Biostatistics (2007), **8**, 2, *pp.* 485–499
doi:10.1093/biostatistics/kxl042
Advance Access publication on December 22, 2006

Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data

BENILTON CARVALHO

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

HENRIK BENGTSSON

Department of Statistics, University of California, Berkeley, CA, USA

TERENCE P. SPEED

Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute, Melbourne, Australia and Department of Statistics, University of California, Berkeley, CA, USA

RAFAEL A. IRIZARRY*

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
ririzarr@jhsph.edu

SUMMARY

In most microarray technologies, a number of critical steps are required to convert raw intensity measurements into the data relied upon by data analysts, biologists, and clinicians. These data manipulations, referred to as preprocessing, can influence the quality of the ultimate measurements. In the last few years, the high-throughput measurement of gene expression is the most popular application of microarray technology. For this application, various groups have demonstrated that the use of modern statistical methodology can substantially improve accuracy and precision of the gene expression measurements, relative to ad hoc procedures introduced by designers and manufacturers of the technology. Currently, other applications of microarrays are becoming more and more popular. In this paper, we describe a preprocessing methodology for a technology designed for the identification of DNA sequence variants in specific genes or regions of the human genome that are associated with phenotypes of interest such as disease. In particular, we describe a methodology useful for preprocessing Affymetrix single-nucleotide polymorphism chips and obtaining genotype calls with the preprocessed data. We demonstrate how our procedure improves existing approaches using data from 3 relatively large studies including the one in which large numbers of independent calls are available. The proposed methods are implemented in the package oligo available from Bioconductor.

Keywords: Affymetrix; Genotyping; High-throughput; Microarrays.

*To whom correspondence should be addressed.

1. INTRODUCTION

The genotyping platform provided by Affymetrix interrogates hundreds of thousands of human single-nucleotide polymorphisms (SNPs) on a microarray. A simple description of the method is the following: DNA is obtained and fragmented at known locations so that the SNPs are far from the ends of these fragments, the fragmented DNA is amplified with a polymerase chain reaction (PCR), and the sample is labeled and hybridized to an array containing probes designed to interrogate the resulting fragments. There are currently 3 products available from Affymetrix: an array covering approximately 10 000 SNPs (GeneChip Human Mapping 10K), a pair of arrays covering approximately 100 000 SNPs (GeneChip Human Mapping 50K Xba array and Hind Array), and a pair of arrays covering approximately 500 000 SNPs (GeneChip Human Mapping 250K Nsp Array and Sty Array). These are referred to as the 10K, 100K, and 500K chips, respectively. The 100K chips have become widely used (Uimari *and others*, 2005; Nannya *and others*, 2005; Huang *and others*, 2006). The main application of this technology is genotyping SNPs at a high-throughput rate. However, various groups have used the arrays for other applications such as copy number estimation (Huang *and others*, 2006; Nannya *and others*, 2005). In this paper, we focus on preprocessing algorithms that can improve downstream analysis for any of these applications. We illustrate these using the main application of this technology, genotyping.

We start this section with a short description of the SNP chip feature-level data. A detailed description is available from Kennedy *and others* (2003). Each SNP on the array is represented by a collection of probe quartets. In the 100K arrays, SNP chips probe sets are composed of 40 features. As with expression arrays, the features are defined by 25-mer oligonucleotide molecules referred to as probes. There are 20 perfect match (PM) paired with 20 mismatch (MM) probes. As in expression arrays, these are created by changing the middle base pair. A difference with expression arrays is that the PM features differ in 3 important ways: First, 2 alleles are interrogated (for most SNPs only 2 alleles are observed in nature). These are denoted by A and B and divide the probes into 2 groups of equal size. For each PM probe representing the A allele, there is an allele B that differs by just 1 base pair (the SNP). Second, features are included to represent the sense and antisense strands. This difference divides the probes into 2 groups that are not necessarily of the same size. Finally, for each allele/strand combination, various features are added by changing the position of the SNP within the probe. In summary, we have 4 discriminating characteristics: PM or MM, allele A or B, sense ($-$) or antisense ($+$), and SNP location. Our methodology makes no use of the MM features mainly because we see a trend in the company no longer to use this type of probe. Note that an array with no MMs can accommodate features for twice as many SNPs.

The general goal of preprocessing for SNP arrays is to normalize and summarize feature intensities and predict the genotype AA, AB, or BB. These predictions will be referred to as "genotype calls." A measure of confidence is also desired. Typically, samples not achieving a specific confidence cutoff at a given SNP receive no calls at that SNP. In this paper, we propose a preprocessing methodology that greatly improves the accuracy of genotyping calls over existing methods. We propose a modular approach in which preprocessing is done in a first step, and a genotyping algorithm is defined for preprocessed data. To illustrate this and to motivate our methodology, we use 3 100K data sets: 1) The HapMap (CEPH) Trio data set, consisting of 30 trios, which is also part of the International HapMap Project and, therefore, has precise genotype calls that can be used as "gold standard," 2) a data set comprised of the same DNA hybridized to 53 arrays, and 3) a data set consisting of 22 randomly selected samples from the data described in Slater *and others* (2005). We will refer to these data sets as the Lab 1, Lab 2, and Lab 3 data sets. The Lab 1 data set will also be referred to as the HapMap data.

The paper is organized as follows: Section 2 describes the previous work in preprocessing and genotyping methods, while Section 3 describes how we normalize and summarize the feature-level data. In Section 4, we show how the normalization we use motivates a useful genotyping algorithm, while in Sections 5 and 6 we present and discuss our results.

2. Previous work and motivation

The principal goal of preprocessing is to summarize the feature intensities into quantities that can be used to discriminate genotype classes. We use a general notation in which θ_A and θ_B are the logarithms (base 2) of quantities proportional to the amount of DNA in the target sample associated with alleles A and B, respectively. Note that if the PCR produced $\mathcal{X}$ copies of the DNA fragments, these quantities should, up to an additive constant, equal the logarithm of 0, $\mathcal{X}$, or $2\mathcal{X}$. Thus, a naive approach to genotyping would be to set thresholds and call genotypes based on the θs being above or below these thresholds. For example, to call an AA genotype, one might require that $\theta_A > C_A$ and $\theta_B < C_B$. However, already the most basic data exploration shows that such an approach will not work well in general. Figure 1 illustrates the problem. Given what we have learned from expression arrays about optical background noise, nonspecific binding, and probe effects, it is no surprise that such naive methods do not perform well. We begin this section by describing some of the more sophisticated existing genotyping algorithms.

