J Hum Genet (2005) 50:53–61
DOI 10.1007/s10038-004-0220-x

ORIGINAL ARTICLE

Takahiro Nakamura · Akira Shoji · Hironori Fujisawa
Naoyuki Kamatani

Cluster analysis and association study of structured multilocus genotype data

Received: 13 September 2004 / Accepted: 5 November 2004 / Published online: 5 February 2005

Abstract We propose an algorithm for testing association using structured multilocus genotype data. The algorithm implements the clustering of the data by a hierarchical clustering technique and a *k*-means algorithm. After clustering, the program analyzes all the clusters together using the Mantel–Haenszel (MH) test, by which common associations in the clusters are examined. To use the MH test, the number of subpopulations has to be determined. A method of cross-validation (CV) and the *k*-means algorithm are applied for estimating the number of subpopulations. The algorithm described was implemented in the computer program POPSTRUCT. In the simulation study, we found that when the two groups with different marker allele frequencies were combined, an inflation of the type I errors was observed. The inflation was more marked when the differences in the marker allele frequencies were larger, the difference in the minor allele frequencies at the disease locus was larger, and the genotype relative risk associated with the disease locus was higher. Our simulation study indicated that the MH test was efficient for decreasing type I errors and increasing the power compared with any test performed on each cluster. Then, we compared the results of STRUCTURE, a model-based method, and POPSTRUCT, a distance-based method. When two subgroups with different allele frequencies were mixed together at a high fixed ratio, POPSTRUCT was superior to STRUCTURE in classifying the combined population into the accurate clusters, each of which reflects one of the original groups.

T. Nakamura (✉) · A. Shoji · N. Kamatani
Division of Statistical Genetics, Institute of Rheumatology, Tokyo Women's Medical University, 10-22 Kawada-cho, Shinjuku-ku, Tokyo 162-0054, Japan
E-mail: nakataka@ior.twmu.ac.jp
Tel.: +81-3-52691725
Fax: +81-3-52691726

H. Fujisawa
The Institute of Statistical Mathematics, Tokyo, Japan

Keywords Hierarchical clustering · *k*-means algorithm · Mantel–Haenszel test · Cross-validation · Population structure

Introduction

One of the most important problems in case-control association studies that use genotypic data is the structuring (or stratification) of the population. If the population is structured, false-positive results are easily obtained by the association study because the distribution of the statistics, such as Pearson's χ^2 statistic for the test of goodness of fit, is based on the assumption that the population is not structured. In association studies that use multilocus genotype data, structuring of the population can be a confounding factor. If the confounding factor is observable, such as gender or age, it is possible to analyze the data by excluding the influences of these factors. However, the structuring of the population is not directly observed for either genotype or phenotype data.

In previous studies, the transmission/disequilibrium test (TDT) was proposed and used as a method for association studies using structured populations (Spielman et al. 1993). The method is little influenced by the presence of population structures. However, this method requires families, each of which has an affected individual. As a method for studies of a structured population, methods have been proposed by Devlin and Roeder (1999), Satten et al. (2001), Pritchard and Rosenberg (1999), and Hoggart et al. (2003). Especially, the algorithm used in the software STRUCTURE (Pritchard et al. 2000a; Falush et al. 2003) was proposed. The degrees of membership to original subpopulations can be obtained by this method for each individual. An association-mapping method using the results of STRUCTURE, STRAT, was proposed (Pritchard et al. 2000b). However, favorable results may not be achieved by STRUCTURE when a few individuals are in one population and the other individuals are in another population.

The purpose of this article is to propose a new method for analyzing structured multilocus genotype data. This method can suggest the association between the candidate locus and phenotype by multilocus genotype data that may include the population structure. This algorithm consists of two parts: obtaining the population structure by the clustering algorithm and testing the association between the genotype and phenotype, excluding population structure. As the clustering technique, we used the hierarchical clustering technique and the *k*-means algorithm (Everitt 1993). The advantages and disadvantages of both methods are discussed. As the method for testing association excluding the population structure, we used the Mantel–Haenszel (MH) test (Cochran 1954; Mantel and Haenszel 1959). This latter test is intended to decrease the number of false-positive results and increase the power. The true number of subpopulations is required to use the MH test. However, this true number is unknown in the real data analysis. We estimated the number by the method of cross-validation (CV) (Stone 1974).

The algorithms described in this article were implemented in the computer program POPSTRUCT. This program contains three modes: analysis, simulation, and generate. The analysis mode is to analyze the data, as described above. We can simulate the occurrence of false-positive results in a structured population with the simulation mode. Furthermore, the effects of our method can be examined. We can generate the structured population by using the generate mode. In this article, simulations for false-positive results are shown. Finally, we compared the results of STRUCTURE and POPSTRUCT to describe the importance of the model-based and distance-based methods. The results of comparisons are shown.

Materials and methods

To test associations between genotype and phenotype, it is very important to examine the structure of the population. False-positive results may appear in the association mapping when using a combined population whose allele frequencies are different between subpopulations. Therefore, we need to inspect whether the population of the sample data is the combination of subpopulations or not. Testing the association by excluding the population structure is also required. In this section, we illustrate algorithms for estimating the population structure using clustering methods. Various clustering techniques have been proposed in previous studies. The hierarchical method, the optimization method, and fuzzy clustering are well known. In this article, we use the hierarchical clustering technique (neighbor-joining method) and *k*-means algorithm, which is a kind of an optimization method (Everitt 1993). We assumed that all the markers were unlinked. The methods presented in this article are available for both biallelic and multiallelic markers.

