BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 4 2007, pages 408–413
doi:10.1093/bioinformatics/btl133

Genome analysis

Accurate automated clustering of two-dimensional data for single-nucleotide polymorphism genotyping by a combination of clustering methods: evaluation by large-scale real data

Shuichi Takitoh[1,4], Shogo Fujii[1,4], Yoichi Mase[1,4], Junichi Takasaki[1], Toshimasa Yamazaki[1], Yozo Ohnishi[2,5], Masao Yanagisawa[4], Yusuke Nakamura[3,5] and Naoyuki Kamatani[1,6,*]

[1]Laboratory for Statistical Analysis, [2]Laboratory of SNP Analysis and [3]Laboratory of Pharmacogenetics, RIKEN SNP Research Center, [4]Department of Computer Science, Waseda University, Shinjuku, Tokyo, Japan, [5]Institute of Medical Science, University of Tokyo, Tokyo, Japan and [6]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan

Received on September 20, 2005; revised on March 30, 2006; accepted on March 31, 2006

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The Invader assay is a fluorescence-based high-throughput genotyping technology. If the output data from the Invader assay were classified automatically, then genotypes for individuals would be determined efficiently. However, existing classification methods do not necessarily yield results with the same accuracy as can be achieved by technicians. Our clustering algorithm, Genocluster, is intended to increase the proportion of data points that need not be manually corrected by technicians.

Results: Genocluster worked well even when the number of clusters was unknown in advance and when there were only a few points in a cluster. The use of Genocluster enabled us to achieve an acceptance rate (proportion of assay results that did not need to be corrected by expert technicians) of 84.4% and a proportion of uncorrected points of 95.8%, as determined using the data from over 31 million points.

Availability: Information for obtaining the executable code, example data and example analysis are available at http://www.genstat.net/genocluster

Contact: kamatani@ior.twmu.ac.jp

INTRODUCTION

An enormous amount of data has emerged from automated high-throughput single-nucleotide polymorphism (SNP) genotype analysis technologies. Invader assay (Ryan *et al*., 1999, Ohnishi *et al*., 2001) is one such genotyping technology in which the invader oligonucleotide, Cleavase enzyme and fluorescence-labeled probes are used. One of two fluorescence signals is released when an allele is present in the DNA sample. The data of the intensities of two fluorescence signals corresponding to the two alleles are used to determine the genotypes of the subjects from whom the DNA samples have been obtained. Since data for multiple subjects are analyzed simultaneously for the determination of the genotypes, accurate clustering of the samples into three genotype categories is important. Thus, accurate clustering of the two-dimensional intensity data into genotype categories is of central importance in such a genotyping system. In other fluorescence-based systems such as the Taqman assay, the same clustering problem exists, and the data presented herein are likely to be useful for such genotyping technologies as well.

Although a number of automated clustering systems are available, such systems have not yet been perfected, and so manual clustering by expert technicians remains the most accurate method. Based on large-scale real data, we developed a new algorithm for clustering the two-dimensional fluorescence-intensity data obtained from the Invader assay.

A number of clustering technologies are currently in use. For example, the *K*-means algorithm is the method that is most widely used for genotype clustering (e.g. Oliver *et al*., 2002; Ranade *et al*., 2001). This method, however, requires pre-specification of the number of clusters. Since the number of different genotypes for an SNP locus in a sample is often unknown, this algorithm might cause a serious problem for clustering the samples into genotypes.

Recently, model-based clustering methods have been applied to the genotype-clustering procedure (Chen and Kalbfleisch, 1996, 2001; Riva and Kohane, 2002; Yeung *et al*., 2001). Fujisawa *et al*. (2004) introduced a penalized likelihood method, and Kang *et al*. (2004) used a t-mixture model for clustering. Nevertheless, problems in clustering genotype data have not been perfected, even with the introduction of such technologies. Moreover, Fujisawa *et al*. (2004) have suggested that conventional maximum likelihood methods might be better than model-based methods if the number of clusters is known.

