J Hum Genet (2004) 49:669–676
DOI 10.1007/s10038-004-0205-9

ORIGINAL ARTICLE

Yasunori Sato · Hideki Suganami · Chikuma Hamada
Isao Yoshimura · Teruhiko Yoshida · Kimio Yoshimura

Designing a multistage, SNP-based, genome screen for common diseases

Received: 8 June 2004 / Accepted: 18 September 2004 / Published online: 24 November 2004

Abstract A genome-wide linkage equilibrium mapping is an emerging strategy to identify risk-modifying genes for common diseases, despite unsettled controversies upon many aspects, including its premises, designs, marker choices and cost benefits. One large-scale attempt in Japan aims to identify disease-associated single nucleotide polymorphisms (SNPs) for five diseases among the Japanese population: Alzheimer's disease, gastric cancer, diabetes, hypertension and asthma. Following an initial screening of c.a. 100,000 SNPs on 940 subjects (five diseases × 188 patients) to select about 2,000 SNPs, we compared which subsequent screening design is more appropriate, and an additional one or two screens to further narrow down any disease-associated SNPs within a fixed total volume of 15,040,000 typings (2,000 SNPs × five diseases × 1,504 subjects, comprising 752 cases and 752 controls). We employed a Monte Carlo simulation to evaluate the probability of identifying truly disease-associated SNPs. The results suggest the single additional stage design (i.e., total two-stage design including the initial screening of 100,000 SNPs) was more practicable for the simple reason that the gain in probability is considered insufficient relative to an associated increase in study complexity in the three-stage design.

Y. Sato (✉) · H. Suganami · C. Hamada · I. Yoshimura
Faculty of Engineering, Tokyo University of Science,
1-3 Kagurazaka, Shinjyuku-ku,
Tokyo 162-8601, Japan
E-mail: yasu@ms.kagu.tus.ac.jp
Tel.: +81-3-52288350
Fax: +81-3-32605770

H. Suganami
Biostatistics and Data Management Department,
Kowa Co., Ltd., Tokyo, Japan

Y. Sato · T. Yoshida · K. Yoshimura
Genetics Division, National Cancer Center
Research Institute, Tokyo, Japan

Keywords Single nucleotide polymorphisms (SNPs) · Genome-wide approaches · Case-control study · Multistage design · Odds ratio · Sensitivity · Disease association study

Introduction

The development and progression of common diseases apparently occur through interactions of multiple factors that include those for life-style, environment, aging, and genes. One popular theory for the genetic basis of common disease susceptibility is the "common disease–common variant" hypothesis (Wright and Hastie 2001), which predicts that an association study can identify multiple genes with risk-modifying common alleles, each with low-to-moderate genetic relative risks. Unveiling such relatively common but "weak" disease genes may lead to at least three important developments in medicine in the postsequencing era: First, a high-population-attributable risk may aid in the identification of a high-risk population, which would benefit from specific preventive interventions, such as chemoprevention or a life-style modification clinic. Second, once all or a majority of the risk-modifying genes and alleles are disclosed, the disease risk for each individual can be evaluated with good accuracy (Pharoah et al. 2002). Third, but most importantly, identification of disease-associated genes should lead us to the uncovering of the molecular pathogenesis of a disease, which is the ultimate basis for development of targeted therapy, diagnosis, and prevention.

Even though the human genome sequencing per se was declared over, about 40% of the identified or predicted genes are tagged with no known functions (Venter et al. 2001; Okazaki et al. 2002), and it is obvious that our knowledge of the remaining "annotated" genes is far less complete. The realization of this serious dearth of knowledge has prompted many scientists to advocate a genome-wide screen for disease-gene hunting in addition

to the candidate gene approach, which relies upon an a priori selection of small numbers of candidate genes based on the existing information or hypothesis.