Although predefined cutoffs are not useful, for most SNPs the values (θ_A, θ_B) from multiple samples form 3 distinct clusters representing the 3 possible genotypes. Affymetrix's default algorithm for their 10K arrays took advantage of this property and used a modified partitioning around the medoids (MPAM) clustering algorithm to detect the clusters. These clusters were then associated with 3 different genotypes. The summarized data were based on a relative allele signal which is essentially a ratio of allele A intensities to the sum of both allele intensities. The intensities were corrected for background using the MMs (Liu *and others*, 2003). The algorithm worked well when there were enough data in each of the 3 genotypes, but not as well in other cases. With the higher density chips, this algorithm was not satisfactory as many SNPs with low minor allele frequency are included in the 100K and 500K arrays (Di *and others*, 2005). For this reason, with the release of the 100K arrays, Affymetrix changed their default procedure to a "dynamic model" (DM) -based algorithm. In this algorithm, 4 different Gaussian models (NULL, AA, AB, and BB) were considered for the probe intensities for each SNP, and a genotype call was made for each sample based on the likelihoods for each genotype. Note that DM is not a modular procedure: the calls are derived directly from the feature intensities on each array separately.

Various problems have been noted with calls obtained from the DM algorithm. In particular, a higher degree of misclassification for the heterozygous calls was observed when compared to MPAM. This

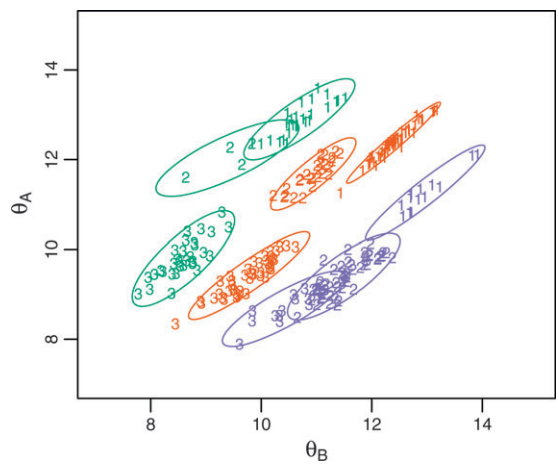

Fig. 1. Genotype regions for 3 SNPs with data from different HapMap samples shown as well. Symbols of different sizes represent the 3 different genotypes and the numbers the 3 different SNPs. The data and regions are obtained with RLMM which is described later in the text.

fact motivated several academic groups to develop their own algorithms (LaFramboise *and others*, 2005; Rabbee and Speed, 2006; Lamy *and others*, 2006). In Rabbee and Speed (2006), the robust linear model with Mahalanobis distance (RLMM) is described and shown to outperform DM on the HapMap data set described above.

RLMM, which is a multiarray procedure, begins by preprocessing the feature-level data using Robust Multiarray Average (RMA) (without background correction), a procedure demonstrated to work well for expression arrays (Irizarry *and others*, 2003). These summarized data are then used to build SNP-specific 'regions' for each genotype using a supervised learning algorithm similar to linear discriminant analysis. To train the algorithm, the HapMap data set was used. This approach is particularly appealing because empirical results demonstrate that different SNPs can produce very different distributions. Figure 1 clearly demonstrates this. Model-based approaches that impose the same (or similar) models on all SNPs as well as algorithms that train on observed data are unlikely to perform. In fact, using cross-validation on the HapMap data set, Rabbee and Speed (2006) demonstrate that RLMM greatly outperforms DM (See Figure 4 in Rabbee and Speed (2006)). However, this classification strategy makes RLMM's genotyping algorithm less useful because SNP-specific feature intensity distributions are different not only across SNP but also within the same SNP across labs/studies. Figure 2(A) clearly shows this. SNPs exhibiting the behavior shown in this figure are common, which implies that regions defined with data from one study/lab will do poorly when applied to data from a different study/lab.

Recently, Affymetrix made a white paper available (Affymetrix, 2006) describing a new preprocessing algorithm based on RLMM. To improve the across-lab compatibility, Bayesian Robust Linear Model with Mahalanobis distance (BRLMM) does not train the classification algorithm on the HapMap data. Instead, BRLMM uses DM calls as initial guesses for class membership and uses these to define genotype regions. The genotype regions are then recalibrated using a Bayesian approach. This algorithm is expected to become their default in the near future. More details are available from Affymetrix (2006).

In this paper, we describe new normalization and summarization methodologies that make across-lab comparison possible. This in turn permits us to use the training algorithm strategy originally implemented by RLMM to create a powerful corrected version. We will refer to our genotyping method as Corrected

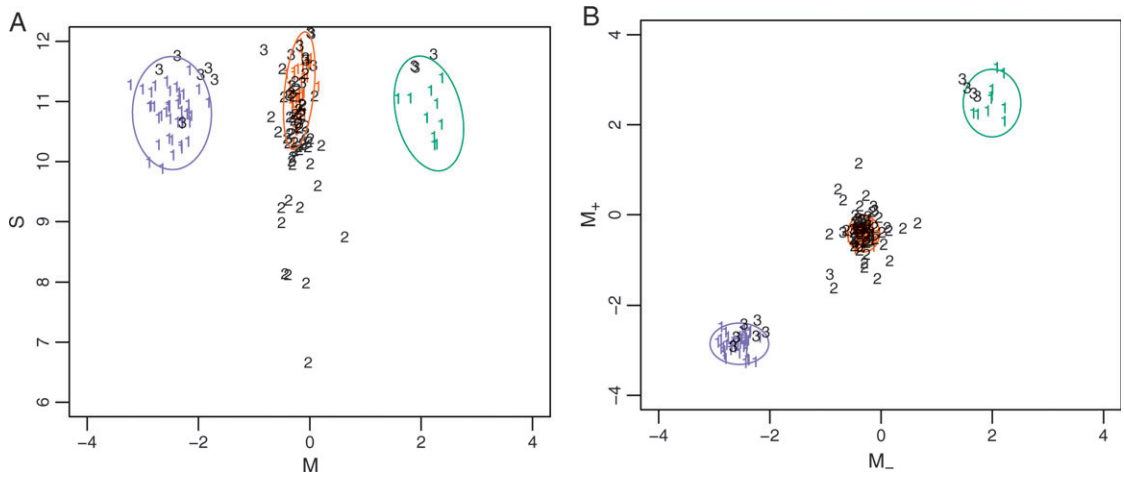

Fig. 2. Genotype regions. (A) RLMM genotype regions obtained using the HapMap data (denoted with 1) and data points from Labs 2 and 3 (denoted by numbers). We plot $S = (\theta_A + \theta_B)/2$ versus $M = \theta_A - \theta_B$ to facilitate comparison with BRLMM and CRLMM. (B) As (A) but for CRLMM. Note that RLMM, BRLMM, and CRLMM are defined in different parts of the text.

