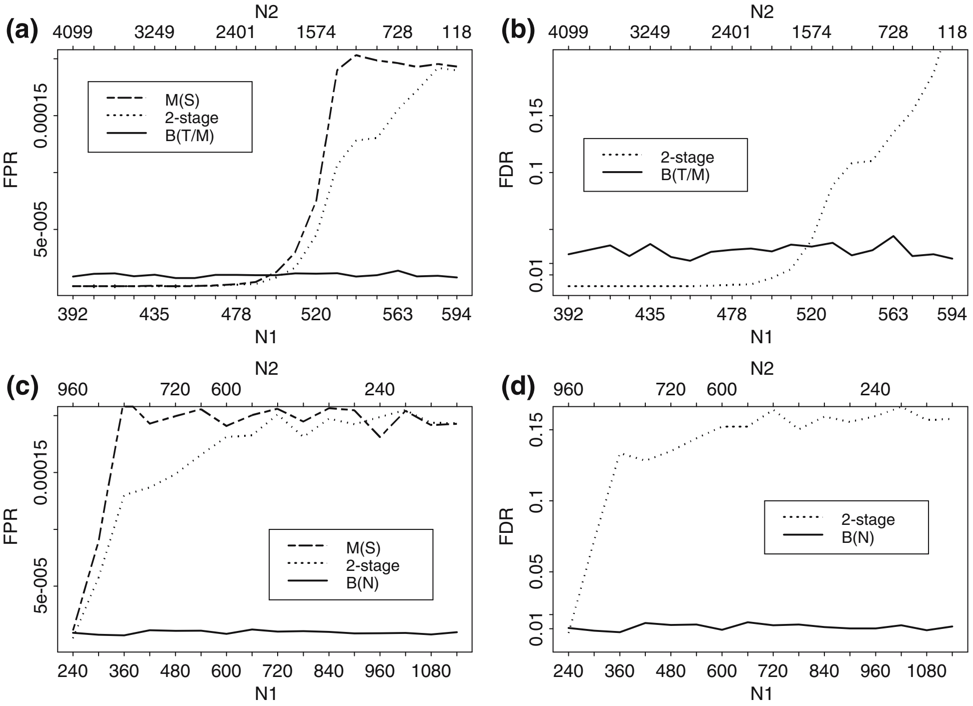

Fig. 1 The curves based on simulations in **a** and **c** are false-positive rate (*FPR*), in **b** and **d** are false discovery rate (*FDR*) with respect to various N_1. In all figures, $M = 5000$, $w = 0.999$, $(\bar{p}, \delta, OR) = (0.5, 0.1, 1.49)$, and $\alpha_1 = 0.05$. In **a** and **b** FTGC with $T = 600$ *M*, and in **c** and **d** FSS with $N = 1{,}200$. Three lines in **a** and **d** denotes M(S) (*dashed line*), Bonferroni method (*solid line*) and two-stage method (*dotted line*)

of allelic odds ratio and the number of markers associated with disease. Moreover, under the two-stage setting, large total sample size did not yield larger power to detect the markers associated with the disease. Hence, we recommend determine the optimal k with easy-to-recruit sample size on the condition that the TPR is manageable or desired for FTGC. For example, the largest TPR(=0.911) occurred at (N_1, k)=(531,2.56) with total sample size $N = 1{,}888$ for $w = 0.999$ and $(\bar{p}, \delta, OR) = (0.5, 0.1, 1.49)$. This also indicated that allocating 87.89% of the total cost in earlier stage would maintain optimal power. The relationship between k and c_1 could be derived as $k = \frac{M}{\mathrm{E}(R)} \times \frac{1-c_1}{c_1}$. Clearly, if one predetermines different allocations of cost ratio $(1–c_1)/c_1$ or markers ratio, $M/\mathrm{E}(R)$, the settings would affect allocations of (N_1, N_2).