Two crucial fundamentals have been developed for the genome screen in Japan: First, a very high-throughput, low-cost, and DNA-saving single nucleotide polymorphism (SNP) typing platform (Ohnishi et al. 2001); and second, a high-quality and comprehensive gene-centric SNP database with allele frequency information among the Japanese population, designated JSNP (Haga et al. 2002; Hirakawa et al. 2002). The combined power of the two fundamentals crystallized in the first success in the whole-genome association study led by RIKEN, identifying functional SNPs that are associated with susceptibility to myocardial infarction (Ozaki et al. 2002), under the auspices of the government-sponsored Millennium Genome Project of Japan. In 2001, another whole-genome association study based on JSNP was launched within the Millennium Genome Project to identify genes associated with Alzheimer's, gastric cancer, diabetes, hypertension, and asthma (Yoshida 2002; Yoshida and Yoshimura 2003). This five-disease, joint, whole-genome association study has been organized as a collaboration of many institutions in Japan, including RIKEN, which provided the high-throughput typing technology (Ohnishi et al. 2001) and a significant portion of the typing itself. The number of SNPs analyzed by the first-stage screening in the whole-genome association study was about 100,000 per person, and a total of 940 patients (188 patients per disease) were genotyped. In the five-disease, whole-genome association study, 188 patients (cases) and 752 patients with other diseases (controls) were compared for each disease calculating sample odds ratios (ORs) for each SNP. In general, a common disease is expected to have more than ten genes with a population OR of 1.5–2.0 (Wright et al. 1999; Pharoah et al. 2002; Ponder 2001).

When 100,000 SNPs are analyzed in the first-stage screening, numerous false-positive results ensue due to the multiplicity of choices. Following the selection of a certain number of SNPs (e.g., 2,000) based on sample ORs in the first screening, one or more succeeding stages of screening are obviously required. However, for the succeeding stage experiments, the number of typings that can be performed is limited by the available resources, such as research funds, project period, and amount of DNA. The objective of the present study is to find the best design for the second and later stages of the whole-genome association study under the given fixed amount of total SNP typings. In this particular whole-genome association study, we assume a total of 3,008,000 typings (2,000 SNPs $\times$ 1,504 people, comprising 752 cases and 752 controls) for each disease.

Satagopan et al. (2002) compared the one-stage design with the two-stage design and showed that if the total number of typings is fixed, the probability for identifying disease-associated SNPs was maximized by a two-stage design. They concluded that 75% of the total cost should be used in the first stage to select the top 10% of SNPs as candidates for the second-stage screening, in which the remaining budget should be spent to select SNPs with a large value of test statistics to maximize the detection rate. However, their conclusion does not give a direct answer to our specific questions for the following reasons: first, Satagopan et al. (2002) compared one- versus two-stage designs for the first-time association study. They assumed a fixed number of candidate SNPs to be analyzed but varying numbers of subjects for the two designs. However, we need to choose one or two additional stages following the finished first stage of the genome screen (i.e., total two stages versus three stages, respectively). Because our "candidate" SNPs are selected from the first-stage experiment on the 100,000 SNPs, we can assign different numbers of SNPs to the two designs. On the other hand, unlike Satagopan et al. (2002), the total number of the subjects is fixed in our case (752 cases and 752 controls) because the subjects ascertained are the most important asset in our analysis. Second, their model assumes test statistics to be approximated to a normal distribution. The most conceivable statistic for such an approximation in a case-control study is a chi-square statistic, as mentioned in their paper, but we would like to evaluate possible designs without restricting the test statistics to the chi-square; for instance, other options include direct comparison of a sample OR and its P value, or P value calculated by Fisher's exact test. Third, they used power P_K, or probability to detect all of the true markers, to judge the performance of the different designs. By contrast, we introduced hit rate (R_h), which we believe to be more practical in the search for the genes implicated in polygenic complex traits.

In this study, to choose the optimal design for a whole-genome association study, a Monte Carlo simulation was undertaken to compare two- versus three-stage designs, which share a common 100,000 genome screen as their first stage. Of the five target diseases in the project, we examined gastric cancer as a representative in this study.

