1-29-2009

QUANTIFYING UNCERTAINTY IN GENOTYPE CALLS

Benilton Carvalho
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics

Thomas A. Louis
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics

Rafael A. Irizarry
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics, ririzarr@jhsph.edu

Quantifying Uncertainty in Genotype Calls

Benilton Carvalho, Thomas A. Louis, Rafael A. Irizarry

January 27, 2009

Benilton Carvalho is a Ph.D. Candidate, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health. Drs. Irizarry (contact author, ririzarr@jhsph.edu) and Louis are Professors in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health. Support provided by NIH grants R01GM083084 from the National Institute of General Medicine, 5R01RR021967 from the National Center for Research Resource, R01 DK061662 from the National Institute of Diabetes, Digestive and Kidney Diseases, R01 HL090577 from the National Heart, Lung, and Blood Institute, R01 GM083084, a CTSA grant to the Johns Hopkins Medical Institutions, and the doctoral scholarship awarded by the Brazilian Funding Agency CAPES (Coordenação de Aprimoramento Pessoal de Nível Superior). We thank the Genetic Association Information Network (GAIN) and James Warram for the GoKind data, Simon Cawley, Affymetrix, and Dan Arking for HapMap data, Nancy Cox, Anuar Konkashbaev, Erin M. Ramos and Lisa J. McNeil for help obtaining and understanding the GoKind data, and Ingo Ruczinksi for helpful comments.

Abstract

Genome-wide association studies (GWAS) are used to discover genes underlying complex, heritable disorders for which less powerful study designs have failed in the past. The number of GWAS has skyrocketed recently with findings reported in top journals and the mainstream media. Mircorarrays are the genotype calling technology of choice in GWAS as they permit exploration of more than a million single nucleotide polymorphisms (SNPs) simultaneously. The starting point for the statistical analyses used by GWAS, to determine association between loci and disease, are genotype calls (AA, AB, or BB). However, the raw data, microarray probe intensities, are heavily processed before arriving at these calls. Various sophisticated statistical procedures have been proposed for transforming raw data into genotype calls. We find that variability in microarray output quality across different SNPs, different arrays, and different sample batches has substantial influence on the accuracy of genotype calls made by existing algorithms. Failure to account for these sources of variability, GWAS run the risk of adversely affecting the quality of reported findings. In this paper we present solutions based on a multi-level mixed model. Software implementation of the method described in this paper is available as free and open source code in the `crlmm` R/BioConductor.

KEYWORDS: Genotyping Uncertainty, Hierarchical Model, Genome-wide Association

1. INTRODUCTION

A single nucleotide polymorphism (SNP) is a single nucleotide DNA variation occurring in the genomes of individuals from the same species. For most SNPs, only two bases are observed. The two possibilities are refereed to as *alleles*. Typically, one is less common and is referred to as the *minor allele*. In this paper we will refer generically to the two alleles in any SNP as allele A and allele B. Because we have have two copies of each autosomal chromosome (maternal and paternal) there are three possible allele combinations at each SNP: AA, AB and BB. These are referred to as *genotypes*. Variations in the DNA sequences of humans can affect how humans develop diseases. Association studies enable testing for association between alleles and phenotypes, e.g. disease status. In the past, association studies would screen through hundreds of SNPs carefully selected to be near candidate genes. Today, microarray technology permits the screening of millions of SNPs across the entire genome and has revolutionized these study which are now referred to as genome wide association studies (GWAS).

Results from large GWAS, for diseases such as bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, Types 1 and 2 diabetes (Wellcome Trust Case Control Consortium 2007), diabetic nephropathy (Mueller, Rogus, Cleary, Zhao, Smiles, Steffes, Bucksa, Gibson, Cordovado, Krolewski, Nierras and Warram 2006), and kidney disfunction (Bash, Erlinger, Coresh, Marsh-Manzi, Folsom and Astor 2008), have received much attention. During the last two years, we have seen a large increase of these studies and many more are in the works. Currently, the typical data analysis procedure is to genotype a large number (thousands) of cases and controls using microarrays and search for SNPs that are statistically associated with disease. However, the process of converting raw intensities into genotype calls consists of complicated statistical manipulation of noisy data and many genotype calls are uncertain. A common analysis approach is simply to perform χ^2-tests to evaluate the association of the declared genotypes and disease, without accounting for uncertainty. As shown by Ruczinski, Li, Carvalho, Fallin, Irizarry and Louis (2009) via simulation, this failure in properly accounting for genotype uncertainty can produce inefficient or invalid associations. Of course, a valid quantification of uncertainty is a prerequisite to using it in association studies and we focus on this aspect.