In the present manuscript, we show that an algorithm created by integrating more than one method [nearest neighbor method and Markov-chain Monte Carlo (MCMC) method] combined with optimization using large-scale real data performs extremely well in the clustering of real two-dimensional genotype data. To handle a large size of data in a high-throughput system, short calculation time is preferred. A model-based method is considered to be useful to perform the clustering without defining the number of clusters beforehand. MCMC method tends to obtain accurate results after a limited number of iterations if the initial clusters are close to the

*To whom correspondence should be addressed.

final results (Gamerman, 1997). Nearest neighbor method tends to make clusters with chain-like appearance (Anderberg, 1973). As observed in our manuscript, the data generated by the Invader assay tend to have the chain-like appearance. This is why we combined the hierarchical clustering method nearest neighbor and the model-based method MCMC. We implemented this algorithm as a computer software application called Genocluster. Using this software, we performed an extensive analysis to determine the performance of this algorithm.

METHODS

Structure of data obtained from a multiple-PCR based Invader assay using a 384-well card system

An output from the Invader assay for a subject is expressed as a list of two real values, each of which indicates the intensity of the fluorescence signal corresponding to an allele at a single SNP locus. Throughout the present study, we used the real data obtained from the card system based on the Invader assay (Ohnishi *et al*., 2001) developed by Ohnishi and Nakamura. The present card system is a very efficient genotyping system in which a 384-well card is used. Based on such a high-throughput system, the disease-related genes for myocardial infarction (Ozaki *et al*., 2002, 2004), rheumatoid arthritis (Suzuki *et al*., 2003; Tokuhiro *et al*., 2003), diabetes mellitus (Kanazawa *et al*., 2004) and osteoarthritis (Kizawa *et al*., 2005) have been discovered. A set of two-dimensional data for ~333 subjects is obtained from one card. Using ~333 points on a two-dimensional plane, each corresponding to a pair of real values for a genotype, we develop an algorithm to cluster the points into one to four clusters: two clusters for two different homozygotes, one cluster for heterozygotes, and one cluster for undetermined genotypes. Figure 2a shows an example of the output from the Genocluster algorithm. The unclassified two-dimensional data represented by the points are classified into the following categories: allele 1 (1), allele 2 (3), both alleles or heterozygote (2), undetermined (4) and NTC (0). These points are usually clustered manually by expert technicians, and it has been shown that such manual clustering works quite well (Ohnishi *et al*., 2001).

Adjustment of the raw data

Let x_i and y_i denote, respectively, the fluorescence intensities obtained for the two alleles for the i-th subject at an SNP. Thus, a point (x_i, y_i) on a two-dimensional plane represents the data for the i-th subject. A set of data is composed of ~333 points for different subjects. Let n denote the number of points for different subjects in a set of data.

The following adjusted values were calculated using the raw data for all i, $i = 1, 2, \ldots, n$:

$$x'_i = \frac{x_i - x_0}{x_{\max} - x_0}, \tag{1}$$

$$y'_i = \frac{y_i - y_0}{y_{\max} - y_0}, \tag{2}$$

where $x_{\max}$ and $y_{\max}$ denote the maximum values of the data for the x and y axes, respectively, while (x_0, y_0) denotes the starting point obtained from the average of eight points for the non-template controls (NTCs). Thus, the origin was determined by the data for the wells to which an NTC was added. The points for which the values of x'_i and y'_i value were either ≤ 0 or were within an ellipsoid with the center at the origin, were classified as 'undetermined'. The ellipsoid was defined as

$$\frac{x'^2}{a^2} + \frac{y'^2}{b^2} = 1, \tag{3}$$

where a and b were 0.08 and 0.06, respectively.

When the number of clusters was judged to be one or was unable to be estimated based on the distribution of all of the points, then all of the clusters were classified as 'undetermined' or as being of one genotype.