Materials and methods

Two-screening designs

Two- and three-stage whole-genome association studies are defined as follows in this study (Table 1).

Two-stage design (A)

A-1 (first stage) In 940 patients (188 gastric cancer cases and 752 people with four other diseases combined as "controls"), 100,000 SNPs per person were typed to determine a sample OR for each SNP to select a candidate SNP (n_1 SNP) in descending order of magnitude of sample ORs.

Table 1 Number of SNPs and subjects (number of cases, controls) analyzed in this whole-genome association study

Stage of experiment	Two-stage design	Three-stage design
Initial stage	100,000 SNPs (188, 752)	100,000 SNPs (188, 752)
Second stage	2,000 SNPs (752, 752)	n_2 SNPs (376, 376)
Third stage		n_3 SNPs (376, 376)

A-2 (second stage) A separate panel of 752 gastric cancer cases and 752 controls (individuals without gastric cancer) were chosen, and n_1 SNPs per person were genotyped and sample ORs were calculated. SNPs with a sample OR above a cutoff value, c, are selected as disease-associated SNPs.

Three-stage design (B)

B-1 (first stage) The experiment was conducted in the same manner as the two-stage design to select candidate SNPs with high sample ORs (n_2 SNPs).

B-2 (second stage) The second panel of 376 gastric cancer cases and 376 nongastric cancer controls were chosen, and the sample OR for the n_2 SNPs was calculated to further narrow down the number of candidate SNPs to n_3.

B-3 (third stage) In the third panel of 376 cases and 376 controls, the n_3 SNPs were genotyped, and the sample OR was calculated by combining data collected in steps B-3 and B-2 to identify the gastric-cancer-associated SNPs with sample ORs above the cutoff value, c.

Framework of simulation experiment

In the present study, a Monte Carlo simulation (Metropolis et al. 1953) experiment was used to compare the performance of the two designs described above. At each SNP, the association between its genotypes and disease status was tested by a 2×2 contingency table (Table 2). Here we assume an either dominant or recessive model, and for an SNP with allele A and allele a, the genotype AA combined with genotype Aa was compared with the genotype aa. An OR was defined as:

$$\psi = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}. \quad (1)$$

Table 2 Number of genotypes for cases and controls

Genotype	Case	Control
AA and Aa	n_{11}	n_{12}
aa	n_{21}	n_{22}
Total	$n_{\cdot 1}$	$n_{\cdot 2}$

When each cell in Table 2 was large, $\log \psi$ approximately had a normal distribution with true $\log \psi$ (mean) and $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$ (standard deviation) (Agresti 1990). The population OR for the gastric-cancer-associated SNPs was ψ ($\psi > 1.0$), while the population OR for SNPs unrelated to the disease was 1.0.

The number of the true SNPs (SNPs truly associated with the disease) was designated n_p, which is given as a simulation parameter. The sample log ORs of those n_p SNPs was generated separately from that of the non-disease-associated SNPs, the number of which is 100,000-n_p. We regarded SNPs experimentally positive when their sample ORs were above a cutoff value, another simulation parameter. The number of the experimentally positive SNPs was then designated N, which contains N_p disease-associated (true-positive) SNPs out of the original n_p SNPs. The number of false-positive SNPs (non-disease-associated SNPs) was therefore $N-N_p$. While N and N_p represent random variables controlled by experimental errors, n_p is a constant assigned as a simulation parameter.

As performance indicators for each design, two indicators defined by formulas 2 and 3, R_h (hit rate) and R_d (detection rate) were used. Hereinafter, these indicators will be expressed as percentages.

$$R_h = \frac{N_p}{N} \quad (2)$$

$$R_d = \frac{N_p}{n_p} \quad (3)$$

R_h is a proportion of true disease-associated SNPs among positive SNPs (positive predictive value) while R_d is the proportion of positively detected SNPs among true disease-associated SNPs (sensitivity). Ideally, both values should approach 100%, but since these two parameters are opposing in nature, an optimal balance between them must be assessed.