Robust Linear Model with Maximum Likelihood Distance (CRLMM). Because our preprocessing method is an adaptation of RMA and can be used with other genotyping algorithms, we will refer to it as SNP-RMA. Below we give a summary of the algorithm, and in the remaining sections we motivate and give further details for each step.

1) For each array, we estimate probe sequence and DNA fragment length effects and remove them from the log feature intensities.
2) We use quantile normalization against a reference sample to remove some of the unwanted array-to-array variation.
3) For each SNP, each of the 2 alleles, and each of the 2 strands, we form a summary over the 20 PM features using a linear model as in RMA.
4) For each strand, the log differences between the A and B allele intensities are calculated. We then remove probe sequence, fragment length, and total intensity effects on the log-ratios. Because these effects are genotype dependent, we use a mixture model which assumes that each unknown genotype results in a different Gaussian distribution.
5) Using the HapMap data as training set, where the genotypes are (for most SNPs) known, we estimate for each SNP means and variances for the log-ratios corrected for the effects estimated in the previous step. A mixed-effects model is used to obtain empirical Bayes estimates.
6) For a new sample and for each SNP, we predict the genotype as the one maximizing the likelihood calculated as though the means and variances derived above are known. Likelihood ratios are used as uncertainty measures.

3. Normalization

A likely explanation for the across-lab differences in cluster distributions seen in Figure 2(A) is the sample preparation effect. In particular, the amplification of DNA through PCR is unique to each sample. In this section, we describe procedures based on observable covariates that can be used to assess and correct the PCR effect: probe sequence and fragment length. Similar corrections have been described by Nannya *and others* (2005). However, these corrections are done to improve the precision of copy number estimates. Here, we demonstrate that effects can still be observed for the allele log-ratio values even after correcting the log intensities. We propose normalization strategies that correct for these log-ratio biases with the goal of improving genotype calls.

3.1 *Correcting for sequence and fragment length*

Supplementary Figures 1 and 2 (available at *Biostatistics* online) demonstrate that fragment length has a strong negative effect on probe intensity, with longer fragments resulting in weaker feature intensities. These figures also demonstrate that the effects are different from sample to sample (seen through the confidence bands in Supplementary Figure 1 available at *Biostatistics* online) and from lab to lab (seen in Supplementary Figure 2 available at *Biostatistics* online), with the lab difference being greater. Nannya *and others* (2005) have also pointed out that the chemical composition of the probe has a strong effect on feature intensity. We have noticed that the sequence effect is position dependent, something that has previously been observed in expression arrays (Wu *and others*, 2004). Figure 3 shows the position-dependent effects of each of the 4 bases for data from 3 different labs. This figure demonstrates that the effects are large and that they change from lab to lab. A particularly important consequence of the sequence effect is that when comparing feature intensities representing the different alleles, one can see relatively large differences due only to sequence. Figure 4(A) shows that the sequence effect can cause relatively large differences between alleles A and B.

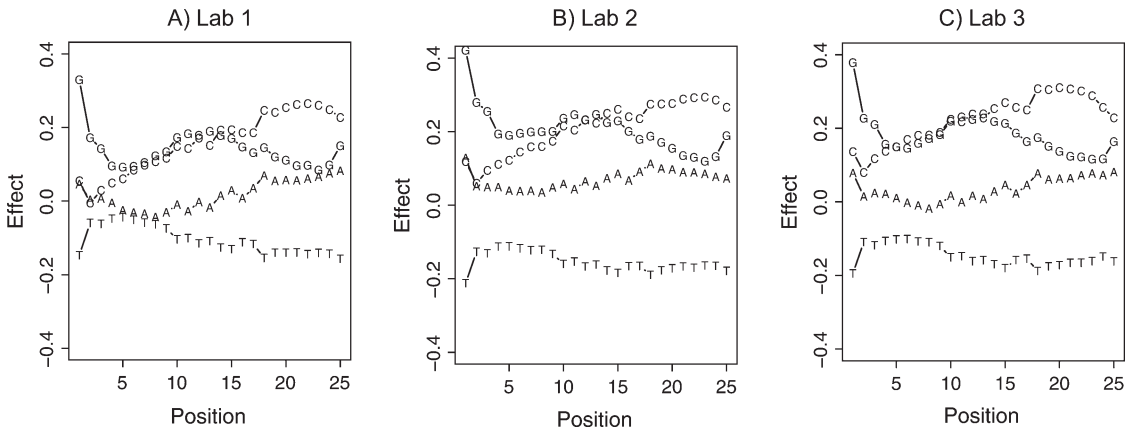

Fig. 3. Position-dependent sequence effects. For a typical array from the HapMap study, the effect of each base at each position is shown. To estimate this effect, we fit Model (3.1) to all PM intensities for all SNPs on the array with no smoothness assumptions on $h_b(t)$. The different bases are denoted with the respective initials. (B) As (A) but for Lab 2, and (C) Lab 3.

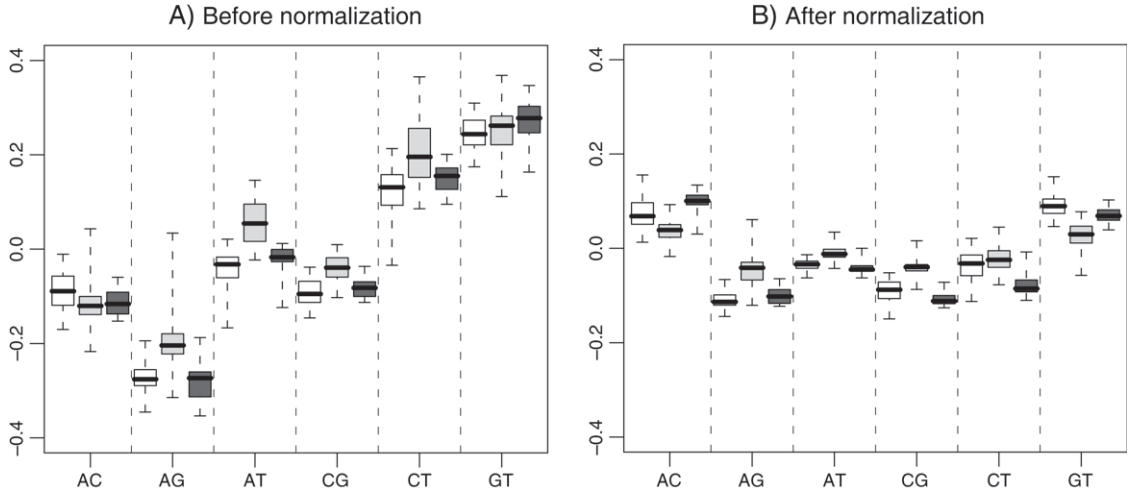

Fig. 4. Sequence effect on allele A to allele B log-ratios for the 6 different base pairs by which the 2 allele probes can differ. The plot shows median effects. Specifically, for each array we obtain the median of these log-ratios across all SNPs with the same pair of bases distinguishing the 2 allele probes. In this plot, we show box plots of those medians stratified by base pair and lab. The 3 different colors represent the 3 different labs: (A) before normalization and (B) after normalization.

