A GENOTYPE CALLING ALGORITHM FOR AFFYMETRIX SNP ARRAYS

Nusrat Rabbee[1*], Terence P. Speed[1,2]

[1]Department of Statistics, University of California, Berkeley, CA, USA;
[2]Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
[*]To whom correspondence should be addressed

ABSTRACT

Motivation: A classification algorithm, based on a multi-chip, multi-SNP approach is proposed for Affymetrix SNP arrays. Current procedures for calling genotypes on SNP arrays process all the features associated with one chip and one SNP at a time. Using a large training sample where the genotype labels are known, we develop a supervised learning algorithm to obtain more accurate classification results on new data. The method we propose, RLMM, is based on a robustly fitted, linear model and uses the Mahalanobis distance for classification. The chip-to-chip non-biological variance is reduced through normalization. This model-based algorithm captures the similarities across genotype groups and probes, as well as across thousands of SNPs for accurate classification. In this paper, we apply RLMM to Affymetrix 100K SNP array data, present classification results and compare them to genotype calls obtained from the Affymetrix procedure DM, as well as to the publicly available genotype calls from the HapMap project.
Availability: The RLMM software is implemented in R and is available from the first author at nrabbee@stat.berkeley.edu.

1 Introduction

Genomic research using SNP microarrays is attempting to identify DNA sequence variants in specific genes or regions of the human genome that are responsible for a variety of phenotypic traits, such as disease risk or variable drug response. The Affymetrix genotyping platforms are providing thousands of SNPs from the human genome on a single chip, to this end. The GeneChip® Human Mapping 10K array interrogates well over 10,000 SNPs by probe sets on one chip, the GeneChip® Human Mapping 100K array is available on two chips and the Mapping 500K array is planned for release in the near future. Once the SNPs are accurately genotyped, the interesting, high-level, biological questions can be more reliably answered.

Kennedy *et al.,* 2003 describe the technology behind building SNP arrays, which is known as Whole Genome Sampling Analysis (WGSA). The arrays contain probe sets to interrogate the two alleles for all the SNPs. The alleles are conventionally referred to as allele *A* and allele *B*. The technology involves synthesizing 25-mer oligonucleotide probes corresponding to a perfect match for the allele *A* sequence (PMA) and to a perfect match for the allele *B* sequence (PMB). In addition, a mismatch probe is synthesized for each allele (MMA and MMB) to detect non-specific binding. This probe quartet is the basic unit for detecting different genotype groups: *AA*, *AB* or *BB* (see Affymetrix Data Sheet for details). For the Mapping 10K array, the MPAM genotyping algorithm is based on clustering chips for each SNP by modified partitioning around medoids (see Liu *et al.*, 2003). Only SNPs with 2 or 3 clearly separated clusters are selected by MPAM and SNPs exhibiting a high degree of misclassification were discarded from the 10K array. This was possible since Affymetrix started with more than 3,000,000 SNPs from the Perlegen database. As the demand for higher density SNP arrays increased, MPAM faced challenges in making correct calls for SNPs with missing genotype groups or low minor allele frequency and required large sample sizes for clustering. Therefore, Affymetrix released a new dynamic model-based algorithm DM for the Mapping 100K array (see Di *et al.*, 2005). The DM method assumes normality for the pixel intensities for a given feature and calculates a log-likelihood for each probe quartet independently, under 4 different models: *Null*, *AA*, *AB* and *BB*. For each probe quartet, the likelihoods are combined to produce a score. Next, these scores on different probe quartets are combined and the Wilcoxon signed rank test is applied to test each model likelihood to produce four p-values. The algorithm then decides which model is best supported by all the 10 probe quartets based on the minimum p-value, for each SNP and each chip. DM is generally very accurate, but exhibits a higher degree of misclassification for known heterozygous bases than for known homozygous bases.