Hierarchical cluster analysis for obtaining the population structure

For the hierarchical clustering technique, we need to define the distances between individuals and the method for joining clusters. In general, Euclidean, L1, and Mahalanobis distance are used. Nonsimilarity is also used instead of these distances for the clustering. In addition, the method to recalculate distances after joining the clusters needs to be specified. In general, the single-linkage clustering technique, group-averaging clustering technique, centroid clustering technique, or Ward's clustering technique was used. The clustering was performed according to the distances defined by one of these methods. Let X denote the multilocus genotype data in n individuals, L denote the number of loci, and K denote the number of subpopulations. The distance between individuals i and j, $D(i, j)$, is defined as

$$D(i,j) = \frac{1}{L}\sum_{l=1}^{L} d_l(i,j),$$

where

$$d_l(i,j) = \begin{cases} 0 & \text{if individual } i \text{ and } j \text{ share two alleles at the } l\text{th locus,} \\ 1 & \text{if individual } i \text{ and } j \text{ share one allele at the } l\text{th locus,} \\ 2 & \text{if individual } i \text{ and } j \text{ do not share any alleles at the } l\text{th locus.} \end{cases}$$

This distance was used in Bowcock et al. (1994) to obtain human evolutionary trees. The distance matrix is given by the above definition. To join clusters, we used Ward's clustering technique. The algorithm for the clustering is as follows:

Step 1: Each of n individuals is given a number to become a cluster.
Step 2: A distance matrix D is made after calculating D_{ij} for the genetic data.
Step 3: Search the smallest entry in matrix D to find the pair with the minimum distance. The rule for cases where there are pairs with the same distance must be decided in advance.
Step 4: Join the pair and decrease the number of clusters from n to $n - 1$
Step 5: Calculate the distances again to update D.
Step 6: Repeat these steps until the number of clusters becomes K.

The choice of distance is arbitrary. If another definition of the distance between individuals and another clustering technique is used, the results may change. The advantage of this algorithm is the rather short time required for the calculation. Therefore, a population with a large number of individuals can be handled. However, this is a distance-based method, but it is not based on genetic rules, which is obviously a

disadvantage. Even if we use multiallelic markers, the distance mentioned above is available. However, we can define another distance based on the similarity of alleles; for example, numbers of repeats of microsatellite markers.

k-means algorithm for obtaining the population structure

For using the *k*-means algorithm, we need to define the criteria for the evaluation of clusters. We defined a likelihood function based on Hardy–Weinberg's equilibrium.

$$L(\mathbf{X}|C) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \prod_{l=1}^{L} \prod_{a=1}^{2} p_{klx_l^{(i,a)}},$$

where $C=(C_1, \ldots, C_K)$ denotes a partition of the individuals and $x_l^{(i,a)}$ is the allele of individual i in chromosome a at the locus l. p_{klj} denotes the relative frequency of allele j in the k th cluster on the locus l. There is no loss of generality, even if we omit the constant 2 for heterozygote loci. We used a log-likelihood function as the criteria.

$$\log L(\mathbf{X}|C) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{l=1}^{L} \sum_{a=1}^{2} \log\left(p_{klx_l^{(i,a)}}\right).$$

The algorithm for the clustering is as follows:

Step 1: Let C be an initial partition.
Step 2: Estimate allele frequencies in each subpopulation and calculate the log-likelihood log $L(\mathbf{X}|C)$.
Step 3: Assign an individual to another cluster and calculate the estimates of allele frequencies and the log-likelihood.
Step 4: Search for the partition with the maximum log-likelihood.
Step 5: Repeat steps 3 and 4 until the log-likelihood converges.

The partition with the maximum likelihood can be obtained by the above method. This is a model-based method and is based on genetic rules (Hardy–Weinberg's equilibrium), which is, obviously, the advantage of this algorithm. However, this method consumes much time for the calculations. Therefore, it cannot be applied to the data that include large numbers of samples. For saving calculation time, it is better to use the results of hierarchical clustering for the initial condition for the partition.

Estimation of the number of subpopulations, *K*

The clustering algorithms mentioned above require the number of subpopulations, K, in advance. However, K is unknown in most of cases in real data analysis. Therefore, the estimation of the number of subpopulations is important. It is difficult to estimate K by comparing the distances or the likelihood. We use a method of CV (Stone 1974) for the problem. CV is a criterion for choosing good models that is an approximately unbiased estimator for the risk function. It is difficult to define the risk function on the basis of the hierarchical clustering technique. Therefore, we apply the k-means algorithm for the estimation. The risk function is defined by a log-likelihood. The criterion CV corresponding to $\hat{K}$ is defined by the equation

$$\begin{aligned} \mathrm{CV} &= -\frac{1}{n}\sum_{i=1}^{n} \max_{k=1,\ldots,\hat{K}} \log L(x_i | x_i \in C_k, G_{-i}) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \max_{k=1,\ldots,\hat{K}} \left(\sum_{l=1}^{L} \sum_{a=1}^{2} \log\left(p_{klx_l^{(i,a)}}\right) \right), \end{aligned}$$

where $\hat{K}$ is the estimate of K. Thus, CV can be calculated by maximizing the log-likelihood over $p_{klx_l^{(i,a)}}$ on the basis of genotype data, except for individual i, G_{-i}. Hardy–Weinberg's equilibrium is assumed in each subpopulation.