In our normalization procedure, our first step is to correct for both sequence and fragment length effects. To do this, we fit a linear model to the log PM intensities:

$$\log_2(\mathrm{PM}) = \mu + g(L) + \sum_{b\in\{\mathrm{A,C,G,T}\}} \sum_{t=1}^{25} h_b(t)\mathbb{I}(b_t = b) + \xi. \tag{3.1}$$

Here, $b_t \in \{\mathrm{A, C, G, T}\}$ represents the base at location t, $h_b(t)$ are smooth functions of the location (each base b is represented by a different function), $\mathbb{I}(b_t = b)$ is 1 when the base at position t is b

and 0 otherwise, $g(L)$ is a smooth function of fragment length L, and ξ is a zero-mean random error which we assume is normally distributed. Supplementary Figures 1 and 2 (available at *Biostatistics* online) and Figure 3 demonstrated that the effects are well described with smooth functions which we model with cubic splines with 5 degrees of freedom. With these assumptions in place, we can estimate μ, $g(\cdot)$, and $h_b(\cdot)$ using least squares. The corrected PM intensities are obtained by subtracting the estimated sequence and fragment length effects for $\log_2(\text{PM})$. Nannya *and others* (2005) demonstrate that corrections such as these reduce unwanted variability substantially. However, in Section 3.4, we demonstrate that sequence and length effects remain for the quantity that is most informative for genotyping, the log-ratio. For example, Figure 4(B) shows that the effect of sequence is reduced but can be further improved.

3.2 *Across-array normalization*

An important lesson learned from analyzing expression data is that across-array normalization is almost always needed. Figure 5 demonstrates that even after the correction described in Section 3.1, array intensity distributions are substantially different. As expected, differences are seen across arrays and even bigger differences across labs. In the case of SNP arrays, it is safe to assume that the theoretical distributions of the target DNA we are measuring should be equal since the total amount of DNA should be the same across samples. Exceptions might come from cases for which a DNA sample has large pieces with extra or deleted copies of chromosome. For all other cases, we can make array intensities comparable across arrays using quantile normalization (Bolstad *and others*, 2003). However, instead of normalizing each study separately, as is commonly done in gene expression experiments, we normalize all array intensities to a reference distribution created with the HapMap data.

3.3 *Summarization*

We summarize the feature intensities within each probe quartet to produce 4 values for each SNP. Specifically, we follow the RLMM approach to fit a linear model (using median polish) to the normalized log PM intensities (Rabbee and Speed, 2006). The linear model includes a term related to sample-specific DNA amount and a term for the probe effect. However, here we fit a separate model to each strand/allele combination instead of combining the strands as done by RLMM. We therefore produce 4 numbers per SNP which we denote by $(\theta_{A,-}, \theta_{B,-})$ and $(\theta_{A,+}, \theta_{B,+})$. In Section 3.4, we describe why we keep sense and antisense values separate.

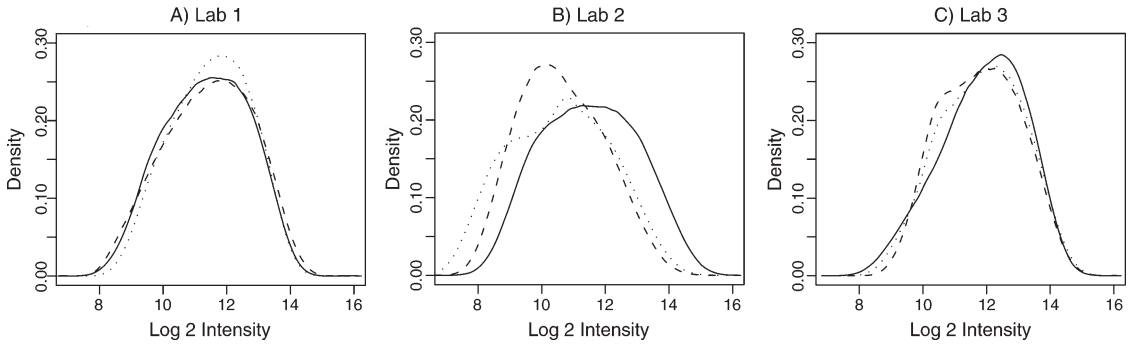

Fig. 5. Empirical densities for log (base 2) intensities from 3 randomly chosen arrays from (A) Lab 1, (B) Lab 2, and (C) Lab 3. The intensities have been corrected for sequence and fragment length.

3.4 *Remaining log-ratio biases*

Figures 1 and 2(A) show that most of the information available for separating the clusters associated with the 3 genotypes are in the upper-left-to-lower-right diagonal direction, that is the log-ratios. The same plots for other SNPs look similar. In fact, it is difficult to find cases where the sum of the intensities provides useful information. For this reason, we consider the log-ratios $M = \theta_A - \theta_B$ as the quantity used for genotyping. Furthermore, there are many instances where one of the 2 strands appears to provide no information. We refer to these as the "noninformative strands." Figure 6 demonstrates that considering the log-ratios for the 2 strands, M_+ and M_-, instead of a summary that contains both, permits us correctly to call genotypes in cases in which the features for one of the strands are noninformative. We have observed roughly 100 SNPs such as the one presented in Figure 6. For this reason, we propose strand-specific log-ratios as the summarized quantity to be used by genotyping algorithms. We denote the log-ratio for SNP i and sample j by $M_{i,j,s}$ with sense and antisense strands denoted by $s \in \{-, +\}$. We code the genotypes by $k = 1, 2, 3$ for AA, AB, and BB, respectively.

Careful data exploration demonstrated that, in general, these M values have powerful discrimination ability. However, we noticed that in some arrays the overall separation in M is better than in others, see Figure 7. We also noticed that, within arrays, SNPs with inferior separability were associated with long fragment lengths or high/low average intensity, $S \equiv (\theta_A + \theta_B)/2$, values, as illustrated in Figure 8. Furthermore, Figure 4 demonstrates that, although much reduced, a sequence effect is still present for log-ratios. In the remainder of this section, we describe our final preprocessing step which estimates these remaining biases.