The estimation of the number of clusters was performed as follows. The two-dimensional plane was scanned by an origin-centered sector with a 5° interior angle, using lower-angle shifting in increments of one degree from −10 to 100°, and the number of points within the vector was counted at each degree. A spline interpolation of the data obtained by plotting the number of points against the angle specifies the density distribution as a function of θ_i, where $\theta_i = \tan^{-1}(y'_i/x'_i)$. The number of maximums of this function corresponds to the number of clusters.

Nearest neighbor method

To the adjusted two-dimensional data, the nearest-neighbor method was applied to broadly classify the points. Thus, excluding the points classified as 'undetermined', all of the points were submitted to the following procedures for the clustering. In the following description, a node refers to a set of points generated as a result of clustering by the nearest neighbor method. Note that a node may have only one point as a member. A cluster, on the other hand, refers to a node obtained as a final result of the clustering by the nearest neighbor method.

The procedures used for the nearest neighbor method were as follows:

(1) Each point was considered as an individual node, and the two nodes with the lowest unsimilarity were combined to construct a single node. The unsimilarity between node p and node q is given by

$$\mathrm{unsim}(p, q) = \min_{i \in N_p, j \in N_q} |\theta_i - \theta_j|, \tag{4}$$

where N_p and N_q denote the sets of points that are the members of nodes p and q, and θ_i and θ_j denote the angles of points i and j, respectively.

(2) Step 1 was repeated until the number of nodes became six.

(3) The nodes consisting of single points were interpreted as 'undetermined' points.

(4) When more than four nodes each consisted of more than two points, the following procedures were required. The average angle $\hat{\theta}$ was calculated for each node. The average angle of node d is

$$\theta_d = \frac{1}{n_d} \sum_{i \in N_d} \theta_i, \tag{5}$$

where n_d denotes the number of points in node d and N_d denotes the set of points that are the members of node d. The angular distance between any two nodes, as defined by the absolute difference of the averages between the two nodes, was calculated. The two nodes with the smallest angular distance were combined.

(5) Step 4 was repeated until the number of nodes became three. Through the above procedures, the number of nodes should have become 1, 2 or 3.

Treatment after the application of the nearest neighbor method

After the above-described nearest neighbor method was performed, the data were treated as follows:

(1) Pairs of nodes with angular distances below a threshold were combined. The importance of the angular distances was shown in Mein *et al*. (2000). The threshold was set at 0.15 rad.

We also defined $\theta_{d\_\max}$, the largest θ_i in node d and $\theta_{d\_\min}$, the smallest θ_i in node d. If three nodes remained, they were tentatively assigned as $d1$, $d2$ and $d3$, so that the average angles of the nodes became $\theta_{d1} < \theta_{d2} < \theta_{d3}$. Then, nodes $d1$, $d2$ and $d3$ were considered to correspond to 'allele1', 'both (or heterozygote)' and 'allele2', respectively. If $\theta_{d3\_\min} - \theta_{d2\_\max} < 0.15$ and $\theta_{d2\_\min} - \theta_{d1\_\max} < 0.15$, then none of the nodes were judged to be combined and the number of clusters was judged not to be determined. If two nodes remained, then the number of clusters was judged not to be determined when

$\theta_{d3\_min} - \theta_{d2\_max} < 0.15$ or $\theta_{d2\_min} - \theta_{d1\_max} < 0.15$. When only one node remained, the number of clusters was judged to be one if $\theta_{d1\_max} - \theta_{d1\_min} < 0.30$ or $\theta_{d3\_max} - \theta_{d3\_min} < 0.30$, otherwise it was judged not to be determined.

(2) Each node was assigned to a certain genotype as follows: if $0 \leq \theta_d < 0.65$, then node d was assigned to 'allele 1'; if $0.65 \leq \theta_d < 1.25$, then node d was assigned to 'both'; if $1.25 \leq \theta_d < 1.85$, then node d was assigned to 'allele 2'; otherwise, node d was assigned to 'undetermined'.

