Title
A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays

Authors
Xiao, Yuanyuan
Segal, Mark R
Yang, Jean YH
et al.

Publication Date
2007-04-25

Peer reviewed

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 12 2007, pages 1459–1467
doi:10.1093/bioinformatics/btm131

Genome analysis

A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays

Yuanyuan Xiao[1,*], Mark R. Segal[1], Y.H. Yang[2] and Ru-Fang Yeh[1,*]

[1]Department of Epidemiology and Biostatistics, Center for Bioinformatics and Molecular Biostatistics, University of California, 185 Berry Street, Lobby 4, Suite 5700, San Francisco, CA 94107, USA and [2]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

Received on August 1, 2006; revised on March 30, 2007; accepted on March 31, 2007

Advance Access publication April 25, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Modern strategies for mapping disease loci require efficient genotyping of a large number of known polymorphic sites in the genome. The sensitive and high-throughput nature of hybridization-based DNA microarray technology provides an ideal platform for such an application by interrogating up to hundreds of thousands of single nucleotide polymorphisms (SNPs) in a single assay. Similar to the development of expression arrays, these genotyping arrays pose many data analytic challenges that are often platform specific. Affymetrix SNP arrays, e.g. use multiple sets of short oligonucleotide probes for each known SNP, and require effective statistical methods to combine these probe intensities in order to generate reliable and accurate genotype calls.

Results: We developed an integrated multi-SNP, multi-array genotype calling algorithm for Affymetrix SNP arrays, MAMS, that combines single-array multi-SNP (SAMS) and multi-array, single-SNP (MASS) calls to improve the accuracy of genotype calls, without the need for training data or computation-intensive normalization procedures as in other multi-array methods. The algorithm uses resampling techniques and model-based clustering to derive single array based genotype calls, which are subsequently refined by competitive genotype calls based on (MASS) clustering. The resampling scheme caps computation for single-array analysis and hence is readily scalable, important in view of expanding numbers of SNPs per array. The MASS update is designed to improve calls for atypical SNPs, harboring allele-imbalanced binding affinities, that are difficult to genotype without information from other arrays. Using a publicly available data set of HapMap samples from Affymetrix, and independent calls by alternative genotyping methods from the HapMap project, we show that our approach performs competitively to existing methods.

Availability: R functions are available upon request from the authors.

Contact: yxiao@itsa.ucsf.edu and rufang@biostat.ucsf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are sites in the genome where individuals differ in DNA sequence by a single base pair. There are ~10 million common SNPs that constitute 90% of the variation in the current human population (The International HapMap Consortium, 2003). While most SNPs have, to date, no characterized role in cell function, select SNPs associated with altered proteins or phenotypic traits have been found. SNPs result from single historical mutation events and, as nearby variants on the ancestral chromosome harboring the new allele tend to segregate together (as a *haplotype*), positional correlations [termed *linkage disequilibrium* (LD)] ensue. LD is fundamental to much of human genetic research: since sequence variants located at, or near, the causal mutation(s) for an inherited disease should still carry the disease association, a strategy for mapping disease loci can be based on testing genome-wide associations between a clinical trait and such (here SNP) variants. The success of such a global search strategy for eliciting genetic influence on disease relies on examining large numbers of SNPs in large numbers of affected individuals and controls, and is only possible due to recently devised high-throughput technologies. These SNP genotyping technologies present many statistical and informatic challenges, forefront of which is the development of a genotyping algorithm that is highly accurate, scalable, efficient and inexpensive. The goals of this article are to develop and illustrate such an algorithm, specific to Affymetrix SNP microarrays.

Affymetrix chips use short oligonucleotide probe quartets to interrogate each dimorphic site and include up to 260 000 SNPs. Each quartet consists of a perfect match (PM) and a mismatch (MM) 25-mer, corresponding to both alleles (arbitrarily named allele A and allele B) of a known SNP, yielding four different probes — PMA, PMB, MMA and MMB — that form the basic unit for quantifying allele-specific hybridization. Each SNP has multiple quartets querying different strands and shifts surrounding the polymorphic site. The primary question for data analysis is, then, how to map the intensities of these probe quartets to a genotype call (AA or AB or BB) for each SNP represented on the array.

Affymetrix developed the clustering-based MPAM algorithm (Liu *et al.*, 2003) for their first-generation (10 K) SNP microarrays, and the model-based dynamic model (DM) algorithm (Di *et al.*, 2005) for subsequent 100 K and 500 K arrays. MPAM uses modified partition-around-medoids to cluster samples (arrays) into different genotypes for each SNP.

*To whom correspondence should be addressed.

It employs numerous fine tunings and heuristic rules in order to cope with SNPs with low minor allele frequencies (MAF) and/or sub-optimal hybridization signals. The DM algorithm uses probe-level log likelihoods to select the best of the four genotype models (*AA*, *AB*, *BB*, and Null) for each probe quartet, followed by a SNP-level aggregation to generate genotype calls for each SNP. It is important to note that these genotype calling algorithms were used and tuned in the probe design and selection phase, so as to optimize performance from a much larger initial pool of probes from known SNPs. For example, the 100 K array contains 116 204 SNPs that were selected based on their preferential hybridization and prediction performance *using the DM algorithm* from a data set comprising the ~ 535 000 known SNPs in *Xba*I and *Hind*III restrictive digestion fragments. Thus, the genotyping performance of DM will be optimistically biased when applied to the 100 K array.