As the number of subpopulations increases, the log-likelihood also increases. Therefore, comparing likelihood is not available for the estimation. The maximum likelihood estimator of the risk is not unbiased. The bias depends on the number of parameters. However, the estimator of the risk using CV is approximately unbiased (Stone 1977). Therefore, we compare values of CV with arraying value of $\hat{K}(\hat{K} = 1, 2, \ldots)$. When the subpopulations could be separated clearly and no outliers were presented, CV decreased. However, CV increased when the subpopulations were similar and the clustering was not efficient. The estimate of K is the number with minimum CV.

The method of CV mentioned above is called leave-one-out CV. The method using leave-one-out CV consumes much time for the calculation. Therefore, we also use K-fold CV (Efron and Tibshirani 1993) for samples with large sizes. We split samples of individuals into M roughly equally sized groups. The minimum risk for each individual within m th part was calculated on the basis of genotype data, except for m th group G_{-m}. This method consumes less time for the calculation. Our simulation studies were performed by K-fold CV.

Test for decreases in the false-positive rate

After the estimation of clusters and the number of subpopulations K was done, we tested the association between marker loci and an affection locus on the basis of the information of the stratification. The results of the association test using the data sampled from the combined population may include false-positive results, even if those using the original subpopulations do not. This phenomenon may occur when the ratio of sample size of cases to controls or the ratio of exposures is biased. If gender is the confounding factor, we can analyze two

sets of data after dividing the entire data set by gender. However, if the difference in allele frequencies between subpopulations is the confounding factor, we cannot divide individuals into some groups by the genotypes easily because the population structure cannot be observed directly. The subgroups reflecting the different subpopulations can be estimated by the hierarchical clustering method. The test of independence is performed in each estimated cluster, and MH tests using the information in clusters are carried out to decrease false-positive results and increase the power. We can exclude the influence of the confounding factor, i.e, the difference in the allele frequencies between subpopulations, by the algorithm. The k clusters are obtained by the clustering method mentioned above. The contingency table for each cluster can be obtained. Let y_{ijk} ($i=1,2;j=1,2;k=1, \ldots, K$) denote the element of i th row and j th column for the k th cluster's contingency table. We tested the hypothesis that common odds ratio equals to 1 by the test statistic $S_{\mathrm{MH}}(Y)$:

$$
\begin{aligned}
S_{\mathrm{MH}}(Y) &= (|y_{11\cdot} - E(Y_{11\cdot})| - 1/2)^2 / V(Y_{11\cdot}), \\
E(Y_{11\cdot}) &= \sum_{k=1}^{K} E(Y_{11k}), \\
V(Y_{11\cdot}) &= \sum_{k=1}^{K} V(Y_{11k}), \\
E(Y_{11k}) &= y_{1\cdot k} y_{\cdot 1k} / y_{\cdot\cdot k}, \\
V(Y_{11k}) &= y_{1\cdot k} y_{2\cdot k} y_{\cdot 1k} y_{\cdot 2k} / \{ y_{\cdot\cdot k}^2 (y_{\cdot\cdot k} - 1) \}.
\end{aligned}
$$

The variables that contain dots in subscription denote the marginal frequencies. This statistic follows χ^2 distribution with 1 degree of freedom, allowing tests that exclude the influence of the confounding factor.

Results

Simulation of the structured data and the estimation of type I error rates

The program POPSTRUCT simulates the appearance of false-positive results by combining the two different subpopulations. To initiate the simulation, the following parameters were given:

1. Numbers of individuals and loci
2. Frequencies of minor alleles at disease locus for the two subpopulations
3. The values a, r, and m, so that a, ar and ar^m are the penetrances for the subjects with the genotypes 1/1, 1/2, and 2/2, respectively, at disease locus
4. Frequencies of minor alleles at marker loci for two subpopulations. There are three options for giving the frequencies of minor alleles at marker loci in the two subpopulations. In the first mode, two different minor allele frequencies were given to the two subpopulations so that all the marker loci for a subpopulation were the same (fixed mode). In the second mode, two different minor allele frequencies, p_0 and p_1, are required. The allele frequencies in subpopulations 1 and 2 were p_0 and p_1, respectively, for the odd-numbered loci. The allele frequencies in subpopulations 1 and 2 were p_1 and p_0, respectively, for the even-numbered loci (flip-flop mode). In the third mode, a real value drawn at random from a uniform distribution in a given range was assigned as a minor allele frequency at each marker locus (random mode).
5. Number of repeats