Table 2 gives the simulation results of TPR, and FPR of two-stage method, Bonferroni method (denoted as B(N)) and M(S) for FSS ($N = 1{,}200$). Column 4 shows the percentage of cost saving, namely the reduction in cost of a two-stage method relative to a single-stage design, $(1-\pi)\left(1-\frac{\mathrm{E}(R)}{M}\right)$. By comparing columns 5–7, the TPR of B(N) was the worst among the three methods. Over a plausible range of $\pi \in(0.55,0.95)$, the proposed method yielded a power comparable to that of the more genotyping-effort single-stage design with similar FPR. Hence, we suggest selecting the optimal range of π(=0.55) while the TPR is satisfied and the reduction of cost is significant for FSS. It is worth noting that at small values of $\pi(\leq 0.25)$, the markers associated with disease in earlier stage were less likely to be chosen for further testing, and the overall power was lower than single-stage method. Alternatively, we presented the simulation results in Figs. 1 and 2. The false-positive results (FPR or FDR) of two-stage method were stable small (in Fig. 1a–d). For FTGC, the TPR of the two-stage setting varied dramatically with (N_1, N_2) and the optimal k appeared to be the maximum TPR, as well as the corresponding sample size was easy-to-recruit (in Fig. 2a–b). Also, it corresponded to deposit at least 80% of the total cost in stage one given the same unit typing cost in both stages. For FSS, the optimal π was around 0.55 while the corresponding design was at minimum cost on condition that the overall power is near-optimal (in Fig. 2d). These results for FSS are consistent with that in Wang et al. (2006). However, under FTGC, such as given total or minimum cost, there are still many options for (N_1, N_2) with desired power, say 80%. Among those choices, the one with easy-to-recruit total sample size conditional on maximum TPR would be the optimal.

We further evaluated the performance of the optimal two-stage method with some existing alternatives. Under

Table 2 Simulation results for cost saving, TPR and FPR of two-stage method and two single-stage methods under FSS. The number of replication is 1,000 in simulation (N=1,200, M=5,000, w=0.999, 0.995, and α_1=0.05)

$(\bar{p}, \delta, OR)$	N_1	π	cs (%)	TPR			FPR*10^3		
				B(N)	M(S)	Two-stage	B(N)	M(S)	Two-stage
(0.5, 0.1, 1.49)[a]	300	0.25	71.15	0.686	0.848	0.632	0.007	0.090	0.057
	480	0.40	57.01	0.685	0.887	0.809	0.011	0.200	0.149
	660	0.55	42.68	0.698	0.893	0.872	0.012	0.200	0.183
	900	0.75	23.72	0.683	0.881	0.881	0.008	0.200	0.192
	1140	0.95	4.74	0.677	0.878	0.881	0.010	0.193	0.193
(0.5, 0.1, 1.49)[b]	300	0.25	70.95	0.683	0.835	0.626	0.012	0.091	0.050
	480	0.40	56.81	0.681	0.881	0.809	0.007	0.200	0.147
	660	0.55	42.53	0.683	0.879	0.864	0.011	0.176	0.161
	900	0.75	23.62	0.685	0.877	0.876	0.008	0.184	0.190
	1140	0.95	4.72	0.682	0.877	0.877	0.009	0.169	0.175
(0.3, 0.1, 1.62)[a]	300	0.25	71.23	0.826	0.927	0.742	0.012	0.091	0.050
	480	0.40	56.94	0.826	0.949	0.897	0.008	0.185	0.159
	660	0.55	42.69	0.837	0.954	0.948	0.009	0.193	0.181
	900	0.75	23.73	0.824	0.949	0.949	0.010	0.202	0.195
	1140	0.95	4.74	0.832	0.949	0.952	0.007	0.185	0.192
(0.3, 0.1, 1.62)[b]	300	0.25	70.99	0.826	0.926	0.737	0.011	0.088	0.044
	480	0.40	56.72	0.824	0.946	0.892	0.011	0.181	0.147
	660	0.55	42.54	0.827	0.947	0.938	0.007	0.176	0.179
	900	0.75	23.63	0.823	0.946	0.947	0.008	0.167	0.180
	1140	0.95	4.73	0.826	0.949	0.948	0.011	0.170	0.171

cs Cost saving as compared with cost under Bonferroni method with N subjects. B(N): Bonferroni method with N subjects. M(S): single-stage method with the same FPR with two-stage method

[a] w = 0.999, [b] w = 0.995

FTGC, an alternative approach using 75% of the cost in stage one to screen all markers and evaluate promising 10% of the markers with the remaining cost in stage two was proposed by Satagopan et al. (2002) (denoted as M(1)). While the sample size was the primary constraint, Satagopan et al. (2004) (denoted as M(2)) advocated that evaluating all the markers on 50% of the subjects in stage one and selecting the most promising 10% of the markers on the remaining individuals in the second stage yielded near-optimal power. We let the total number of markers to be selected at the end of the study is five for M(1) and M(2).