Algorithm of simulation experiment

The algorithm of the two-stage design, whole-genome association study is as follows:

A-1 (first stage)

1. We generated log ORs randomly for the n_p "true" markers (i.e., SNPs truly associated with the disease) in the first stage according to a normal distribution with mean $\log \psi$ and standard deviation $\sqrt{\frac{1}{94} + \frac{1}{94} + \frac{1}{376} + \frac{1}{376}}$. Here, n_p, the number of such true SNPs, was set at 100 based on inference from the literature. The population OR, ψ, was variably set from 1.3 to 1.9 with an increment of 0.2. The standard deviation was chosen as the minimum possible

value, which corresponds to a disease-associated genotype frequency of 50% (see below in *Design parameters*).

2. We generated log ORs randomly for the (100,000-n_p) non-disease-associated SNPs according to a normal distribution with mean 0 and standard deviation $\sqrt{\frac{1}{94}+\frac{1}{94}+\frac{1}{376}+\frac{1}{376}}$.
3. The log ORs generated above were combined and sorted in descending order of the values and then were selected up to the n_1th SNPs. Here, n_1 was fixed at 2,000.

A-2 (second stage)

4. We generated log ORs randomly for the n_p' true markers in the second stage according to a normal distribution with mean log ψ and standard deviation $\sqrt{\frac{1}{376}+\frac{1}{376}+\frac{1}{376}+\frac{1}{376}}$. Here, n_p' is the number of SNPs that are truly associated with the disease and were chosen in the first-stage screening (the starting number of the true SNPs was n_p). The number of the subjects analyzed in the second stage was 752 cases and 752 controls, who are individuals different from those analyzed in the first stage.
5. We generated log ORs randomly for the n_1-n_p' non-disease-associated SNPs according to a normal distribution, with mean 0 and standard deviation $\sqrt{\frac{1}{376}+\frac{1}{376}+\frac{1}{376}+\frac{1}{376}}$.
6. The log ORs generated in (4) and (5) for the second-stage typing were combined and sorted in descending order. The SNPs were considered disease-associated when their ORs exceeded the cutoff value (c). The value of "c" was variably set from 1.0 to 2.5, with an incremental of 0.1.
7. We calculated the hit rate R_h and detection rate R_d, as defined above.
8. Steps (1)–(7) were repeated 10,000 times.

The algorithm of the three-stage design is essentially identical to the two-stage design except for parameters such as standard deviation and the number of SNPs analyzed at each stage; standard deviation of log ψ used in the simulation for the second and third stages in the three-stage design is $\sqrt{\frac{1}{188}+\frac{1}{188}+\frac{1}{188}+\frac{1}{188}}$.

Design parameters

The two designs in this whole-genome association study include constants that should be set as simulation parameters: c, n_1, n_2, and n_3. The value of "c" is identical for both designs and is set from 1.0 to 2.5, with an incremental pitch of 0.1.

In this whole-genome association study, the first-stage screening was already finished, and n_1 (the number of SNPs analyzed in the second stage of the two-stage design of the whole-genome association study) was fixed at 2,000. In the case of the three-stage design, the total 752 cases and 752 controls need to be divided into second- and third-stage typings. Because our typing system requires sample numbers to be equal to or a multiple of 188, we initially compared the following combinations of case or control number for the second stage/third stage: 188/564, 376/376, and 564/188. We chose the 376/376 combination because it gave the best R_h and R_d for the SNPs for ψ (population OR) ≥1.5 (Table 3). It follows, then. that n_2 and n_3 were determined so that their sum equaled 4,000 ($n_2 > n_3$) to keep the total typing cost the same between the two- and three-stage designs. In the three-stage design, n_2, was set variably from 2,500 to 3,500, with an incremental pitch of 500, and n_3 was automatically set for each n_2. Table 1 shows the relationship between the number of subjects and the number of SNPs for the two designs. The number of simulations was set at 10,000.