In Section 2, we outline the statistical problem and describe previous work. In Section 3, we

outline the model and describe estimation procedures. In Section 4, we demonstrate the utility of our methodology with three datasets. Finally, in Section 5, we summarize and discuss our findings.

2. CONVERTING RAW INTENSITIES TO GENOTYPE CALLS

The first step, referred to as preprocessing, converts raw microarray intensities into quantities proportional to the amount of DNA in the target sample associated with each allele A and B for each SNP. We denote these summarized intensities by I_A and I_B. We do not consider this first step and refer the reader to Carvalho, Bengtsson, Speed and Irizarry (2007), Affymetrix (2006), Affymetrix (2007) and Korn, Kuruvilla, McCarroll, Wysoker, Nemesh, Cawley, Hubbell, Veitch, Collins, Darvishi, Lee, Nizzari, Gabriel, Purcell, Daly and Altshuler (2008) for details. We focus on the second step (*genotype calling*): mapping the observed intensities, (I_A, I_B), into posterior probabilities of the three possible genotypes (AA, AB, and BB) and thereby providing a confidence measure that can be used to decide which calls to omit or to introduce the appropriate genotype uncertainty when assessing association.

2.1 Current Approaches

A naive approach to genotyping is to set confidence thresholds and call genotypes based on the Is being above or below these thresholds. For example, to call an AA genotype one might require that $I_A - I_B > C$. Unfortunately, the probe effect, described in detail in the microarray literature (Li and Wong 2001*a*; Li and Wong 2001*b*; Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf and Speed 2003; Naef and Magnasco 2003; Wu, Irizarry, Gentleman, Martinez-Murillo and Spencer 2004), requires a different cut-off for each SNP. This requirement stems from the fact that the abundance of each SNP allele is measured with different probes, having different sequences and therefore different hybridization properties, resulting in large SNP to SNP variability in the distribution of intensities I_A and I_B (see Figure 1). Competing genotype calling algorithms use different strategies for determining these SNP-specific cutoffs. Many use unsupervised clustering, for example Di, Matsuzaki, Webster, Hubbell, Liu, Dong, Bartell, Huang, Chiles, Yang, mei Shen, Kulp, Kennedy, Mei, Jones and Cawley (2005) with the Dynamic Model (DM) based algorithm and Wellcome Trust Case Control Consortium (2007) with CHIAMO. The

more successful algorithms train on data for which genotypes are known, for example BRLMM (Affymetrix 2006), CRLMM (Carvalho et al. 2007), BRLMM-P and (Affymetrix 2007), and Birdseed (Korn et al. 2008). For most SNP's on these training arrays we have independent genotype calls for 270 HapMap samples (The International HapMap Consortium 2003). These calls are based on consensus results from various technologies and are considered a gold-standard.

RLMM, BRLMM and CRLMM The major manufacturers of SNP microarrays are Affymetrix and Illumina. We focus on the Affymetrix platform as this company has provided access to raw data in ways that greatly facilitate method and software development. (Illumina has agreed to make their raw data available, but in our experience their file formats and description of data are somewhat clumsy and difficult to manage. A beta version of our methods for Illumina is currently being developed). The manufacturer provides a default algorithm and during the last five years, Affymetrix has upgraded their SNP array product four times, at an amazing pace for the last upgrade, as Table 1 shows. Changes to their default algorithm have come along with changes in technology.

The default algorithm for the Affymetrix's 100K and 500K products initially was BRLMM (Affymetrix 2006), a Bayesian version of the Robust Linear Model with Mahalanobis distance (RLMM) algorithm (Rabbee and Speed 2006). But, Carvalho et al. (2007) noticed that this algorithm did not perform well across data from different laboratories and so developed a procedure that corrected for various batch-related effects resulting in the algorithm termed the *corrected* RLMM or CRLMM. Subsequently, Affymetrix has upgraded their product and their algorithm twice. The latest product is referred to as the SNP 6.0 array with Birdseed (Korn et al. 2008) as the default algorithm. However, Lin, Carvalho, Cutler, Arking, Chakravarti and Irizarry (2008) found that Birdseed has limitations similar to BRLMM in cross-lab instability of their calls. They also found CRLMM provided confidence measures that better correlated with observed accuracy. Therefore, we treat CRLMM as the leading genotyping algorithm and use the CRLMM model as a starting point for our work.