Table 3 lists the TPR, FPR, FDR, total sample size or cost saving of optimal two-stage method, M(1), and M(2) under several parameter configurations for M=100, and 5,000. By comparing the FPR and FDR, the proposed optimal design produced less false-positives than that of M(1) and M(2) regardless of allelic odds ratio and the total number of markers. Actually, the expected number of false-positive results for proposed method was less than one at various M. But for M(1) or M(2), it was larger than one false alarm. Besides, looking at the TPR, the power of the optimal design was consistently larger than that of M(1) and M(2). In practice, the optimal two-stage design was also superior in terms of total sample size or cost-efficiency. For example, the optimal design recruited fewer individuals than M(1) for FTGC under different allelic odds ratio. For FSS, the optimal design produced similar cost reduction, but the power was obviously larger, as well as less false-positive results.

Discussion

We propose an optimal two-stage design in genetic association studies under the constrained FTGC or FSS. Different from Wen et al. (2006), the optimization is related to limited resources and focuses on efficient allocation of subjects. To accomplish the purpose of maintaining good power when detecting truly relevant markers, we suggest a grid-search algorithm for optimal cost-efficient strategies. Briefly, the concept can be applied to other two-stage settings where the factors of the overall power interact differently. Our proposal has several advantages. First, the (N_1, k) or (N_1, π) can be determined analytically with optimal TPR, bearable FPR and satisfied cost. When the total resources are limited, there are many possible allocations of N_1 and N_2. The impact of allocations on TPR is

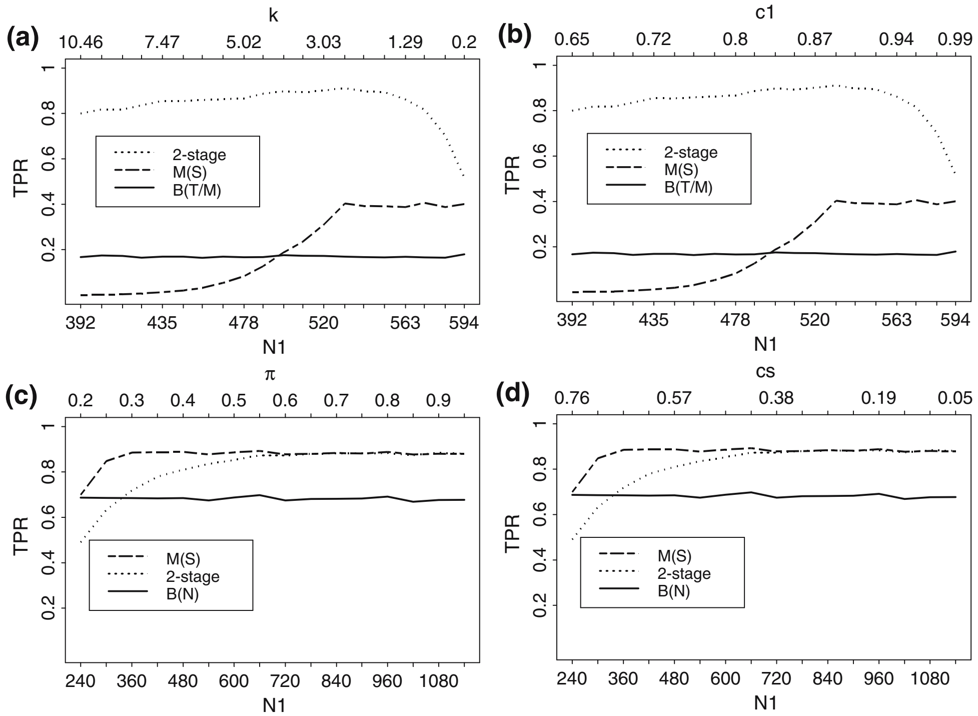

Fig. 2 The curves based on simulations in **a**–**d** are true-positive rate (*TPR*) with respect to various N_1. In all figures, $M = 5000$, $w = 0.999, (\bar{p}, \delta, OR), (\bar{p}, \delta, OR) = (0.5, 0.1, 1.49)$ and $\alpha_1 = 0.05$. In **a** and **b** $T = 600\ M$, and in **c** and **d** $N = 1{,}200$. Distinct lines correspond to different methods. The *solid line* is for Bonferroni method with the same resources, the *dashed line* is for M(S) and the *dotted line* is for two-stage method

Table 3 Simulation results for TPR, FPR and FDR of optimal two-stage method, M(1) and M(2) under FTGC and FSS. The number of replication is 1,000 in simulation