A possible effect of genotype frequencies on the simulation was first evaluated by changing the disease-associated genotype frequencies in the range of 10–50%. Because the two- and three-stage designs showed essentially the same relative pattern at each genotype frequency (Table 5), here we show the result of the simulation with the minimum standard deviation of a sample log OR, as described above in *Algorithm of simulation experiment*; the difference between the two designs should best be illustrated with the minimum standard deviation, which corresponds to the assumption of 50% disease-associated genotype frequencies.

Table 3 Hit rates R_h (upper figures) and detection rates R_d (lower figures) for various combinations of the number of subjects in the three-stage design[a]

Number of cases (= number of controls) Second stage/third stage	ψ (population OR)			
	1.3	1.5	1.7	1.9
188/564	69.67	97.78	99.19	99.45
	1.65	21.10	61.39	87.01
376/376	50.01	99.92	99.98	99.98
	0.69	17.07	62.89	91.85
564/188	62.35	99.22	99.74	99.82
	1.06	17.29	58.47	89.28

[a] R_h and R_d were calculated for varying numbers of n_3, ranging from 500 to 2,000 with an incremental pitch of 500, and $n_3 = 500$ showed the best R_h and R_d, which are shown in this table

Results and discussion

Comparison of the two designs

Table 4 shows R_h and R_d (%) when the cutoff value, c, is set at 1.6. Taking into account the number of simulations, digits after the second decimal point are considered irrelevant. From the perspective of comparing the two experimental designs, the column with $\psi = 1.7$ is the most

Table 4 Hit rates R_h (upper figures) and detection rates R_d (lower figures) in percent

Design		Number of SNPs	ψ (population OR)			
			1.3	1.5	1.7	1.9
Two-stage design	Second stage	$n_1 = 2{,}000$	39.79	99.93	99.98	99.99
			0.51	14.88	59.07	89.95
Three-stage design	Second (mid) stage	$n_2 = 2{,}000$, $n_3 = 2{,}000$	41.80	88.38	95.73	97.15
			1.81	18.36	54.14	83.18
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	39.15	86.48	94.90	96.54
			2.04	19.59	55.65	84.05
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	36.86	84.73	93.94	95.86
			2.22	20.54	56.75	84.55
		$n_2 = 3{,}500$, $n_3 = 500$	35.31	83.22	93.08	95.18
			2.38	21.21	57.61	84.91
	Third (final) stage	$n_2 = 2{,}000$, $n_3 = 2{,}000$	41.22	99.95	99.99	99.99
			0.53	14.86	59.16	89.92
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	44.72	99.94	99.98	99.99
			0.59	15.75	60.76	90.79
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	47.51	99.92	99.98	99.99
			0.64	16.48	61.97	91.41
		$n_2 = 3{,}500$, $n_3 = 500$	50.01	99.92	99.98	99.98
			0.69	17.07	62.89	91.85

important, since R_d was about 60%. Under the three-stage design, when n_2 is set at 2,000, R_h and R_d are almost identical to those of the two-stage design ($n_1 = 2{,}000$). When n_2 is increased, R_d is also elevated in the three-stage design, but the gain in R_d is modest—only about 3.8% higher than that of the two-stage design.

Table 4 also shows the R_h and R_d at the second (mid) stage of the three-stage design. Comparing R_h of the second stage with that of the final third stage, it is noted that the third stage is necessary to suppress false positives in the three-stage design. On the other hand, when the population OR (ψ) is less than 1.5, R_d was actually more decreased in the third stage than in the second stage.

According to the study by Satagopan et al. (2002), the two-stage design is more advantageous than the one-stage design, but our analysis revealed only a marginal gain in R_d. Since sample ORs for SNPs unrelated to the disease did not exceed the cutoff value of 1.6, R_h was almost 100%; only 60% were identified in both designs (Figs. 1, 2). Moreover, the typing laboratory is burdened with the additional chore of having to divide the whole typing job into three stages instead of two, such as an increase in the complexity of the system and in the difficulty in finding the optimal combination of multiplex PCR. Overall, the limited extent of the predicted increase in R_d does not warrant the adoption of the more complicated three-stage design in the whole-genome association study in the Millennium Genome Project of Japan, which favors the simple two-stage scheme.

