J Hum Genet (2002) 47:360–365

ORIGINAL ARTICLE

Akira Saito · Naoyuki Kamatani

Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method

Received: March 20, 2002 / Accepted: April 11, 2002

Abstract Recently, the use of genome-wide linkage disequilibrium (LD) analysis to localize traits has attracted much attention because of the introduction of high-throughput genotyping systems. However, a limitation of such studies is often the total cost of genotyping in addition to sample size. Therefore, it is important to estimate optimal conditions for such a study given the total cost of genotyping. In the present study, we have introduced the "stepwise focusing method," in which candidate markers are selected in a stepwise fashion. In the first focusing step, samples from both case and control groups are genotyped at a certain number of single-nucleotide polymorphisms (SNPs) (for example, 50000), and the markers that exhibit significant intergroup differences by a χ^2 test are selected. In the first step, the risk of type I error is set rather high (for example, 0.1), and, therefore, most of the selected markers are false positives. In the second step, the markers selected in the first step are tested by using samples obtained from a different set of case–control samples. We performed extensive simulation studies to estimate both the type I error and the power of the test by changing parameters such as genotype relative risk, disease allele frequency, and sample size. If the total number of genotypings was limited, the stepwise focusing method yielded optimal conditions and was more powerful than conventional methods.

Key words Association studies · Study design · Simulation · Statistical power · SNPs

A. Saito (✉) · N. Kamatani
Division of Statistical Genetics, Institute of Rheumatology, Tokyo Women's Medical University, 10-22 Kawada-cho, Shinjuku-ku, Tokyo 162-0054, Japan
Tel. +81-3-5269-1725; Fax +81-3-5269-1726
e-mail: asaito@ior.twmu.ac.jp

A. Saito
Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan

Introduction

The detection of susceptibility genes for complex traits has attracted much interest in human genetics. There is growing interest in the use of single-nucleotide polymorphisms (SNPs) for the analysis of complex human diseases (Collins et al. 1998; Kruglyak 1999). For complex diseases, genome-wide association studies using SNPs have been suggested as more appropriate for detection of susceptibility genes than genome-wide linkage analyses (Risch and Merikangas 1996; Lander 1996; Morton and Collins 1998; Risch and Teng 1998). The study of genome-wide associations in complex traits requires a large sample size because the expected effect of each gene is small and extremely low significance levels need to be adopted (Risch 2000; Ohashi and Tokunaga 2001). Actually, however, in addition to the large sample size required, the cost of genotyping is frequently one of the most important considerations. If the total budget for genotyping is limited, it is very important to choose the optimal study design for the budget available (Amos and Page 2001; Gu and Rao 2001). In the present study, we introduce a stepwise focusing method to reduce the total cost of genotyping. By this method, the candidate markers are selected in a stepwise fashion. We evaluated the validity of the stepwise focusing method under several conditions by using Monte Carlo simulations. We then searched for the optimal conditions for the stepwise focusing method when the total numbers of both typings and markers to be examined were fixed.

Methods

Models

We consider the case in which only one biallelic marker is associated with a disease (denoted a disease-associated locus). A disease-associated locus is assumed to consist of two alleles, A and a, the former being positively associated with the disease. Let k be the probability that an individual with

the genotype *aa* develops the disease, and *r* the genotype relative risk, so that the probability that an individual with the genotype *Aa* develops the disease is kr ($0 < k < 1$, $r \geq 1$). We assume that the probability that an individual with genotype *AA* exhibits the disease is kr^m ($m \geq 1$). Then, the prevalence of the disease in the population will be

$$K = p_A^2 kr^m + 2p_A(1 - p_A)kr + (1 - p_A)^2 k,$$

where p_A is the frequency of allele *A* in the population. We assume Hardy-Weinberg equilibrium (HWE) in the population.

