Johns Hopkins University
Johns Hopkins University, Dept. of Biostatistics Working Papers

Year 2006 *Paper* 111

EXPLORATION, NORMALIZATION, AND GENOTYPE CALLS OF HIGH DENSITY OLIGONUCLEOTIDE SNP ARRAY DATA

Benilton Carvalho [*] Terence P. Speed [†]

Rafael A. Irizarry [‡]

[*]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[†]Division of Genetics & Bioinformatics, WEHI, Melbourne, Australia, Department of Statistics, UC Berkeley

[‡]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, rafa@jhu.edu

EXPLORATION, NORMALIZATION, AND GENOTYPE CALLS OF HIGH DENSITY OLIGONUCLEOTIDE SNP ARRAY DATA

Benilton Carvalho, Terence P. Speed, and Rafael A. Irizarry

Abstract

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 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 methodology useful for preprocessing Affymetrix SNP chips and obtaining genotype calls with the preprocessed data. We demonstrate how our procedure improves existing approaches using data from three relatively large studies including one in which large number independent calls are available. Software implementing these ideas are avialble from the Bioconductor oligo package.

Exploration, Normalization, and Genotype Calls of High Density Oligonucleotide SNP Array Data

Benilton Carvalho
Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205

Terence P. Speed
Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia
Department of Statistics, UC Berkeley, CA

Rafael A. Irizarry*
Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205, rafa@jhu.edu

Abstract

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 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 methodology useful for preprocessing Affymetrix SNP chips and obtaining genotype calls with the preprocessed data. We demonstrate how our procedure improves existing approaches using data from

*To whom correspondence should be addressed

three relatively large studies including one in which large number independent calls are available. Software implementing these ideas are avialble from the Bioconductor `oligo` package.

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 amplifi ed with a polymerase chain reaction (PCR) reaction, and the sample is labeled and hybridized to an array containing probes designed to interrogate the resulting fragments. There are currently three 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 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 for a handful of different applications (Uimari et al., 2005; Nannya et al., 2005; Huang et al., 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 et al. (2006), Nannya et al. (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 et al. (2003). Each SNP on the array is represented by a collection of probe quartets. In the 100K arrays, SNP chips probesets are composed of 40 features. As with expression arrays, the features are defi ned by 25-mer oligonucleotide molecules referred to as perfect match (*PM*) probes. The probesets also include a mismatch (MM) pair for each PM feature. As in expression arrays, these are created by changing the middle base pair. A difference with expression arrays is that PM features differ in three important ways:

First, two alleles are interrogated (for most SNPs only two alleles are observed in nature). These are denoted by A and B and divide the probes into two groups of equal size. For each *PM* probe representing the *A* allele there is an allele *B* that differs by just one base pair (the SNP). Second, features are included to represent the sense and antisense strands. This difference divides the probes into two 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 four 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 to no longer use this type of probe. Notice that an array with no *MM*s can accommodate features for twice as many SNPs.

The general goal of preprocessing for SNP arrays is to normalize and summarize feature intensities into genotype calls (AA, AB, BB). A measure of confidence is also desired. Typically, samples not achieving a specific confidence cut-off at a given SNP receive no calls at that SNP. In this paper we propose preprocessing methodology that greatly improves accuracy of genotyping calls over existing methods. We propose a modular approach in which preprocessing is done in a first step and a genotyping algorithms is defined for preprocessed data. To illustrate this, and to motivate our methodology, we use three datasets: 1) The HapMap Trio Dataset, consisting of 30 trios analyzed on the 100K Mapping platform, which are also part of the International HapMap Project and, therefore, have precise genotype calls that can be used as "gold-standard", 2) a dataset comprised of the same DNA hybridized to 53 arrays, and 3) a dataset consisting of 22 samples as described in Slater et al. (2005). We will refer to these datasets as the lab 1, lab 2, and lab 3 datasets. Lab 1 dataset will also be referred to as the Hapmap data.

The paper is organized as follows: Section 2 describes 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. Notice that if the PCR produced x copies of the DNA fragments, these quantities should the log of 0, x, or $2x$. 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 $\theta_A > C_1$ and $\theta_B < C_2$. However, the most basic data exploration demonstrates that such an approach will not work well in general. Figure 1A illustrates the problem. Given what we have learned from expression arrays about optical background noise, non-specific 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 cut-offs are not useful, for most SNPs the θ_A and θ_B do form three distinct clusters representing the three possible genotypes. Affymetrix's default algorithm for their 10K arrays took advantage of this property and used a modified partitioning around the mediods (MPAM) clustering algorithm to detect the clusters. These clusters were then associated with the three different genotypes. The summarized data was based on a relative allele signal (RAS) which is essentially a ratio of allele A intensities to the sum of both alleles intensities. The intensities were corrected for background using the *MM* (Liu et al., 2003).The algorithm worked well when there was enough data in each of the three 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 et al., 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 four different Gaussian models (NULL, AA, AB, and BB) were considered for the probe intensities for each SNP, and a genotype call made for each sample based on the likelihoods for each genotype. Notice that DM is not a modular procedure: the calls are derived directly from the feature intensities.

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 fact motivated several academic groups to develop their own algorithms (LaFramboise et al., 2005; Rabbee and Speed, 2006). In Rabbee and Speed (2006) the robust linear model with Mahalanobis distance (RLMM) is described and shown to outperform DM on the Hapmap dataset described above.

RLMM begins by preprocessing the feature-level data using RMA, a procedure demonstrated to work well for expression arrays Irizarry et al. (2003). These summarized data are then used to build a SNP-specific *regions* for each genotype using a supervised learning algorithm similar to linear discriminant analysis (LDA). To train the algorithm, the Hapmap dataset was used. This approach is particularly appealing because empirical results demonstrate that different SNPs can produce very different distributions. Figure 1A clearly demonstrates this. Model-based approaches that impose the same (or similar) models on all SNPs are unlikely to perform as well as algorithms that train on observed data. In fact, using cross-validation on the Hapmap dataset, Rabbee and Speed (2006) demonstrate that RLMM greatly outperforms DM (See Figure 4 in Rabbee and Speed (2006)). However, this preprocessing strategy make RLMM's genotyping algorithm less useful because SNP-specific feature intensity distributions are different not only across SNP but within the same SNP across labs/studies. Figure 1B 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 calling data from a different study/lab.

Recently, Affymetrix has made a white paper document available Affymetrix (2006) describing a new preprocessing algorithm based on RLMM. To improve the across-lab compatibility, 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 re-calibrated using a Bayesian calculation. This algorithm is expected to become the default in the near future. More details are available here 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

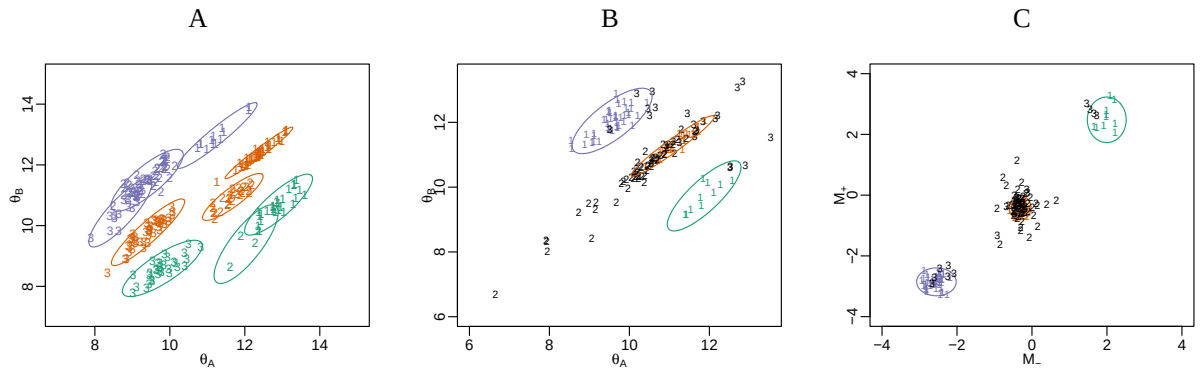


Figure 1: Genotype regions. A) RLMM genotype regions for three SNPs with data from different samples shown as well. Different colors represent the three different genotypes and the numbers the three different SNPs. B) RLMM genotype regions obtained using the Hapmap data and data points from labs 2 and 3 (denoted by numbers). C) As B) but for CRLMM.

