



Algorithms for large-scale genotyping microarrays

Wei-min Liu*, Xiaojun Di, Geoffrey Yang, Hajime Matsuzaki, Jing Huang, Rui Mei, Thomas B. Ryder, Teresa A. Webster, Shoulian Dong, Guoying Liu, Keith W. Jones, Giulia C. Kennedy and David Kulp

Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

Received on March 18, 2003; revised on May 30, 2003; accepted on June 6, 2003

ABSTRACT

Motivation: Analysis of many thousands of single nucleotide polymorphisms (SNPs) across whole genome is crucial to efficiently map disease genes and understanding susceptibility to diseases, drug efficacy and side effects for different populations and individuals. High density oligonucleotide microarrays provide the possibility for such analysis with reasonable cost. Such analysis requires accurate, reliable methods for feature extraction, classification, statistical modeling and filtering.

Results: We propose the modified partitioning around medoids as a classification method for relative allele signals. We use the average silhouette width, separation and other quantities as quality measures for genotyping classification. We form robust statistical models based on the classification results and use these models to make genotype calls and calculate quality measures of calls. We apply our algorithms to several different genotyping microarrays. We use reference types, informative Mendelian relationship in families, and leave-one-out cross validation to verify our results. The concordance rates with the single base extension reference types are 99.36% for the SNPs on autosomes and 99.64% for the SNPs on sex chromosomes. The concordance of the leave-one-out test is over 99.5% and is 99.9% higher for AA, AB and BB cells. We also provide a method to determine the gender of a sample based on the heterozygous call rate of SNPs on the X chromosome. See <http://www.affymetrix.com> for further information. The microarray data will also be available from the Affymetrix web site.

Availability: The algorithms will be available commercially in the Affymetrix software package.

Contact: wliu@cs.iupui.edu

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) may cause changes of gene expressions or may be close to more complicated and

unknown genetic variation sites and serve as useful markers to genetic studies. They may influence the susceptibility of individuals to diseases and the efficacy or side effects of drugs. The advances of oligonucleotide microarray technology make it possible to determine genotypes of many thousands of SNPs on all chromosomes in parallel (Fodor *et al.*, 1993; Chee *et al.*, 1996; Lipshutz *et al.*, 1999; Cargill *et al.*, 1999; Mei *et al.*, 2000; Dong *et al.*, 2001).

The HuSNP microarray was successfully used to study the loss of heterozygosity related to lung carcinoma by Lindblad-Toh *et al.* (2000). Cutler *et al.* (2001) proposed the ABACUS (Adaptive Background genotype Calling Scheme) algorithm for the Variation Detection Arrays (VDA). Here, we describe the algorithms for the Mapping 10k array applicable to much more complicated assays. The current Mapping 10k arrays interrogate over 10 000 SNPs. The algorithms include feature extraction, classification, formation of statistical models and application of pre-classification and post-call filters. The crucial part is the classification of features with our modified partitioning around medoids (MPAM). MPAM is based on the robust classification method called partitioning around medoids (PAM, Kaufman and Rousseeuw, 1987, 1990). MPAM includes terms in the objective function to penalize the small between-group distances. The average silhouette width (Rousseeuw, 1987) is a useful measure of classification quality. Once a large training data set is well classified, we can build statistical models to facilitate the genotype calls of new data. Moreover, we find that the heterozygous call rate (HCR) of SNPs on the X chromosome can determine the gender of a sample if it is not known.

2 EXPERIMENTAL DESIGN AND METHODS

The experimental part of the DNA microarray genotyping strategy includes *in silico* fractionation, synthesis of predicted oligonucleotide fragments on microarrays, biochemical fractionation, and allele specific hybridization. The critical step in biochemical fractionation is the fragment selection by PCR (FSP) proposed by Kennedy *et al.* (2003).

*To whom correspondence should be addressed at Department of Computer and Information Science, Indiana University Purdue University Indianapolis, 723 W. Michigan St., Indianapolis, IN 46202-5132, USA.