Neither DM and MPAM algorithms explicitly uses the available *truth* for training purposes, even though a large number of reference genotype calls are available on more than 3 million SNPs. The algorithms instead use this data for verification, SNP selection and tuning purposes. Furthermore, DM does not make use of the available data from multiple chips. Neither algorithm exploits the similarities across thousands of SNPs. Here we propose a classification algorithm which uses the robust, multi-chip average (RMA) method to combine the intensities across probes and chips (Irizarry *et al.*, 2003) and produce allele-based summaries. This is a supervised learning procedure, which takes advantage of the large number of publicly available calls on the SNPs in defining regions for each genotype group. These improvements lead to more accurate classification results on a subset of SNPs from the Mapping 100K array – Xba set. For this subset of SNPs, genotype calls are publicly available for 90 Centre du Etude Polymorphisme Humain (CEPH) individuals, from the HapMap project (see HapMap, 2003). These HapMap reference genotypes were derived from sources other than Affymetrix, Inc. or Perlegen, Inc., to make the comparisons as independent as possible.

2 Algorithm

The RLMM algorithm, based on a multi-chip model with the Mahalanobis distance classification, consists of three parts: (i) *robustly fitting a linear model* – which reduces non-biological variability from the probe data, for each allele (ii) *forming decision regions for each genotype class* – which are bivariate Gaussian or Mahalanobis regions and are formed by making efficient use of training data available to inform the algorithm of the centers and spread of the intensities for each genotype groups of every SNP; (iii) *classifying new data* – which calls genotypes on samples on new chips according to their Mahalanobis distances to the three groups formed for that SNP.

Multi-chip Robust Linear Model

First, we *pre-process* the data by applying quantile normalization to the probe intensities (see Bolstad *et al.*, 2003), in order to minimize chip-to-chip non-biological variability. Normalization is essential for implementing a multi-chip model to the probe intensities. This normalization method assumes the same underlying distribution of intensities across chips.

Second, we $\log_2$ transform the normalized intensities and robustly fit a linear model to estimate the chip and probe effects. The details and benefits of the robust multi-array average model (RMA) of probe intensity measures have been discussed by Irizarry *et al.*, 2003. Let I denote the total number of chips present either in the training or test sample and J denote the number of allele A or allele B perfect-match probe intensities in the data set.

For SNP n, the model we fit to the allele A probe intensities is:

$$\log_2(y_{A,ij}^{n}) = \theta_{A,i}^{n} + \beta_{A,j}^{n} + e_{ij} \qquad \text{where } i=1,\ldots,I;\ j=1,\ldots,J$$

where $y_{A,ij}^{n}$ is the normalized probe intensity for chip i, allele A probe j and SNP n, and $\theta_{A,i}^{n}$ is the chip effect determined from the A probes, and $\beta_{A,j}^{n}$ is the probe effect, and e_{ij} is an error term with mean zero, assumed independent, identically distributed. We repeat this step by fitting the above model separately to the allele B probes.

For each SNP n, the multi-chip model reduces produces 2-dimensional estimates of $\boldsymbol{\theta}_i^{n} = (\theta_{A,i}^{n}, \theta_{B,i}^{n})$, which are summary measures of the allele A and B intensities for chip i. The model is applied to the training set and test set separately. Note that RLMM only uses the perfect-match intensities for the model. Preliminary investigation showed that including the mismatch probes in the model did not yield better (i.e., more readily separated) estimates for $\boldsymbol{\theta}$. We are continuing to explore the topic of using mismatch probes in our analysis.

Mahalanobis regions

The second central part of our algorithm takes the summary A and B intensities as input and forms the decision regions. The regions for RLMM are characterized by bivariate, Gaussian distributions. Since the θ_A

and θ_B values are correlated, the regions formed by these 2-dimensional points are ellipses and the Mahalanobis distance will be used as the decision metric.