To determine the empirical type I error rates, two subpopulations were generated under the assumption of no association between marker loci and the disease locus. Each individual becomes affected at the probability of the individual's penetrance that was determined by the genotype at the disease locus. The parameter a was set at 0.02, m at 1, and r between 1 and 5. After generating the two subpopulations, the same number (fixed at 150) of affected and nonaffected subjects were selected from each of the two subpopulations, and the test of the association was performed. There were two options for the test of the association. Thus, allele frequencies were compared between the affected and nonaffected groups using the χ^2 test (allele-frequency mode). In the other mode, the number of the subjects with the minor allele was compared between the affected and nonaffected groups (dominant mode). We used allele-frequency mode in this study. Adjustment by Yates's correction was not performed for the analysis. The proportion of the marker loci that exhibited the significant association at the given α value as judged by the χ^2 test was recorded as the empirical type I error rate. The alleles at marker loci should not be associated with the disease because the simulation was performed under the assumption that there was no association between the marker loci and the disease locus. Therefore, the results of the positive association were judged as type I errors. The above association test was done for each subpopulation. Then, the two subpopulations were combined, and both affected and nonaffected individuals were similarly selected. Then, the similar test of the association was performed on the combined population. The combined population was subjected to the clustering procedure, and the two final clusters were obtained. The test of the association was performed on each cluster, and the MH test was performed on the two clusters. The empirical type I error rates were determined from the proportions of the marker loci that exhibited the significant association at the given α value. In a single step of the simulation and the analysis, six different empirical type I error rates of four kinds were obtained: two rates for the two subpopulations, one for the combined population, two for the two clusters, and one when the MH test was used on two clusters.

The empirical power was determined by performing the test of the association using the genotype data at disease locus. It was determined in each of the groups in which the determination of the empirical type I error rate was performed.

The simulation of the two subpopulations followed by the analysis was repeated 1,000 times, and the distributions of the six type I error rates of four different kinds were determined. For the simulation, the number of the marker loci was set at 199 and the number of the disease locus at 1.

If the final cluster that each subject belonged to was different from his (or her) original subpopulation, the subject was judged as misclassified. As a method for clustering, we used hierarchical clustering in these simulations.

Table 1 illustrates the parameters used for the following five separate simulations and misclassification rate, which are the results of the simulations. Figure 1a shows the results of one such simulation, followed by the analysis. In this simulation, frequencies of minor alleles at the disease locus were 0.1 and 0.3 in the two subpopulations, and the random mode was used for the marker loci in which the range of the minor allele frequencies was 0.1–0.4. After the simulation, the rate of the misclassification was determined, which turned out to be 0.05. The threshold value for the χ^2 statistic for the test of the association was set at 3.84, where the cumulative distribution function for χ^2 distribution with 1 degree of freedom is 0.95. This means that α value was set at 0.05. Figure 1a and Table 2 indicate that the empirical type I error rates were about the same as the α value of 0.05 in the original subpopulations, two final clusters, and the MH test. However, the number of type I errors increased when the two subpopulations were combined. The type I error rate increased in the combined population as the genotype relative risk increased (Fig. 1a).

Figure 1b and Table 3 show the results of another simulation. The difference between the simulations for Fig. 1a, b is that in the latter simulation, the range of the minor allele frequencies for the marker loci was between 0.1 and 0.2. This means that the difference in the minor allele frequencies between the two subpopulations was not large. The rate of the misclassification was 0.42, indicating that the clustering procedure did not work efficiently because the difference in the minor allele frequencies between the two subpopulations was small. Figure 1b shows that the type I errors did not increase significantly in the combined population. In these cases, the population was not significantly structured, and the clustering analysis was not efficient.

In Fig. 1c, the range of the minor allele frequencies at marker loci was between 0.4 and 0.5. Therefore, the difference in minor allele frequencies between the two subpopulations at each marker locus was not very much different. Other conditions were the same as in Fig. 1a, b. The type I error rate did not increase significantly, even in the combined population. Rather, the type I errors inflated when the data in the clusters were analyzed by the MH test. The method does not work because of a high misclassification rate, 0.46.

Figure 1d, e show the results of similar simulations. For Fig. 1d, the frequencies of minor alleles at the disease locus were 0.1 and 0.2 in the two subpopulations. The range of the minor allele frequencies at marker loci was 0.1–0.4. For Fig. 1e, however, the minor allele frequencies at the disease locus were 0.1 and 0.5 in the two subpopulations, and the range of the minor allele frequencies at marker loci was 0.1–0.4.

In both simulations, the type I errors were inflated for the combined populations, and the inflation was more evident when the genotype relative risks were high. Such inflations in the type I errors were not evident in the original subpopulations, in each cluster, or in the result of the MH test after the clustering procedure (Fig. 1d, e). When the results of the simulations in Fig. 1d, e were compared, the inflation of the type I errors was more evident in the simulation in Fig. 1e. This difference is attributed to the fact that in the simulation in Fig. 1d, the frequencies of the minor alleles at the disease locus were 0.1 and 0.2 in the two subpopulations whereas they were 0.1 and 0.5 in the simulation in Fig. 1e. Therefore, the inflation of the type I errors was augmented by the increase in the difference of the minor allele frequencies at the disease locus. The latter factor, of course, is related to the incidence of the affected individuals in the subpopulations.

Figure 2 shows the change of the power as a function of the genotype relative risk. As expected, the power increased when the genotype relative risk increased. After the clustering procedure, the power for each cluster was significantly lower than the power for the combined population. However, when the data from the two separate clusters were analyzed together by the MH test, the power was comparable to that for the combined population. This is as expected because the sample size for each cluster was approximately half of the size in the combined population.