Since the majority of SNP markers can be considered as biallelic, we use a probe quartet as the basic unit for detecting different genotypes. A probe quartet includes a probe pair for allele *A* and a probe pair for allele *B*. A probe pair includes a perfect match cell and a mismatch cell. To make our genotyping calls more reliable, we use multiple probe quartets with the polymorphic nucleotide having different shifts from the center of the 25-mer probe sequence. For seven probe quartets, the shifts are set to be $-4, -2, -1, 0, +1, +3$ and $+4$. We tile multiple probe quartets for both sense and antisense strands. Therefore, seven probe quartets per strand mean 56 probe cells per SNP. We also examined the possibility of using fewer probe quartets and fewer or none of mismatch cells.

3 ALGORITHMS

Our algorithms include feature extraction, classification, model formation, call making and filtering. The steps of classification and model formation are only applied to training data. Once models are formed, the classification and model formation are not used in routine call making.

3.1 Detection filter

A detection filter can block weak or unreliable signals before the classification. We use the discrimination scores (DS) for target detection. The DS of a probe pair is the ratio of $(PM - MM)/(PM + MM)$, where PM is the intensity of the perfect match cell and MM is the intensity of the mismatch cell. Let us denote the DS of the *i*th probe pair of *A* and *B* alleles in the sense strand by $d_i^{(sA)}$ and $d_i^{(sB)}$, and those in the antisense strand by $d_i^{(tA)}$ and $d_i^{(tB)}$. The DS of *A* allele in the sense strand is defined to be $d^{(sA)} = \text{median}(d_i^{(sA)})$. Similarly, we can define $d^{(sB)}$, $d^{(tA)}$ and $d^{(tB)}$. The DS of a SNP can be defined as $d = \max(\min(d^{(sA)}, d^{(tA)}), \min(d^{(sB)}, d^{(tB)}))$. The superscripts *A* and *B* denote quantities for alleles *A* and *B*, and superscripts *s* and *t* denote quantities for sense and antisense strands. Let *D* be the threshold with the default value 0.03. If $d < D$, the signal is regarded as weak and the SNP of this particular sample is not used in the classification procedure. If too many data for a SNP do not pass the detection filter, no particular model will be formed. However, we can use the average of quality models of other SNPs to make genotype calls as explained later.

3.2 Feature extraction

We define the relative allele signal (RAS) for the *i*th probe quartet of the sense strand as

$$s_i^{(s)} = A_i^{(s)} / (A_i^{(s)} + B_i^{(s)}), \quad (1)$$

where $A_i^{(s)} = \max(\text{PM}_i^{(sA)} - \text{MM}_i^{(s)}, 0)$, $B_i^{(s)} = \max(\text{PM}_i^{(sB)} - \text{MM}_i^{(s)}, 0)$, and the average mismatch signal is

$\text{MM}_i^{(s)} = (\text{MM}_i^{(sA)} + \text{MM}_i^{(sB)})/2$. Because both $\text{MM}_i^{(sA)}$ and $\text{MM}_i^{(sB)}$ are mismatch intensities, their average, $\text{MM}_i^{(s)}$, is a more robust contrast to the perfect match signals than each individual mismatch intensity. Since RAS is essentially a ratio of intensities, it is not sensitive to microarray variability. Of course, Equation (1) is meaningful only when the denominator is positive. For the probe quartets with RAS defined, we further define the RAS for the sense strand as $s^{(s)} = \text{median}(s_i^{(s)})$. Similarly, we can define $s^{(t)}$ for antisense strand. The pair $(s^{(s)}, s^{(t)})$ is a point in the unit square of the feature space. Intuitively, if the point is close to the corner (1, 1), the genotype should be *AA*; if it is close to the origin (0, 0), the genotype should be *BB*; if it is near the center (0.5, 0.5), the genotype should be *AB*. However, because the affinity of target and probe is sequence dependent, and because the existence of cross hybridization is the rule rather than the exception, the locations, shapes and sizes of the genotype clusters are usually SNP dependent. Therefore, we estimate their distributions with classification for every SNP separately.

3.3 Classification with MPAM

The PAM (Kaufman and Rousseeuw, 1987, 1990; Struyf *et al.*, 1997) is a robust classification method based on a dissimilarity matrix. It can well classify most SNPs, but sometimes produces questionable results. We propose the MPAM, which includes PAM as a special case when the parameter $\lambda = 0$. Our usage of MPAM can be considered as unsupervised clustering because MPAM itself does not assign the genotypes. However, immediately after MPAM, we assign the genotypes based on the median coordinates of clusters. Moreover, the number of clusters (2 or 3) is pre-determined. Therefore, we still call it classification.