SNPs with well-defined Genotype Groups

First, for each SNP *n*, we obtain the mean vectors and covariance matrices for the 2-dimensional points (θ_A,θ_B), in each of the three-genotype groups (i.e., *AA*, *AB* and *AB*), from the chips in the training set. Let $\mathbf{m}=(m^{AA}{}_A, m^{AA}{}_B, m^{AB}{}_A, m^{AB}{}_B, m^{BB}{}_A, m^{BB}{}_B)$ denote the 6x1 vector of group centers and $\mathbf{S}=((s^2{}_A)^{AA}, (s^2{}_B)^{AA}, (r)^{AA}, (s^2{}_A)^{AB}, (s^2{}_B)^{AB}, (r)^{AB}, (s^2{}_A)^{BB}, (s^2{}_B)^{BB}, (r)^{BB})$ denote the 9x1 vector of group dispersion parameters. For SNPs with sufficient sample size in each genotype group, the parameters of these two vectors can be easily estimated from the training data and the three decision regions formed. The decision region for genotype group *g* is characterized by $\mathbf{m_g}$, the 2x1 row vector of means and $\mathbf{S}_g$, the 2x2 covariance matrix.

Next, we robustly fit the linear model described in section 2.1, to the test data set and obtain estimates of $\boldsymbol{\theta}=(\theta_A,\theta_B)$ for each chip in this set. Using the decision regions formed above by the training set, we compute the Mahalanobis distance of each chip in the test set, from the center of genotype group *g*: $D^2{}_g(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{m_g})\mathbf{S}^{-1}{}_g(\boldsymbol{\theta} - \mathbf{m_g})^T$. Subsequently, each chip with allele estimate, $\boldsymbol{\theta}$, is assigned to a genotype class using the Mahalanobis distances as a minimum distance classifier.

SNPs with low minor allele frequency

When a SNP has a low minor allele frequency or a missing genotype group, the **m** and **S** parameters cannot be estimated reliably from the training set for that SNP. In this case, we use the multivariate normal (MVN) distribution theory to estimate these parameters from thousands of SNPs, where the groups are well-defined. Since the elements of **m** are correlated with each other, we use regression to predict the center of each genotype group for the SNPs where the training data do not provide sufficient information. We take a similar approach to estimating the elements of **S**, although we use normalizing transformation for the elements.

We assume that for each SNP *n*, the vector **m** is normally distributed with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. First, the vector parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, are estimated from a random sample of 5,000 SNPs present in the 100K data set with well-defined groups. Let *g* denote the missing or sparse genotype group and *g'* and *g"* denote the other two groups. Second, we compute the parameters for the conditional distribution of the center of group *g*, $\mathbf{m}_{g|(g',g")}$, given the other two groups centers, by estimating the mean vector $\boldsymbol{\mu}_{g|(g',g")}$ and partial covariance matrix, $\boldsymbol{\Sigma}_{g|(g',g")}$. Here, we assume that $\mathbf{m}_{g|(g',g")} \sim \text{MVN}(\boldsymbol{\mu}_{g|(g',g")}, \boldsymbol{\Sigma}_{g|(g',g")})$. Third, the matrix of regression coefficients, **B**, is formed, where $\mathbf{B} = \boldsymbol{\Sigma}^T_{g,(g',g")} \boldsymbol{\Sigma}^{-1}_{(g',g"),(g',g")}$, from the multi-SNP data. Finally, the predicted value of $\mathbf{m}_{g|(g',g")}$ is calculated as $\mathbf{B}\mathbf{m}_{(g',g")} + \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = \boldsymbol{\mu}_g - \mathbf{B}\boldsymbol{\mu}_{(g',g")}$. We repeat the process for **S,** where each group's variance-covariance matrix is predicted, in a manner similar to the group's center, from the other two groups' variance-covariance matrix. Once the matrices of regression coefficients are calculated from the multi-SNP data for each group's center and covariance, RLMM uses these estimated parameters to predict a group's center or covariance matrix, when that group is missing or sparse.

Classification

Once the group centers, **m**, and the dispersion parameters, **S**, are determined either from the training data or by prediction, RLMM is ready for classification. For each chip in the test set, the allele summary estimate, $\boldsymbol{\theta}=(\theta_A,\theta_B)$, is assigned genotype group g^*, if the minimum Mahalanobis distance $D^2{}_g$ occurs for $g=g^*$. The minimum distance, $\min(D^2{}_g)$, also provides a quality score δ, for each call. Since under bivariate normality, the distances in each group *g* follow a χ^2 distribution with 2 degrees of freedom, we computed the quantiles of the empirical distribution of the distances (δ) to determine cutoffs for the quality score. Decreasing the cutoff value for making calls, usually increases accuracy of the calls. Thus, RLMM is able to adjust the percentage of calls it makes at a user-specified level, thereby increasing its accuracy level.