We describe these effects with a simple mixture model. To simplify the fitting procedure, we estimate the model separately on each array and treat the sense and antisense features as independent and identically distributed. We therefore drop the j and s notations and write

$$[M_i | Z_i = k] = f_k(\mathbf{X}_i) + \varepsilon_{i,k}, \tag{3.2}$$

where the Z_i represent the unknown true genotype of SNP i with possible values $k = 1, 2, 3$ (AA, AB, BB), $\mathbf{X}_i$ represent covariates known to cause bias, f_k describe the effect associated with these covariates, and $\varepsilon_{i,k}$ an error term which we assume to be a normal random variable with mean 0 and variance τ_k^2. We constrain the model such that $f_{j,2} = 0$ and assume that $f_1 = -f_3$ and $\tau_1^2 = \tau_3^2$. We assume that this is a mixture model with mixing probabilities $\Pr(Z_i = k) = \pi_k$ across all SNPs. The π_k are estimated but not predicted.

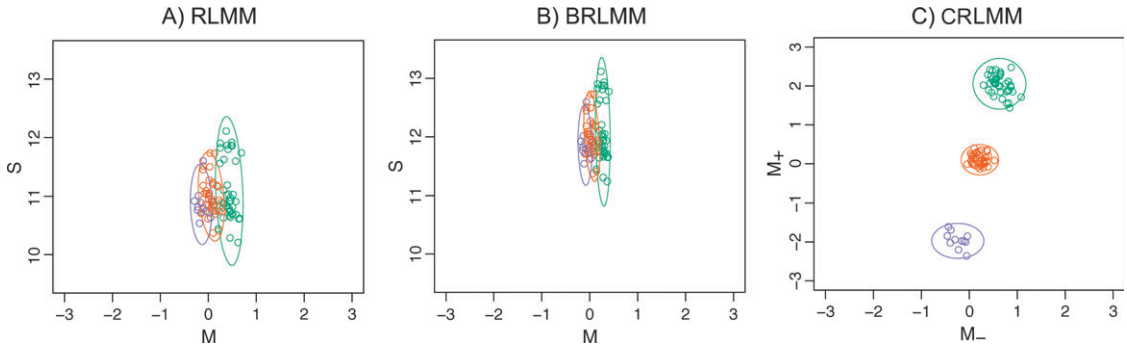

Fig. 6. Effect of the probes from the noninformative strands. The HapMap data were used. (A) RLMM, which averages sense and antisense strands, genotype regions for SNPs for which the sense strand does not differentiate. (B) As (A) but for BRLMM which also averages sense and antisense strands and (C) CRLMM, which keeps sense and antisense information separate.

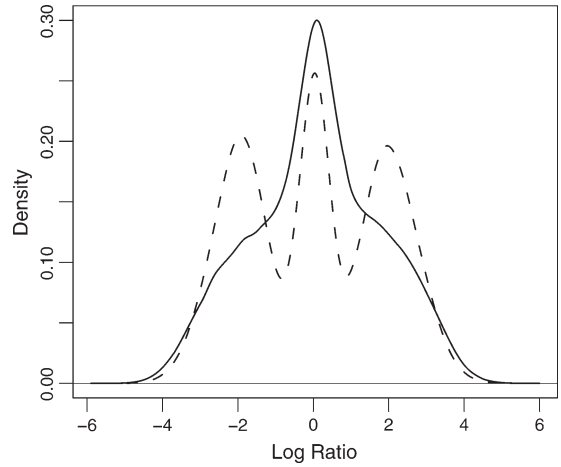

Fig. 7. Empirical density distribution of the across-SNP M values for the array with the best (dashed) and worst (solid) SNR ratios. The best data set came from the HapMap data and the worst came from Lab 2.

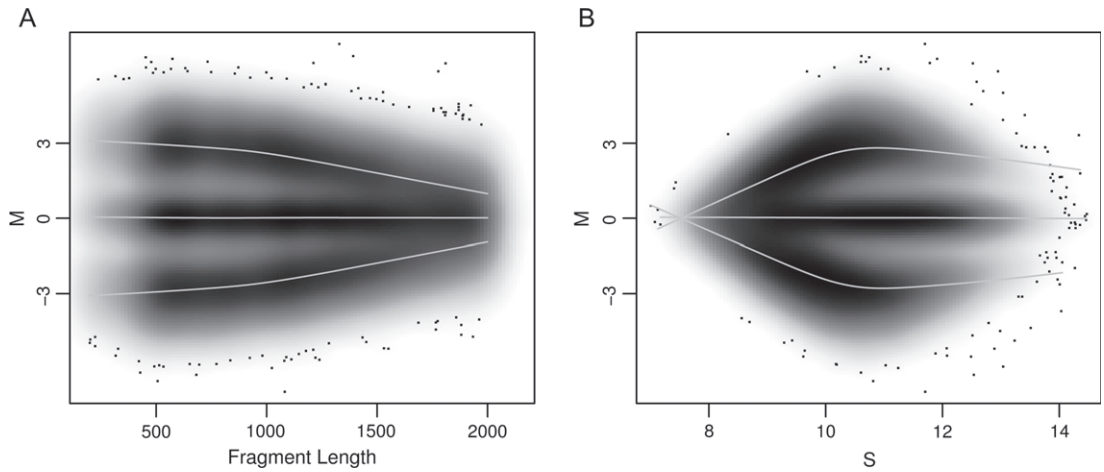

Fig. 8. (A) M values (both strands are included) from a typical array from the HapMap data plotted against fragment length. Instead of plotting the points we show, with different shades, the data density. The solid lines are the estimated f values from Model (3.2). (B) As (A) but for intensity instead of fragment length.

In this section, we have demonstrated that we should include at least the following 3 covariates in (3.2): fragment length L_i, average intensity S_i (treated as a fixed covariate), and a factor coding the base pair bp_i at the SNP. We therefore define $X = (L, S, \text{bp})$. Furthermore, Figures 4(B) and 8 suggest that we can model $f_1 = f_3$, in Model (3.2), as

$$f_1(L_i, S_i, \text{bp}_i) = \mu_{\text{bp}_i} + f_L(L_i) + f_S(S_i)$$

with μ_{bp_i} being a mean level that differs for each SNP base pair $\text{bp}_i \in \{\text{AC}, \text{AG}, \text{AT}, \text{CG}, \text{CT}, \text{GT}\}$, f_L a cubic spline with 3 degrees of freedom, and f_S a cubic spline with 5 degrees of freedom. This model has 16 parameters, and since we have thousands of observations, we obtain very precise estimates of f. With these assumptions in place, we fit Model (3.2) using the expectation maximization (EM) algorithm. Examples of the estimated f_L and f_S are included in Figure 8.