Although an accurate estimation of the number of disease-associated SNPs may not be possible, Sing et al. (1996) suggested that about 100 SNPs or about 30 genes determine susceptibility to coronary artery disease. Wright et al. (1999), Pharoah et al. (2002) and Ponder (2001) also documented the involvement of similar numbers of disease-associated genes in other common diseases. Setting the number of disease-asso-

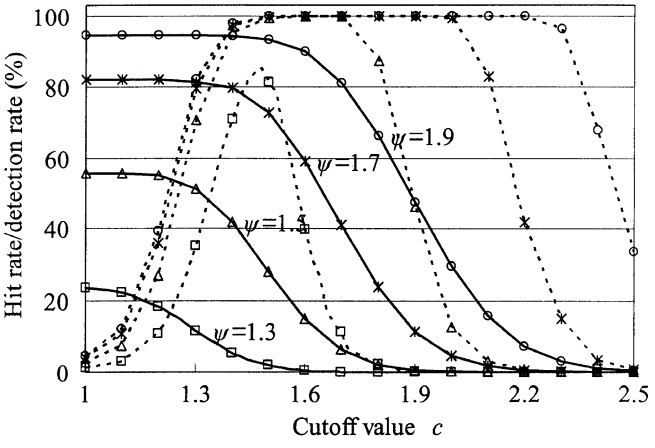

Fig. 1 Hit rate R_h and detection rate R_d against cutoff value c for the two-stage design. *Broken line* R_h, *solid line* R_d

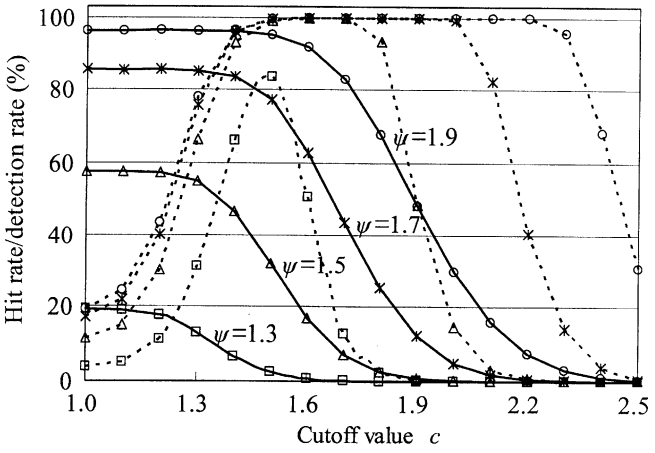

Fig. 2 Hit rate R_h and detection rate R_d against cutoff value c for the three-stage design. *Broken line* R_h, *solid line* R_d

ciated SNPs at 100 was thus not completely without grounds. However, even when the number of disease-associated SNPs, n_p, is set at 10, 30, 50, or 70, simulations showed results similar to those using an n_p of 100 (data not shown). It should be pointed out, however, that the present study is based on simulation, and that it is also necessary to seek theoretically optimal conditions.

To assess the effect of genotype frequencies, the simulation was repeated with the genotype frequency varying from 10 to 50%. Although the frequencies influenced power significantly, the comparison between the two- and three-stage designs showed essentially the same relative pattern (Table 5), confirming that our final conclusion is not significantly affected by the difference in genotype frequency.