By Bayes' theorem, the probabilities that affected individuals have the *AA*, *Aa*, or *aa* genotypes will be $\Pr(AA \mid \text{Affected}) = p_A^2 kr^m/K$, $\Pr(Aa \mid \text{Affected}) = 2p_A(1 - p_A)kr/K$, and $\Pr(aa \mid \text{Affected}) = (1 - p_A)^2 k/K$, respectively. Therefore, the expected frequencies of allele *A* and *a* among the affected individuals will be

$$\Pr(A \mid \text{Affected}) = \left(p_A^2 kr^m + p_A(1 - p_A)kr\right)\Big/K$$

and

$$\Pr(a \mid \text{Affected}) = \left(p_A(1 - p_A)kr + (1 - p_A)^2 k\right)\Big/K,$$

respectively.

Similarly, the probabilities that unaffected individuals have *AA*, *Aa*, or *aa* genotypes are $\Pr(AA \mid \text{Unaffected}) = p_A^2(1 - kr^m)/(1 - K)$, $\Pr(Aa \mid \text{Unaffected}) = 2p_A(1 - p_A)(1 - kr)/(1 - K)$, and $\Pr(aa \mid \text{Unaffected}) = (1 - p_A)^2(1 - k)/(1 - K)$, respectively. Then, the expected frequencies of allele *A* and *a* among the unaffected individuals will be

$$\Pr(A \mid \text{Unaffected}) = \frac{1}{1 - K}\left(p_A^2(1 - kr^m) + p_A(1 - p_A)(1 - kr)\right)$$

and

$$\Pr(a \mid \text{Unaffected}) = \frac{1}{1 - K}\left(p_A(1 - p_A)(1 - kr) + (1 - p_A)^2(1 - k)\right),$$

respectively.

Statistical tests

In population-based case–control studies, the aim is to detect those alleles more frequently observed in the cases than in the control groups. Generally, the statistical test used in association-based case–control study design is the χ^2 test for independence to detect significant intergroup differences (e.g., Sasieni 1997). We assume biallelic (SNP) markers and have employed the χ^2 test without Yates' correction by using 2×2 allele number tables (df = 1) to characterize marker–disease associations. In the present study, we assume no population stratification. When the numbers of allele *A* and *a* among *n* affected individuals are N_A and N_a, respectively, and the numbers of alleles *A* and *a* among *n* unaffected individuals are M_A and M_a, respectively, then

$$\chi^2 = \frac{(M_A N_a - M_a N_A)^2}{n(M_a + N_a)(M_A + N_A)}.$$

Stepwise focusing method

The stepwise focusing method comprises two (or more) steps. In the first focusing step, the samples from both case and control groups are genotyped at a certain number of SNP markers, and the markers that exhibit significant intergroup differences with the χ^2 test using 2×2 allele number tables at a relatively high significance level α (for example, $\alpha = 0.1$) are selected. This means that most of the markers selected in the first step are likely to be false positives. In the second step, the markers selected in the first step are tested by using a different set of case–control samples. In the second step, a larger number of individuals and a lower α value are used for marker selection. For example, suppose we carry out a case–control study using 50000 SNP markers and the criterion for the selection of markers is $\alpha = 0.1$ in the first step; then, the average number of false-positive markers selected in the first step will be 5000. If those 5000 markers are tested at $\alpha = 0.0001$ in the second step, then, on average, less than one false-positive marker will be selected.

Simulations and power calculations

To calculate the power to detect disease susceptibility alleles with the stepwise focusing method compared with that using the conventional method, we performed extensive simulations. First, we calculated the expected frequencies of alleles *A* and *a* in the case and control groups [$\Pr(A \mid \text{Affected})$ and $\Pr(a \mid \text{Affected})$, and $\Pr(A \mid \text{Unaffected})$ and $\Pr(a \mid \text{Unaffected})$, respectively] with various values of *r*, p_A, *k*, and *m* in the population. Second, we sampled n_1 affected and n_1 unaffected individuals by using a Monte Carlo simulation according to the probabilities of *A* and *a* alleles in each population assuming HWE. χ^2 tests for independence were then performed on the simulated samples. We repeated these tests 10000 times, and the proportion of tests in which the null hypothesis (no association) was rejected was determined.