$FTGC(T/M = 600)(\bar{p}, \delta, OR)$		TPR		FPR × 10^2		FDR		Total sample size		Optimal two-stage design	
		2-stage	M(1)	2-stage	M(1)	2-stage	M(1)	2-stage	M(1)	N_1	k
(0.5, 0.1, 1.49)	(a)	0.911	0.862	0.011	0.014	0.089	0.138	1888	1950	531	2.555
	(b)	0.895	0.862	0.155	0.720	0.027	0.138	1276	1950	531	1.402
(0.3, 0.08, 1.47)	(a)	0.797	0.734	0.011	0.027	0.104	0.266	1797	1950	536	2.352
	(b)	0.792	0.775	0.346	1.183	0.068	0.225	999	1950	561	0.780

$FSS(N = 1200)(\bar{p}, \delta, OR)$		TPR		FPR × 10^2		FDR		Cost saving		Optimal two-stage design	
		2-stage	M(2)	2-stage	M(2)	2-stage	M(2)	2-stage (%)	M(2) (%)	N_1	π
(0.3, 0.1, 1.62)	(a)	0.948	0.695	0.018	0.031	0.138	0.305	42.69	45	660	0.55
	(b)	0.974	0.866	0.371	0.705	0.058	0.134	40.63	45	660	0.55
(0.5, 0.1, 1.49)	(a)	0.872	0.603	0.018	0.040	0.152	0.397	42.68	45	660	0.55
	(b)	0.928	0.824	0.319	0.925	0.053	0.176	45.19	45	600	0.50

(a) $M = 5000$, $w = 0.999$; (b) $M = 100$, $w = 0.95$. M(1) rule-of-thumb two-stage design in Satagopan et al. (2002) under FTGC, the corresponding k is 10/3. M(2): rule-of-thumb two-stage design in Satagopan et al. (2004) under FSS, the corresponding π is 50%

more obvious than that on false-positive results. One would also use the algorithm to examine adequate total cost or sample size before studies. The rule of thumb is to identify the mode of TPR curve against k (or π such as the case of cost savings for FSS) to find the optimal condition.

Second, under FTGC, we show that the optimal design k is about 2.5 with moderate total sample sizes and this translates to a design where M(=5,000) markers are screened with approximately 88% of total cost in earlier stage, and then R selected markers are tested with the