Let *n* be the number of distinct points, and we consider the problem of classifying them into *k* groups ($1 < k < n$). In the case of genotyping, we may have $k = 1, 2$ or 3 . Classification is done for $k = 2$ and 3 . If the results of classification for $k = 2$ and 3 are of low quality, the data are considered as from one group. Let $d(x_i, x_j)$ be the Euclidian distance between points x_i and x_j . PAM minimizes the objective function

$$f = \sum_{i=1}^n \min_{j=1, \dots, k} d(x_i, m_j) \quad (2)$$

for a subset (m_1, \dots, m_k) of (x_1, \dots, x_n) , and m_1, \dots, m_k are called the medoids of groups G_1, \dots, G_k . PAM minimizes the sum of distances of all points to the nearest medoids without consideration of the distances between groups. When there are significantly more points in a group than those in another group, PAM tends to separate the large group into two small groups to reduce the total sum of distances of all points to the nearest medoids (Fig. 1a). MPAM penalizes the small between-group distances. Figure 1b shows that MPAM

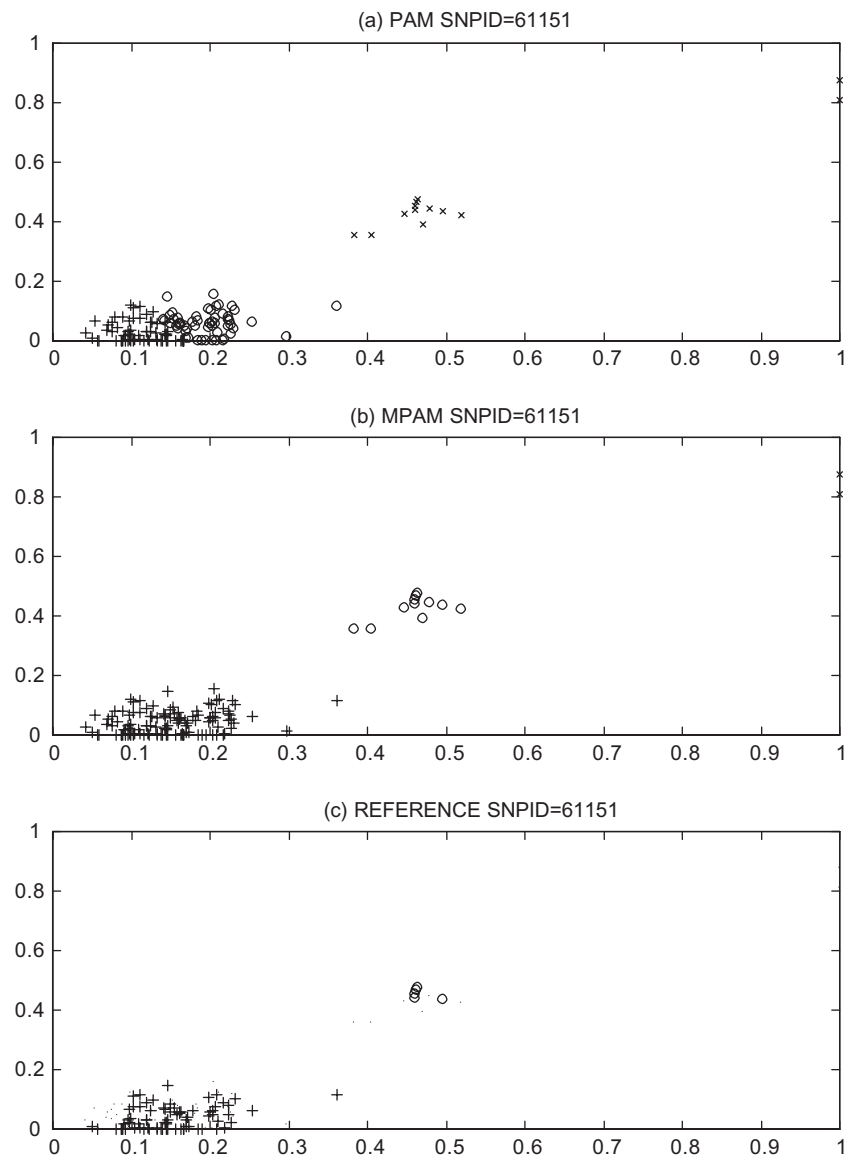


Fig. 1. The results of (a) PAM, (b) MPAM with $\lambda = 5$ and (c) reference calls. The horizontal axis is for the relative allele signal in the sense strand, and the vertical axis is for the relative allele signal in the antisense strand. The marks ‘x’ denote AA calls, the circles denote AB calls, the marks ‘+’ denote BB calls, and the dots in (c) indicate that these points have no reference calls. In this example, the traditional PAM splits a large cluster, and MPAM can keep a large cluster and is consistent with the reference calls. Note that MPAM results in (b) show three groups. The AA group has points on the boundary.

improves the genotyping call results. MPAM minimizes the new objective function

$$g = f - \lambda \sum_{j=1}^k D_j, \quad \text{where } D_j = \min_{x_a \in G_j, x_b \notin G_j} [d(x_a, x_b)] \quad (3)$$

is the smallest distance of group G_j to any point in other groups. The non-negative coefficient λ can adjust the penalty of small between-group distances. We tried parameter λ

from 0 to 20 000 and found that $\lambda = 5$ produced satisfactory results.

3.4 Quality of classification

The proper measures for the quality of classification help us make the decision whether to use the 2- or 3-group classification result to form a statistical model, or to take other approaches. Three quality measures are implemented.

1. *Average silhouette width.* The silhouette width of a point is a relative measure of the difference between the distance

of the point to the nearest neighbor group and the distance of the point to other points in the same group (Rousseeuw, 1987). The larger the silhouette width, the better the point is classified.