In a similar vein, GEL (genotype calling using empirical likelihood) proposed by Nicolae *et al*. (2006), employs likelihood calculations at the quartet level based on preliminary genotype calls as supplied (for example) by DM. Improvements over DM are furnished by weighting information from each quartet according to its genotyping quality. Both GEL and DM have a higher genotyping accuracy than MPAM, yet they do not account for probe-specific effects or incorporate multi-array information, leaving room for improvement. Accordingly, Rabbee and Speed (2006) proposed the classification-based RLMM (robust linear model with Mahalanobis distance) algorithm that takes advantage of the large number of publicly available SNP calls from the HapMap project in order to define genotyping rules. They show improved genotyping accuracies, compared to the DM algorithm, for a set of HapMap individuals using 100 K arrays. Another recent development, SNiPer-HD (SNiPer-High Density, (Hua *et al*., 2006), employs an expectation-maximization (EM) algorithm with parameters based on a training sample set, also exhibited superior performance to DM. However, both RLMM and SNiPer-HD rely crucially on the availability of good training data sets which most projects lack. For instance, in the case of SNiPer-HD (Hua *et al*., 2006), the authors took advantage of a set of 900 arrays with known genotyping information. The PLASQ algorithm proposed by LaFramboise *et al*. (2005) seeks to infer allele-specific copy number changes along with genotype calls via linear models on the probe intensities using an EM algorithm. It also requires calibration from a set of at least 8–15 normal diploid samples. Among these above-mentioned methods, DM, GEL, SNiPerHD and PLASQ operate mainly within each SNP, and do not exploit similarities of allele-specific hybridization patterns across the thousands of available SNPs. MPAM and RLMM do attempt to incorporate between SNP information, albeit only in cases where MAF are concerned.

Realizing DM's limitations, we and others have independently and simultaneously developed improved genotyping algorithms for mapping 100 K and 500 K arrays. BRLMM (Affymetrix, 2006), advanced by Affymetrix, is an extension of the RLMM model that removes RLMM's dependence on training data. It uses stringent DM calls as initial genotyping seeds to derive a prior distribution for typical genotype regions. Each SNP is then visited and its genotyping regions re-calibrated using an *ad hoc* Bayesian procedure. CRLMM, proposed by Carvalho *et al*. (2006), is also in line with the principles of RLMM. The genotyping component of CRLMM is largely similar to BRLMM, but it employs refined normalization and summarization methodologies to facilitate cross-lab data comparison and integration.

Here, we propose an algorithm that integrates allele specific information via a multi-array, multi-SNP (MAMS) approach. Our algorithm starts with model-based clustering of the multitude of SNPs located on the same array, using a resampling scheme for computational efficiency. Based on the derived genotypes, a subset of SNPs showing unique hybrization kinetics is identified, and subject to a within-SNP, across-array clustering approach. Our method requires no fine tuning or training data. We evaluate the performance of MAMS on both the 100 K and 500 K SNP array platforms and show that the accuracy and efficiency of MAMS compare favorably to other genotyping methods.

2 METHODS

Our algorithm consists of four components as further detailed below: (i) preprocessing: summarizing probe-level intensities into SNP-level indices for the two alleles; (ii) single-array, multi-SNP (SAMS) genotype calls: applying model-based clustering to the SNP-level indices within an *array*, and making genotype calls based on model-based inference; (iii) multi-array, single-SNP (MASS) genotype calls: employing hierarchical clustering on the indices within each *SNP* and (iv) MAMS genotype calls: aggregate (ii) and (iii) by evaluating quality scores of SAMS and MASS calls.

2.1 Data

We evaluated the MAMS algorithm on both the Affymetrix 100 K public dataset [90 Centre d'Etude du Polymorphisme Humain (CEPH) samples] and the 500 K public dataset (39 CEPH samples). For both datasets, we used reference calls from the International HapMap Project (The International HapMap Consortium, 2003) that were generated using other genotyping techniques. Of the 116 204 SNPs on the *Xba*I and *Hind*III (100 K) arrays, 27 049 SNPs were available from HapMap release 14. Among the 262 264 SNPs on the *Nsp*I (500 K) arrays, 54 540 SNPs have independent calls in HapMap release 20. These SNPs form the HapMap reference set that we use to benchmark our algorithm.

2.2 Preprocessing

Unless otherwise noted, all intensities are on the log2 scale. Let PMA_{ijk} and PMB_{ijk} be the probe intensities for array i, SNP j and probe quartet k, for allele*s* A and B, respectively. Similarly, MMA_{ijk} and MMB_{ijk} are their MM counterparts. We use $\mathrm{MM}_{ijk} = \max(\mathrm{MMA}_{ijk}, \mathrm{MMB}_{ijk})$ as an arguably more robust summary of non-specific binding and cross-hybridization for probe quartet k. We summarize the $4k$ probe intensities by taking medians of background-corrected signals for alleles A and B (θ_A and θ_B, respectively), and their relative signal intensity (θ_{AB}):

$$\theta_{A,ij} = \mathrm{median}_k(\mathrm{PMA}_{ijk} - \mathrm{MM}_{ijk}), \quad (1)$$

$$\theta_{B,ij} = \mathrm{median}_k(\mathrm{PMB}_{ijk} - \mathrm{MM}_{ijk}), \quad (2)$$

$$\theta_{AB,ij} = \mathrm{median}_k(\mathrm{PMA}_{ijk} - \mathrm{PMB}_{ijk}). \quad (3)$$

These indices form the basis of all downstream analysis.

2.3 SAMS: single-array multi-SNP classification with mixture models

Let $x_j = (\theta_{AB}, \theta_A, \theta_B)$ denote the intensity indices for the jth SNP (suppressing the subscript for array i; $j = 1, \ldots, n$). We adopt a normal mixture model-based approach to cluster the SNP data so that each observation is from a mixture of G multivariate normal distributions with proportions $\pi_1, \ldots, \pi_g$. We take $G = 3$ corresponding to the three genotypes: AA, AB and BB. The likelihood for the mixture model is:

$$L(\tau_g, \pi_g) = L(\tau_{AA}, \tau_{AB}, \tau_{BB}; \pi_{AA}, \pi_{AB}, \pi_{BB}) = \prod_{j=1}^{n} \sum_{g=1}^{G} \pi_g f_g(x_j; \tau_g),$$

where, f_g is the multivariate normal density function of the gth component, and τ_g denotes the corresponding parameters: mean μ_g and covariance matrix Σ_g. Maximum likelihood estimates of model parameters can be obtained via the EM algorithm (Dempster *et al.*, 1977). The E-step computes a matrix, z_{jg}, which is an estimate of the conditional probability that the jth SNP belongs to the gth component of the mixture given the current parameter estimates. The M-step computes the mixing proportions, means and covariance matrices given the current probabilities z_{jg}.