RLMM to create a powerful corrected version. We will refer to our genotyping method as CRLMM. Because our preprocessing method is an adaptation of RMA and can be used with other genotyping algorithm, we will refer to it as SNPRMA.

3 Normalization

A likely explanation for the across-lab differences in cluster distributions seen in Figure 1A is the sample preparation effect. In particular, the effect of DNA polymerase chain reaction (PCR) which is used to amplify each DNA is 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 et al. (2005). However, these corrections are done on the log intensity scale and we fi nd that effects can still be observed for the log-ratio values. We propose normalization strategies that correct for these log-ratio biases as well.

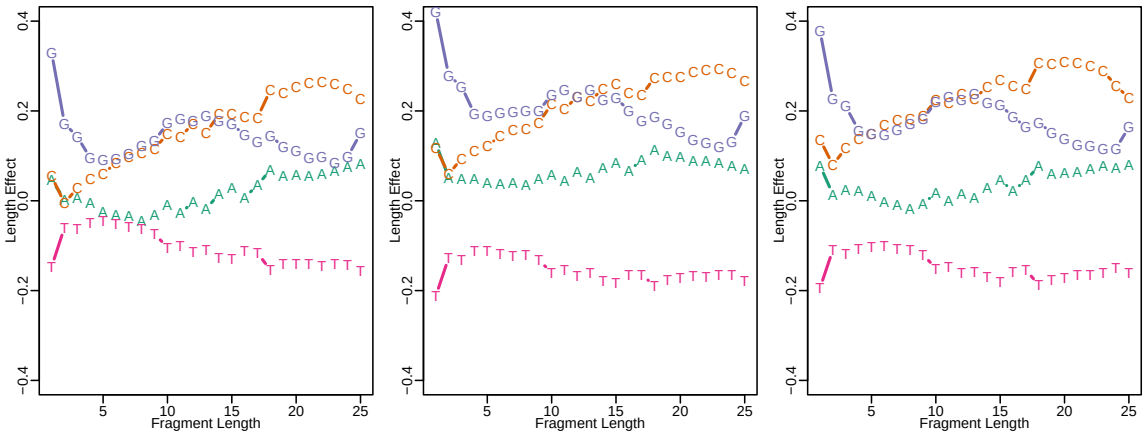


Figure 2: Position dependent sequence effect. A) For a typical array from the Hapmap study, the effect of each base at each position is shown. The different bases are denoted with different colors. B) As A) for lab 2, and C) lab 3.

3.1 Correcting for sequence and fragment length

Nannya et al. (2005) noticed that fragment length has a strong effect on probe intensity, with longer fragments resulting in weaker feature intensities (Supplemental Figures 1-2). These figures demonstrate that the effects are different from sample to sample and from lab to lab, with the lab difference being greater. Nannya et al. (2005) have also pointed out that GC content has a strong effect on feature intensity. We have noticed that the sequence effect is actually position dependent, something that has previously been observed in expression arrays (Wu et al., 2004). Figure 2 shows the position dependent effects of each of the four bases for three different labs. This figure demonstrates that the effects are large, and that they change from sample to sample and 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 3A shows that the sequence effect can cause relatively large differences between alleles A and B.

In our normalization procedure, our first step is to correct for both sequence and fragment length effects.

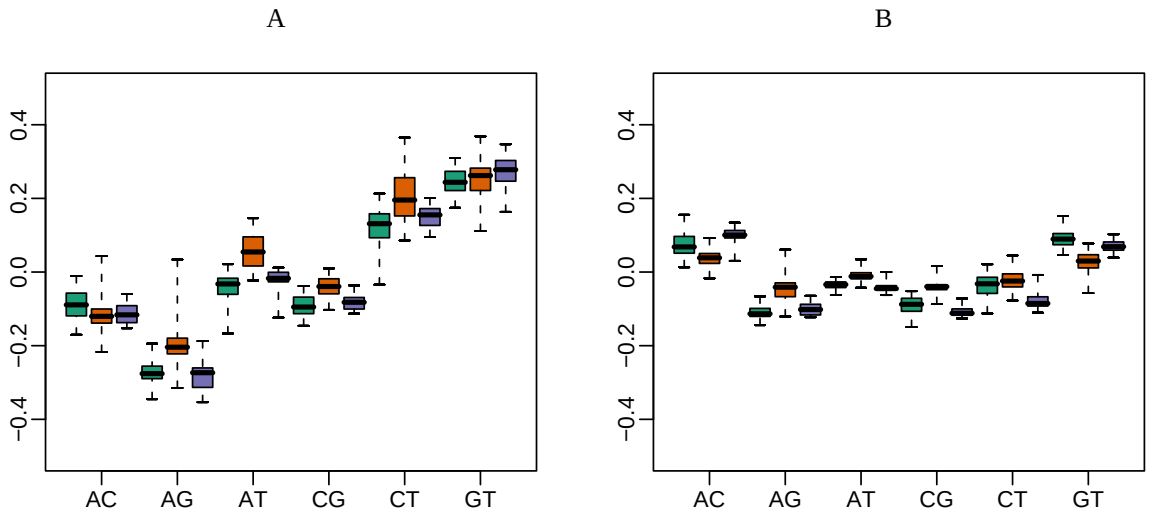


Figure 3: Sequence effect on allele A to allele B log-ratio for the six different possible base pairs. For each array we obtain the median of the log-ratios for all SNPs with the same base-pair at the SNP. In this plot we show boxplots of those medians for each of the six possible base pairs. The three different colors represent the three different labs. A) Before normalization and B) after normalization.