3 Results

Multi-Chip Model

The first step in the RLMM algorithm is to normalize the probe intensities and apply the robust, linear model to the transformed and normalized probe-level intensities, in order to obtain the estimated $\boldsymbol{\theta} = (\theta_A, \theta_B)$ values for each chip, for any given SNP. Plotting the 2-dimensional $\boldsymbol{\theta}$ vector for each chip shows the clear ellipses formed for the three different genotype classes in the following figure. The estimated $\boldsymbol{\theta}$ values are referred to as allele *A* and allele *B* values. The ellipse in the bottom right is for genotype group *AA* , the one in the center is for genotype group *AB,* and the one in the top left is for group *BB*. The residual plots indicate that the linear model fits the data reasonably well.

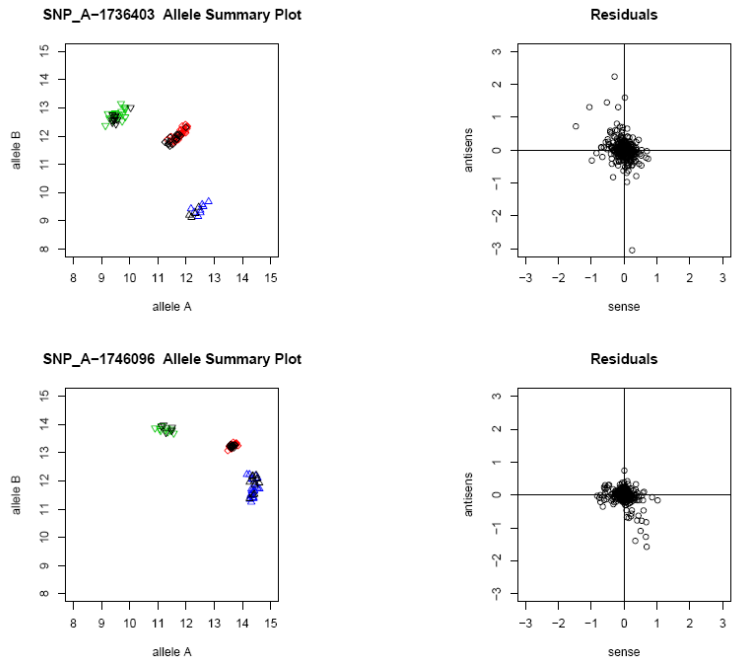

Figure 1 – Side-by-side Allele Summary Plot and Residual Plot for two typical SNPs in the Mapping 100K - Xba set.

Decision Regions from Training Data

The second step in RLMM is to compute the mean and covariance matrix of points from each of the three genotype groups from the training data, provided there are no missing groups or very sparse groups (number of observations in a group ≤ 5). Then the algorithm proceeds to call the genotype on each (θ_A, θ_B) pair of the test data based on the minimum Mahalanobis distance from each group mean. The minimum Mahalanobis distance also serves as a quality score for each call made on the test set. The algorithm can reduce the desired percentage of calls (e.g., 90%) in order to possibly increase the accuracy of calls. The appropriate thresholds are easily obtained from the empirical distribution of the minimum Mahalanobis distances. We note that the empirical distribution closely follows the theoretical distribution of χ^2 with 2 degrees of freedom, as shown in the following figure.