(3) Genocluster interprets the data not to be analyzed due to the major departure from HWE. In case of three clusters, all the points are judged to be undetermined when $N_{d1} > 2 \times N_{d2}$ and $N_{d3} > 2 \times N_{d2}$ since this indicates a major departure from HWE, where N_d denotes the number of points in node d. In case of two clusters, all the points are judged to be undetermined when there are two different homozygotes in the absence of heterozygotes. In case of one cluster, all the points are judged to be undetermined when there are only heterozygotes.

Excluding the points classified as 'undetermined' in adjustment of the raw data, all of the points were submitted to the following steps of the MCMC procedure.

The MCMC method

In the following clustering process, in which the MCMC method was used, the points in each cluster were assumed to follow a two-variate normal distribution with the following probability density function:

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}, \quad (6)$$

where μ_X and μ_Y denote the means for the x and y axes, respectively, σ_X and σ_Y denote the standard deviations for the x and y axes, respectively, and ρ denotes the correlation coefficient.

Clustering by the Gibbs sampler was performed as follows:

(1) The following calculation was iterated a sufficient number of times (e.g. 100 000):

(a) Let C_j denote the set of indices of the points that belong to the j-th cluster. Given the observed points (x_i, y_i), $i \in C_j$ belonging to a single cluster j, the maximum likelihood estimates of the means, the standard deviations, and the correlation coefficient were calculated as follows:

$$\hat{\mu_X} = \frac{1}{n_j}\sum_{i\in C_j} x_i, \quad (7)$$

$$\hat{\mu_Y} = \frac{1}{n_j}\sum_{i\in C_j} y_i, \quad (8)$$

$$\hat{\sigma_X} = \sqrt{1/n_j \sum_{i\in C_j}(x_i - \hat{\mu_X})^2}, \quad (9)$$

$$\hat{\sigma_Y} = \sqrt{1/n_j \sum_{i\in C_j}(y_i - \hat{\mu_Y})^2}, \quad (10)$$

$$\hat{\rho} = \frac{\sum_{i\in C_j}(x_i - \hat{\mu_X})(y_i - \hat{\mu_Y})}{\sqrt{\sum_{i\in C_j}(x_i - \hat{\mu_X})^2}\sqrt{\sum_{i\in C_j}(y_i - \hat{\mu_Y})^2}}, \quad (11)$$

where n_j denotes the number of points in the j-th cluster.

(b) All points other than those classified as 'undetermined' were submitted to the following procedure based on the Gibbs sampler. After the parameters were estimated, the probability density function for each cluster was updated by substituting the estimates for the parameters in Equation (6). The state space was the set of all possible clustering states in which all of the points were involved. For the transition, one of the points was selected as a candidate for a point that might change the cluster to which it belongs. In Genocluster, a candidate point was selected randomly from the points classified as 'undetermined' in the nearest neighbor method step.

(c) Let j denote the cluster number to which the selected point belongs. The candidate point was moved to the cluster index $j*$ with a probability of

$$\alpha_{j*} = \frac{f_{j*}(x_i, y_i)}{\sum_{k=1}^{3} f_k(x_i, y_i) + t_0}, \quad (12)$$

where i denotes the point number for the point selected in Step b, f_k denotes the probability density function for the k-th cluster updated in Step b, and t_0 denotes a small fixed real value.

(2) After steps a–c of the Gibbs sampler were repeated a sufficient number of times, the values obtained from not all, but rather from a number of the steps of the MCMC method were used for the statistical calculation. Thus, the values obtained from the steps in the burn-in phase (1000 steps) were abandoned, and the values obtained from every 1000 steps thereafter were used.

For the i-th point, the proportion of the steps in which that point was classified into the j-th cluster was calculated as

$$p_{ij} = \frac{M_j(i)}{\sum_{k=1}^{4} M_k(i)}, \quad (13)$$

where $M_k(i)$ denotes the number of steps in which the i-th point was classified into the k-th cluster. In this case, the cluster for 'undetermined' points was considered to be the fourth cluster. If $p_{ij} > 0.9$ for any j, then the i-th point was determined to be classified into the j-th cluster; otherwise, the i-th point was classified into the 'undetermined' cluster.