To do this we simply fi t a linear model to the log *PM* intensities:

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

Here b_t represents the base at location t, $h_b(t)$ are smooth functions of location (each base b is represented by a different function), $1(b_t = b)$ is 1 when the base at position t if b, and 0 otherwise, and $g(L)$ is a smooth function of fragment length L. Supplemental Figures 1-2 and Figure 2 demonstrate that the effects are well described with a smooth functions which we model with a 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 from subtracting the estimated sequence and fragment length effects for $\log_2(PM)$. Nannya et al. (2005) demonstrates 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 3B shows that the effect of sequence is reduced but can be

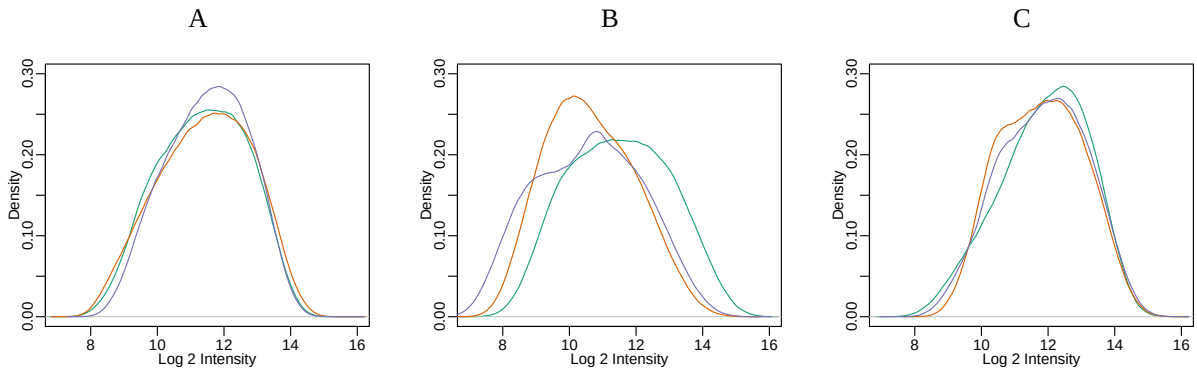


Figure 4: Empirical densities for log (base 2) intensities from three randomly chosen arrays from A) lab 1, B) lab 2, and C) lab 3. The intensities have been corrected for sequence and fragment length.

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 4 demonstrates that even after the correction described in the previous section, 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 sample. 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 et al. (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.

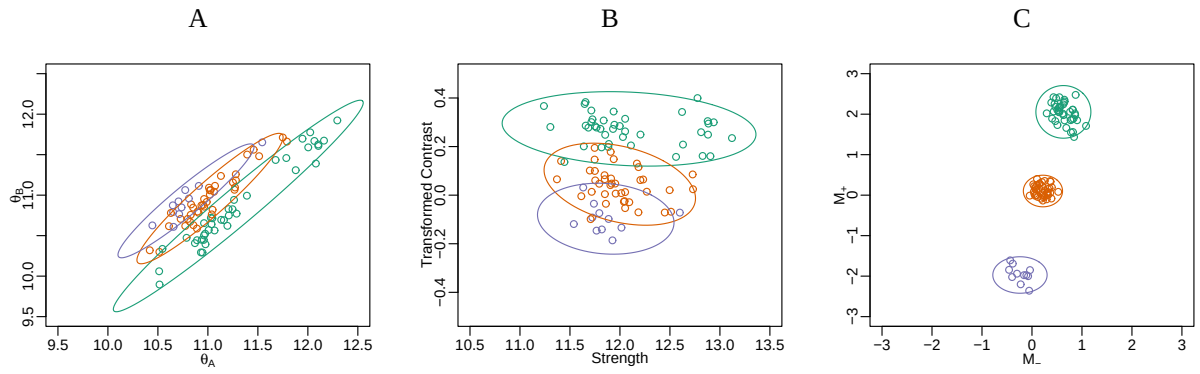


Figure 5: Effect of the "broken" probe effect. A) RLMM genotype regions for SNPs for which the sense strand does not differentiate. B) As A) but for BRLMM and C) CRLMM.

3.3 Summarization

We summarize the feature intensities within each probe quartet to produce four values for each SNP. Specifically, we follow the RLMM approach of using median polish to fit a linear model 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, we actually fit a separate model to each strand/allele combination instead of combining the strands as done by RLMM. We therefore produce four numbers per SNP which we can denote with: $\theta_{A,-}, \theta_{A,+}, \theta_{A,-}, \theta_{B,+}$. In Section 3.4 we describe why we keep sense and antisense values separate.

3.4 Remaining log-ratio biases

Figures 1A and 1B show that most of the information available for separating the clusters associated with the three genotypes is in the upper-left-to-lower-right diagonal direction, i.e. the log ratios. The same plot 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 two strands appears to provide no information. Figure 5 demonstrates that considering the log-ratios for the two strands, M_+ and M_-, instead of a summary

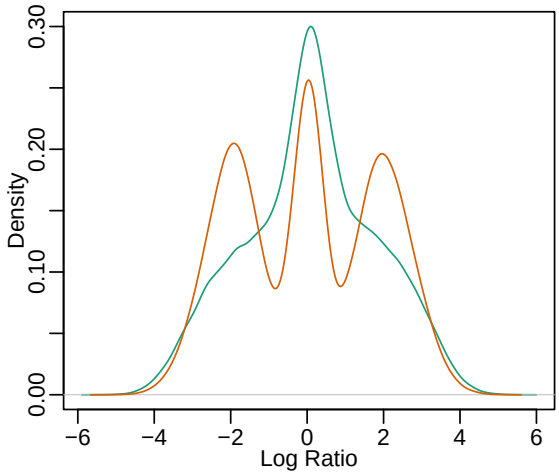


Figure 6: Empirical density distribution of the across SNP M values for the array with the best (red) and worst (green) SNR ratios.

that contains both, permits us to correctly call genotypes in cases in which the features for one of the strands is not informative. We have observed roughly 100 cases such as the one presented in Figure 5. For this reason we propose these strand specifi c log-ratios as the summarized quantity to be used by genotyping algorithms. We denote the log-ratio for SNP i, sample j by $M_{i,j,s}$ with sense and antisense strands denoted by $s = +, -$. 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 there was better separation than others, as demonstrated in Figure 6. 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. Figure 7 shows this very clearly. Furthermore, Figure 3 demonstrates that, although much reduced, a sequence effect is still present for log-ratios. In the remainder of this Section we describe our fi nal preprocessing step which estimates these remaining biases.

We describe these effects with a simple mixture model. To simplify the fi tting procedure we estimate the model separately on each array and treat the sense and antisense feature as exchangeable. We therefore drop

the j and s notation and write:

$$[M_i|Z_i = k] = f_k(X_i) + \varepsilon_{i,k}, \tag{2}$$

where the X_i represents covariates known to cause bias, f_k describes the effect associated with these covariates, and $\varepsilon_{i,k}$ an error term which we assume to be a normal random variables with mean 0 and variance τ_k^2. We assume that $f_{j,2} = 0$, and $f_{j,1} = -f_{j,3}$ and that $\tau_1^2 = \tau_3^2$.

In this section we have demonstrated that we should include at least the following three covariates in (2): fragment length L_i, the average intensity S_i (treated as a fixed covariate), and a factor coding the base pair at the SNP. Figures 3B and 7 suggest that we can use the following model: Let $f_1(L_i, S_i, b_i) = \mu_{b_i} + f_L(L_i) + f_S(S_i)$ with μ_{b_i} a mean level that differers for each SNP base pair (b_i =AC, AG, AT, CG, CT, or GT), f_L a cubic spline with three degrees of freedom, and f_S a cubic spline with five degrees of freedom. This model has 16 parameters and, since we have thousands of observations, we obtain very precise estimates of f. We fit the model using the EM algorithm. Examples of the estimated f_L and f_S are included in Figure 7.