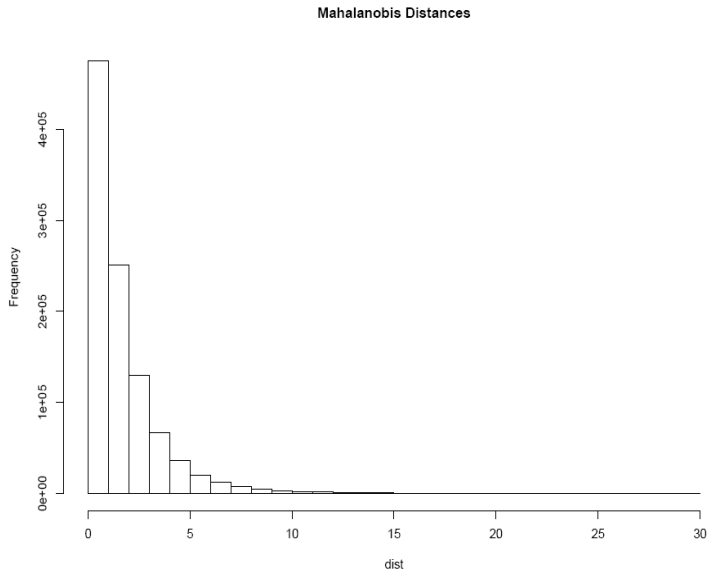

Figure 2 – Histogram of minimum Mahalanobis distances pooled over 5,000 SNPs

Decision Regions by Prediction

The third step in the RLMM procedure is to estimate the mean and covariance matrix (**m**,**S**) either when there is a missing genotype group or when there are very sparse genotype groups in the training sample. RLMM predicts the missing or sparse group mean and covariance matrix by regression, from the other two groups with sufficient data. The regression parameters are obtained from the multi-SNP, multivariate normal model of the group means and covariance matrix elements. We illustrate the motivation behind the regression approach by plotting the different group means against each other, across a random sample of 5,000 SNPs (supplemental figure A), which shows a strong correlation between the group means. We do a principal component decomposition of the sample covariance matrix, **V,** of the 6x1 vector of the group centers, **m**. The percentages of variance explained by the 6 components were, in decreasing order: (88, 7, 3, 1, .2, .04). While the first principal component is considered a measure of the *size* of group centers, our empirical investigation revealed that the second principal component reflects the *position* of the group centers, and the third reflects the *relative position* of the *AB* group center from the closer homozygous group.

.

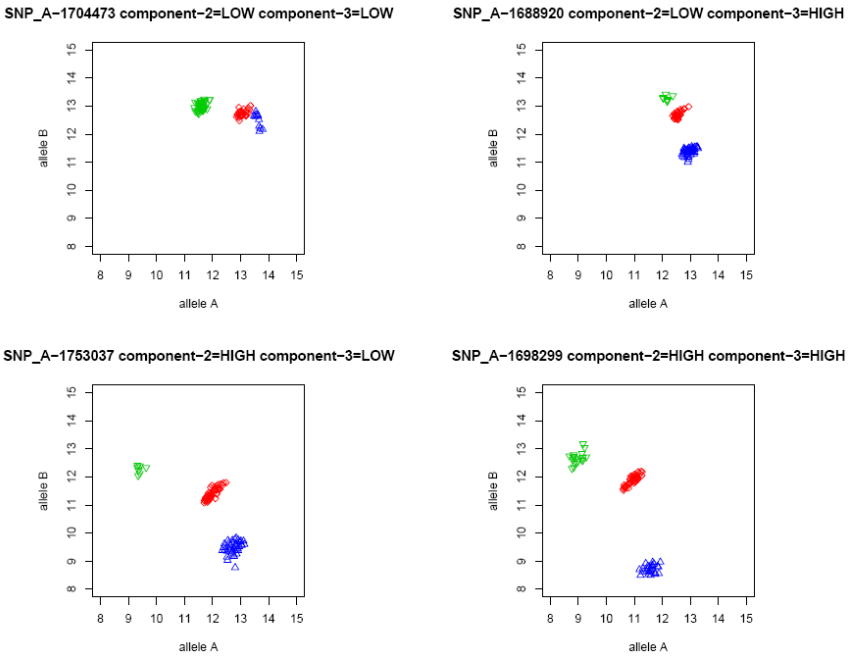

Figure 3 – Allele summary plots of SNPs whose means have difference combinations of HIGH and LOW values of the second and third principal components of the mean vectors

We have thus found the features derived from the second principal component $(.35m^{AA}_{A} - .61m^{AA}_{B} + .2m^{AB}_{A} - .12m^{AB}_{B} - .62m^{BB}_{A} + .26m^{BB}_{B})$ and the third principal component $(-.51m^{AA}_{A} - .24m^{AA}_{B} - .41m^{AB}_{A} + .43m^{AB}_{B} + .13m^{BB}_{A} + .56m^{BB}_{B})$, to be useful indicators for assessing SNP-level classification quality.