Results and discussion

Power of the stepwise focusing method

When we set the parameter values to $k = 0.03$, $r = 2$, $m = 1$, and $p_A = 0.3$, then we obtained $K = 0.045$, $\Pr(A \mid \text{Affected}) = 0.40$, $\Pr(a \mid \text{Affected}) = 0.60$, $\Pr(A \mid \text{Unaffected}) = 0.30$, and $\Pr(a \mid \text{Unaffected}) = 0.70$. The first step of

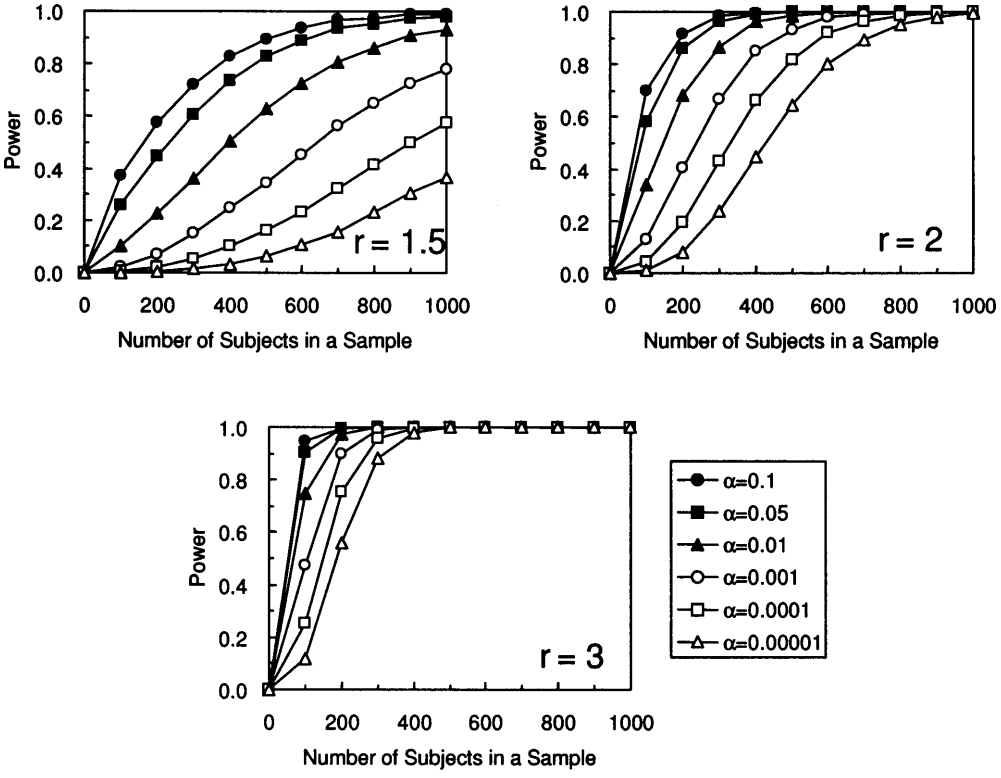

Fig. 1. Relationship between numbers of affected individuals examined and power. Parameters were $p_A = 0.3$, $k = 0.03$, and $m = 1$, and r as shown in the figure, where p_A denotes the frequency of allele A in the population, k is the probability of disease for a subject with genotype aa, and m and r are parameters such that the probabilities of the disease for the subjects with the genotypes Aa and aa are kr and kr^m, respectively. The probabilities of alleles A and a in affected and unaffected individuals were calculated from the above parameters, and the numbers of affected and unaffected individuals shown were simulated. The simulation was repeated 10000 times, and the proportion of repeats with significant differences (with varying α value) is shown as the power

focusing was simulated by randomly selecting the alleles from both affected and unaffected populations according to the probabilities described above. Thus, if the number of subjects in each population was 250 ($n_1 = 250$), we selected either allele A or a according to the probabilities described above for 250 affected subjects and 250 unaffected subjects, assuming HWE. The number of allele A should follow the binomial distribution for each of the two groups. We then made a 2×2 table, each cell of which contained the number of either allele A or allele a. The rows of the table corresponded to affected and unaffected groups, and the columns corresponded to alleles A and a. We then calculated the sample χ^2 value from the data in the table. This simulation (sampling of the alleles and calculating the χ^2 value) was repeated 10000 times, and the results showed that 0.90 of all sampling attempts resulted in χ^2 values of more than 3.84, which is the threshold χ^2 value at $\alpha = 0.05$. The relationship between the number of affected individuals examined and the power is shown in Fig. 1. Thus, with the parameters assigned as above, about 90% of the disease-associated markers can be expected to be selected in the first step. However, when markers with χ^2 values larger than this threshold are selected, an average of 5% of all markers selected will not be related to the disease. Therefore, if 50000 markers not associated with the disease are tested, the expected number of false-positive markers will be 2500. In the stepwise focusing method, the 2500 markers selected in the first step are subjected to the second focusing step by using a different set of samples consisting of a certain number (n_2) of affected and unaffected individuals.