The orientation, volume and shape of the component distributions are determined by the covariance matrix Σ_g, which can be parametrized in a variety of ways (Banfield and Raftery, 1993). In the simplest case, where $\Sigma_g = \lambda I$, all clusters are spherical and of the same size, and only one parameter needs to be estimated. Alternatively, when all geometric features are allowed to vary between clusters, $G(d(d+1)/2)$ parameters need to be estimated, where d is the data dimension. We consider all combinations of (constant, variable) × (orientation, volume, shape) parameterizations for modeling fitting. To select the best-fitting model in mixture model clustering, Fraley and Raftery (2002) used the Bayesian information criterion (BIC) for its appropriateness and good performance; we also adopt this approach.

Model fitting and selection make recourse to the `R` library `mclust` (Fraley and Raftery, 2002). Because of the large data set size we randomly sample 2000 data points and apply model-based clustering to this smaller set. This procedure is repeated 10 times, and final estimates of means and covariance matrices are obtained by averaging over these sets. Posterior conditional probabilities of genotype class membership for each SNP in the entire data set are then computed using these parameters. The SAMS call for SNP i on array j, $\mathrm{cl}_{\mathrm{SAMS},ij}$, is just the genotype cluster with the highest posterior probability and its genotyping uncertainty can be obtained by subtracting the highest posterior probability from 1.

2.4 MASS: multi-array single-SNP hierarchical clustering

Clustering/classification based on within-SNP intensities forms the basis of the (Liu *et al.* 2003) and Rabbee and Speed (2006) methodologies. We also integrate this approach into our algorithm. For a given SNP j, let $x_{1j}, \ldots, x_{Ij}$ denote the I ($I = 90$ for 100 K arrays and $I = 39$ for 500 K arrays) samples in the 3D space $(\theta_A, \theta_B, \theta_{AB})$. To assign the I samples into genotype clusters, we apply agglomerative hierarchical clustering using Euclidean distance and average linkage, and cut the dendogram according to the number of clusters determined by $\mathrm{cl}_{SAMS,ij}$. $\mathrm{cl}_{\mathrm{MASS},ij}$ is subsequently determined based on the cluster membership.

2.5 MAMS by comparative quality scores

A well-classified SNP can be characterized as follows: distances between different clusters that define genotypes are large and distances among samples within each cluster are small. The tightness of each cluster indicates that the SNP exhibits consistent hybridization signals across the multiple arrays (samples), whereas the separation of clusters shows that the PM probes successfully detect signals from both alleles and there is no substantial cross-hybridization in the MM signals. The silhouette width has been frequently used to assess the quality of clustering (Rousseeuw, 1987), and for a point l in a cluster, is defined as

$$s(l) = [b(l) - a(l)]/\max(a(l), b(l)),$$

where $a(l)$ is the average distance from l to all points within the cluster, and $b(l)$ is the average distance from l to all data points outside the cluster. The smaller the within-cluster distance and the larger the between-cluster distance, the better the point is classified. Therefore, a large and positive silhouette width signals that the point is well classified, whereas a negative silhouette width signals the point is arbitrarily or wrongly classified. We calculate silhouette scores for the clustering results from both the SAMS (Section 2.3) and MASS (Section 2.4) procedures, denoted $s_{SAMS,ij}$ and $s_{MASS,ij}$, respectively. The resultant MAMS genotype cl_{ij} is then determined by competitive calls of $\mathrm{cl}_{SAMS,ij}$ and $\mathrm{cl}_{\mathrm{MASS},ij}$:

$$cl_{ij} = \begin{cases} \mathrm{cl}_{SAMS,ij} & \text{if } s_{SAMS,ij} > 0, \\ \mathrm{cl}_{\mathrm{MASS},ij} & \text{if } s_{MASS,ij} > 0 \text{ AND } s_{SAMS,ij} < 0. \end{cases} \quad (4)$$

To assign genotype confidence to MAMS calls, we calculate the *post hoc* posterior probabilities of a SNP belonging to the best-fitting genotype models. The model parameters required for this purpose are SNP-specific mean θ matrices and the genotype-specific variance–covariance matrices estimated in the M step of SAMS .

3 RESULTS

Figure 1 plots median-summarized intensities for all SNPs on one of the CEPH *Hind* III arrays, with color labeling according to known HapMap genotypes (obtained independently). It is apparent that the genotype groups fall into three distinct clusters (Fig. 1a, PMA versus PMB), and demonstrates that the majority of the SNPs exhibit commensurate allele-specific hybridization patterns. The top cloud (green) is for genotype group BB, the one along the diagonal (blue) is AB, and that at the bottom is AA. Interestingly, genotype group clusters can be recovered using MM intensities alone (Fig. 1b, MMA versus MMB), but with no clear separation between clusters. This indicates that the MM probes retain allele-specific hybridization signal. Figure 1c and d display background-adjusted indices θ_A and θ_B and the relative allele index θ_{AB} (Equation 1–3). The enhanced separation suggests that accounting for MM signals aids discrimination. This contrasts with what is observed in expression arrays, where incorporating MM signals often has adverse effects on predicting transcript abundance (Irizarry *et al.*, 2003).

3.1 SAMS

We applied model-based clustering to the points in 3D space $(\theta_{AB}, \theta_A, \theta_B)$ for each array. Colinearity among the three indices is less of a concern in this classification setting than it is in regression contexts. Since there is extra information in the third dimension (results not shown), we use all three indices for downstream classification analysis. Model-based clustering yields the results shown in Figure 2. The model selected is the unconstrained model with variable volume, orientation and