Simulation study for estimating K

We evaluated the accuracy of the method using CV by simulation study. The method using leave-one-out CV

Table 1 Parameters used in simulations of type I errors and percentage of misclassification, which are the results of clustering

Simulation number	1	2	3	4	5
Range of frequencies of minor alleles	0.1–0.4	0.1–0.2	0.4–0.5	0.1–0.4	0.1–0.4
Frequencies of minor alleles at affected locus in two subpopulations	0.3, 0.1	0.3, 0.1	0.3, 0.1	0.2, 0.1	0.5, 0.1
Percentage of misclassifications	4.98	40.683	45.627	5.118	4.566

Table 2 Empirical type I error rates with standard deviations (SDs) for simulation 1–5. *MH* Mantel–Haenszel test

Genotype relative risk						
Simulation number	Data	1	2	3	4	5
1	Original cluster 1	0.0508 ± 0.0159	0.0503 ± 0.015	0.0513 ± 0.0151	0.0502 ± 0.0153	0.0497 ± 0.0152
	Original cluster 2	0.0509 ± 0.0154	0.0504 ± 0.0163	0.0491 ± 0.0152	0.0504 ± 0.0156	0.0501 ± 0.0157
	Combined cluster	0.0525 ± 0.0172	0.0584 ± 0.0219	0.0667 ± 0.0248	0.0735 ± 0.0275	0.0803 ± 0.0306
	Estimated cluster 1	0.0505 ± 0.0156	0.0509 ± 0.0156	0.0505 ± 0.0157	0.0516 ± 0.0156	0.051 ± 0.0157
	Estimated cluster 2	0.0508 ± 0.0154	0.0517 ± 0.0163	0.0507 ± 0.0158	0.052 ± 0.0156	0.051 ± 0.0155
	MH	0.0498 ± 0.016	0.0505 ± 0.0159	0.0516 ± 0.0154	0.0518 ± 0.0161	0.0503 ± 0.0158
2	Original cluster 1	0.0503 ± 0.0157	0.0503 ± 0.0153	0.05 ± 0.0152	0.0502 ± 0.0157	0.0509 ± 0.0152
	Original cluster 2	0.0504 ± 0.0152	0.0517 ± 0.0163	0.0494 ± 0.0155	0.0503 ± 0.0158	0.0499 ± 0.015
	Combined cluster	0.051 ± 0.0154	0.0517 ± 0.0161	0.052 ± 0.0161	0.0537 ± 0.0164	0.0545 ± 0.0168
	Estimated cluster 1	0.051 ± 0.0156	0.0517 ± 0.0163	0.0512 ± 0.0153	0.0512 ± 0.0154	0.0522 ± 0.0165
	Estimated cluster 2	0.0501 ± 0.0153	0.0514 ± 0.0157	0.0523 ± 0.0162	0.0519 ± 0.016	0.0518 ± 0.0164
	MH	0.0494 ± 0.0162	0.0496 ± 0.0167	0.05 ± 0.0166	0.0512 ± 0.0167	0.0515 ± 0.0173
3	Original cluster 1	0.054 ± 0.0158	0.0533 ± 0.0158	0.0533 ± 0.0163	0.0529 ± 0.0164	0.0537 ± 0.0158
	Original cluster 2	0.0542 ± 0.0163	0.0531 ± 0.0157	0.054 ± 0.0167	0.0539 ± 0.0157	0.0528 ± 0.0157
	Combined cluster	0.053 ± 0.0156	0.0546 ± 0.0165	0.0553 ± 0.0156	0.0557 ± 0.016	0.0569 ± 0.0166
	Estimated cluster 1	0.0507 ± 0.0159	0.0509 ± 0.0153	0.0517 ± 0.0156	0.0522 ± 0.0158	0.052 ± 0.0154
	Estimated cluster 2	0.0509 ± 0.0163	0.051 ± 0.0156	0.0504 ± 0.0158	0.0511 ± 0.0154	0.0524 ± 0.016
	MH	0.0695 ± 0.124	0.0653 ± 0.1145	0.0696 ± 0.0119	0.0735 ± 0.1284	0.0671 ± 0.1182
4	Original cluster 1	0.0503 ± 0.0157	0.0506 ± 0.0152	0.05 ± 0.0154	0.0528 ± 0.0152	0.0503 ± 0.0158
	Original cluster 2	0.0496 ± 0.0156	0.0506 ± 0.0157	0.0499 ± 0.0157	0.0505 ± 0.0154	0.0501 ± 0.0153
	Combined cluster	0.0538 ± 0.0168	0.0541 ± 0.0186	0.0582 ± 0.0204	0.0602 ± 0.023	0.0634 ± 0.0234
	Estimated cluster 1	0.051 ± 0.0159	0.0505 ± 0.0148	0.0507 ± 0.0156	0.05 ± 0.0153	0.0505 ± 0.0158
	Estimated cluster 2	0.0515 ± 0.0156	0.0507 ± 0.0152	0.0511 ± 0.0155	0.0509 ± 0.0152	0.0509 ± 0.0152
	MH	0.0515 ± 0.0146	0.0506 ± 0.0157	0.0504 ± 0.0154	0.0506 ± 0.0158	0.0508 ± 0.0157
5	Original cluster 1	0.0496 ± 0.0149	0.0506 ± 0.0154	0.0501 ± 0.015	0.0488 ± 0.0155	0.0502 ± 0.0152
	Original cluster 2	0.0502 ± 0.0159	0.0496 ± 0.016	0.0497 ± 0.0147	0.0499 ± 0.0152	0.0507 ± 0.0146
	Combined cluster	0.0525 ± 0.0163	0.0672 ± 0.0256	0.0869 ± 0.0346	0.1021 ± 0.0365	0.1134 ± 0.0407
	Estimated cluster 1	0.0498 ± 0.015	0.0509 ± 0.0154	0.051 ± 0.0154	0.0501 ± 0.0159	0.0508 ± 0.0158
	Estimated cluster 2	0.0508 ± 0.0155	0.0509 ± 0.0164	0.0504 ± 0.0158	0.0507 ± 0.0159	0.0517 ± 0.0158
	MH	0.0501 ± 0.0153	0.0501 ± 0.0161	0.0503 ± 0.0158	0.0501 ± 0.0159	0.0512 ± 0.0158