Although the main reason for fitting (3.2) is to obtain estimates of f, 2 other useful summaries can be derived. The first is an estimate of the probability of membership of sample j in genotype k for SNP i given $(M_{i,j,k,-}, M_{i,j,k,+})$. We denote these estimates by $\hat{\pi}_{i,j,k}$ and note that they are readily available from the EM algorithm as they are the weights used by the M step. In Supplementary Figure 3(A) (available at *Biostatistics* online), we compare the predicted probabilities to the actual error rates (computed using the HapMap data). The figure confirms that they are useful. Furthermore, we can use $\text{argmax}_k \hat{\pi}_{i,j,k}$ as a genotype call for SNP i on sample j. In Section 4, we describe how we sometimes use these calls as "initial guesses." Second, after fitting Model (3.2) for each array, we can compute the quantity $\text{SNR} = \text{median}(\hat{f}_1^2)/\text{avg}_k \hat{\tau}_k^2$, with the median calculated across SNPs. If data from different genotype classes are well separated, this signal-to-noise ratio (SNR) quantity will be large. For example, if this quantity is close to 0, it will be impossible to distinguish between heterozygous and homozygous. Thus, we can use SNR as an array-specific quality measure. In Supplementary Figure 4 (available at *Biostatistics* online), we demonstrate the utility of the SNR summary by showing plots like those in Figure 8 for the arrays producing the best and worst SNR. This figure shows that for the second array, information about genotypes is probably lost. We conjecture that a cutoff threshold C_{SNR} can be defined so that removing arrays with SNRs lower than C_{SNR} improves the overall performance of the analysis.

Note that even after fitting (3.2), we cannot correct the M values by subtracting f because we do not know genotype Z. In Section 4, we describe a genotyping algorithm that incorporates the estimated f.

4. GENOTYPE CALLING

As mentioned above, we use a supervised learning approach for genotype calling. For most SNPs on the arrays, we have independent genotype calls for all the samples in the HapMap data. These calls are based on consensus results from various technologies and are considered a gold standard. We use HapMap calls to define "known" genotypes which in turn permits us to define a training set. However, for the 100K data, these calls are not available for about 4% of the SNPs. For these, we use the initial guesses described in Section 3.4 to define the known classes. With the training data in place, we use a 2-stage hierarchical model and give likelihood-based closed-form definitions of the genotype regions as described below.

For each SNP, we define two-dimensional genotype regions based on the sense and antisense M values. However, even with 90 samples, there are genotype groups for some SNPs for which we have a very small number of observations available at the training step. For these cases, the hierarchical model presented in this section becomes very useful. Using empirically derived priors for the centers and scales of the other genotype regions, we give a closed-form empirical Bayes solution to predict centers and scales for cases with few or no observations.

4.1 *The model*

Let $Z_{i,j}$ be the unknown genotype for SNP i on sample j. As above, we code the genotypes by $k = 1, 2, 3$ for AA, AB, and BB, respectively. Figure 1 suggests that genotype regions are SNP-specific when considering (θ_A, θ_B) as the quantity of interest. Similar pictures for (M_-, M_+) (data not shown) demonstrate that the same is true for the log-ratios. Furthermore, these pictures suggest that the behavior of the log-ratio pairs can be modeled by bivariate normal distributions. We therefore propose a 2-level hierarchical multichip model with the first level describing the variation seen in the location of genotype regions across SNPs and the second, the variation seen across samples within each SNP. The model can be written out as follows:

$$[M_{i,j,s}|Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(\mathbf{X}_{i,j,s}) + m_{i,k,s} + \epsilon_{i,j,k,s}, \tag{4.1}$$

where $\mathbf{X}_{i,j,s}$ and $f_{j,k}$ are as in Section 3.4 but with the j and s notations reintroduced, $m_{i,k,s}$ is the SNP-specific shift from the typical genotype region centers, and $\epsilon_{i,j,k,s}$ represent the measurement error. As mentioned in Section 3.2, we expect different samples to have different biases, thus the effect function f now depends on j. Note that the SNP-specific covariates $\mathbf{X}$ also depend on the sample because the average signal S may vary from sample to sample. The m values represent the cluster center shifts not accounted for by the covariates included in $\mathbf{X}$.

To define the first level of our model, we denote the vector of SNP-specific region centers with $\mathbf{m}_i = (m_{i,1,+}, m_{i,2,+}, m_{i,3,+}, m_{i,1,-}, m_{i,2,-}, m_{i,3,-})'$. Data exploration shows that we can model the distribution of this vector with a multivariate normal distribution (Supplementary Figure 5 available at *Biostatistics* online). We will denote the variance–covariance matrix of $\mathbf{m}$ by V. Note that by definition, $\mathbf{m}$ is centered at $\mathbf{0}$ since the mean levels of the 3 genotypes are absorbed into f. This mean level, $J^{-1}\sum_j I^{-1}\sum_i f_{j,1}(\mathbf{X}_{i,j,s})$, is roughly 3.

The second level of the model, the variability seen within the genotypes for each SNP, is described by the ϵ values. We assume these to be independent (conditioned on the genotype Z) normals across samples and SNPs with SNP/strand-dependent variance $\sigma^2_{i,k,s}$. We use an inverse χ^2 prior to improve estimates when not enough data are available, that is

$$\frac{1}{\sigma^2_{i,k,s}} \propto \frac{1}{d_{0,k}s^2_{0,k}}\chi^2_{d_{0,k}},$$

where $d_{k,0}$ are the degrees of freedom of the χ^2-distribution and $s^2_{0,k}$ represent the variance of a typical SNP.

4.2 *The training step*

Because the large number of SNPs permits us to estimate the f_js precisely, for simplicity, we treat them as known. With this estimate of f_j in place for each sample, all we need to make our likelihood-based genotype calls are estimates of the m's and σ's in (4.1). In this section, we describe our proposed supervised learning approach. The key idea is to consider the HapMap calls as known genotypes and use this information to obtain maximum likelihood estimates (MLEs) of $\mathbf{m}$ and the σ values. A second step is to update these estimates with posterior means derived from the hierarchical model. Below, we describe the details.

Because we are treating $Z_{i,j}$ and f as known, we can define the MLEs for $\mathbf{m}$ and the σ values in closed form:

$$\hat{m}_{i,k,s} = N^{-1}_{i,k}\sum_{j\in\mathcal{J}_{i,k}}\{M_{i,j,s} - f_{j,k}(\mathbf{X}_{i,j,s})\}, \tag{4.2}$$

$$\hat{\sigma}^2_{i,k,s} = (N_{i,k}-1)^{-1}\sum_{j\in\mathcal{J}_{i,k}}\{M_{i,j,s} - f_{j,k}(\mathbf{X}_{i,j,s}) - \hat{m}_{i,k,s}\}^2. \tag{4.3}$$

Here, $\mathcal{J}_{i,k}$ is the set of indexes associated with samples of genotype k on SNP i and $N_{i,k}$ is the number of indexes in $\mathcal{J}_{j,k}$. Note that we may also use robust versions of (4.2) and (4.3).