Although the main reason for fitting (2) is to obtain estimates of f, two 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 $M_{i,j,k,+}$ and $M_{i,j,k,+}$. We denote these estimates as $\hat{\pi}_{i,j,k}$ and notice that they are readily available from the EM algorithm as they are the weights used by the M step. In supplemental Figure 5 we compare the predicted probabilities to the actual error rates (computed using the Hapmap data). The figure confirms that they are in fact useful. Furthermore, they provide excellent first guesses as demonstrate by Supplemental Figure 5C. Notice that even with a no call rate of 0% we achieve concordance (with Hapmap calls) rates of 99%. Second, the quantity $median(f)^2/\text{avg}_k\tau_k^2$ can be used as a quality measure for the array since it gives us a general sense of separability between genotype classes. In supplemental Figure 6 we demonstrate the utility of the SNR summary by showing plots like those in Figure 7 for the arrays producing the best and worst SNR. This figures shows that for the second array, information about genotypes is likely lost. We conjecture that a cut-off C can be defined so that removing arrays with SNRs lower than C improves the overall performance of the analysis.

Notice that even after fitting (2) we can not correct the M values by subtracting f because we do not know

Z. In the next Section we describe a genotyping algorithm that incorporates the estimated f.

4 Genotype Calling

As mentioned above, we use a supervised learning approach to 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 these calls to define *known* genotypes which in turn permits us to define a training set. However, these calls are not available for about 4% of the SNPs on the array. 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 two-stage hierarchical model and give likelihood-based closed-form definitions of the genotype regions. Details follow.

For each SNP, we define two dimensional genotype regions based on the sense and antisense M values. However, even with 90 samples, there are 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 sections becomes very useful. Using empirically derived priors for the centers and scales of the genotype regions, we give a closed form empirical Bayes solutions 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 1A suggests that genotype regions are SNP-specific when considering θ_A and θ_B as the quantities of interest. Similar pictures for M_+ and 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 a bivariate normal distributions. We therefore propose a two-level hierarchical multi-chip 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:

$$[M_{i,j,s}|Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(X_{i,j,s}) + m_{i,k,s} + \varepsilon_{i,j,k,s}. \quad (3)$$

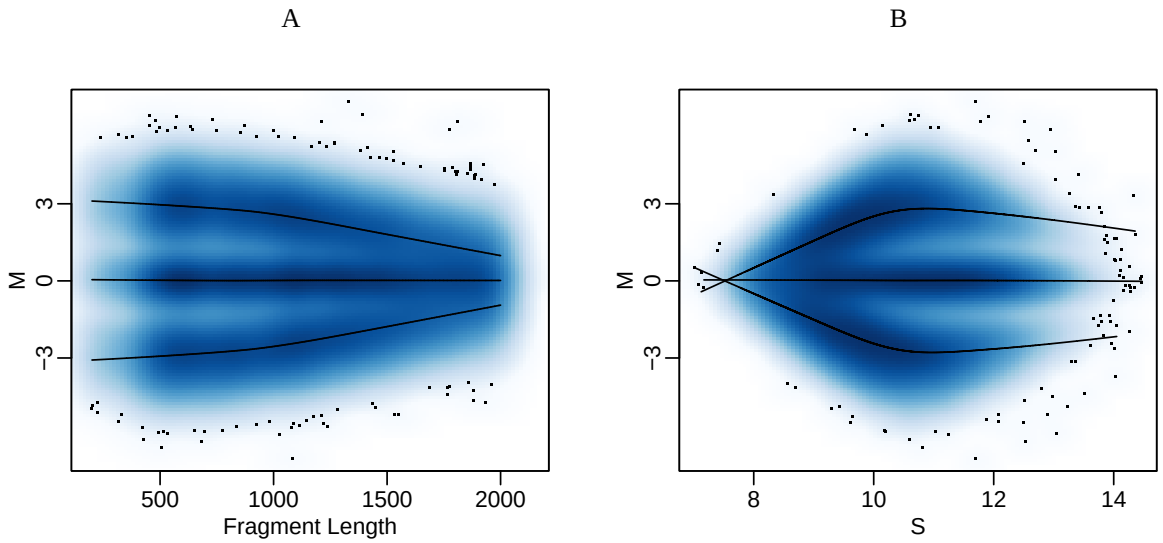


Figure 7: A) M values plotted against fragment length. Instead of plotting the points we show, with shades of blue, the data density. The solid lines are the estimated fs. B) As A) but for intensity instead of fragment length.

where $X_{i,j,s}$ and $f_{j,k}$ are as in Section 3.4 but with the j and s notation re-introduced, $m_{i,k,s}$ is the SNP-specific shift from the typical genotype region centers, and $\varepsilon_{i,j,k,s}$ represents measurement error. As mentioned in the previous section we expect different samples to have different biases thus the effects function f now depends on j. Notice that the SNP-specific covariates X also depend on sample because the average signal S may vary from sample to sample. The ms represent the cluster center shifts not accounted for by the covariates included in X.

To define the first level of our model we denote the vector of SNP-specific region centers with $\mathbf{m}=(m_{i,1,+},m_{i,2,+},m_{i,3,+},m_{i,1,-},m_{i,2,-},m_{i,3,-})'$. Data exploration shows that we can model the distribution of this vector with a multivariate normal distribution (Supplemental Figure 4). We will denote the variance-covariance matrix of $\mathbf{m}$ by V. Notice that by definition, $\mathbf{m}$ is centered at 0, since the mean levels of the three genotypes are absorbed into f. This mean level, $J^{-1}\sum_j I^{-1}\sum_i f_{j,1}(X_{i,j,s})$ is roughly 3.

The second level of the model, the variability seen within the genotypes for each SNP, is described by the εs. We assume these to be independent (conditioned on 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 is available, i.e.

$$\frac{1}{\sigma^2_{i,k,s}} \propto \frac{1}{d_{0,k}s^2_{0,k}}\chi^2_{d_{0,k}} \tag{4}$$

where $d_{k,0}$ are the degrees of freedom of the χ^2 distribution and $s^2_{0,k}$ represents 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 ms and σs in (3). 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 of $\mathbf{m}$ and the σs. 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 maximum likelihood estimates (MLE) for the $\mathbf{m}$ and σ in closed form:

$$\hat{m}_{i,k,s} = N_{i,k}^{-1}\sum_{j\in G_{i,k}}\{M_{i,j,s} - f_{j,k}(X_{i,j,s})\} \text{ and } \hat{\sigma}^2_{i,k,s} = N_{i,k}^{-1}\sum_{j\in G_{i,k}}\{M_{i,j,s} - f_{j,k}(X_{i,j,s}) - \hat{m}_{i,k,s}\}^2. \tag{5}$$

Here, $G_{i,k}$ is the set of indexes associated with samples of genotype k on SNP i and $N_{i,k}$ is the number of indexes in $G_{j,k}$. Notice that we may also use robust versions of (5).

As mentioned above there are various cases for which not enough data is 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 takes care of cases with few observations. The shrinkage step is defined as follows:

$$\tilde{\mathbf{m}}_i = (V^{-1} + \mathbf{N}_i\Sigma^{-1})^{-1}\mathbf{N}_i\Sigma^{-1}\hat{\mathbf{m}}_i \tag{6}$$

$$\tilde{\sigma}^2_{i,k,s} = \frac{(N_{i,k}-1)\hat{\sigma}^2_{i,k,s} + d_{0,k}s^2_{0,k}}{(N_{i,k}-1)+d_{0,k}}, \text{ for } N_{i,k} > 1. \tag{7}$$

For $N \leq 1$, there is no sample variance to use in equation (7) and we simply use $\tilde{\sigma}^2_{i,k,s} = s^2_{0,k}$ Here $\hat{\mathbf{m}}$ is the vector of sample means: $(\hat{m}_{i,1,+}, \hat{m}_{i,2,+}, \hat{m}_{i,3,+}, \hat{m}_{i,1,-}, \hat{m}_{i,2,-}, \hat{m}_{i,3,-})'$, Σ is a 6×6 diagonal matrix with

$\Sigma_{k,k} = \Sigma_{k+3,k+3} = s^2_{0,k}$, and $\mathbf{N}_i$ is a 6×6 diagonal matrix with entries $(N_{i,1}, N_{i,2}, N_{i,3}, N_{i,1}, N_{i,2}, N_{i,3})$. To apply equations (6) and (7) 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).