RESULTS

The above algorithm, implemented in the computer software Genocluster, was applied to the data from the Invader assay card system. Sets of data obtained in this system were composed of the two-dimensional data from ~333 subjects concerning a single SNP locus. Each set of data was applied to Genocluster and the results were expressed as a two-dimensional graph in which the points of different clusters were expressed using different colors.

We compared the accuracy of the clustering between the three clustering methods, i.e. the nearest neighbor method alone, the MCMC method, and the Genocluster method (the combination of the nearest neighbor method and the MCMC method) (Fig. 1). The results of the clustering were validated by the results obtained by expert technicians. As shown in Figure 1, the results differed between different clustering methods.

Figure 1a shows an example of clustering using the nearest neighbor method in Genocluster. However, each one of the plots was judged to be undetermined although it should be classified in each cluster (Allele1 or Allele2). Figure 1b shows the results of the classification by MCMC method different from the MCMC method used in Genocluster. In the method used in Figure 1b, all the plots are the candidates of the transition, and three plots that should be classified as Allele1 were judged as undetermined. In Figure 1c,

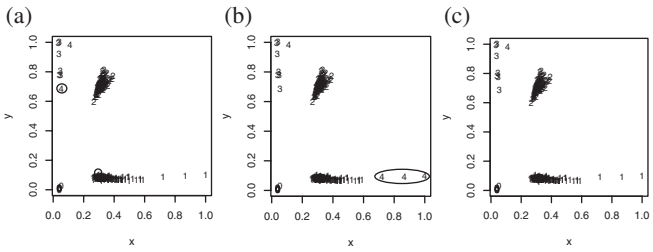

Fig. 1. Outputs from three different clustering methods. Each panel shows the result of the clustering for a set of data (data for a single SNP from 333 subjects) using one of the three clustering methods (nearest neighbor method alone, the MCMC method, or Genocluster). As described in the METHODS Section, 333 points indicating the intensities of two different alleles at an SNP were plotted on a two-dimensional plane. Using one of the clustering methods, the points were classified into allele 1 (1), allele 2 (3), both or heterozygote (2) and undetermined (4) categories. (0 indicates NTC.) Panels (**a–c**) show the results for the same set of data obtained using different clustering methods. The clustering methods used were the nearest neighbor method alone (a), the MCMC method alone (b), and Genocluster (c).

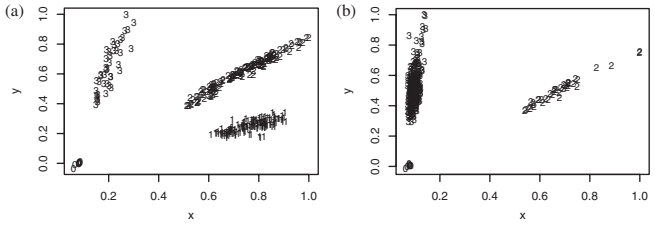

Fig. 2. Outputs from Genocluster. The panels show the results obtained using different sets of data. Each panel indicates the results of the clustering by Genocluster. Typical outputs from Genocluster, one for three clusters (**a**) and the other for two clusters (**b**), are shown.

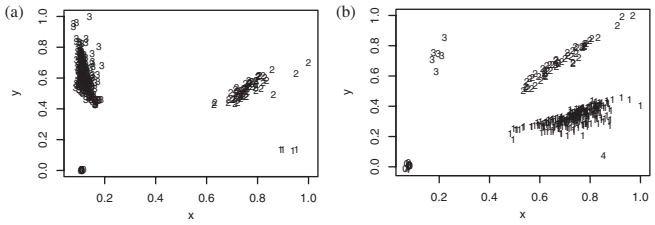

Fig. 3. Outputs from Genocluster. Both of the panels indicate the presence of three clusters. In panel (**a**), there are only a few minor homozygotes. In panel (**b**), there are a few outliers.