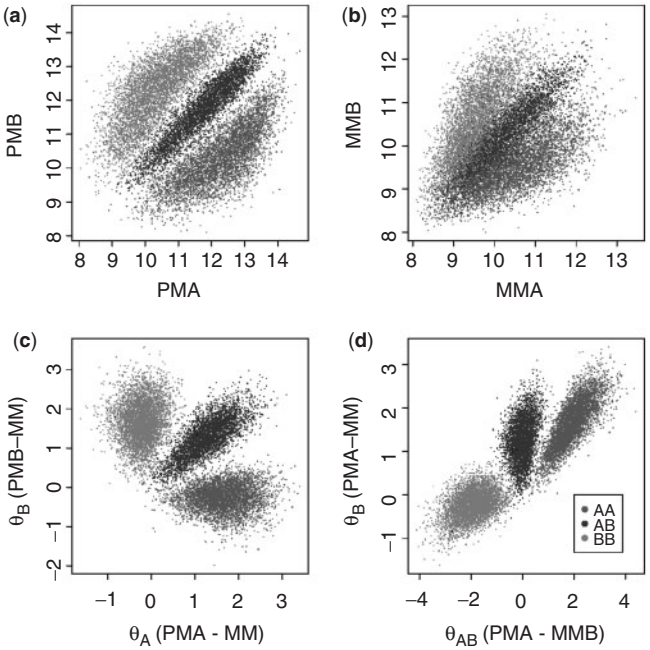

Fig. 1. Scatter plots of intensity indices for one of the CEPH *Hind*III arrays. Each point represents a SNP with color labeled according to its known HapMap genotype. Panels (**a**) and (**b**) plot median summarized intensities for the two alleles. Panels (**c**) and (**d**) plot background-adjusted indices, θ_A, θ_B and θ_{AB}.

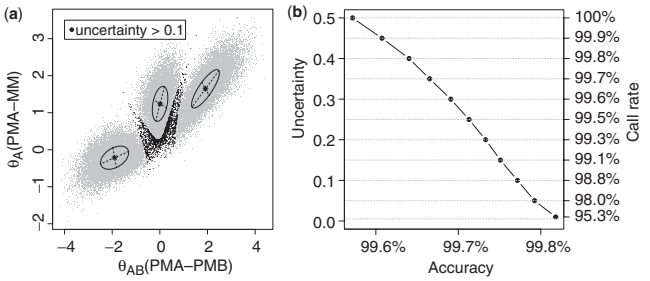

Fig. 2. Application of SAMS model-based clustering to SNPs from a CEPH *Hind*III array. Panel (**a**) plots θ_A versus θ_{AB} with the means and SDs of the fitted model indicated by the superimposed ellipses. Black points correspond to SNPs that have uncertainty ≥ 0.1. Panel (**b**) illustrates the relationship between uncertainties and genotyping accuracies as measured by concordance with HapMap reference genotypes. Each point represents the accuracy for SNPs with uncertainty less than or equal to the corresponding threshold. Call rates are calculated as the percentages of SNPs that have uncertainties less than or equal to the associated thresholds.

shapes for the three component normal distributions. The superimposed ellipses in panel (a) are projections of the fitted model, with axes indicating SDs. Figure 2a also highlights the points that have uncertainty >0.1 which, as expected, fall between clusters. Figure 2b confirms that the points that have higher uncertainty values are also the ones that are most likely to be misclassified. The overall accuracy rates for the *Xba*I and *Hind* III sets of 100 K arrays, treating the HapMap genotypes as ground truth, are 99.50 and 99.57%, respectively, slightly lower than Affymetrix's DM algorithm.

3.2 MASS

The second step in the MAMS procedure takes advantage of the availability of multiple arrays and investigates the distribution pattern of the θ vectors for each SNP separately. This proves informative for SNPs exhibiting idiosyncratic hybridization properties. Such SNPs are difficult to reliably genotype with any approach (including SAMS) that relies on combining information from multiple SNPs. Figure 3 underscores the motivation behind the MASS component. The θ vector for all 90 *Xba*I arrays from two SNPs are displayed, color coded to indicate genotypes as assigned by SAMS. Both SNPs show well-formed genotype clusters among the 90 arrays, however, only the first SNP possesses a high confidence (uncertainty <0.05) in genotypes assigned by SAMS. For the second SNP, half of the arrays in the heterozygous group are misclassified as *BB* homozygotes. Close inspection reveals that the misclassification is due to a collective shift of the θ_{AB} signals from zero to negative values. This shift, consistent among all arrays of the heterozygous genotype, stems from an intensity bias between the PM for the two alleles. This collective shift, along with the well-delineated clusters of this SNP (as characterized by small within- and large between-cluster distances), suggest that a SNP-dependent clustering approach may yield more accurate classifications. Indeed, using MASS as operationalized above, including inheriting the number of clusters from SAMS, correctly genotypes all 90 arrays for this SNP.

3.3 MAMS

Even though MASS is more effective with some SNPs than SAMS as illustrated above, it faces challenges in correctly classifying SNPs when separation between genotype groups is insufficient, or when the number of genotype groups is not evident. To identify and characterize such probes, we employed silhouette width to quantify within-SNP clustering quality. Based on the scatterplots used to assign genotypes, we and others observe that satisfactory classification can be obtained when the silhouette width is >0.65. Employing this cutoff, 2.6% of the *Hind* III SNPs do not form good-quality clusters, and $\sim$ 5.1% of the SNPs cannot be confidently assigned to genotypes. For such cases, multiple-array, single-SNP methods are inferior to multi-SNP based algorithms in terms of genotype accuracy. The MPAM algorithm, as proposed by Affymetrix for 10 K SNP arrays, invokes many heuristic rules, including visual inspection to overcome this problem. To avoid making such arbitrary decisions, and to build a more flexible, robust and scalable algorithm, we base our genotype calls on SAMS and correct these only in instances where such calls are apparently wrongly assigned, as indicated by negative SAMS silhouette widths in Equation (4). An example is showcased in Figure 3b. For this SNP, the within-cluster average silhouette widths from SAMS genotypes are 0.92 (*BB*), 0.14 (*AB*) and 0.82 (*AA*), whereas they are 0.92 (*BB*), 0.75 (*AB*) and 0.82 (*AA*)