[TABLE 1 ABOUT HERE]

[FIGURE 1 ABOUT HERE]

CRLMM uses HapMap calls to define *known* genotypes, which in turn permit us to define a training set. With the training data in place, Carvalho et al. (2007) describe a supervised learning approach based on a two-stage hierarchical model. Unlike other algorithms, CRLMM models $M \equiv \log_2(I_A) - \log_2(I_B)$ instead of the intensity pair. This choice makes CRLMM more robust to probe effects because the probe effects of the two allele probes have similar additive effects and so partially cancel. This is demonstrated by Figure 2. To account for a well described dependence of M on the overall intensity $S \equiv \{\log_2(I_A) + \log_2(I_B)\}/2$, (Carvalho et al. 2007; Affymetrix 2006; Affymetrix 2007), Carvalho et al. (2007) fit splines using a mixture model and correct the bias with the fitted curves. Then, for a given SNP, the distribution of M, conditioned on genotype, is modeled as Gaussian. To account for the remaining probe effect, each SNP $i = 1, \ldots, I$ has a different mean μ_i and standard deviation (s.d.) σ_i. Sample means and standard deviations from the training data are used to estimate the μ_is and σ_is. However, due to low minor allele frequencies, even this large training dataset provides relatively few data points for the rare genotype in some SNPs.

A hierarchical model is used to improve the precision of the model parameters for these SNPs. Carvalho et al. (2007) make use of an empirical Bayes approach in which the means, conditioned on genotype, follow a multivariate normal distribution and the variances a inverse gamma distribution. The approach permits CRLMM to borrow strength from other SNPs. To make calls, CRLMM treats the estimated parameters as known and computes posterior probabilities for each genotype given the observed log-ratio M. The posteriors are then used as a confidence measure. Lin et al. (2008) found that the confidence measures provided by CRLMM were not optimal and propose an ad-hoc adjustment based on a training approach. Currently, CRLMM uses these adjusted confidence measures.

[FIGURE 2 ABOUT HERE]

3. THE ENHANCED CRLMM MODEL

3.1 The need for an enhanced approach

The current approach to determine association between SNPs and disease is to perform an association test between genotypes and outcomes, e.g., a χ^2-test for discrete outcomes. The SNPs

with too low a confidence score are "set aside," but the confidence cutoff is quite arbitrary and can affect results. Importantly, because there is more uncertainty associated with heterozygous calls (AB) than with homozygous calls (AA, BB), specifying a single cutoff for both (the current practice) can lead to bias due to informative missingness. Since CRLMM, and other calling algorithms are model based as are assessments of association, a natural extension is to develop association tests based on genotype probabilities rather than hard calls. Marchini, Howie, Myers, McVean and Donnelly (2007) and Plagnol, Cooper, Todd and Clayton (2007) use such probability based calls to combine results across different platforms. Ruczinski et al. (2009) demonstrate that using probability based calls improves the power of GWAS.

3.2 Problems with probabilities based on CRLMM

The posterior probabilities currently provided by CRLMM have three crucial limitations:

1. The posteriors are overly optimistic in favor of the genotype attaining the highest probability. The main reason for this is that the actual tails of the conditional distributions of M are longer than predicted by the Gaussian assumption. Figure 3 shows one example in which one observation has posterior of almost 1 and, yet, the call is wrong. This occurs because the posterior is a very small number divided by an even smaller one.

[FIGURE 3 ABOUT HERE]

2. The statistical uncertainty of estimates from the training step is ignored, resulting in overconfident calls for minor alleles.

3. We have observed that the genotype parameters shift from batch to batch and these batch effects are not in the current CRLMM model. As a result, batches of questionable quality are not detected by the CRLMM algorithm.