If n_2 is 550, and the results are analyzed equivalently by using the χ^2 test for independence with $\alpha = 0.0001$ as the significance level, then the expected number of false-positive markers will be $2500 \times 0.0001 = 0.25$. The Monte Carlo simulation performed as in the first step revealed that 0.89 of the positive markers would be selected. Therefore, the probability that a positive marker will be selected through the two steps is $0.9 \times 0.89 \simeq 0.80$. Thus, if the population parameters are as stated above, the numbers of each group for the first and second steps are 250 and 550, respectively, and the thresholds for selection of the first and second steps are χ^2 values at $\alpha = 0.05$ and $\alpha = 0.0001$, respectively, the expected overall false-positive rate and power will be 0.000005 and 0.80, respectively. If 50000 markers are examined, the risk of excluding the null hypothesis of independence will be $50000 \times 0.000005 = 0.25$ considering Bonferroni's correction for multiple testing. The total number of genotypings necessary for this study will be $50000 \times 500 + 2500 \times 1100 = 27750000$.

How many genotypings are needed to obtain a similar type I error rate and power if the number of focusing steps is only one? For the expected type I error rate to be 0.000005, the threshold χ^2 value at $\alpha = 0.000005$ is required for the first (and only) step. The number of affected (and unaffected) individuals required to obtain a power of 0.80 with the parameters given above will be 635 (according to a Monte Carlo simulation as described above). In this case, the total number of typings required will be $50000 \times 1270 = 63500000$, which is much larger than that required with the stepwise focusing method (27750000 typings, as stated above). Alternatively, we can fix both the total number of genotypings and the overall type I error rate, and observe differences in the power. If the total number of typings is 27750000, the number of affected (and unaffected) indi-