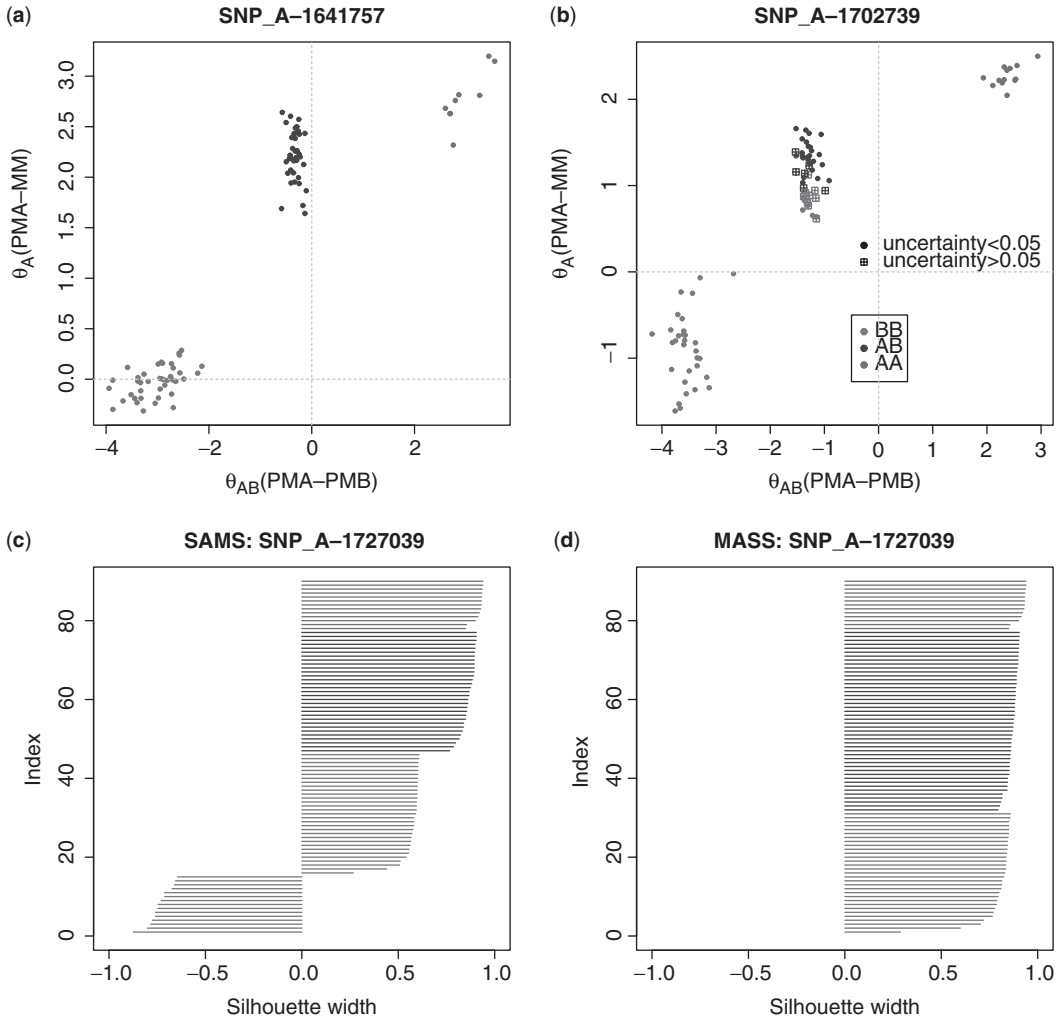

Fig. 3. Motivation for MASS within-SNP clustering. Shown in panels (**a**) and (**b**) are the 90 CEPH arrays for two exemplary SNPs. Colors and shapes indicate SAMS genotypes and uncertainties, respectively. While both SNPs show well-formed genotype clusters, the second SNP (panel b) is difficult for SAMS to genotype correctly due to allele signal imbalance. Panels (**c**) and (**d**) display the silhouette widths for the second SNP based on SAMS (c) and MASS (d) genotypes.

for MASS genotypes. Figure 3c and d display the silhouette widths for each individual array according to SAMS and MASS calls, respectively. As noted, the silhouette widths are much improved for 15 heterozygous samples under MASS genotyping.

3.4 Accuracy of MAMS genotyping

We measure the accuracy of MAMS genotypes by computing its concordance with HapMap reference calls and compare it to DM for 100 K arrays. Excluding the HapMap no calls, the concordance between MAMS and HapMap genotyping is 99.70 and 99.62% for *Hind*III and *Xba*I arrays, respectively. The corresponding concordance between DM and HapMap is slightly lower: 99.63 and 99.61%. The improvement in accuracy was achieved by reducing misclassification for known heterozygous bases while maintaining classification accuracy for known homozygous bases (see Table 1). For both *Xba*I and *Hind*III (100 K) arrays, the DM algorithm was integral to probe design and selection. Hence, performance results for DM are overly optimistic. To obtain an unbiased comparison and, more importantly, to test the extensibility of MAMS to a much larger data set, we evaluate the performance of MAMS using the recently released 500 K data set consisting of 39 CEPH samples and compare it to both DM and BRLMM. The genotype concordance between MAMS and HapMap is 99.57%, whereas BRLMM is 99.59% and DM has a notably worse accuracy of 99.21%. Table 1 displays the relation between concordance and the call rates calculated using default DM and BRLMM settings for the three algorithms. We chose a threshold for MAMS confidence metric that would lead to comparable call rate and accuracy to those given by default DM and BRLMM settings.

3.5 SNP quality measurements

Even though during the development of the 100 K SNP arrays Affymetrix conducted probe screening at both the SNP and quartet level, some SNPs still perform consistently worse than

Table 1. Comparison of MAMS, DM and BRLMM genotyping concordance with HapMap reference calls based on 90 *Hind*III and 39 *Nsp*I arrays

Array	Method	Cutoff	Call rate (%)	Concordance (%)		
				Overall	Hom	Het
Nsp I 250K	DM	1	100	99.21 (0.28)	99.32 (0.27)	98.95 (0.43)
	BRLMM	1	100	99.59 (0.11)	99.65 (0.11)	99.47 (0.14)
	MAMS	1	100	99.57 (0.11)	99.67 (0.10)	99.31 (0.16)
	DM	0.33	96.27 (1.19)	99.66 (0.08)	99.72 (0.08)	99.50 (0.15)
	BRLMM	0.5	99.70 (0.14)	99.69 (0.06)	99.74 (0.06)	99.55 (0.11)
	MAMS	0.3	99.37 (0.23)	99.67 (0.07)	99.77 (0.05)	99.45 (0.14)
Hind III 50K	DM	1	100	99.63 (0.15)	99.81 (0.05)	99.20 (0.27)
	MAMS	1	100	99.70 (0.11)	99.77 (0.13)	99.53 (0.18)
	DM	0.33	99.50 (0.32)	99.74 (0.07)	99.84 (0.03)	99.50 (0.23)
	MAMS	0.3	99.58 (0.23)	99.74 (0.07)	99.83 (0.06)	99.59 (0.15)

Number in parentheses are standard errors.

others. Of course, one would like to identify such SNPs and to this end we propose two SNP quality measurements.