Notice that (6) and (7) 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 equations (6) and (7).

These estimated parameters, $\tilde{\mathbf{m}}$ and $\tilde{\sigma}^2$, are stored and used to call genotypes in other datasets. This is described in the next section.

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: $M_{i,j,+}, M_{i,j,-}$. Notice that these M values can come from any study and we will use the centers and scales, defined by (6) and (7), estimated from the Hapmap data. We do this by forming a likelihood based distance function δ defined by:

$$\delta_{i,k} \equiv \sum_{s \in \{+,-\}} \left\{ \log(\tilde{\sigma}_{i,k,s}) + \left(\frac{M_{i,j,s} - f_{j,k}(X_{i,j,s}) - \tilde{m}_{i,k,s}}{\tilde{\sigma}_{i,k,s}} \right)^2 \right\} \tag{8}$$

Our prediction is the genotype k that minimizes $\delta_{i,k}$. Furthermore, the log likelihood ratio tests serves as a 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. Supplemental figure 5D 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.

4.4 Recalibration

Although our pre-processing procedure greatly improves comparability across lab/studies, some slight differences in cluster centers appear to persist (data not shown). For this reason we re-calibrate the centers and scales to the new clusters in the following manner: 1) After obtaining genotype calls, use those achieving log-

likelihood ratios associated with 99% concordance rates and recalculate the centers and scales by repeating steps (5), (6), and (7). Then we compute calls using these new centers and scales.

5 Results

In this section we demonstrate that using our methodology provides better separability of cluster, call rates, and across-lab agreement than RLMM and BRLMM.

To assess the separability of clusters we compare the silhouette widths (Rousseeuw, 1987), a standard approach used in the unsupervised learning literature, for RLMM, BRLMM and CRLMM. Figure 8A shows the empirical cumulative distribution function between the RLMM and CRLMM clusters. In particular notice that the 99% worst distance is almost 3 times better for CRLMM over RLMMM. The improvements are dramatic. Similar improvements over BRLMM are observed.

In Rabbee and Speed (2006) cross-validation was used to estimate the error rates. However, Figure 1 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. Supplemental Figure 5C shows correct call rates for the initial guesses provided by the mixture model fit described in Section 3.4. Notice that the initial guesses, which are not based on a supervised learning approach, slightly outperforms RLMM. Supplemental Figure 5D 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% for heterozygotes and 99.92% from homozygotes.

Figure 1 demonstrates how CRLMM provides predictions that are useful across labs/studies. In Figure 1C the ellipses were obtained from the training data. Notice how only for CRLMM the data for other two studies fall in, or are close to, the regions defined by training on the Hapmap data. Thousands of other SNPs show behavior similar to the one shown in Figure 1. Figure 8B is a particularly interesting example. For this SNP the Hapmap data had no AA gold-standard calls. Notice how the prediction defined by (6) and (7) create

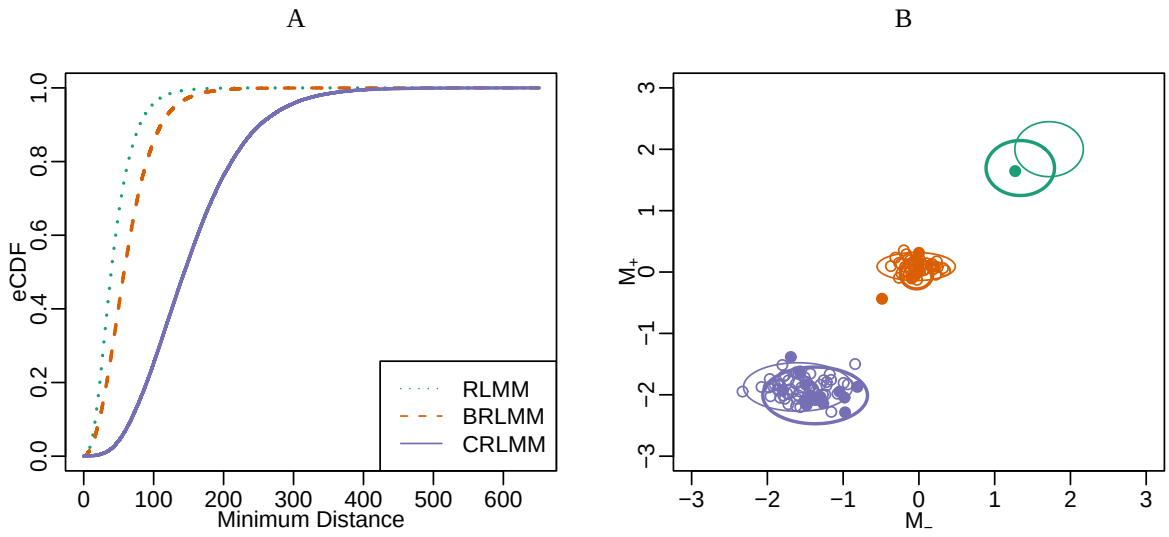


Figure 8: Illustration of usefulness of CRLMM. A) Empirical cumulative distribution function (eCDF) of the silhouette widths for the SNP-specific genotype regions for RLMM, BRLMM, and CRLMM. B) For a particular SNP, data from Hapmap are shown with empty circles and the data from lab 3 with solid circles. Notice that for the Hapmap data, there are no AA samples. The thin-line ellipses are defined using the Hapmap data. The green thin-line ellipse is the AA region predicted with the Bayesian correction (6). The thiker lines are the regions derived after recalibration for the lab 3 data. For the lab 3 data it appears that we have one AA sample and it is predicted correctly.

a region for which data from another lab, that appears to come from an AA, falls 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 four 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 SNPRMA.

The summarized data, sequence information, fragment lengths and intensity effects can then be used to make genotyping calls. Notice 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 define 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 cluster centers.

7 Acknowledgments

We thank Dan Arking, Ben Bolstad, Henrik Bengtsson, 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 BISTI grant R33HG002708-04, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil), and the National Institutes of Health Specialized Centers of Clinically Oriented Research (SCCOR) translational fund (212-2494). The work of Terence Speed is partially funded by NIH grant 2-P50-MH060398-06.