Let i be a point in group G . If i is the only point in G , we define its silhouette width to be $s(i) = 0$. Otherwise, $s(i)$ is defined in terms of $a(i)$ and $b(i)$. Here, $a(i)$ is the average distance from the point i to other points in the same group G : $a(i) = \sum_{j \in G, j \neq i} d(i, j) / (|G| - 1)$, and $|G|$ is the number of points in group G . Let the distance between the point i and another group C be $d(i, C) = \sum_{j \in C} d(i, j) / |C|$. The distance of i to the nearest neighbor group is $b(i) = \min_{C \neq G} d(i, C)$. The silhouette width of i is

$$s(i) = [b(i) - a(i)] / \max[b(i), a(i)]. \quad (4)$$

The average silhouette width is $s = \sum_{i=1}^n s(i) / n$. We found that if s is larger than 0.65 the data set looks well classified. We use a more conservative value 0.7 as a condition for a classification to be considered as good.

2. *Separation of groups.* The separation of groups indicates the absolute distances between groups in the feature space. We first calculate the medians of sense and antisense RAS of every group (AA, AB or BB). The sense separation is defined to be the minimum of the distance between AA and AB sense medians and the distance between the AB and BB sense medians. Similarly, we can define the antisense separation. We define the minimum of sense and antisense separations as the separation for a SNP.

3. χ^2 -test for the Hardy–Weinberg equilibrium. A χ^2 -test can tell the deviation from the Hardy–Weinberg equilibrium. Since our sample sizes are usually small and the conditions for this equilibrium may not hold, we only report its p -values for the convenience of further exploration.

3.5 Formation of statistical models

Classification algorithms take long time to find the optimal solution. We establish statistical models based on the classification results of training data, and use these models to make calls for new data. Our models are stable, and we only make new models when we use different labeling techniques or study reduced number of probes.

There are many models for classification. We use a robust model modified from the classical multivariate normal model with equal prior probabilities and the covariance matrices equal to the same multiple of the identity matrix. Under these assumptions, the probability of a point in a group is consistent with its proximity to the group center, and we can use Fisher's linear discriminants (e.g. Johnson and Wichern, 2002). We call the models robust because we use sample medians to estimate the group centers. Let us consider k groups with multivariate normal distributions $N(\mathbf{x}_i, \sigma^2 I)$ ($i = 1, \dots, k$). The linear discriminant $d_i(\mathbf{y}) = \mathbf{x}'_i(\mathbf{y} - 0.5\mathbf{x}_i)$. The point \mathbf{y} is classified to group j if $j = \arg \max_i [d_i(\mathbf{y})]$. The variance

σ^2 can be estimated with $\text{median}(r^2) / (2 \ln 2)$, where r^2 is the squared distance of a classified point to the corresponding distribution center.