3.5.1 Relative signal strength (RSS) Research in expression arrays indicates that low intensity probes are, in general, not reliable (Huber *et al*., 2002). To investigate the effect of signal strength on genotyping confidence, we devised a measure 'Relative Signal Strength' (RSS) for MM-adjusted average allele signal strength. It is defined as $\mathrm{RSS} = \max(\theta_A, \theta_B)$, the maximum of the two MM adjusted allele signals. Due to copy number differences between homozygous bases and heterozygous bases, signals tend to be lower for heterozygous genotypes. Therefore, we examined the relationship between RSS and MAMS genotype accuracy separately for homozygotes and heterozygotes in Figure 4a. We divided SNPs according to their RSS percentiles and calculated the accuracy within these subsets. For instance, the leftmost points in both curves show that accuracies of the SNPs whose RSS values are among the lowest 1% are 94.2 and 97.4% for AA/BB and AB, respectively. For AA/BB, genotyping accuracy improves monotonely with increasing RSS and reaches 99.5% for the top 90% of the SNPs whose PM signals are > 2-fold as strong as MM signals. Most impressively, the accuracy exceeds 99.9% for the top 60% SNPs. The trend for the heterozygous bases, however, differs in two ways. First, accuracy peaks at a lower level (99.6%) reflecting the intrinsic difficulty in classifying heterozygous genotypes. Second, accuracy drops to an average of 99.3% for the top 10% of SNPs that have the highest RSS.

3.5.2 Signal bias Further investigation revealed that the decrease in accuracy for heterozygous SNPs with high RSS is linked to disparate signal intensities between alleles A and B. We plotted the absolute values of θ_{AB} against MAMS genotyping accuracies for the heterozygous bases in Figure 4b. As expected, when the difference between the intensity of the two alleles is more than 2-fold, accuracy decreases appreciably. It is particularly strikingly that few correct calls are made when $\theta_{AB} \geq 1.5$, which translates to a 3-fold difference in signal intensity between alleles A and B. To characterize the SNPs that display strong signal bias between the two alleles we conducted a two-sided test on the vector $\theta_{AB,ik}$ for the ith SNP under the null hypothesis, $H_0 : \mathrm{mean}(\theta_{AB,ik}) = 0$. Controlling the false discovery rate at 0.1, there are 103 SNPs showing strong signal bias and hence prone to misclassification. The identification of these SNPs might have important consequences for other applications of SNP arrays, for instance, estimation of allele frequencies from pooled DNA (Meaburn *et al*., 2006). These highly allele-imbalanced SNPs might also be the results of copy-number polymorphisms that warrant further investigation (Iafrate *et al*., 2004).

3.6 Array quality measurements

In SNP genotyping experiments, it is important to obtain array quality measures so that decision on whether to repeat an array can be made. The single-array, clustering-based algorithm of SAMS yields a natural quality metric based on the separation of the three genotype clusters in the 3D θ space. If the three clusters are well separated, the uncertainty measures for an overwhelming majority of the SNPs will be small and their associated silhouette widths will be large. Accordingly, we devised two array quality measures: (1) SAMS call rate, which is defined as the percentage of SNPs having uncertainty values smaller than a cutoff; (2) median silhouette width summarizing all SNPs on the array. As all of the 90 *Hind*III and 39 *Nsp*I HapMay arrays provided by Affymetrix are of superior quality, we used one HapMap array (*Nsp*1 in Fig. 5) and three other *Nsp* arrays of suboptimal-quality samples produced at the UCSF Genomics Core Facility (*Nsp*2-4 in Figure 5) to demonstrate the utility of these two metrics. Note that the array *Nsp*4 was used as a negative control by hybridizing a sample to the array that did not contain any genetic material. Figure 5a depicts the distribution of θ_{AB} values for all four arrays and deterioration of the separation of the three genotype clusters is apparent and for *Nsp*4, as expected, there is no discernable groups. The SAMS call rates at uncertainty cutoff 0.3 for these four arrays are depicted in Figure 5b and are, 99.2, 90.4, 75.4 and 46.7%, respectively and are comparable with DM call rates. A SAMS call rate (at uncertainty 0.3) higher than 90% suggests an array