References

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

Bolstad BM, Irizarry RA, Åstrand M, Speed TP (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 TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM,

Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S (2005) Dynamic model based algorithms for screening and genotyping over 100 k snps on oligonucleotide microarrays. Bioinformatics 21(9):1958–1963
URL http://www.hubmed.org/display.cgi?uids=15657097

Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, Shapero MH (2006) Carat: a novel method for allelic detection of dna copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 7:83–83
URL http://www.hubmed.org/display.cgi?uids=16504045

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of affymetrix genechip probe level data. Nucleic Acids Research 31

Kennedy GC, Matsuzaki H, Dong S, min 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 MS, Boyce-Jacino MT, Fodor SP, Jones KW (2003) Large-scale genotyping of complex dna. Nature Biotechnology 21:1233–1237

LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, Harrington D, Sellers WR, Meyerson M (2005) Allele-specifi c amplifi cation in cancer revealed by snp array analysis. PLoS Comput Biol 1(6)
URL http://www.hubmed.org/display.cgi?uids=16322765

Liu WM, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G, Jones KW, Kennedy GC, Kulp D (2003) Algorithms for large-scale genotyping microarrays. Bioinformatics 19(18):2397–2403
URL http://www.hubmed.org/display.cgi?uids=14668223

Lönnstedt I, 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 DK, Kennedy GC, Ogawa S (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65(14):6071–6079
URL http://www.hubmed.org/display.cgi?uids=16024607

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix snp arrays. Bioinformatics 22(1):7–12

URL http://www.hubmed.org/display.cgi?uids=16267090

Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20(1):53–65

Slater HR, Bailey DK, Ren H, Cao M, Bell K, Nasioulas S, Henke R, Choo KH, Kennedy GC (2005) High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 snps. Am J Hum Genet 77(5):709–726

URL http://www.hubmed.org/display.cgi?uids=16252233

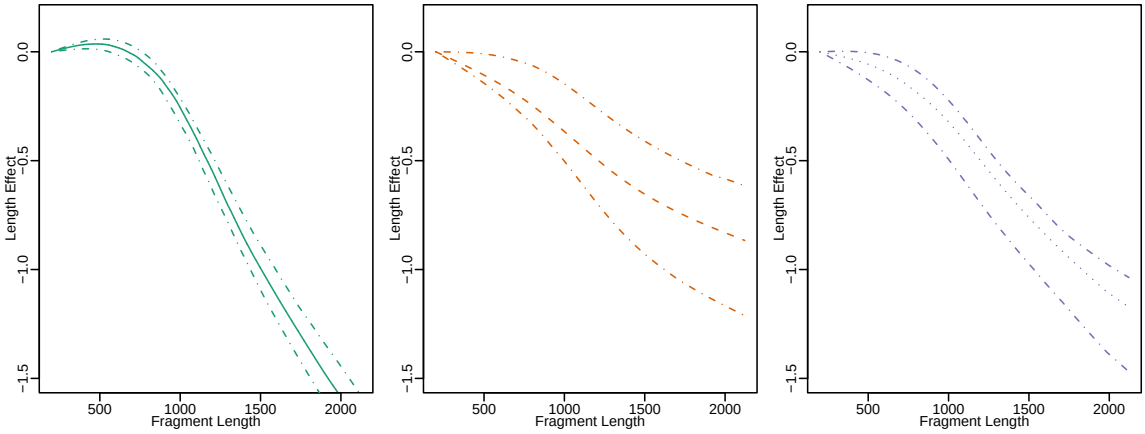
Smyth G (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3:Article 3

Uimari P, Kontkanen O, Visscher PM, Pirskanen M, Fuentes R, Salonen JT (2005) Genome-wide linkage disequilibrium from 100,000 snps in the east finland founder population. Twin Res Hum Genet 8(3):185–197

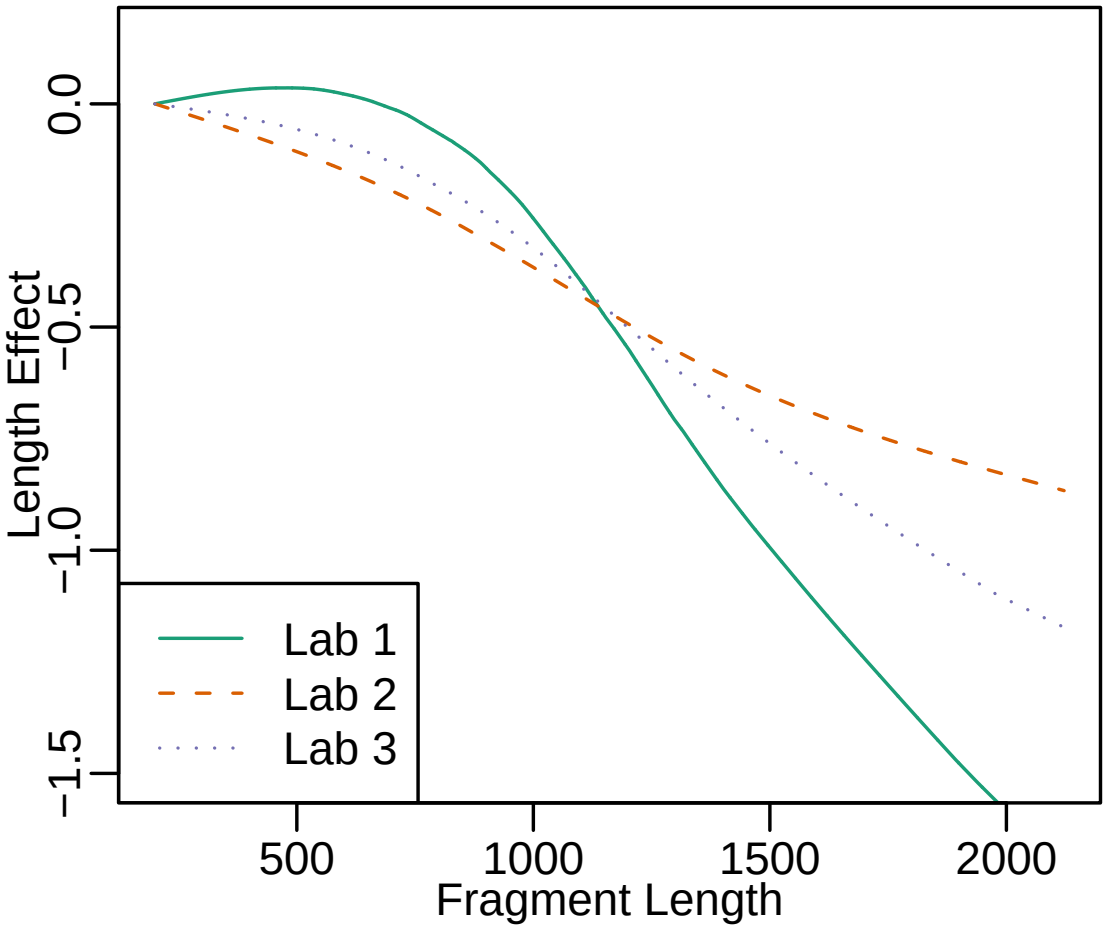
URL http://www.hubmed.org/display.cgi?uids=15989746

Wu Z, Irizarry R, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model based background adjustement for oligonucleotide expression arrays. Journal of the America Statistical Association