Comparison with HapMap calls

Comparison to HapMap calls are made for both RLMM and DM to determine accuracy of the two algorithms. Of the 15,910 SNPs where both DM and HapMap calls were available from the Xba set, we excluded all the monomorphic SNPs (SNPs with 0 or 1 members in two of the genotype groups). Tables 1 and 2 show the concordance between HapMap calls (column) and RLMM calls (rows) or DM calls (rows) for a total of 11,446 SNPs. For each SNP, calls are made for 90 individuals from the 30 CEPH family trios.

HapMap RLMM	AA	AB	BB	NC
AA	339,756	476	12	1440
AB	196	356,575	184	1699
BB	32	498	327,772	1478

Table 1 – RLMM Comparison with HapMap (99.86%concordance; discordant calls, excluding NoCalls=1,398)

HapMap DM	AA	AB	BB	NC
A	339,502	1249	9	1420
AB	457	355,168	544	1745
BB	25	1132	327,415	1452

Table 2 – DM Comparison with HapMap (99.67%concordance; discordant calls, excluding NoCalls=3,416)

For Table 1, RLMM used the HapMap calls for training itself via a leave-one-out cross-validation approach. Note that the HapMap calls include some NoCalls(NC), whereas RLMM and DM are making calls for every chip. DM does not explicitly avail itself of the known HapMap calls for training. However, the 100K arrays are comprised only of SNPs which show a high degree of concordance with available calls from HapMap or Perlegen. For RLMM, most of the 1,398 calls discordant with HapMap varied only for one or two chips per SNP. In fact, the 1,398 discordant calls were spread across 656 SNPs, with 412 SNPs having only 1 discordance, 117 having just 2, and so on, while a few SNPs had a high number of discordances (22,23,49). We visually investigated 50 random SNPs from the 656 SNPs, where the RLMM and HapMap calls had any discordance. It appears that in 36 of those SNPs, RLMM calls were correct; calls were ambiguous in 9; RLMM calls were incorrect in 3; and RLMM was likely trained with wrong labels in the remaining 2.

We also compared RLMM with DM directly for the SNPs above and obtained a 99.7% concordance. In fact, the diagonal entries are larger than the corresponding entries in the previous two tables. However, this is attributable to the fact that neither RLMM nor DM are making NoCalls. Therefore, there are more chips to make calls on. The overwhelming majority of the calls are in agreement between the two algorithms.

RLMM achieves higher accuracy in genotype calling, when compared with DM in the set of SNPs we investigated, using the leave-one-out test on HapMap calls. In supplemental Figures B and C, we show instances of RLMM correctly making genotype calls, whereas the calls produced by the DM and sometimes by the HapMap algorithm appear to be incorrect. In the figure below, we show an overall accuracy curve.

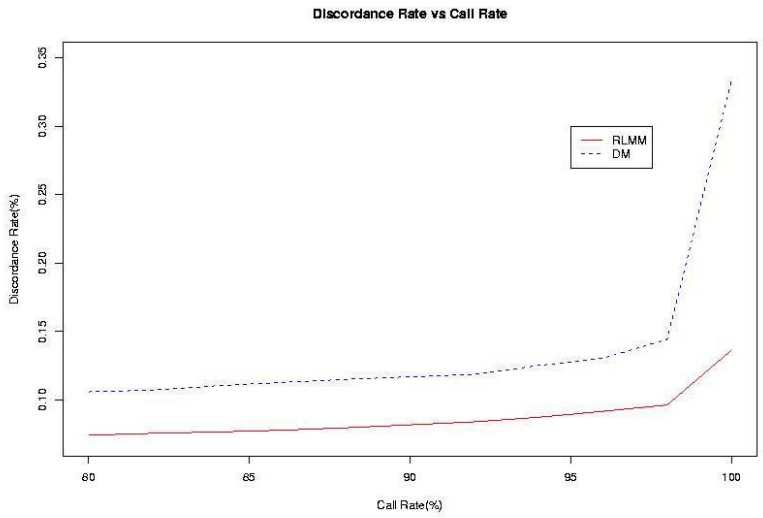

Figure 4 – Percentage of Calls versus Percentage of Discordance with HapMap calls for RLMM and DM (*n*=11,446 SNPs).