The third point is particularly troublesome. A logistics problem with these large GWAS is that hybridizations need to be processed in batches. Because DNA samples are stored in 96 well plates and robots make it convenient to run all samples in a plate at once, plates are usually confounded with hybridization times. To make matters worst, it is rarely the case that a GWAS randomizes

or controls for plate when storing samples. Therefore, it is common that plate and outcome of interest are at least partially confounded. Therefore, if genotyping algorithms do not appropriately assess these batch effects, it will be difficult if not impossible, to distinguish real from artifactual associations. The new methods presented in this paper, successfully detect problematic batches, by simply inspecting some of the estimated model parameters.

To address these deficits, we have developed an enhancement to the CRLMM model that provides much improved posterior probabilities and a powerful probability-based approach to detecting problematic SNPs and batches.

3.3 The Enhanced Hierarchical Model

We structure our analysis via the hierarchical model,

$$\begin{aligned}
Z_{ij} &\quad iid \quad \text{trinomial}\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) \\
[\boldsymbol{\mu}_i|Z_{ij}=g] &\quad iid \quad N_3(\mathbf{0},\mathbf{V}) \\
[\boldsymbol{\lambda}_{ij}|\boldsymbol{\mu}_i,Z_{ij}=g] &\quad iid \quad N_3(\mathbf{0},\mathbf{U}_j) \\
[M_{ijk}|\mu_{ig},\lambda_{ijg}] &\quad = \quad f_{jkg}(S_{ijk})+\mu_{ig}+\lambda_{ijg}+\sigma_{ig}\epsilon_{ijkg} \qquad (1) \\
[\epsilon_{ijkg}|\boldsymbol{\mu},\boldsymbol{\lambda}] &\quad iid \quad t_6(0) \\
\sigma^2_{ij} &\quad iid \quad d_g s_g^2 \frac{1}{\chi^2_{d_g}}.
\end{aligned}$$

Index $i=1,\ldots,I$ represents SNP, $j=1,\ldots,J$ represents the batch, $k=1,\ldots,K$ represents the sample, and $g=AA,AB$ or BB is the genotype. The Z's are unobserved, true genotypes, the M's the observed log ratios, $\boldsymbol{\mu}_i=(\mu_{iAA},\mu_{iAB},\mu_{iBB})'$, $\boldsymbol{\lambda}_{ij}=(\lambda_{ijAA},\lambda_{ijAB},\lambda_{ijBB})'$, the d_g are the degrees of freedom. The s_g^2 are the variance of a typical SNP and are estimated from the training data.

Data exploration demonstrates that, for large and small intensity values, M for the AA and BB genotypes are shrunken towards 0 (Carvalho et al. 2007, Figure 8). As done by Carvalho et al. (2007), we account for this intensity-dependent bias with the deterministic function f_{jkg}, requiring that $f_{jkAB}=0$ and $f_{jkAA}=-f_{jkBB}$ for all j,k. Differences across genotypes (e.g., M's for AA are on average larger than Ms for AB) are absorbed into f. These functions are estimated in a separate step, as described in detail by Carvalho et al. (2007), and are treated as known.

Introduction of the $\boldsymbol{\lambda}_{ij}$ provides the first improvement to the CRLMM model. A second improvement is to make the measurement error level more robust to outliers by assuming that the ϵ_{ijkg} are $iid\ t_6(0)$ random variables. Use of 6 degrees of freedom was selected empirically and can be changed. The σ^2_{ig} account for different SNPs having different scales of error.

3.4 Estimating parameters

Note that model (1) has $I \times (2+J) \times 3 + 12$ parameters. With $I = 906,600$ SNPs, these are too many for a global estimation procedure to be practical. In this section we describe an effective approximate modular procedure. In a first step, we take advantage of the existing training data to estimate the $\boldsymbol{\mu}$'s. Then, for each new batch j, we treat the $\boldsymbol{\mu}$'s as known and estimate the $\boldsymbol{\lambda}_j$. Both steps implement a two-stage approach wherein robust least squares parameter estimates are produced, along with their standard errors, and then these are fed into a second stage that shrinks to improve precision. Our approach permitted us to produce priors without the need of a non-linear algorithm. This was an important feature given the size of the typical datasets: one million SNPs and several hundred samples distributed across dozens of batches. This approach resulted in a powerful software tool that outperforms the default algorithm in computation speed.