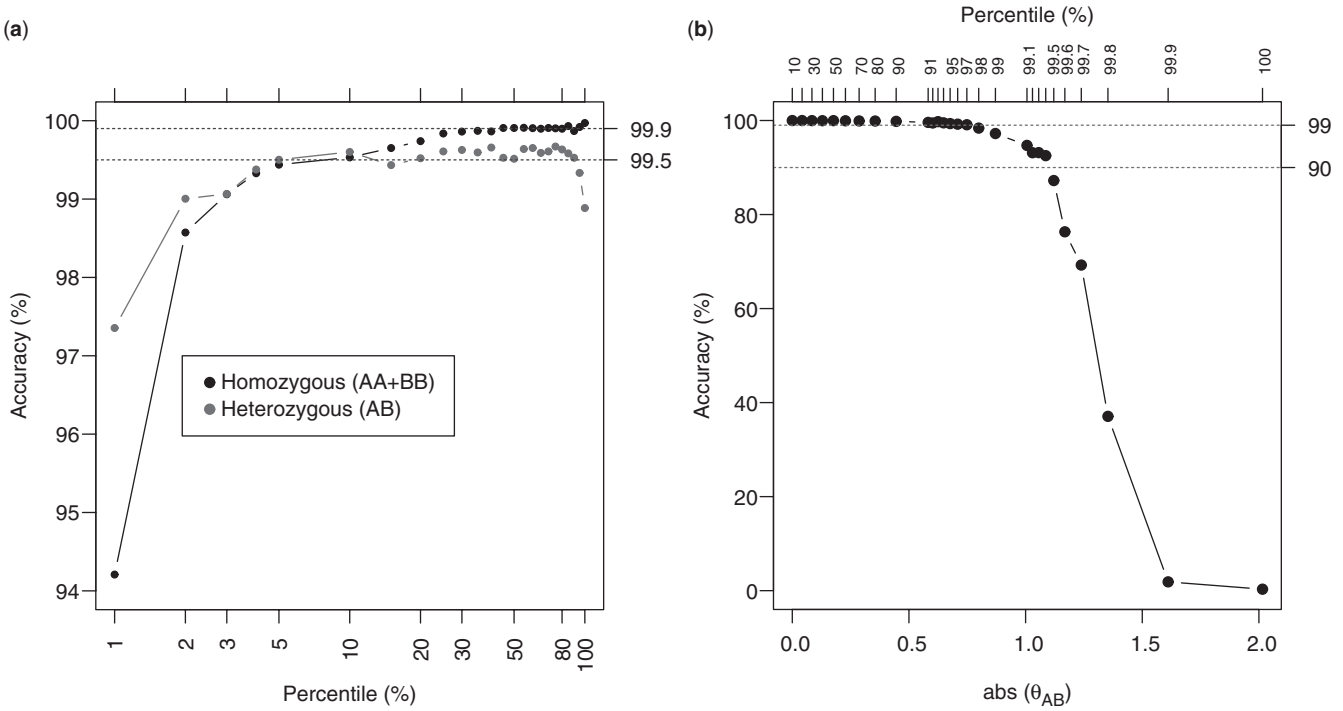

Fig. 4. SNP quality measurements. Panel (**a**) depicts the relationship between RSS and accuracy. SNPs are categorized according to RSS and accuracies are measured within each subset. Panel (**b**) depicts the relationship between signal bias and accuracy for heterozygous genotypes.

is of high quality. Median silhouette widths in Figure 5 are not as sensitive as SAMS call rates, but can be used as a reference.

3.7 Prediction accuracy versus number of quartets

The 100 K SNP array uses 10 probe quartets with the polymorphic nucleotide having different shifts from the center of the probe sequence. We investigated the possibility of reducing the number of quartets for each SNP since so doing enables enlargement of the pool of SNPs that could be queried on an array with the same density. Our strategy was to randomly choose 1, 3 or 5 quartets for each SNP and then subject the reduced data to the same MAMS procedure. Figure 1 in Supplementary Material displays the distribution of θ_{AB} and θ_A in panels (a) and (b), respectively. With only one quartet, the parameter θ_{AB} from the three genotypes are still able to form distinctive, albeit overlapping distributions. However, θ_A has become much less discriminatory in separating genotype *BB* from *AB* and *AA*. The overall accuracies using the 90 CEPH *Hind*III arrays from MAMS are 96.73, 98.84, 99.23 and 99.69% for 1, 3, 5 and 10 quartets, respectively. It is therefore conceivable to use only 5 quartets for genotyping arrays, as accuracy remains high.

4 DISCUSSION

The MAMS genotying algorithm consists of two components, which operate at different data levels. The first component is the SAMS approach. By employing a sampling method, SAMS is scalable to large SNP arrays, and by operating within a single array, it does not require large sample sizes to make genotyping feasible. The algorithm takes about an hour to run 90 *Hind*III arrays on a 1.4 GHz and 1 GB RAM Dell PC. The running time can be shortened by skipping the model selection step and directly estimating parameters for the most complex model. This can be done safely owing to the abundance of data points. The second component of MAMS employs a MASS clustering algorithm, analogous to similar predecessor genotyping algorithms. By adding this multi-array approach and by employing silhouette scores as an objective assessment of genotype quality of SAMS and MASS, MAMS is able to adaptively handle subsets of SNPs that exhibit idiosyncratic hybridization behavior while retaining genotying accuracies for the multitude of well-behaved SNPs. Similar operational procedures are also employed in BRLMM and consequently both algorithms display superior genotyping accuracy compared to DM. Unlike BRLMM, which requires specification of parameters for clustering space transformation and weights for the prior distribution, MAMS relies on practically no tuning parameters. One could, however, opt to set a user-determined threshold of the difference in silhouette scores between SAMS and MASS, above which the algorithm accepts genotyping results of MASS. However, the optimal threshold is dependent upon number of samples and is not recommended when sample size is small (< 30).

Our MAMS genotyping algorithm operates at the SNP intensity level by summarizing probe level data into self-normalizing SNP level indices. These θ indices are essentially ratios of intensities, and are therefore not sensitive to between-array variation. This property obviates the need for

normalization. However, when cross-lab arrays need to be integrated, there might be sample preparation effects that differ from lab to lab that can lead to accentuated between-array variation (Nannya *et al*., 2005 and Carvalho *et al*., 2006). In such cases, we strongly recommend investigating if batch effects exist before carrying out a genotyping analysis. Some appropriate strategies for normalization are described in Carvalho *et al*. (2006). We note that these can incur a substantial computational burden.

We also found that, unlike expression arrays, incorporating MM information into the indices is desirable. We applied MAMS using only the θ_{AB} index, which contains no MM information. This led to a decrease in prediction accuracy from 99.69 to 99.30% for the *Hind*III arrays. However, because MM signals retain a large portion of allele-specific hybridization signals, they need to be robustly summarized across alleles so as to account for signal disparities between the two alleles. To illustrate this finding, we compare the densities of various intensity summaries of allele *A* in Figure 6a. PMs alone clearly do not provide sufficient information for genotyping discrimination, which implies that hybridization intensities of the PMs are highly SNP dependent. Comparing MM and MMA adjusted allele *A* signals (θ_A = PMA–MM, black solid line, versus PMA–MMA, black dotted line) indicates that θ_A yields the most discriminatory power. A similar comparison supports the choice of θ_{AB} = PMA–PMB as shown in Figure 6b. Analogous findings on the positive effects of MM probes on SNP genotyping was also reported in LaFramboise *et al*. (2005). Even though we and others showcased the values of MMs in SNP genotyping, MAMS can easily accommodate other indices that do not make use of MM probes, for instance, the 2D indices used in Affymetrix (2006) and Carvalho *et al*. (2006). Our stepwise procedure also makes it flexible with incorporating other methods in one of the steps, for instance, BRLMM's Bayesian component can replace MASS if desired.