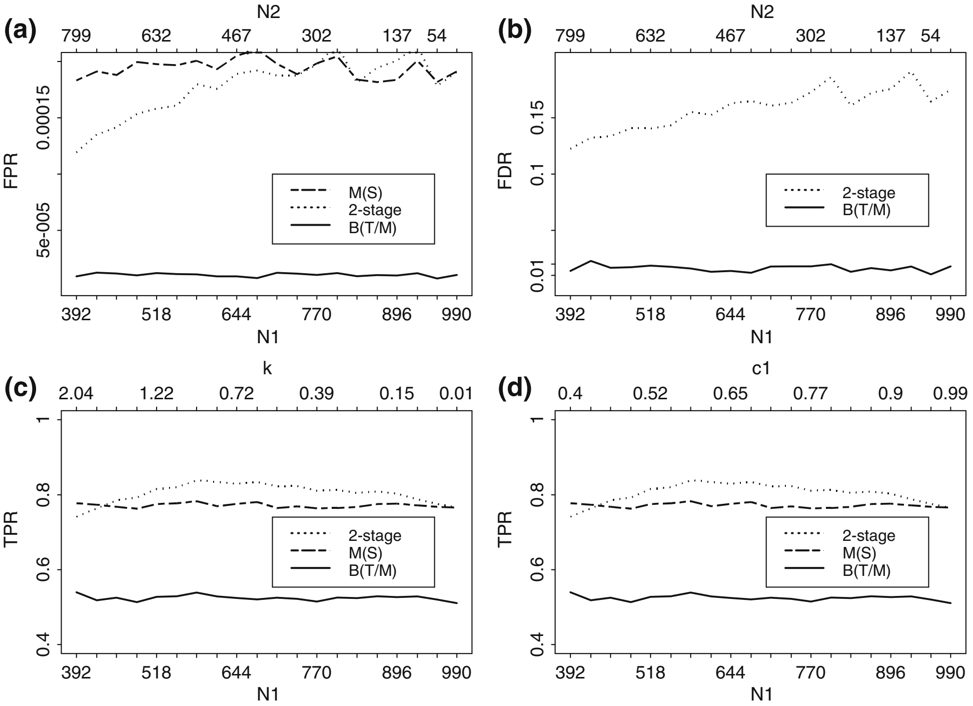

Fig. 3 The curves based on simulations in **a** and **b** are false-positive rate (*FPR*) and false discovery rate (*FDR*), and **c, d** are true-positive rate (*TPR*) with respect to various N_1 for c_g = 15 under FTGC. In all figures, $(\bar{p}, \delta, OR) = (0.5, 0.1, 1.49), T = 1000M, M = 5000$, $w = 0.999$ and α_1=0.05. *Distinct lines* correspond to different methods. The *solid line* is for Bonferroni method, the *dashed line* is for M(S) and the *dotted line* is for two-stage method

remaining cost. Furthermore, the grid-search algorithm can be extended to different per-genotype cost ratio at each stage for FTGC. For instance, using factor c_g for the ratio of per-genotype cost in stage 2 versus that in stage 1, we present in Fig. 3 the relationships between (FPR, FDR, TPR) and (N_1, N_2, k and c_1), based on simulations results under c_g=15 (as suggested in Wang et al. 2006). It is obvious that the optimal k is less than 1 and the corresponding proportion of total cost in stage one is nearly 60–65%. Alternatively, if the sample size is restricted, we recommend π between 0.5 and 0.6 to get a higher overall power and substantial cost reduction. That is, to screen M(=5,000) markers with nearly 55% of total sample size in earlier stage, and then test all individuals with the selected significant R markers. Finally, we investigate the power and false-positive results of alternative two-stage methods. The optimal two-stage method is superior to existing alternatives. The superiority remains when compared in terms of cost-efficiency. The proposed approach provides specific criteria in formal testing with pre-specified significance level for each stage. Satagopan et al. (2002) suggested to determine the number of selected markers prior to a two-stage proposal. This approach is not straightforward since the number of markers associated with the disease is usually unknown.

Our method will provide useful guidelines when planning large-scale association studies. Besides, the scheme of the method does not change with the test statistic used. The same argument applies to the case when more than one locus is considered, though the test may become more complex. Other applications include the association test for tag SNPs or haplotypes. Another issue is if the proportion of cost at the earlier stage c_1 is chosen in advance, the TPR and FPR can be estimated corresponding to N_1=c_1T/M, and $k = \frac{M}{\mathrm{E}(R)} \times \frac{1-c_1}{c_1}$. Moreover, if one sets up a certain proportion of ‘promising’ markers, say 0.1 (i.e. E(R)/M=0.1), we could also perform a grid search over a plausible range of k and find the optimal allocation of (N_1, N_2). Kuchiba et al. (2006) recommended the use of their proposal when the proportion of true associated markers (they called it π_1) is greater than or equal to 0.01. In that case, this grid search algorithm will provide optimal choice of N_1 and N_2 as well.

Appendix

Following the work of Wen et al. (2006), the overall FPR, probability of claiming unassociated SNPs significant in both stages, is approximated by $\hat{\alpha}_2(N_1)$ below,

$$\hat{\alpha}_2(N_1) = \begin{cases} 0.05/E(R)\,, & \text{if } N_1/(N_1+N_2) \geq (z_{\alpha_1}/z_{\alpha_2})^2 \\ (2\times[1-\Phi\left(\sqrt{\frac{N_1+N_2}{N_1}}\times z_{\alpha_1/2}\right), & \text{if } N_1/(N_1+N_2) < (z_{\alpha_1}/z_{\alpha_2})^2 \end{cases}$$

where the expected value of R, $E(R)$, can be estimated by $Mw\alpha_1 + M(1\text{–}w)(1\text{–}\beta_1(N_1))$, with $1-\beta_1(N_1) = \Phi\left(\frac{\sqrt{N_1}\delta-\sigma_0 z_{\alpha_1/2}}{\sigma_1}\right)$ as the power in the first stage, Φ is the cumulative density function of standard normal distribution, and $z_{\alpha 1/2}$ represents the $100(1\text{–}\alpha_1\ /2)$-th quartile of the standard normal distribution. Similarly, the other index $\mathrm{TPR} \cong 1-\beta_2 = \Phi\left(\frac{\sqrt{N_1+N_2}\delta-\sigma_0 z_{\hat{\alpha}_2(N_1)/2}}{\sigma_1}\right)$, the probability of declaring truly associated SNPs as significant in both stages, is approximately $(1\text{–}\beta_2)$ with respect to $\hat{\alpha}_2(N_1)$ and N_1+N_2. Where σ_0 and σ_1 denote the standard deviation of difference in mean allele frequencies under the null and alternative hypothesis, respectively.