consumes much time for the calculation. Therefore, we used *K*-fold CV. We created simulations under two conditions. The conditions for generating sample data are shown in Table 3. In these simulations, 100 structured samples were generated at random. We calculated CV for each sample for the conditions $\hat{K} = 2, 3, 4$. We counted the number with the minimum CV in each case. The results are shown in Table 4. The frequency of the exact estimation was 0.91 when the true number of subpopulations was 2. The frequency of the exact estimation was 0.73 when the true number of subpopulations was 3. These results indicated that CV is a good indicator to estimate the validity of the results.

Comparison between the distance-based and model-based clustering methods

We compared the results from the analysis by two separate programs, POPSTRUCT with hierarchical clustering, and STRUCTURE, both of which analyze the data from the structured populations. STRUCTURE uses the Markov chain Monte Carlo algorithm to obtain the proportions of alleles that belong to some original subpopulations. STRUCTURE can be used in both admixture and no admixture models. In contrast, a subject can belong to a single subpopulation in POPSTRUCT. These methods are based on different approaches. Hierarchical clustering is a distance-based model. In contrast, STRUCTURE uses a model-based approach. We compared the two algorithms by the bootstrap method.

Two groups of data each consisting of the genotype data at 200 unlinked loci for 1,000 individuals were generated using the conditions for the two subpopulations. Then, n_1 and n_2 subjects were randomly selected as two subpopulations from each of the groups by the bootstrap sampling method. Thereafter, the two subpopulations were combined, followed by the analysis of the combined data either by POPSTRUCT or by STRUCTURE. After the combined population was classified into two clusters, the misclassified subjects were counted. The above steps were repeated 1,000 times.

We tested two sets of conditions: $n_1 = n_2 = 100$ and $n_1 = 196$, $n_2 = 4$. No admixture model was used in STRUCTURE. When POPSTRUCT was used, clusters were joined until the number of the clusters became two.

In the first set of conditions, i.e., $n_1 = n_2 = 100$, the misclassification rate was 0.0002 ± 0.0009 by POPSTRUCT whereas there was no misclassification by STRUCTURE. These results indicate that both algorithms are equally satisfactory as accurate clustering machineries.

When the second set of conditions was used, however, the misclassification rate by POPSTRUCT was 0.0019 ± 0.027 whereas that by STRUCTURE was