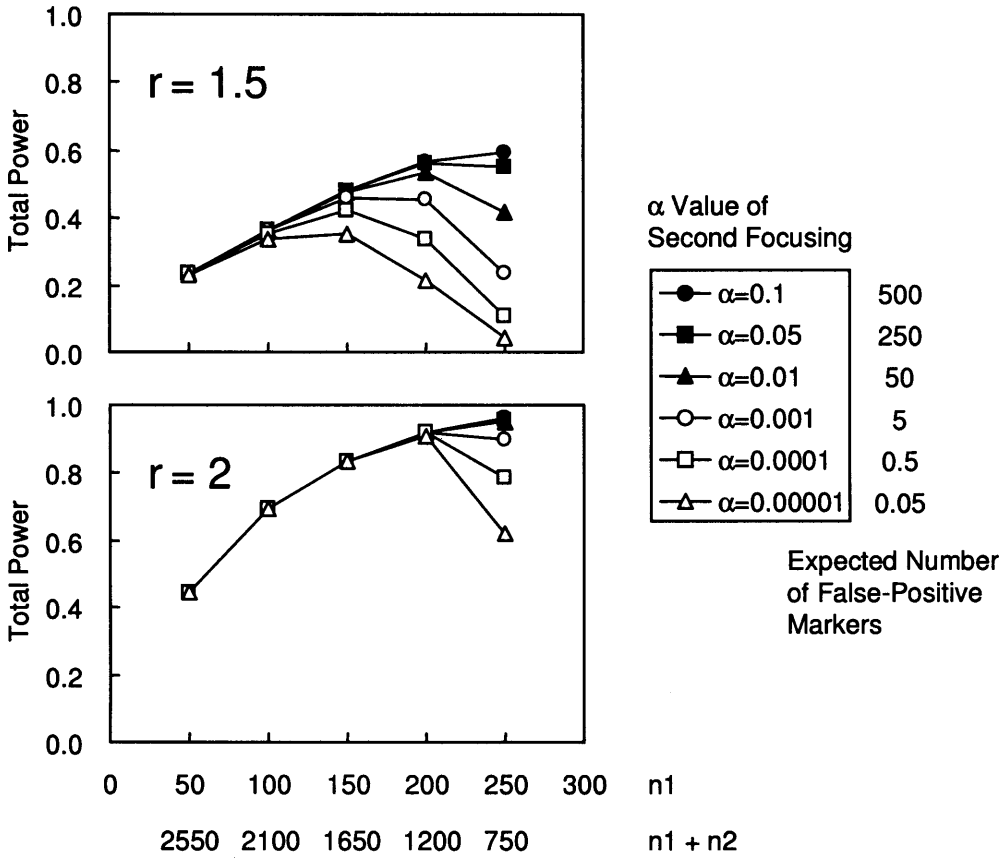

Fig. 2. Overall power in detecting association between a marker and the disease by the stepwise focusing method. The relationship between the number of subjects in each sample (affected and unaffected) in the first (n_1) and second (n_2) steps and the overall power. The total number of genotypings and the number of markers examined were set at 30000000 and 50000, respectively. The fixed values ($p_A = 0.3$, $k = 0.03$, and $m = 1$) were as in Fig. 1. The simulation was performed for two r values. The significance value in the first focusing step was set at 0.1, and, therefore, the expected number of false-positive markers after the first step was 5000. The significance value in the second step was varied from 0.1 to 0.00001, and the expected number of false-positive markers at each significance value is shown in the *right panel*. The *left panels* show the overall power under various conditions

viduals whose genes can be typed will be 277, and the power will be 0.15. Therefore, a much higher power can be attained with the stepwise focusing method than with the conventional method.

Optimization of study design using the stepwise focusing method

We searched for the optimal conditions for the stepwise focusing method when the total numbers of both genotypings and markers to be examined were fixed. These conditions are realistic for genome-wide linkage disequilibrium (LD) analyses using high-throughput typing systems. In such studies, the total number of typings limits the study because it affects the total cost most strikingly. Let N_T be the total number of typings allowed and N_M be the number of markers examined, and let the number of affected subjects examined in the first step be n_1. Then, the total number of typings in the first step will be $2N_M n_1$, and the number of typings for the second step will be $N_T - 2N_M n_1$. If the number of affected subjects examined in the second step is n_2, then the upper limit of the number of markers examined in the second step will be $(N_T - 2N_M n_1)/n_2$. Figure 2 shows the relationship between the total number of samples and the power of the stepwise focusing method when the population parameters are fixed at $p_A = 0.3$, $k = 0.03$, and $m = 1$. The threshold χ^2 value for the first focusing step was fixed at $\alpha = 0.1$. Therefore, the expected number of false-positive markers at the first step will be 5000. If we can collect 750 affected (and unaffected) individuals, and set the numbers of each group for the first and second steps at 250 and 500, respectively, the expected number of false-positive markers and overall power will be 0.5 and 0.79, respectively for $r = 2$ and an α value of the second focusing step of 0.0001. If we can collect 1200 affected (and unaffected) individuals, we can attain an overall power of 0.92, as the expected number of false-positive markers is 0.5.

The relationship between n_1, n_2, and the overall power under several values of p_A and k is shown in Fig. 3. The overall power is higher when the probability of disease for a subject with genotype *aa* is large. When the frequency of the susceptible allele in the population is small, the power is much lower, even if the stepwise focusing method is adopted.

In addition to the numbers of genotypings and markers, the number of affected (and unaffected) individuals is sometimes limited. Figure 4 shows the relationship between the significance value at the first and second steps, n_1, n_2, and the overall power when the total number of subjects in the affected and unaffected sample ($n_1 + n_2$) and the total number of genotypings are limited. The optimal number of subjects to be selected at the first and second steps varies depending on the value to which the significance level at the first step is set. In the case shown in Fig. 4, we can attain an overall power of 0.91 (the highest under such conditions) when we set the numbers of each group for the first and second steps at 200 and 800, respectively, and set the significance value at the first step at 0.1. Almost the same