As mentioned above, there are various cases for which not enough data are available to trust $\hat{m}$ and $\hat{\sigma}^2$ as reliable estimates of a region center and scale. The hierarchical model described in Section 4.1 provides closed-form solutions for the posterior means which can be viewed as a useful shrinkage of the estimates

that automatically take care of cases with few observations. The shrinkage step is defined as follows:

$$\tilde{\mathbf{m}}_i = (V^{-1} + \mathbf{N}_i \boldsymbol{\Sigma}^{-1})^{-1} \mathbf{N}_i \boldsymbol{\Sigma}^{-1} \hat{\mathbf{m}}_i, \tag{4.4}$$

$$\tilde{\sigma}^2_{i,k,s} = \frac{(N_{i,k} - 1)\hat{\sigma}^2_{i,k,s} + d_{0,k} s^2_{0,k}}{(N_{i,k} - 1) + d_{0,k}}, \quad \text{for } N_{i,k} > 1. \tag{4.5}$$

For $N \leqslant 1$, there is no sample variance to use in (4.5) and we simply use $\tilde{\sigma}^2_{i,k,s} = s^2_{0,k}$. Here, $\hat{\mathbf{m}}$ is the vector of sample means $(\hat{m}_{i,1,+}, \hat{m}_{i,2,+}, \hat{m}_{i,3,+}, \hat{m}_{i,1,-}, \hat{m}_{i,2,-}, \hat{m}_{i,3,-})'$, $\boldsymbol{\Sigma}$ is a 6×6 diagonal matrix with $\boldsymbol{\Sigma}_{k,k} = \boldsymbol{\Sigma}_{k+3,k+3} = s^2_{0,k}$, and $\mathbf{N}_i$ is a 6×6 diagonal matrix with entries $(N_{i,1}, N_{i,2}, N_{i,3}, N_{i,1}, N_{i,2}, N_{i,3})$. In order to apply (4.4) and (4.5), we need prior parameters $d_{0,k}$, $s^2_{0,k}$, and V. We use the empirical Bayes-type approach described in Lönnstedt and Speed (2002) and Smyth (2004).

Note that (4.4) and (4.5) are simply weighted averages of the prior and observed means, with the weights controlled by sample size and the prior means for the variances. In Section 6, we give an example of the utility of the update defined by (4.4) and (4.5).

These estimated parameters, $\tilde{\mathbf{m}}$ and $\tilde{\sigma}^2$, are stored and used to call genotypes in other data sets. This is described in Section 3.

4.3 *Likelihood-based calls*

The final step is to make a genotype call for any given pair (sense and antisense) of observed log-ratios, $(M_{i,j,-}, M_{i,j,+})$. Note that these M values can come from any study, and we will use the centers and scales, defined by (4.4) and (4.5), estimated from the HapMap data. We do this by forming a likelihood-based distance function δ defined by

$$\delta_{i,k} \equiv \sum_{s \in \{-,+\}} \left\{ \log(\tilde{\sigma}_{i,k,s}) + \left(\frac{M_{i,j,s} - f_{j,k}(\mathbf{X}_{i,j,s}) - \tilde{m}_{i,k,s}}{\tilde{\sigma}_{i,k,s}} \right)^2 \right\}.$$

Our prediction is the genotype k that minimizes $\delta_{i,k}$. Furthermore, the log-likelihood ratio (LLR) tests serve as useful measures of confidence. Specifically, our measure of confidence is $\delta_{i,2} - \delta_{i,k}$ for homozygous calls and $\min(\delta_{i,1} - \delta_{i,2}, \delta_{i,3} - \delta_{i,2})$ for heterozygous calls. Supplementary Figure 3(D) (available at *Biostatistics* online) demonstrates that if we apply this method to the HapMap data (the training data), we obtain an impressive concordance rate as described in more detail in Section 5.

5. RESULTS

Most algorithms provide a measure of uncertainty for each call, in our case its the LLR described in Section 4.3, and define a cutoff for this limit. If this limit is not reached for a given SNP at a particular sample, no call is made. The proportion of cases where no calls are made is referred to as the "no-call rate." A common way to assess genotyping algorithms is to compare their concordance rates with the HapMap project calls to their no-call rate. In this section, we demonstrate that using our methodology provides better separability of cluster, no-call rates, and across-lab agreement than RLMM and BRLMM.

To assess the separability of clusters, we compare the median silhouette widths (Rousseeuw, 1987), a standard approach used in the unsupervised learning literature to measure cluster tightness (across-cluster distance to within-cluster distance ratio). Figure 9(A) shows the empirical cumulative distribution function for the RLMM, BRLMM, and CRLMM clusters. In particular, note that the 99% worst distance is almost 3 times better for CRLMM over RLMM. The improvements are dramatic. Similar improvements over BRLMM are observed.

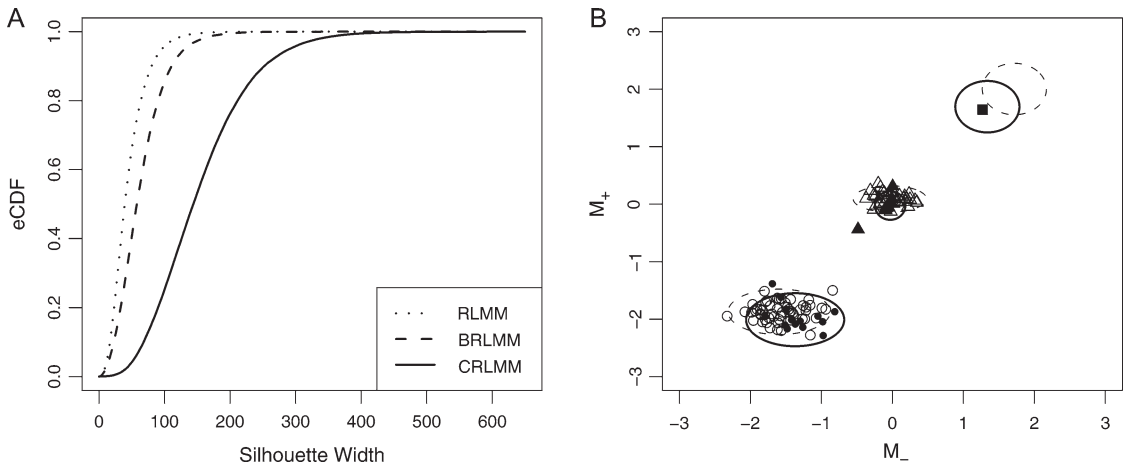

Fig. 9. Illustration of usefulness of CRLMM. (A) Empirical cumulative distribution function (eCDF) of the silhouette widths for the SNP-specific genotype regions for RLMM, BRLMM, and CRLMM. (B) For a particular SNP, data from HapMap are shown with empty symbols and the data from Lab 3 with solid symbols. Note that for the HapMap data, there are no AA samples. The dashed ellipses are defined using the HapMap data. The top dashed ellipse is the AA region predicted with the Bayesian correction (4.4). The solid lines are the regions derived after recalibration for the Lab 3 data. For the Lab 3 data, it appears that we have 1 AA sample and it is predicted correctly.