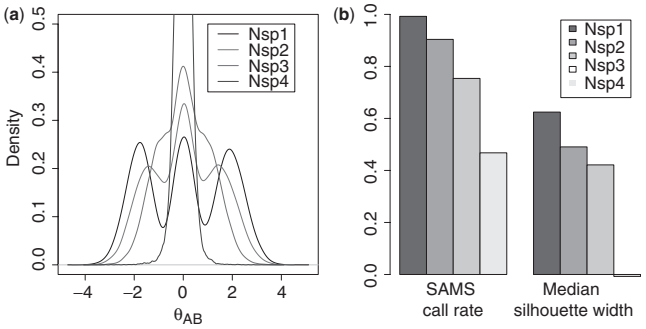

Fig. 5. Array quality measurements. Panel (**a**) plots the distribution of θ_{AB} and panel (**b**) plots the two quality metrics for the four arrays. SAMS call rate was calculated for an uncertainty cutoff at 0.3.

The performance of MAMS is superior to DM and comparable to BRLMM. The improvement over DM is primarily due to substantial gain in accuracy of heterozygous bases. However, this improvement still leaves room for further enhancement, especially for SNPs with MAF. We found that for SNPs that have a lower than 10% MAF, genotyping accuracy for MAMS is 0.55% lower than the other MAF categories (for which accuracies are essentially constant), whereas for BRLMM, it is 0.35% lower (see Fig. 2 in Supplementary Material). Even though minor bias against SNPs with low MAF could be ameliorated by increasing sample size for both BRLMM and MAMS, how to improve performance without making recourse to larger samples is an area of future research.

Two important by-products of MAMS are the two SNP quality measures that quantify the prediction performance on a per SNP basis. SNiPer-HD Hua *et al*. (2006) also furnishes a quality control metric for each assayed SNP based on silhouette widths. Such use is best suited to large (hundreds of samples), genome-wide SNP association studies since the magnitudes of silhouette scores is highly dependent upon sample size, and is also biased unfavorably toward SNPs with low MAF. Our use of silhouette scores differs in that they are employed to identify potentially misclassified SNPs by SAMS. Further, our SNP quality measures can identify SNPs that exhibit an inequality of allele intensities, which renders heterozygous genotypes more susceptible to misclassification. The biochemical basis for this

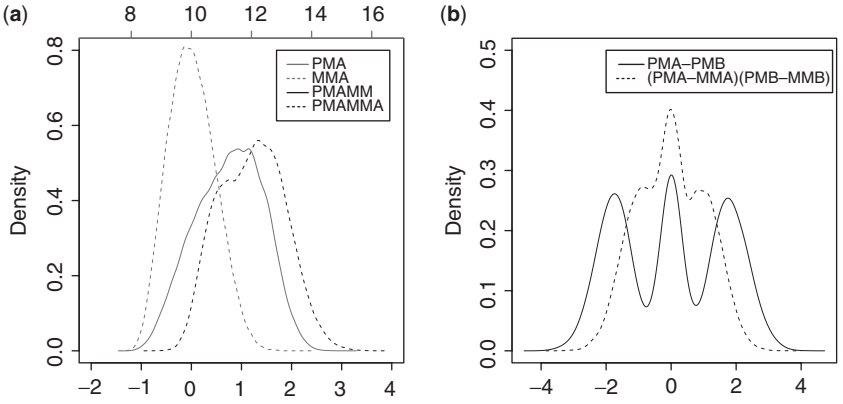

Fig. 6. Density plots of various intensity summaries from a *Hind*III array. Panel (**a**) plots the distribution of PM and MM signals (PMA and MMA, respectively), and MM and MMA adjusted signals (PMA–MM and PMA–MMA, respectively). All four summaries based on allele A. Panel (**b**) compares the distributions of relative allele signals with or without allele-specific MM adjustment (PMA–PMB) and (PMA–MMA), (PMB–MMB), respectively.

behavior could include extreme G-C content of the probes, type of the polymorphic base pair and suboptimal PCR amplicon size. We investigated the effect of polymorphic base-pair type on discriminative power as approximated by θ_{AB}. Boxplots of the distribution of θ_{AB} for genotypes AB stratified by the six polymorphic base-pair groups — AC, AG, AT, CG, CT and GT — are displayed in Supplementary Figure 3. It is clear that θ_{AB} displays systemic trends depending on the type of the two alleles. When alleles A and B form base pairs of the same strength, as measured by the number of hydrogen bonds — i.e. for AT and CG — θ_{AB} centers approximately at zero. However, when there is an inequality in the number of hydrogen bonds between the two alleles, θ_{AB} leans in the direction of the base that forms more hydrogen bonds. For instance, when allele A is 'A' and allele B is 'C', θ_{AB} is more likely to be negative. How this influences discriminatory performance is the subject of further investigation. Interestingly, we did not find an enrichment of 'AC','AG','CT' and 'GT' polymorphic pairs in the SNPs that show an imbalance in the signal intensities of the two alleles.

ACKNOWLEDGEMENTS

We thank CBMB, UCSF for funding support and Dr Joseph Wiemels for providing Nsp I arrays.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2006) BRLMM: an improved genotype calling method for the genechip human mapping 500 k array set. *Technical report*. Affymetrix, Inc. White Paper.

Banfield,J.D. and Raftery,A.E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.

Carvalho,B. *et al*. (2006) Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. *Technical report*. Johns Hopkins University.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.

Di,X. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100 k SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *JASA*, **97**, 611–631.

Hua,J. *et al*. (2006) SNiPer- HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics*, **23**, 57–63.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differetial expression. *Bioinformatics*, **1**, 1–9.

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet*., **39**, 949–951.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

LaFramboise,T. *et al*. (2005) Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput. Biol*., **1**, e65.

Liu,W.-M. *et al*. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397–2403.

Meaburn,E. *et al*. (2006) Genotyping pooled dna using 100 k SNP microarrays: A step towards genomewide association scans. *Nucleic Acids Res*., **34**, e28.

Nannya,Y. *et al*. (2005) A robust algorithm for copy number detection for high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*., **65**, 6071–6079.

Nicolae,D.L. *et al*. (2006) GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics*, **22**, 1942–1947.

Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*., **20**, 53–65.

The International HapMap Consortium (2003) The international hapmap project. *Nature*, **426**, 789–796.