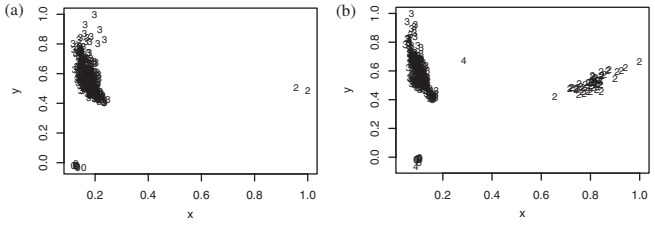

Fig. 4. Outputs from Genocluster. Both the panels indicate the presence of two clusters. In panel (**a**), there are only a few heterozygotes. In panel (**b**), there are a few outliers.

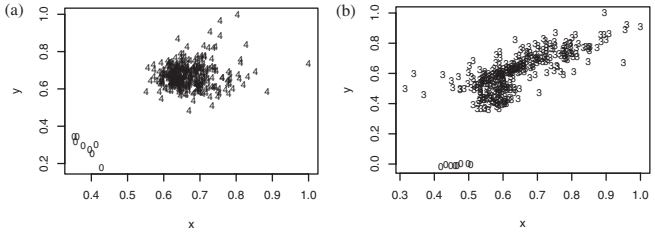

Fig. 5. Outputs from Genocluster. In panel (**a**), all points were judged to be 'undetermined'. In panel (**b**), all points were judged to be members of a cluster with a homozygote genotype.

only the plots that had been judged as undetermined in Figure 1a were the candidates of the transition. According to this strategy, outliers are handled as outliers and the plots that should be classified in the clusters are classified correctly.

We then applied these three methods to 960 sets of data for different SNPs. We counted the number of uncorrected points for each set of data for each method. When the 960 sets of data were applied to Genocluster, the number of uncorrected (by expert technicians) points was not more than 3 in 834 sets. This number is larger than that for the nearest neighbor method (791) and the MCMC method (763), indicating that Genocluster performs better than the other methods.

Excluding the 'undetermined' cluster, three clusters appear in Figure 2a, whereas in Figure 2b, only two clusters appear. Thus, the number of clusters was different between Figure 2a and b. When three clusters other than the cluster of 'undetermined' points appears, the genotype of each cluster is easily determined, as indicated by different colors in Figure 2a. The median of the heterozygote cluster should have an intermediate angle between those of the minor-allele homozygotes and the major-allele homozygotes. However, when only two clusters appear, as in Figure 2b, the genotype of each cluster must be determined carefully. Since the presence of two different homozygotes in the absence of heterozygotes in the sample does not show HWE, the two clusters should be the major-homozygote cluster and the heterozygote cluster. Thus, knowledge of the HWE is important for the assignment of the clusters to the genotypes. Which of the two clusters is the heterozygote cluster was determined by the angle information.

Even if there are three clusters, the number of members in a cluster might be quite low. Thus, in both Figure 3a and b, one of the clusters had only a small number of members, accurate clustering in such cases is more difficult than in cases like that shown in Figure 2a. In addition, Figure 3b includes two 'undetermined' points. Figure 4a shows two clusters, one of which has only two heterozygote points, whereas Figure 4b shows one cluster for homozygotes and one cluster for heterozygotes. In addition, two 'undetermined' points appear in Figure 4b.

In Figure 5a, all of the points were judged to be 'undetermined' points, whereas in Figure 5b, all of the points were judged to belong to a single homozygote cluster.

The presence of outliers was one of the greatest problems encountered during the clustering. In Figures 3b and 4b, some points were judged to belong to 'undetermined' because nodes with outliers had single members and had not merged with other nodes by the nearest neighbor method, and the degrees of membership to any of the

clusters given by the MCMC method were low. Although the presence of clusters with small numbers of members is a problem, Genocluster could correctly classify such points, as shown in Figures 3a and 4a. This is because, in the transition phase of the MCMC method, a candidate point was selected from the points that were classified as 'undetermined' in the nearest neighbor method step. Therefore, even if there were only a few points in a cluster, the parameters of the distribution could be reliably estimated.