In Rabbee and Speed (2006), cross-validation was used to estimate the error rates. However, Figure 2 demonstrates that within-lab/study error rates are not necessarily accurate. This is due to the fact that supervised learning procedures may over adapt to results from one lab which may result in poor performance when we switch to data from other labs/studies. For this reason, we do not use cross-validation to evaluate the methods. Supplementary Figure 3(C) (available at *Biostatistics* online) shows correct call rates for the initial guesses provided by the mixture model fit described in Section 3.4. Note that the initial guesses, which are not based on a supervised learning approach, slightly outperform RLMM. Supplementary Figure 3(D) (available at *Biostatistics* online) shows how call rates, within the training data, increase close to perfection. Even with a no-call rate of 0%, calling every single SNP on every array, we obtain concordance rates of 99.85% from heterozygotes and 99.92% from homozygotes.

Figure 2 demonstrates how CRLMM provides predictions that are useful across labs/studies. In Figure 2(C), the ellipses were obtained from the training data. Note how only for CRLMM do the data for other 2 studies fall in, or are close to, the regions defined by training on the HapMap data. Thousands of other SNPs show behavior similar to that shown in Figure 1. Figure 9(B) is a particularly interesting example. For this SNP, the HapMap data had no AA calls. Note how the prediction defined by (4.4) and (4.5) creates a region for which data from another lab, that appears to come from an AA, fall close enough to be called AA.

6. Discussion

We have described a preprocessing algorithm for Affymetrix SNP arrays that greatly improves upon existing methods. The procedure is based on 4 steps: 1) Feature intensities are corrected for fragment length and sequence effects. 2) We then quantile normalize using a predefined reference distribution. 3) Next, median polish is used to summarize feature intensities into one number for every allele keeping sense and antisense summaries separate. 4) As a final step, a mixture model is used to correct for fragment length

and intensity-dependent biases on the log-ratio of the summarized intensities. We refer to this approach as SNP-RMA.

The summarized data, sequence information, fragment lengths, and intensity effects can then be used to make genotyping calls. Note that at this stage, one can use MPAM-, RLMM-, or BRLMM-like procedures to make genotype calls. We demonstrate that the supervised approach used by RLMM works very well in conjunction with a correction based on a posterior mean derived from a carefully derived hierarchical model. Although we use HapMap calls to define known classes and a training set, these calls could be avoided entirely and the preliminary calls from our mixture model could be used in their place to give a set of high-quality calls for determining the cluster centers.

ACKNOWLEDGMENTS

We thank Dan Arking, Ben Bolstad, Simon Cawley, Aravinda Chakravarti, James MacDonald, Shin Lin, Tom Louis, Hua Ren, and Howard Slater for advice, help with code, and/or sharing data. The work of Benilton Carvalho and Rafael Irizarry was partially funded by the Bioconductor Biomedical Information Science and Technology Initiative grant R33HG002708-04, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil), and the National Institutes of Health (NIH) Specialized Centers of Clinically Oriented Research (SCCOR) translational fund (212-2494). The work of Terence Speed was partially funded by NIH grant 2-P50-MH060398-06. *Conflict of Interest:* None declared.

REFERENCES

AFFYMETRIX (2006). BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set. *Technical Report, White Paper*. Santa Clara, CA: Affymetrix, Inc.

BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M. AND SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185–193.

DI, X., MATSUZAKI, H., WEBSTER, T. A., HUBBELL, E., LIU, G., DONG, S., BARTELL, D., HUANG, J., CHILES, R., YANG, G. *and others* (2005). Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* **21**, 1958–1963.

HUANG, J., WEI, W., CHEN, J., ZHANG, J., LIU, G., DI, X., MEI, R., ISHIKAWA, S., ABURATANI, H., JONES, K. W. *and others* (2006). CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics* **7**, 83.

IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. AND SPEED, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

KENNEDY, G. C., MATSUZAKI, H., DONG, S., MIN LIU, W., HUANG, J., LIU, G., SU, X., CAO, M., CHEN, W., ZHANG, J. *and others* (2003). Large-scale genotyping of complex DNA. *Nature Biotechnology* **21**, 1233–1237.

LAFRAMBOISE, T., WEIR, B. A., ZHAO, X., BEROUKHIM, R., LI, C., HARRINGTON, D., SELLERS, W. R. AND MEYERSON, M. (2005). Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Computational Biology* **1**, e65.

LAMY, P., ANDERSEN, C. L., WIKMAN, F. P. AND WIUF, C. (2006). Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Research* **34**, e100.

LIU, W. M., DI, X., YANG, G., MATSUZAKI, H., HUANG, J., MEI, R., RYDER, T. B., WEBSTER, T. A., DONG, S., LIU, G. *and others* (2003). Algorithms for large-scale genotyping microarrays. *Bioinformatics* **19**, 2397–2403.

LÖNNSTEDT, I. AND SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

NANNYA, Y., SANADA, M., NAKAZAKI, K., HOSOYA, N., WANG, L., HANGAISHI, A., KUROKAWA, M., CHIBA, S., BAILEY, D. K., KENNEDY, G. C. *and others* (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research* **65**, 6071–6079.

RABBEE, N. AND SPEED, T. P. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12.

ROUSSEEUW, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65.

SLATER, H. R., BAILEY, D. K., REN, H., CAO, M., BELL, K., NASIOULAS, S., HENKE, R., CHOO, K. H. AND KENNEDY, G. C. (2005). High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. *The American Journal of Human Genetics* **77**, 709–726.

SMYTH, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**.

UIMARI, P., KONTKANEN, O., VISSCHER, P. M., PIRSKANEN, M., FUENTES, R. AND SALONEN, J. T. (2005). Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland founder population. *Twin Research and Human Genetics*, **8**, 185–197.

WU, Z., IRIZARRY, R., GENTLEMAN, R., MARTINEZ-MURILLO, F. AND SPENCER, F. (2004). A model based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–917.

[*Received June 27, 2006; revised September 18, 2006; accepted for publication October 12, 2006*]