We divide the models into three tiers based on the classification quality and accept or adjust the models accordingly. The model of a SNP belongs to the first tier if it has good three-group classification, i.e. there are at least two points in every group and the average silhouette width and separation are large enough. We accept the first tier models without adjustment. If the three-group classification has large enough average silhouette width and separation but a group has only one point, we categorize the model as in the second tier. If the average silhouette width or the separation of the three-group classification is not large enough, but the two-group classification has large enough average silhouette width and separation, we also rank the two-group model as in the second tier. For models in the second tier, we use the locally weighted regression smoothing (Hastie and Tibshirani, 1990) to estimate the center of distribution for the group with only one or zero point based on the models in the first tier. All other models are categorized as in the third tier, which includes the situation that there is really only one group and both 2- or 3-group classifications are of low quality. We use the average of the first tier models as the model for a SNP in the third tier of classification.

The locally weighted regression smoothing can be described as follows. Let the known good parameters, e.g. the centers of two groups in a second tier model be \mathbf{p}_0 , find K nearest first tier models with corresponding parameters \mathbf{p}_i ($i = 1, \dots, K$). Let the largest distance be $L = \max_{i=1, \dots, K} d(\mathbf{p}_0, \mathbf{p}_i)$, where $d(\mathbf{p}_0, \mathbf{p}_i)$ is the distance between parameter vectors \mathbf{p}_0 and \mathbf{p}_i . For fast computation, we use the 1-distance. The weight function $w(u) = (1 - u^3)^3$, if $u \in [0, 1]$; and $w(u) = 0$, otherwise. We calculate $w_i = w[d(\mathbf{p}_0, \mathbf{p}_i) / L]$. The other parameters \mathbf{q}_0 , e.g. the center of the group with 0 or 1 point, is estimated as $\mathbf{q}_0 = \sum_{i=1}^K w_i \mathbf{q}_i / \sum_{i=1}^K w_i$.

Since male has a single X chromosome and Y chromosome, the genotype of a SNP on the X or Y chromosomes for a male sample can only be homozygous. For male samples, we should use only two-group classification for SNPs on the X or Y chromosomes.

3.6 Post-call filters

To reach high accuracy, we implemented the following post-call filters. The region filter is set as the default filter.

1. *Confidence filter.* The probability of a type i call is proportional to $\exp[d_i(\mathbf{y}) / \sigma^2]$. We denote the largest discriminant as $\max[d_i(\mathbf{y})]$ and the second largest discriminant as $\text{second}[d_i(\mathbf{y})]$. Their rescaled difference $c = [\max(d_i(\mathbf{y})) - \text{second}(d_i(\mathbf{y}))] / (\sigma^2 \ln 10)$ is the logarithm with base 10 of the probability ratio and can be used as a confidence measure. However, for data such as $\mathbf{y} = (1, 0)$, the sense and antisense relative allele signals suggest opposite homozygous calls, and it usually gives very high confidence for a heterozygous call. The region filter can better exclude such unreliable points.

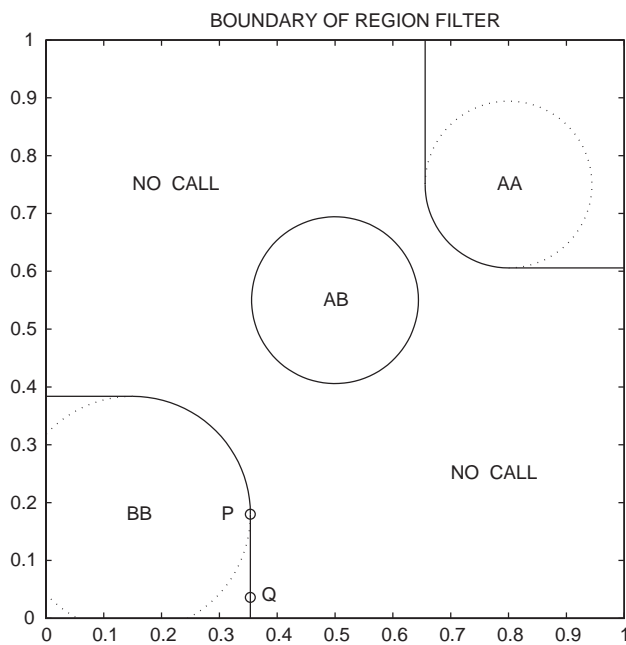


Fig. 2. The post-call region filter. Acceptable regions include three disks and points bounded by the horizontal or vertical line segments tangent to the circles for the two homozygous groups. The radius of a circle equals 0.8 times the radius of the corresponding group as defined in the text.