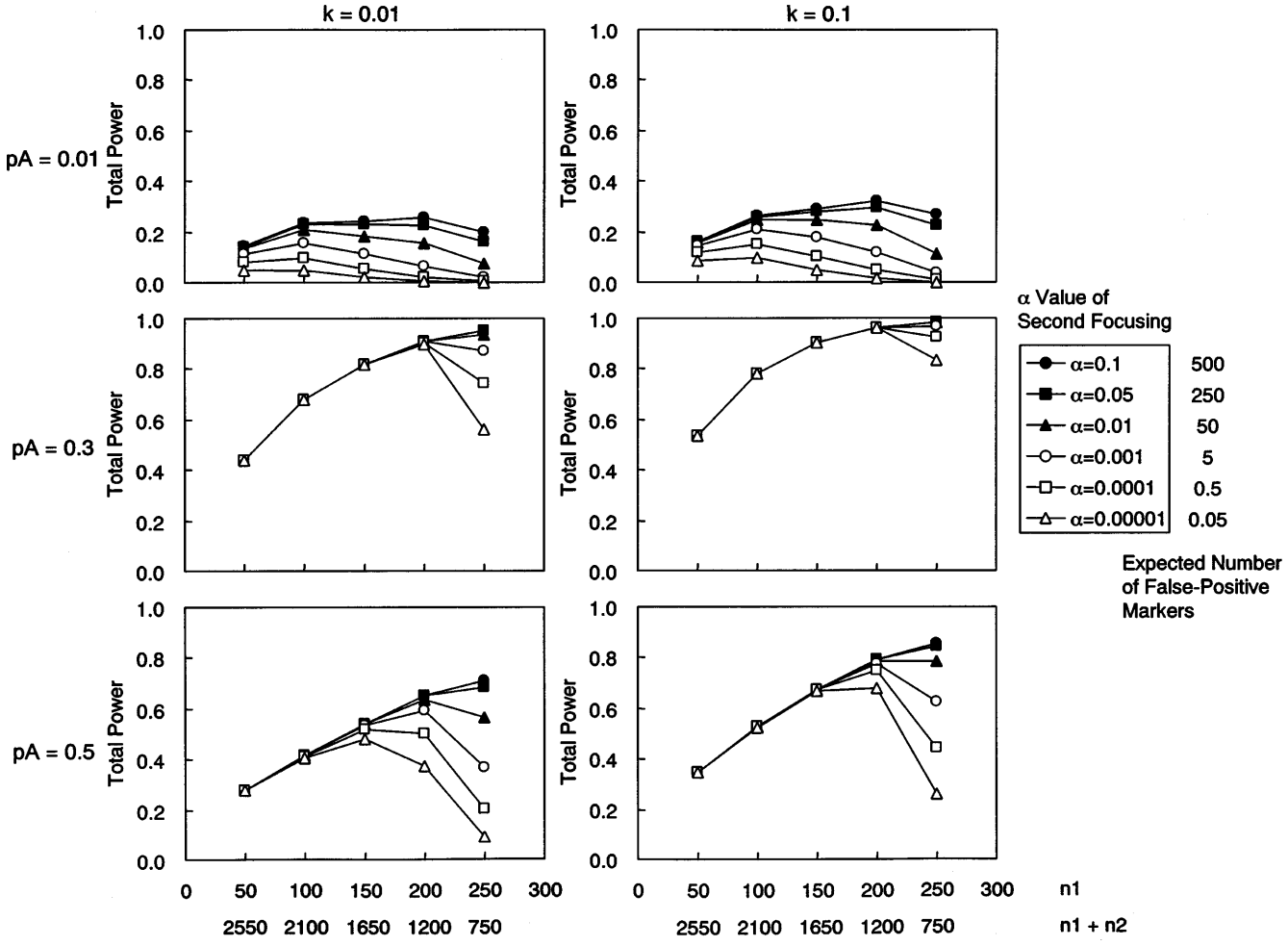

Fig. 3. Relationship between n_1, n_2, and the overall power for several values of p_A and k. The total number of genotypings and the number of markers examined were set at 30 000 000 and 50 000, respectively. The population parameters were set to $r = 2$ and $m = 1$. The significance value in the first step was set at 0.1