Considering fixed total genotyping cost T=MN_1+RN_2 with the same per-genotyping cost in both stages and fixed M and w, the optimal design is to allocate N_1 and N_2 efficiently to achieve maximum TPR. For simplicity, let $N_2 = kN_1$ and R can be replaced with $E(R)$. Therefore, T can be rewritten as T=MN_1+($Mw\alpha_1$+M(1–w)(1–β_1(N_1)))kN_1. In other words, k=(T–N_1M)/($N_1E(R)$) is a function of N_1. The goal is to find the best value of (N_1, k) such that the two-stage method has maximum power. The *TPR* is defined as

$$TPR = 1-\beta_2(N_1) = \Phi\left(\frac{\sqrt{(1+(T-N_1M)/N_1E(R))N_1}\delta-\sigma_0 z_{\hat{\alpha}_2(N_1)/2}}{\sigma_1}\right)$$

where $\hat{\alpha}_2(N_1) = \min(0.05/E(R),\ 2\times[1-\Phi(\sqrt{1+k}$ $\times z_{\alpha_1}\ /2)])$, and Φ, $z_{\alpha\_1/2}$, σ_0 and σ_1 are defined as previous.

It is not straightforward to derive the analytical form of k, but the maximum TPR can be searched through the range of N_1. The upper bound for N_1 is T/M, which implies all resources are allocated in the first stage. By setting the power in the first stage larger than 0.8, we can obtain a reasonable range for N_1 based on $N_{1(0)}$ and T/M, where $N_{1(0)}$ satisfies the equation $1-\beta_1(N_{1(0)}) = \Phi\left(\frac{\sqrt{N_{1(0)}}\delta-\sigma_0 z_{\alpha_1/2}}{\sigma_1}\right) = 0.8$. Given T, M, w, allele frequency $\bar{p}$ and the effect size δ, we perform a grid search of (N_1, k) for the optimal design.

When the total number of participants (=N) is limited and when N_1=π N, the TPR is defined as $\mathrm{TPR} = 1-\beta_2(N_1) = \Phi\left(\frac{\sqrt{N}\delta-\sigma_0 z_{\hat{\alpha}_2(N_1)/2}}{\sigma_1}\right) \leq 1-\beta_1(N_1)$, where $\hat{\alpha}_2(N_1) = \min(0.05/E(R),\ 2\times[1-\Phi(z_{\alpha_1/2}/\sqrt{\pi})])$. Given N, M, and w, the TPR is only affected by $\hat{\alpha}_2(N_1)$ and is smaller or equal to 1–β_1(N_1). If π is smaller, the TPR is bounded from above by 1–β_1(N_1), which is small as well. On the other hand, if π is large, the TPR is likely to be large, but the cost increases dramatically. Hence, one needs to strike a balance between genotyping cost and the TPR. Denoting the cost as

$$\mathrm{T}(\pi) = (M\pi + \mathrm{E}(R)(1-\pi))N = (\pi + (w\alpha_1 + (1-w)$$
$$(1-\beta_1(N_1)))(1-\pi))MN,$$

where T(π) is also a function of π given N, M, and w. Similarly, a grid search of π can be set up to find the maximum TPR and affordable cost analytically.

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33(suppl):228–237

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

Kuchiba A, Tanaka NY, Ohashi Y (2006) Optimum two-stage design in case-control association studies using false discovery rate. J Hum Genet 51:1046–1054

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731

Miller RA, Galecki A, Shmookler-Reis RJ (2001) Interpretation, design, and analysis of gene array expression experiments. J Gerontol A Biol Sci 56A(2):B52–B57

Ohashi J, Clark AG (2005) Application of the stepwise focusing method to optimize the cost-effectiveness of genome-wide association studies with limited research budgets for genotyping and phenotyping. Ann Hum Genet 69:323–328

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

Saito A, Kamatani N (2002) Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. J Hum Genet 47:360–365

Satagopan JM, Verbal DA, Venkatraman ES, 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

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213

Thomas DC (2006) Are we ready for the genome-wide association studies? Cancer Epidemiol Biomarkers Prev 15(4):595–598

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

van den Oord EJ, Sullivan PF (2003) A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. Hum Heredity 56:188–199

Wang H, Thomas DC, Pe'er I, Stram DO (2006) Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368

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