2. *Region filter.* The ratio of the distance between a point and the nearest center of a distribution to the radius of this distribution can be used as an efficient filter to exclude inaccurate calls. We name it the distance-to-radius ratio. The radius of a homozygous distribution is defined to be half of the distance between its center and the heterozygous center. The radius of the heterozygous distribution is defined to be the smaller of the two radii of homozygous distributions. We found that higher than 99% accuracy can be reached if we accept a call when its distance-to-radius ratio is less than 0.8. We can extend the acceptance region to include points closer to the homozygous corners (0, 0) and (1, 1) as shown in Figure 2. We can show that point Q outside the 0.8 distance-to-radius ratio circle in Figure 2 has higher confidence c of type BB than point P on this circle. Therefore, our extension is statistically sound.

3. *Detection filter.* The detection filter can also be used as a post-call filter, e.g. we can reject a call if the discrimination score $d < D_1$. Here the threshold D_1 does not have to be the same as the threshold D used in the pre-classification detection filter.

4 RESULTS

We produced a genotype microarray containing probes for 10043 SNPs. Among them, 5625 probe sets have seven probe quartets for each of the sense and antisense strand, and the

Table 1. Concordance with SBE genotypes and call rates

	C (%)	R1 (%)	R2(%)
PnM_A	99.36	97.6	98.4
PnM_Sex	99.64	97.9	99.6
PnM2_A	99.48	98.1	99.4
PnM2_Sex	99.64	97.3	99.6
P_A	99.46	99.0	99.7
P_Sex	99.64	94.9	99.6

Column C lists the concordance rates of calls on the Early Access genotyping microarray with the SBE reference genotypes. Column R1 lists the overall call rates. Column R2 lists the call rates for the reference data. The first two rows of PnM list the performance of features defined in Section 3.2. The postfix _A indicates the statistics for the SNPs on the autosomes. The postfix _Sex indicates the statistics for those on the sex chromosomes. The rows of PnM2 list the performance of features RAS2 defined in the discussion section with PM and MM probes and $C = 10$. The last two rows of P list the performance of features RAS3 defined in the discussion section with PM probes and $H = 0.1$.

remaining 4418 probe sets have five probe quartets for each strand. The restriction enzyme for fragmentation is XbaI. We did hybridization experiments for 133 individuals (48 African Americans, 26 Asians, 53 Caucasians, 3 Mexican Americans and 3 Native Americans).

To compare our results with the genotypes obtained by other methods, we purchased some single base extension (SBE) reference genotypes. The intersection of the SBE reference calls and our genotyping microarray contains 19566 calls of 469 SNPs. The concordance rates and call rates are listed in Table 1. Of course, the SBE reference genotypes may also contain errors. Therefore, 100% concordance may not be the best goal.

We used the leave-one-out test for cross validation. We take one individual out of the set, and use the remaining 132 individuals to form models and use these models to make calls for the one that is not used in model building, and compare these calls with the calls made with models of 133 individuals. This is done for all 133 possible choices of 132 individual models. We found the concordance rates are 99.98% for SNPs on autosomes and 99.90% for SNPs on sex chromosomes for AA, AB and BB cells.

We also consider the informative Mendelian relationship. Within the individuals in our sample, there is a CEPH (Centre d'Etude du Polymorphisme Humain) family trio. A trio includes two parents and a child. We count the calls that are inconsistent with the Mendelian relationship. If the error in a trio is contributed by one member in the trio, the overall error rates are 0.19% for autosomal SNPs, 0% for SNPs on sex chromosomes.