Selection of cutoff value

In Table 4, the cutoff value was fixed at 1.6. Figures 1 (the two-stage design) and 2 (the three-stage design,

Table 5 Hit rates R_h (upper rows) and detection rates R_d (lower rows) in percent for different genotype frequencies (10–50%)

Genotype frequency (%)	Design	Number of SNPs	ψ (population OR)			
			1.3	1.5	1.7	1.9
10	Two-stage design	$n_1 = 2{,}000$	10.83	57.14	84.53	92.71
			1.01	8.79	30.86	59.62
	Three-stage design	$n_2 = 2{,}000$, $n_3 = 2{,}000$	11.48	58.34	85.66	92.93
			1.04	8.85	31.02	59.61
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	10.87	55.18	83.73	91.68
			1.21	9.67	33.14	62.05
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	10.35	53.09	82.04	90.64
			1.33	10.37	34.62	63.78
		$n_2 = 3{,}500$, $n_3 = 500$	9.88	52.37	81.22	90.20
			1.32	10.47	34.46	63.53
20	Two-stage design	$n_1 = 2{,}000$	40.23	89.74	95.33	96.37
			0.85	11.07	38.61	65.40
	Three-stage design	$n_2 = 2{,}000$, $n_3 = 2{,}000$	38.30	90.11	95.52	96.37
			0.86	11.27	38.62	65.52
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	39.19	88.72	94.74	95.68
			1.00	12.27	40.71	67.90
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	39.14	87.42	94.01	95.08
			1.11	13.08	42.30	69.51
		$n_2 = 3{,}500$, $n_3 = 500$	38.59	86.53	93.52	94.72
			1.17	13.44	42.60	69.59
30	Two-stage design	$n_1 = 2{,}000$	47.53	99.07	99.65	99.70
			0.69	13.31	48.91	79.64
	Three-stage design	$n_2 = 2{,}000$, $n_3 = 2{,}000$	47.73	99.12	99.59	99.67
			0.68	13.49	49.24	79.86
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	50.34	98.94	99.50	99.57
			0.75	14.44	51.26	81.50
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	53.20	98.82	99.43	99.49
			0.85	15.25	52.76	82.69
		$n_2 = 3{,}500$, $n_3 = 500$	54.44	98.71	99.38	99.43
			0.90	15.79	53.64	83.25
40	Two-stage design	$n_1 = 2{,}000$	46.63	99.85	99.91	99.95
			0.62	14.76	55.85	86.30
	Three-stage design	$n_2 = 2{,}000$, $n_3 = 2{,}000$	42.61	99.81	99.92	99.95
			0.58	14.76	55.63	86.30
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	45.62	99.78	99.90	99.93
			0.63	15.70	57.36	87.50
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	48.55	99.72	99.89	99.92
			0.68	16.46	58.73	88.37
		$n_2 = 3{,}500$, $n_3 = 500$	50.75	99.70	99.87	99.91
			0.72	17.06	59.64	88.90
50	Two-stage design	$n_1 = 2{,}000$	39.79	99.92	99.98	99.98
			0.51	14.88	59.07	89.95
	Three-stage design	$n_2 = 2{,}000$, $n_3 = 2{,}000$	41.22	99.94	99.98	99.99
			0.53	14.86	59.16	89.91
		$n_2 = 2{,}500$, $n_3 = 1{,}500$	44.72	99.93	99.98	99.98
			0.59	15.75	60.75	90.78
		$n_2 = 3{,}000$, $n_3 = 1{,}000$	47.51	99.92	99.98	99.98
			0.64	16.48	61.97	91.41
		$n_2 = 3{,}500$, $n_3 = 500$	50.01	99.91	99.97	99.98
			0.69	17.07	62.88	91.85

$n_2 = 3{,}500$) show changes in R_h and R_d with changing cutoff values. The choice of c should primarily depend on the purpose of the genome screen or its role in the overall gene-hunting scheme; some investigators may request a higher positive predictive value (R_h) while sacrificing sensitivity (R_d), yet other researchers may choose the other way around. One simple way to deal with the trade-off between R_h and R_d may be to select c as intersection points of the R_h and R_d curves for each ψ value, i.e., (ψ: c) = (1.3:1.24), (1.5:1.26), (1.7:1.30), and (1.9:1.38). With these cutoff values, the degree of improvement in R_d for the three-stage design is about the same as that mentioned above for $c = 1.6$. However, it is generally recommended that a cutoff value be set a little lower than ψ, and collecting information about population ORs of the disease-associated SNPs is clearly important.