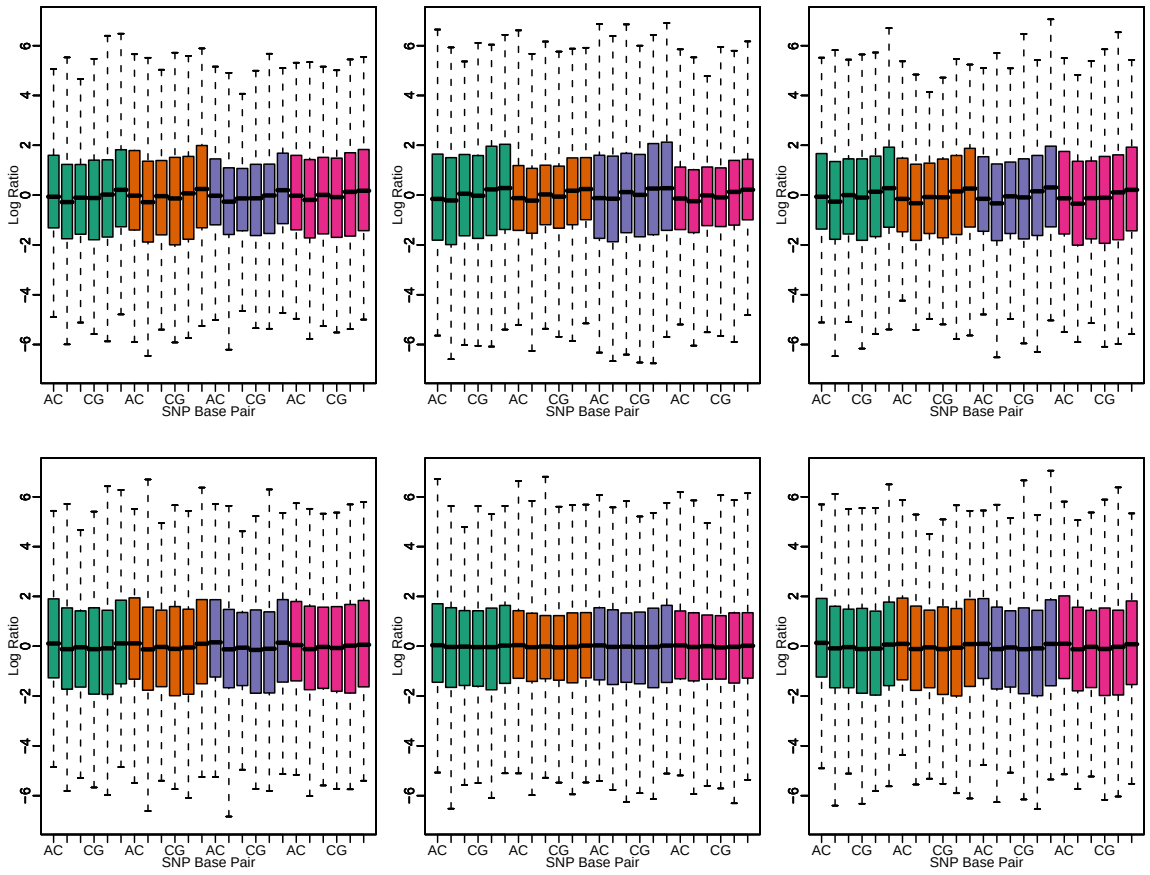
Supplemental Figures



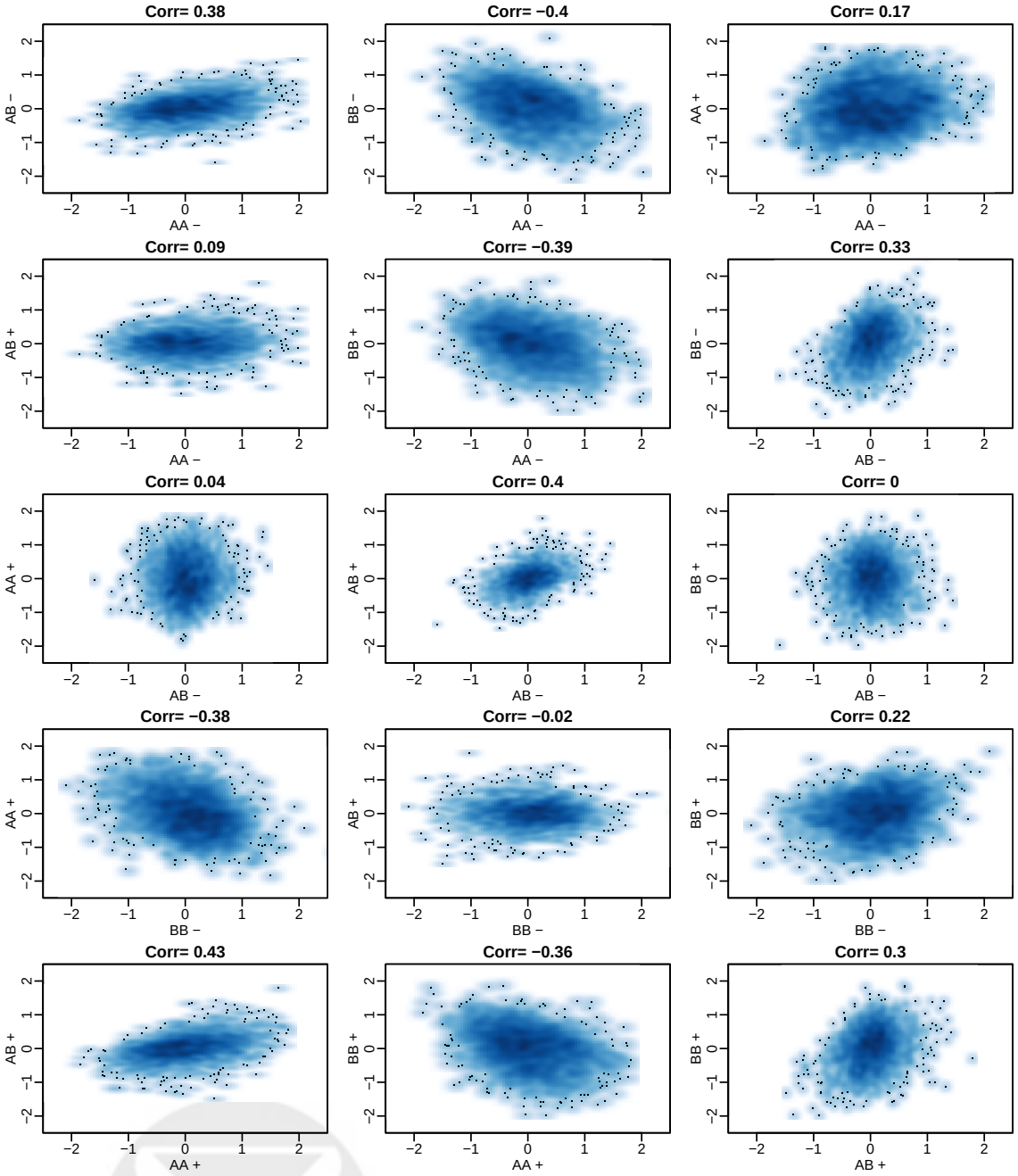
Supplemental Figure 1: A) For the Hapmap data, an estimate of the fragment length effect is obtained by fi tting a smoothing spline to the log intensity data. The average effect across samples is calculated and displayed. Point-wise 95% confi dence inetervals are also shown. B) As A) for lab 2. C) As A) for lab 3.