5 DISCUSSION

We found that the gender of a sample can be accurately estimated by the HCR of the SNPs on the X chromosome. We can

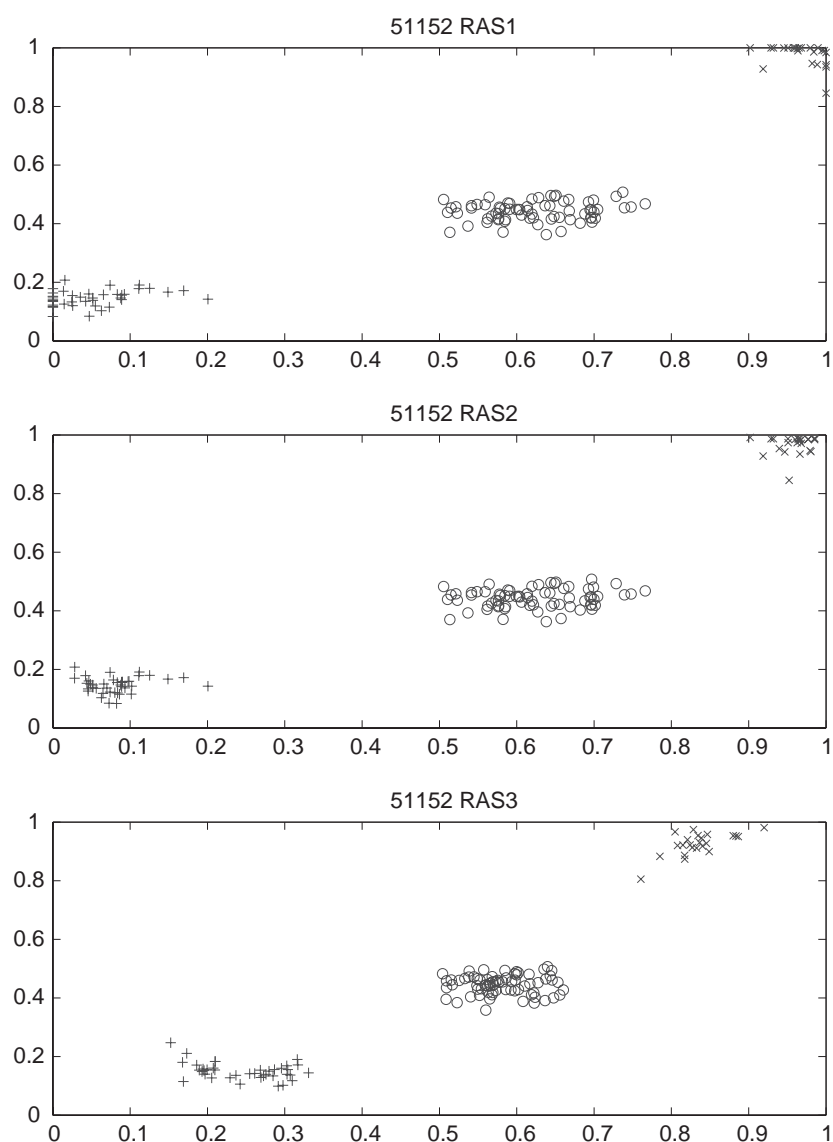


Fig. 3. The scatter plots of (a) the relative allele signals defined in Section 3.2, (b) the features RAS2 with $C = 10$ and (c) the features RAS3 with $H = 0.1$ of SNP 51152 for the data of 133 individuals. The marks 'x' denote AA calls, the circles denote AB calls, and the marks '+' denote BB calls.

form models for the SNPs on the X chromosomes as those on autosomes, i.e. consider all three possible calls. The heterozygous call rate is defined to be the ratio of the number of qualified heterozygous calls to the number of total qualified calls. The qualified calls are made by models with large enough separation and average silhouette width and have the confidence values c larger than 6. Since male sample should have only type A or B for these SNPs, the HCR for male should be low and that for female are usually high. If HCR is larger than a threshold, we estimate that the sample is from a female, otherwise, we estimate it is from a male. The threshold is set to be 0.115.