Population OR for gastric cancer

Using the real experimental data of about 80,000 SNPs on gastric cancer from the first stages of the whole-genome association study, the frequency distribution of sample ORs was calculated and shown in Fig. 3. About 90% of the SNPs fall within the range of a sample OR from 0.6 to 1.7, suggesting that disease-associated SNPs within this range of a population OR would be difficult to detect by the sample size of the current project. In other words, the number of subjects needs to be increased to detect gastric-cancer-associated SNPs with population ORs smaller than 1.7 in this whole-genome association study. It was, therefore, reasonable to compare the performance of the screening designs under the condition in which the population OR, ψ, and cutoff value are set to about 1.7 and 1.6, respectively.

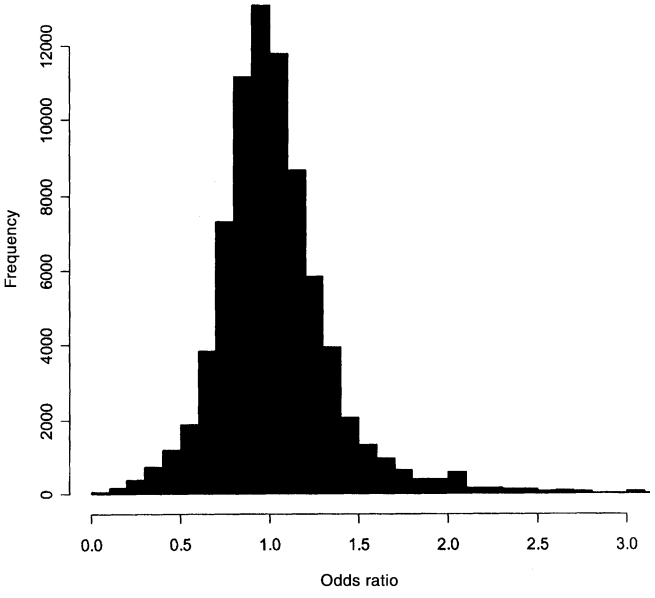

Fig. 3 Observed distribution of genotype ORs for gastric cancer in the first stage of the whole-genome association study in the Millennium Genome Project of Japan

It should be noted, however, that our sample ORs were measured between individuals with gastric cancer and those with the other four diseases combined (“control”) in the present five-disease whole-genome association study. Thus, the control population for the first stage is different from that of the second- or third-stage screening in which the controls were specifically ascertained as those without gastric cancer. As a consequence, some ambiguity may exist in our assessment of the population ORs for common gastric cancer.

Issues to be addressed in future studies

In the present simulation, each SNP is assumed to be distributed independently among the subjects. However, this is not always the case because significant linkage disequilibrium (LD) is often noted over pairs of the SNPs. In the future, correction of the present study may be necessary, taking into account the LD structure of the relevant SNPs.

Acknowledgements We thank N. Takano and R. Nakajima for their technical assistance. This study was partially supported by the Program for Promotion of Fundamental Studies in Health Sciences of the Pharmaceuticals and Medical Devices Agency of Japan.

References

Agresti A (1990) Categorical data analysis. Wiley, New York

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. J Hum Genet 47:605–610

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A,

Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I& II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36

Ponder BA (2001) Cancer genetics. Nature 411:336–341

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. Biometrics 58:163–170

Sing F, Haviland B, Reilly L (1996) Genetic architecture of common multifactorial diseases. In: Variation in the human genome (Ciba Foundation Symposium 197). Wiley, Chichester, pp 211–229

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Wright AF, Hastie ND (2001) Complex genetic diseases: controversy over the Croesus code. Genome Biol 2:2007.1–2007.8

Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. Nat Genet 23:397–404

Yoshida T (2002) SNP project in the Millennium Genome Project, Japan (in Japanese). Gan To Kagaku Ryoho 29:963–967

Yoshida T, Yoshimura K (2003) Outline of disease gene hunting approaches in the Millennium Genome Project of Japan. Proc Jpn Acad 38:168–176