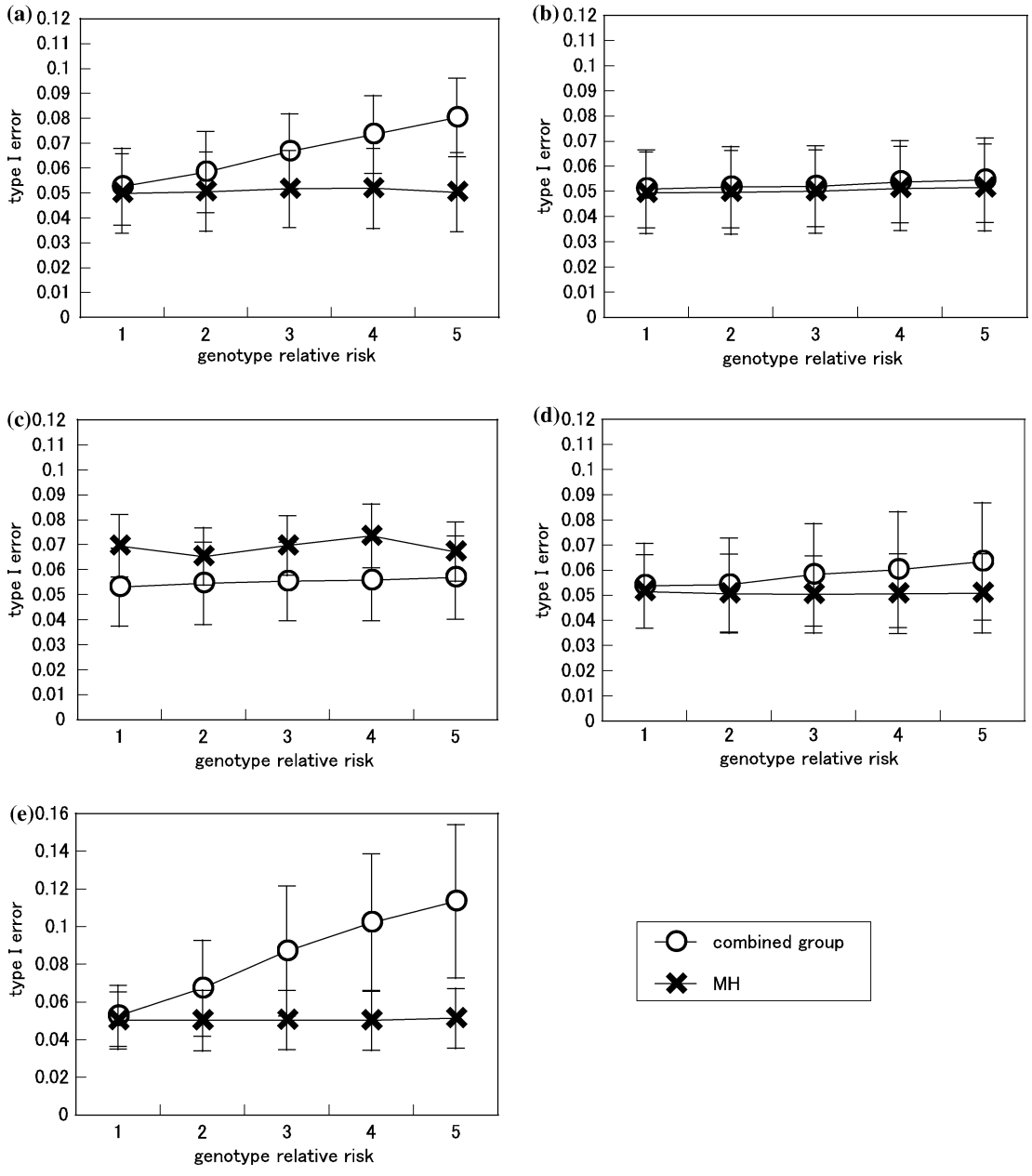

Fig. 1 Mean empirical type I error rates with standard deviations (SDs) calculated from simulations at five different genotype relative risks. After the simulation under the conditions in simulations **1** (**a**), **2** (**b**), **3** (**c**), **4** (**d**), or **5** (**e**) described in Table 1, the empirical type I error rates were calculated by the five different methods. The data in the combined group were used. The Mantel–Haenszel (MH) test was performed on the data from the two clusters. As the genotype relative risk increased, the type I error for combined data increased; the type I error in the MH test did not increase in **a**, **d**, or **e**. In these three conditions, misclassification rates were low. When the clustering does not work, as in simulations 2 and 3, the algorithm does not work

0.22 ± 0.20, which suggests that POPSTRUCT outperforms STRUCTURE under these conditions.

When the size of one of the subpopulations is very small, the estimation of the allele frequencies would be inaccurate. Therefore, the allele frequencies estimated in STRUCTURE would be quite unreliable, and this is likely to be the reason why STRUCTURE could not efficiently infer the subpopulations that the subjects belong to. The same kind of problem is likely to occur when the size of a subpopulation is very small. Our results suggest that both of the methods should be used, i.e., the distance-based and model-based methods for clustering the subjects with genotypic data.

Discussion

In the present study, we attempted to construct an algorithm to reduce the problems of the inflation of the type I errors in the association studies using structured populations. Although there are two separate approaches to this problem—the fuzzy clustering method and the optimization, or hierarchical, method—we used

Table 3 Parameters for simulations of estimating K

	True number of subpopulations (K)	Number of individuals	Number of loci	Range of allele frequencies
Simulation 1	2	120 (60:60)	100	0.2–0.8
Simulation 2	3	120 (40:40:40)	100	0.2–0.8

the latter method, since Pritchard et al. (2000a) have published an algorithm based on the former method.

Thus, using the clustering method, we constructed an algorithm to classify the subjects in a combined population with structure into clusters, each of which has little structure, and tested the association between a disease locus and a candidate marker using the genotype data. The test of the association was performed by the MH test using data from both of the clusters. We implemented the algorithm in the computer program POPSTRUCT.

Using the program in which the simulation procedure for generating the structured population was also implemented, we found that the combination of the two separate groups with different marker allele frequencies causes the increases of type I errors. This increase was more marked when the differences in the marker allele frequencies were larger and when the difference in the minor allele frequencies at the disease locus was larger. In addition, the increase became larger when the genotype relative risk associated with the disease locus was higher.

Then, we calculated the empirical power (the proportion of the simulations that generated the data with positive association between the disease and the