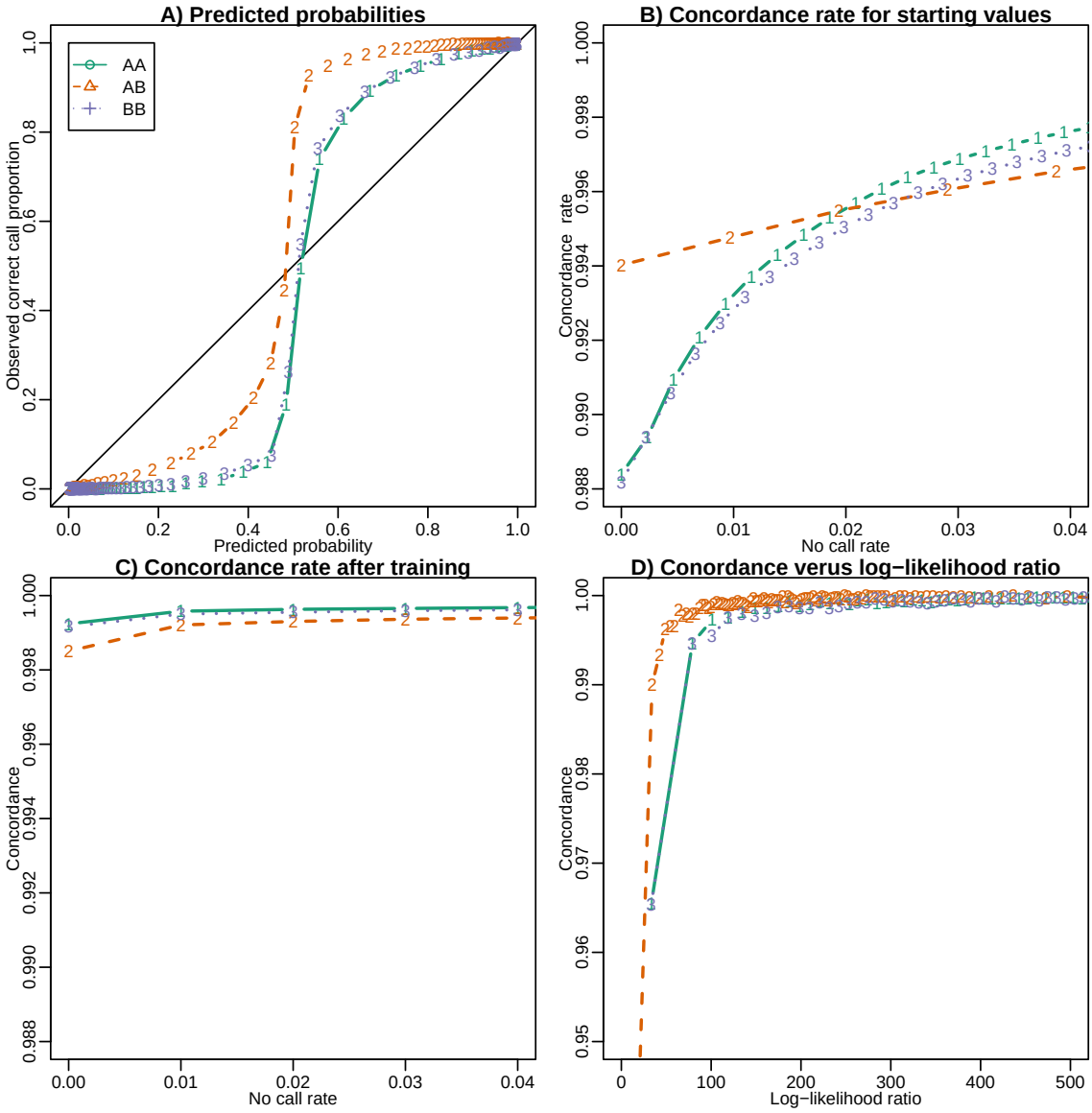
Supplemental Figure 2: The averages shown in Supplemental Figure 1 shown in one Figure.



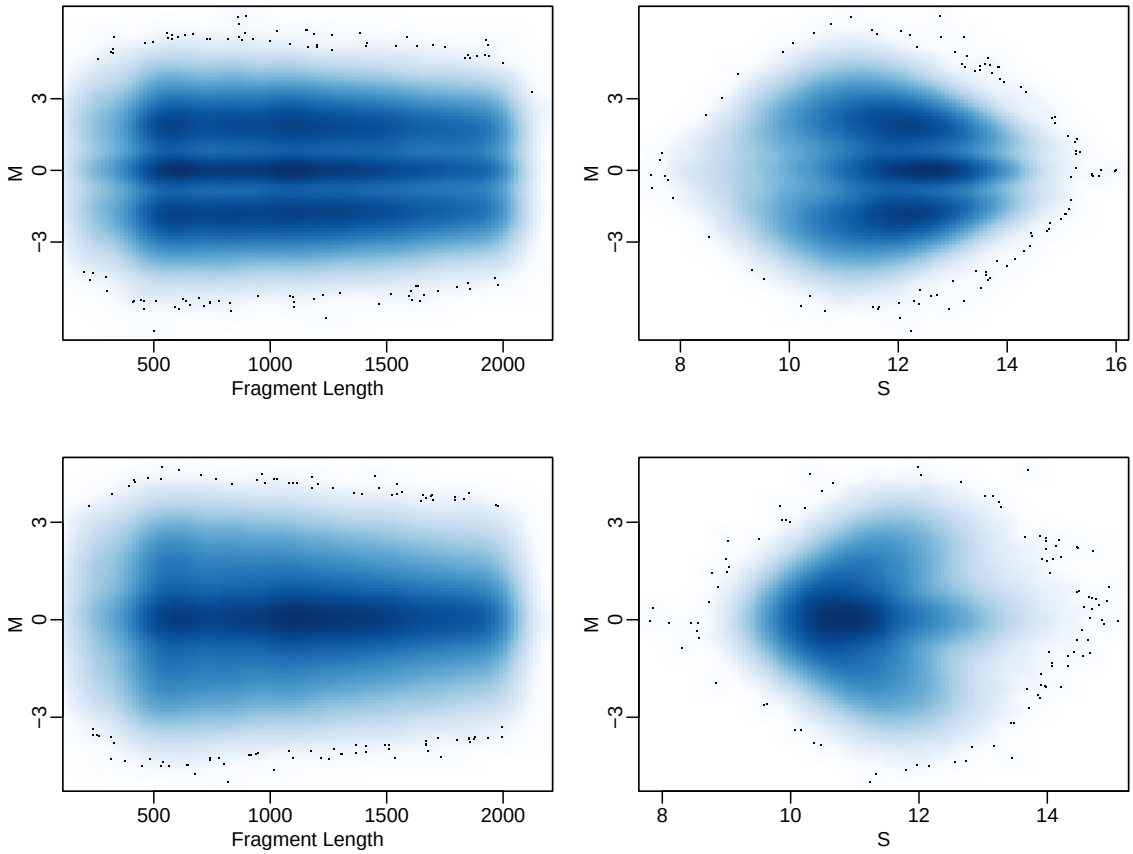
Supplemental Figure 3: Sequence effect by lab. For three randomly selected arrays in each lab we show the *M* values stratified by the SNP base pair. Above is before normalization and below is after.



Supplemental Figure 4: All pairwise scatter plots for the six estimated *m* values. The correlations are shown on top of the figures.



Supplemental Figure 5: A) For all SNPs reaching predictive probability $\hat{\pi}_k$ of being genotype k, obtained from fi tting model 2, we calculate the proportion of those SNPs that are actually k. We plot these proportions against $\hat{pi}_k$. B) Observed concordance between initial guesses, based on $\hat{pi}_k$, and Hapmap calls. C) Observed concordance between CRLMM calls and Hapmap calls. D) Observed concordance plotted against observed log-likelihood ratios.



Supplemental Figure 6: Like Figure 7 but for the arrays with best and worst SNR.