We can also use other methods to form the relative allele signals. For example, when PM is less than MM in only one strand, the feature defined in Section 3.2 tends to become 0 or 1, which may not always be a good indicator of the relative allele signal. As a different approach to handle this situation, we can use $MM = (MM^{(A)} + MM^{(B)})/2$, $A' = PM^{(A)} - MM$, $B' = PM^{(B)} - MM$, where A' and B' may be negative. Let $F = \max[C - \min(A', B'), 0]$, where C is a small positive number, such as 10. Let $A = A' + F$ and $B = B' + F$. We may define the relative allele signal as $RAS2 = A/(A + B)$. Here F is an adaptive pseudocount varying from 0 (if both A' and B' are larger than C) to a positive number so that A and

B are at least C . Figure 3b shows that RAS2 can move certain points on the boundary to the inner region of the unit square.

We may also use only the perfect match cells. An intuitive idea is to introduce the feature $R = \text{PM}^{(A)} / (\text{PM}^{(A)} + \text{PM}^{(B)})$. However, R is more centered at 0.5 than the features defined with the differences of PM and MM. To make it distribute broader on the unit interval, we can introduce the transformation $R' = (R - H) / (1 - 2H)$, where H is non-negative and smaller than 0.5. We then shrink the values of R' outside the unit interval to the nearest boundary, i.e. we can use the feature $\text{RAS3} = \max[\min(R', 1), 0]$. If only PM probes are used, the filter based on discrimination scores will not be available. Figure 3c shows the scatter plot of RAS3 of a SNP. Table 1 lists its performance for $H = 0.1$. The performance of RAS2 and RAS3 on more data is under study.

We can also use fewer probe quartets. We even tried to use only one probe quartet per allele per strand for a SNP, and obtained over 99% concordance rates with the reference genotypes. But for robustness, we still use multiple probe quartets per allele per strand.

6 CONCLUSION

The high density DNA microarrays provide a cost-efficient large scale genotyping tool. Ten thousand SNPs per microarray are currently feasible. Our algorithms are scalable and can be used to microarray with smaller feature sizes, fewer probes and more SNPs.

ACKNOWLEDGEMENTS

We thank Manqiu Cao, Mark Chai, Wenwei Chen, Richard Chiles, Steve Fodor, Amy He, Matthew Ho, Earl Hubbell, Weiwei Liu, Gregory Marcus, Michael Mittman, Carsten Rosenow, Gretchen Schieber, Mei-Mei Shen, Sean Walsh and Jane Zhang for helpful discussion or providing data. We also thank the referees for their constructive comments and suggestions.

REFERENCES

- Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardle, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemes, J. *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.
- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, J., Morris, M.S. and Fodor, S.P. (1996). Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
- Cutler, D.J., Zwick, M.E., Carrasquillo, M.M., Yohn, C.T., Tobin, K.P., Kashuk, C., Mathews, D.J., Shah, N.A., Eichler, E.E., Warrington, J.A. *et al.* (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res.*, **11**, 1913–1925.
- Dong, S., Wang, E., Hsie, L., Cao, Y., Chen, X. and Gingeras, T.R. (2001) Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res.*, **11**, 1418–1424.
- Fodor, S.P.A., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P. and Adams, C.L. (1993) Multiplexed biochemical assays with biological chips. *Nature*, **364**, 555–556.
- Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall, London.
- Johnson, R.A. and Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis*, 5th edn. Prentice-Hall, Upper Saddle River.
- Kaufman, L. and Rousseeuw, P.J. (1987) Clustering by means of medoids. In Dodge, Y. (ed.), *Statistical Data Analysis Based on the L_1 -norm and Related Methods*. North Holland, Amsterdam, pp. 405–416.
- Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.-m., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J. *et al.* (2003) Large-scale genotyping of complex DNA, *Nat. Biotechnol.*, in press.
- Lindblad-Toh, K., Tanenbaum, D.M., Daly, M., Wichester, E., Lui, W.-O., Villapakkam, A., Stanton, S.E., Larsson, C., Hudson, T.J., Johnson, B.E. *et al.* (2000) Loss of heterozygosity analysis of small-cell lung carcinomas using single nucleotide polymorphism arrays. *Nat. Biotechnol.*, **18**, 1001–1005.
- Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21** (Suppl. 1), 20–24.
- Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M., Reid, B. and Lockhart, D.J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, **10**, 1126–1137.
- Rousseeuw, P.J. (1987) Silhouette: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
- Struyf, A., Hubert, M. and Rousseeuw, P.J. (1997) Integrating robust clustering techniques in S-Plus. *Computat. Stat. Data Anal.*, **26**, 17–37.