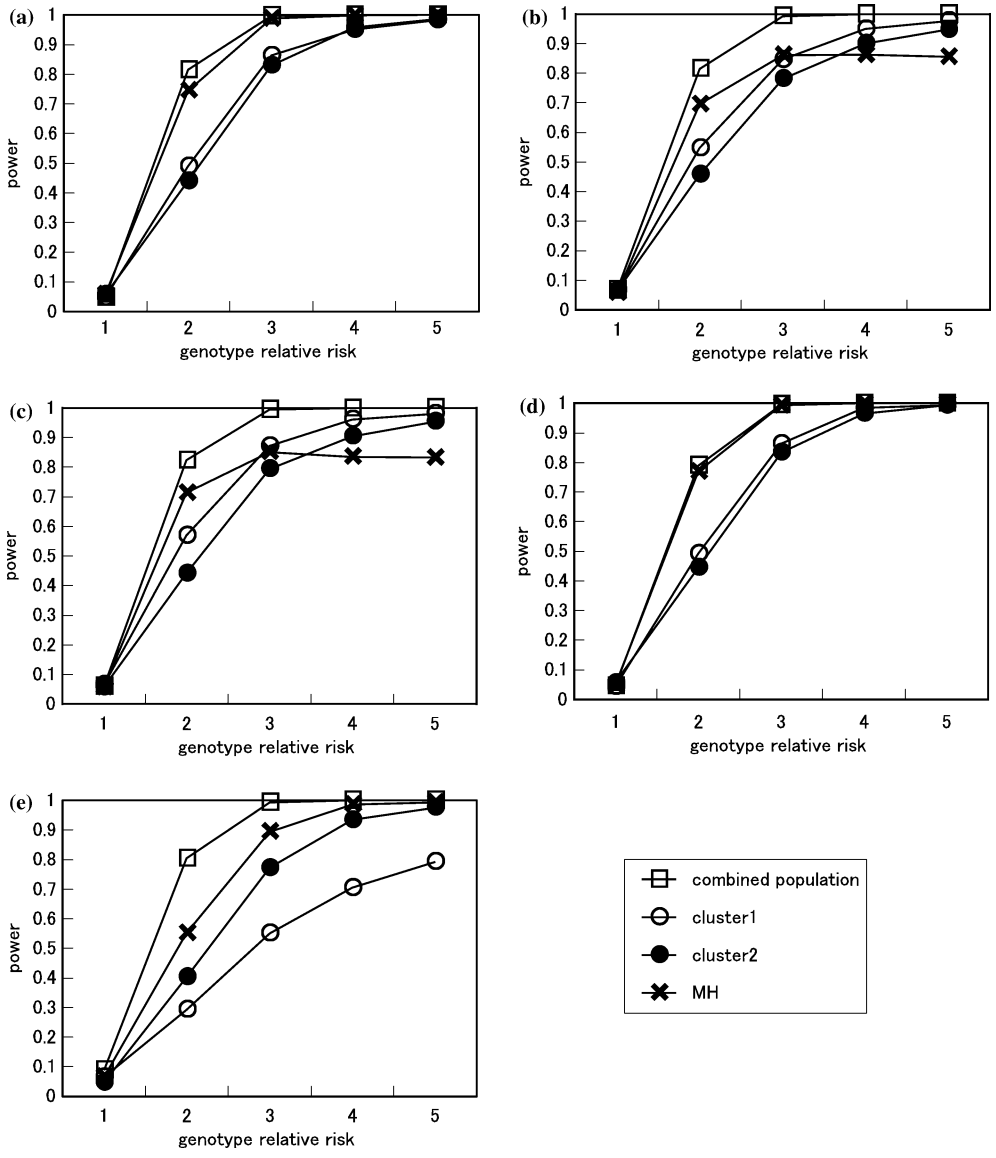

Fig. 2 Empirical powers calculated from five different simulations under the conditions described in Table 1 at five different genotype relative risks. After simulations under the conditions in simulation **1** (**a**), **2** (**b**), **3** (**c**), **4** (**d**), or **5** (**e**) described in Table 1, the empirical powers were calculated by five different methods. The power for each cluster was significantly lower than the power for the combined population. However, when the data from two separate clusters were analyzed together by the Mantel–Haenszel (MH) test, the power was comparable to that for the combined population. When the clustering does not work, as in simulations 2 and 3, the algorithm does not work

Table 4 Frequencies with the lowest cross-valuation (CV). The numbers in bold type are the highest

Estimate of K	Simulation 1	Simulation 2
2	**91**	14
3	9	**73**
4	0	13

minor allele at the disease locus) for the original two subpopulations before the admixture, the mixed population, two clusters obtained after the clustering procedure, and by the MH test. In most of the conditions tested, the MH test showed slightly lower than that for the combined population but higher than that for any test performed on each original subgroup and each cluster. However, when the differences in the allele frequencies between the two original populations were very small, the power for the MH test decreased, even if the genotype relative risk was high. In such cases, the misclassification rate was very high, which indicates that this test is not suitable when the original population is not very structured.

Then, we compared the results of the analyses by both model-based and distance-based methods using the same data. When two subgroups with different allele frequencies were mixed together at a fixed ratio of 1:1, both of the programs classified the subjects into clusters that each reflected one of the original subpopulations. The rate of the misclassification was small, although it differed according to the conditions used. When the ratio of the mixture deviated (i.e., 1:49), hierarchical clustering was superior to STRUCTURE in classifying the combined population into accurate clusters that each reflected one of the original groups.

The algorithm implemented in STRUCTURE is based on the genetic model. The model can be extended for the admixtured population. However, it is difficult to define the distance that can be applied for the admixtured population. If individuals in the sample were admixtures, STRUCTURE was superior to the distance-based model (Pritchard et al. 2000a). Because of this benefit, however, it is not suited to the situation where a small number of subjects from a subgroup is mixed with a large number of those from another. Such a case can be present in real situations. Therefore, both approaches, model-based and distance-based, should be used in real data analysis.

We proposed that the mixture of different subgroups can cause significant inflation of type I errors in association studies and have examined the parameters that play major roles in such increases. We then constructed an algorithm to overcome the structuring of the population and performed association studies in the presence of the structure. In the next step, we should apply the present algorithm to real data and test whether it is useful to reduce the increase in the type I errors in association studies using genetic information.

References

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Cochran WG (1954) Some methods for strengthening the common χ^2 test. Biometrics 10:417–451

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Everitt BS (1993) Cluster analysis, 3rd edn. Edward Arnold, New York

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, Mckeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

Mantel N, Haenszel W (1959) Statistical aspect of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Donelly P (2000a) Inference of population structure using multilocus genotype data. Am J Hum Genet 67:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donelly P (2000b) Association mapping in structured population. Am J Hum Genet 67:170–181

Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using novel latent-class model. Am J Hum Genet 68:466–477

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–513

Stone M (1974) Cross-validation and multinomial prediction. Biometrika 51:509–515

Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J R Stat Soc B 39:44–47