Figure 4 shows the effect of decreasing the call rate on the quality of calls for RLMM and DM. RLMM cutoffs are determined for each call rate % from the empirical distribution of the minimum Mahalanobis distances. DM cutoffs are obtained from the minimum p-values distribution under each of the three models: *AA*, *AB* and *BB*. Overall, the RLMM procedure is less discordant with HapMap for *all* call rates. For a fixed discordance rate, RLMM achieves much higher call rate than DM.

4 Discussion

Probe-level, multi-chip models enable RLMM to obtain accurate summaries of allele *A* and allele *B* intensity measures from only the 20 perfect-match A and B probes. The model, together with quantile normalization, reduces chip-to-chip variability and probe-to-probe variability. Since RLMM has the power to classify SNPs with only the perfect-match probes, we note that a halving the number of probes is possible on the arrays. Since unambiguous decision regions are formed for most of the SNPs we investigated from the Mapping 100K – Xba set, an unsupervised algorithm could be used successfully to classify the $\boldsymbol{\theta}$ vectors. The Mahalanobis distances provide a chip-level quality score for each call. We also extract two important features

from the principal component decomposition of the group centers, which will help identify a priori, the SNPs on the array, for which the probe-level data do not adequately discriminate between the alleles.

Multi-SNP aggregation provides a regression mechanism to predict the group mean and covariance matrix, when a group is absent or sparsely represented in the training data. RLMM uses the correlation present in the group means, across genotype groups, to predict the missing group's center. This gives RLMM increased classification accuracy for making calls in these SNPs. RLMM achieves a higher overall accuracy rate than DM, as shown in Tables 1 and 2 in the previous section, when compared with the HapMap calls on a given set of SNPs from the Mapping 100K array. At call percentages, RLMM also shows reduced discordance with HapMap calls, relative to DM (see Figure 4). RLMM achieves higher call rate than DM for the same level of accuracy. For example, for the same level of accuracy , RLMM achieves above 98% call rate, whereas DM achieves about 90%.

RLMM, which is based on a proven, probe-level statistical model (RMA) and standard classification theory gains considerably in accuracy in making calls on new data, by making efficient use of the training data. In the case of SNP arrays, a large amount of training data is available from the public domain. In the near future, we plan to extend this algorithm to work with SNP data where no training data is available, as well as to identify copy number polymorphisms.

ACKNOWLEDGEMENTS
We acknowledge the generous assistance of Simon Cawley and Earl Hubbell of Affymetrix, Inc. in providing the SNP data and valuable feedback. We wish to thank Henrik Bengtsson, David Clayton, Francois Collin, Jon McAuliffe, and Benjamin Rubinstein for providing thoughtful comments.

REFERENCES

Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M.S., Boyce-Jacino, M.T., Fodor, S.P., Jones, K.W. (2003) Large-scale genotyping of complex DNA. Nature Biotechnology, 21, 1233-1237.

Affymetrix, Inc. (2005) GeneChip® Human Mapping 100K Set.
http://www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf

Di, X., Matsuzaki, H., Webster, T.A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang G., Shen, M., Kulp, D., Kennedy, G.C., Mei, R., Jones, K.W., Cawley, S. Kruglyak, L. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. Bioinformatics, 21, 1958-1963

Liu, W., Di, X., Yang, G., Matsuzaki, H., Jing, H., Mei R., Ryder T., Webster, T.A., Dong, S., Liu G., Jones, K., Kennedy, G., Kulp, D. (2003) Algorithms for large-scale genotyping microarrays. Bioinformatics, 19, 2397-2403.

HapMap(2003) The international hapmap consortium. Nature, 426, 789-796.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19, 185-193.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U.,
Speed, T.P. (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics, 4, 249-264.

Rao, C.R (2002) *Linear Statistical Inference and Its Applications*, 2nd edn. Wiley, NY