In Figure 5a, all the points were judged to be 'undetermined' points because the number of clusters was either judged to be one or was unable to be estimated based on the distribution of all of the points and $\theta_{d\_\max} - \theta_{d\_\min} \geq 0.30$, whereas in Figure 5b, all of the points were judged to belong to a single homozygote cluster because $\theta_{d\_\max} - \theta_{d\_\min} < 0.30$ and $1.25 \leq \theta_d < 1.85$. Even if the number of clusters was not judged to be one or was unable to be estimated based on the distribution of all of the points, the following rule was applied. When three nodes remained, and $\theta_{d3\_\min} - \theta_{d2\_\max} < 0.15$ and $\theta_{d2\_\min} - \theta_{d1\_\max} < 0.15$, all of the nodes were combined into one node. Moreover, although the shape of the cluster in Figure 5b is somewhat similar to that of a heterozygote cluster, the presence of only heterozygotes without any homozygotes does not accord with HWE.

The points that were judged to be 'undetermined' could in fact belong to specific clusters; however, the clusters of some points may remain 'undetermined' even if the classification is performed using the best clustering algorithm or by an expert technician. Rather than mis-classifying such points, the algorithm should leave such points as 'undetermined' and recommend that the samples should be submitted to re-genotyping.

DISCUSSION

The laboratory of Ohnishi and Nakamura has genotyped thousands of SNPs using the Invader-assay card system (Ohnishi *et al*., 2001; Ozaki *et al*., 2002, 2004; Suzuki *et al*., 2003; Tokuhiro *et al*., 2003; Kanazawa *et al*., 2004; Kizawa *et al*., 2005). In all of our experiments, the classification of the points obtained from the Invader-assay card system was performed by expert technicians. This is because genotyping by expert technicians has outperformed the results obtained by previous automated clustering systems (e.g. Auto Caller; Applied Biosystems Inc.).

However, according to our experience, the time for the clustering of the points for a SNP using the data from the Invader Assay should be shorter than 6 s, so that a large-scale genotype system proceeds without delay.

Therefore, an automated clustering system is required to deal with various complex problems so as to provide accurate clustering results. To achieve this goal, careful optimization of the treatments before the data were submitted to the Gibbs sampler was important. Accurately estimating the number of clusters was especially important. If the number of estimated clusters was the same as the true number of clusters, then the clustering algorithm performed quite well. Although the angular information was useful for estimating the number of clusters in the treatment before the Gibbs sampler step, it was not useful enough because individual clusters could not necessarily be separated by the lines on the origin. One of the reasons why Genocluster works well is that Genocluster uses the angular information after the initial nearest neighbor method is performed. Genocluster interprets the node composed of a single plot as an outlier. However, one of the problems in Genocluster is that it cannot detect outliers when there are many of them. When one hopes to classify scattered data properly, the addition of an algoritm to detect outliers before the classification is likely to be useful. However, the aim of Genocluster is to classify typical data generated by well-controled experiments.

Although EM algorithm has been proposed, the program based on the EM algorithm is likely to consume a lot of calculation time since the iteration continues until it converges. MCMC method tends to obtain accurate results after a limited number of iterations if the initial clusters are close to the final results. In addition, Genoculster can perform the calculation in even shorter times because only undetermined plots after the nearest neighbor method are the candidates of the MCMC clustering method in the algorithm.

Extensive evaluation of the performance using real data has not yet been done. In the present manuscript, we have constructed an algorithm to cluster two-dimensional data into genotypes and have performed extensive comparisons between the results of the clustering by expert technicians and those obtained using the Genocluster algorithm.

When the expert technicians judged that a set of data required corrections, they changed the clusters to which some points belonged. The percentage of uncorrected points denotes the percentage of points belonging to clusters that were not changed by expert technicians. We determined the percentages of both the acceptance and the uncorrected points using a large number of datasets. Data from 10 different sets of experiments, each of which had 9600 different SNPs, were examined. Each SNP examination included 333 subjects. Therefore, the entire set of data included $333 \times 9600 \times 10 = 31\,968\,000$ points. This experiment clarified that Genocluster exhibited a high percentage of acceptance (84.4%) and a high percentage of uncorrected points (95.8%).