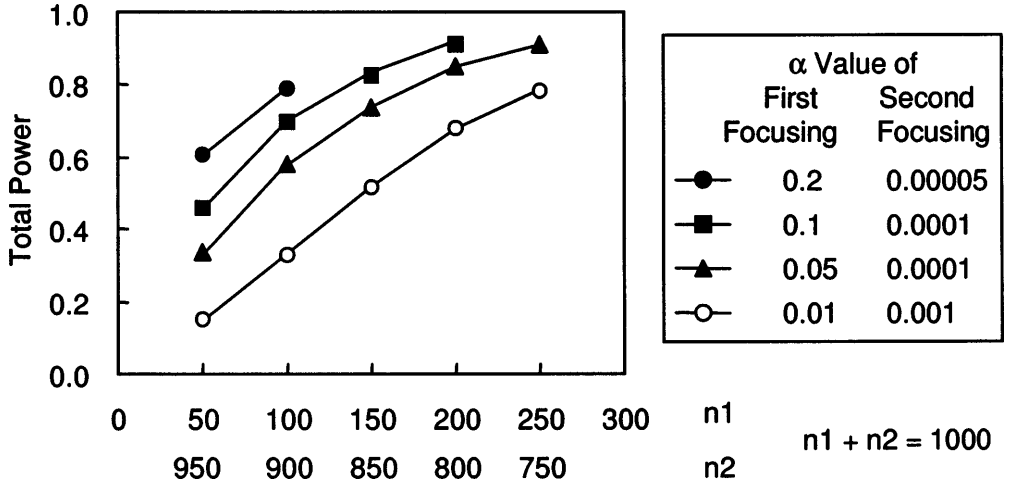

Fig. 4. Relationship between the significance value at the first and second steps, n_1, n_2, and the overall power when the total number of subjects in the affected and unaffected sample ($n_1 + n_2$) and the total number of genotypings are limited. $n_1 + n_2$ was fixed at 1000 and the total number of genotypings was fixed at 30 000 000. The number of markers examined was set at 50 000. The population parameters were set to $r = 2$, $p_A = 0.3$, $k = 0.03$, and $m = 1$. The significance values at the first and second steps were set so that the expected number of false-positive markers was less than 1. Results are plotted only for possible combinations in which the total number of genotypings is under 30 000 000

power can be attained when $n_1 = 250$, $n_2 = 750$, and the significance value at the first step is set at 0.05. The total numbers of typings necessary in each case are 28000000 and 28750000, respectively. Thus, we can determine the optimal values for the significance value at the first step, n_1, and n_2 under limited conditions using the stepwise focusing method.

In the present study, we used a Monte Carlo simulation method to infer alpha and beta statistics. To calculate the statistical power or required sample sizes, a classical equation using normal distribution is often applied. However, when one of the ratios (the expected frequencies of allele *A* and *a* in the case and control groups) in the 2×2 tables to test is too small (for example, when $p_A = 0.01$ in Fig. 3), it cannot be assumed that the difference in the two ratios follows a normal distribution. The Monte Carlo simulation method enables us to calculate almost exact statistical power or required sample sizes even in such cases. In almost all cases examined in this study (except for $p_A = 0.01$), the calculation results from the Monte Carlo simulation and from the classical equation are in good agreement (data not shown).

Thus, the stepwise focusing method presented here may be very useful for attaining a much higher power when a study is limited by the total cost of genotyping. By using the stepwise focusing method, the optimal study design given the total budget for genotyping can be estimated.

Acknowledgments This study was supported by the Research for the Future Program of the Japan Society for the Promotion of Science.

References

Amos CI, Page G (2001) Cost of linkage versus association methods. In: Rao DC, Province MA (eds) Genetic dissection of complex traits. Academic, San Diego, pp 213–221

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229–1231

Gu C, Rao DC (2001) Optimum study designs. In: Rao DC, Province MA (eds) Genetic dissection of complex traits. Academic, San Diego, pp 439–457

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci U S A 95:11389–11393

Ohashi J, Tokunaga K (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46:478–482

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

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

Risch N, Teng J (1998) The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8:1273–1288

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261