Although the results achieved by the Genocluster algorithm were not superior to the clustering by expert technicians, the performance of the proposed automated system was sufficient to be helpful for technicians in clustering two-dimensional data into genotypes. Thus, the data can be clustered by Genocluster and the resulting clustering can later be examined by technicians. When the technicians judge that the clustering is sufficient (84.4%), the data are not submitted to further manual clustering procedures by the technicians. When the technicians judge that the clustering is not sufficient, when necessary, the technicians correct the clusters to which a limited number of points (4.2%) belong. Although the proposed method does not eliminate the need for expert technicians, Genocluster will be very helpful for technicians in genotype clustering of large-scale data.

CONCLUSION

We have constructed the Genocluster algorithm for accurate automated clustering of two-dimensional data for single-nucleotide polymorphism genotyping by combining the nearest neighbor method and the MCMC method. The present algorithm, implemented by computer software, was optimized using real data from the Invader Assay. Evaluation of the algorithm using large-scale real data has shown the performance of Genocluster to be quite good (acceptance rate 84.4% and percentage of

uncorrected points 95.8%), although the Genocluster algorithm did not outperform expert technicians. Moreover, Genocluster was found to be useful for helping technicians to cluster genotypes.

ACKNOWLEDGEMENTS

The authors would like to acknowledge S. Shibata, S. Kato, K. Nakazono, N. Miyagawa and H. Higuchi for their excellent suggestions and comments. This study was supported by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of Interest: none declared.

REFERENCES

Anderberg,M. (1973) *Cluster Analysis for Applications*. Academic Press, New York.

Chen,J. and Kalbfleisch,J.D. (1996) Penalized minimum-distance estimates in finite mixture models. *Can. J. Stat.*, **24**, 167–175.

Chen,J. and Kalbfleisch,J.D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Statist. Soc. B*, **63**, 19–29.

Fujisawa,H. *et al*. (2004) Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics*, **20**, 718–726.

Gamerman,D. (1997) *Marcov chain Monte Carlo*. Chapman & Hall, New York.

Kanazawa,A. *et al*. (2004) Association of the gene encoding wingless-type mammary tumor virus integration-site family member 5B (WNT5B) with type 2 diabetes. *Am. J. Hum. Genet.*, **75**, 832–843.

Kang,H. *et al*. (2004) Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphism. *Am. J. Hum. Genet.*, **74**, 495–510.

Kizawa,H. *et al*. (2005) An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nat. Genet.*, **37**, 138–144.

Mein,C.A. *et al*. (2000) Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res.*, **10**, 330–343.

Ohnishi,Y. *et al*. (2001) A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.*, **46**, 471–477.

Olivier,M. *et al*. (2002) High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic. Acids. Res.*, **30**, e53.

Ozaki,K. *et al*. (2002) Functional SNPs in the lymphotoxin–alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.*, **32**, 650–654.

Ozaki,K. *et al*. (2004) Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion *in vitro*. *Nature*, **429**, *72–75*.

Ranade,K. *et al*. (2001) High-throughput genotyping with single nucleotide polymorphisms. *Genome Res.*, **11**, 1262–1268.

Riva,A. and Kohane,S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.

Ryan,D. *et al*. (1999) Non-PCR-dependent detection of the factor V Leiden mutation from genomic DNA using a homogeneous invader microtiter plate assay. *Mol. Diagn.*, **4**, 135–144.

Suzuki,A. *et al*. (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.*, **34**, 395–402.

Tokuhiro,S. *et al*. (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.*, **35**, 341–348.

Yeung,K.Y. *et al*. (2001) Model-based clustering and data transformations for gene expressions data. *Bioinformatics*, **17**, 977–987.