Estimating SNP-specific shifts To estimate $\mathbf{V}$ we use an empirical Bayes approach (Carlin and Louis 2009). We start by obtaining robust versions of the sample means and variances of the training data to estimate the μ's and σ's by $\hat{\mu}_{ig}$ and $\hat{\sigma}_{ig}$. These robust estimates are used to account for the t-distributed errors. Since the training dataset is considered the reference from which batches deviate, we assume $\boldsymbol{\lambda} = 0$, and thus $\hat{\mu}_{ig}$ and $\hat{\sigma}_{ig}$ are unbiased estimates. Then, $\mathbf{V}$ is estimated by the sample variance-covariance of $\hat{\boldsymbol{\mu}}_i \equiv (\hat{\mu}_{iAA}, \hat{\mu}_{iAB}, \hat{\mu}_{iBB})', i = 1, \ldots, I$, producing $\hat{\mathbf{V}}$. Note that some genotypes will have very few points available in the training data to use in estimating the μ's and σ's and the estimate will be imprecise. Now, to borrow strength across SNPs we use $\hat{\mathbf{V}}$ to shrink the $\hat{\boldsymbol{\mu}}_i$ using the posterior distribution formula for a multivariate Gaussian:

$$\tilde{\boldsymbol{\mu}}_i = (\hat{\mathbf{V}}^{-1} + \mathbf{W}_i^{-1})^{-1}\mathbf{W}_i^{-1}\hat{\boldsymbol{\mu}}_i \tag{2}$$

with $\mathbf{W}_i$ a diagonal matrix with entries $s_g^2/N_{ig}, g = 1, \ldots, 3$ and N_{ig} the number of points available in the training data to estimate μ_{ig}. To shrink the variance estimates we use

$$\tilde{\sigma}_{ig}^2 = \frac{(N_{ig}-1)\hat{\sigma}_{ig}^2 + d_g s_g^2}{(N_{ig}-1)+d_g}, \text{ for } N_{ig} > 1. \quad (3)$$

When $N_{ig} \leq 1$, we simply use the posteriors s_g^2. These computations use the training data and most users will not have access to it. Therefore, we save the $\tilde{\boldsymbol{\mu}}_i$'s, $\tilde{\sigma}_{ig}$'s, and N_{ig}'s and include them as part of the software that implements the enhanced CRLMM.

Estimating batch-specific shifts Here we describe the two-stage approach used to estimate $\boldsymbol{\lambda}_j$ for each batch $j = 1, \ldots, J$. The general idea was to use the previously estimated SNP-specific shift parameters, $\tilde{\boldsymbol{\mu}}_i$'s and $\tilde{\sigma}_{ig}$'s, to produce preliminary posteriors for each genotype. These were used to create a *pseudo-training* dataset. The $\boldsymbol{\lambda}_{ij}$ were then estimated following a procedure similar to the one used to estimate $\boldsymbol{\mu}$. Some details follow.

The first step is to obtain starting values for the posteriors by assuming there is no batch specific shift, $\boldsymbol{\lambda} = 0$ and that the SNP-specific shifts $\boldsymbol{\mu}$ are known:

$$p_{ijkg}^{(0)} = \Pr(Z_{ijg} = g | M_{ijk}, \boldsymbol{\mu}_i = \tilde{\boldsymbol{\mu}}_i, \boldsymbol{\lambda}_i = 0, \sigma_{ig} = \tilde{\sigma}_{ig}).$$

We then assign a genotype to each SNP for each sample in the batch by simply maximizing these posteriors:

$$\hat{Z}_{ijk}^{(0)} = \arg\max_g p_{ijkg}^{(0)}.$$

A pseudo-training dataset was created with these calls.

The expected value of M_{ijk} conditioned on $Z_{ijk} = g$ is $f_{jkg}(S_{ijk}) + \mu_{ig} + \lambda_{ijg}$. We therefore assume that the average (in practice we compute a robust average) deviation

$$\hat{\lambda}_{ijg} \equiv \frac{1}{N_{ijg}^{(0)}} \sum_{k \in X_{ijg}} (M_{ijk} - f_{jkg}(S_{ijk}) - \tilde{\mu}_{ig}),$$

with $X_g \equiv \{k \text{ such that } \hat{Z}_{ijk}^{(0)} = g\}$ and N_{ijg} is the number of elements in X_g, is an unbiased estimate of λ_{ijg}.

In the second stage, $\mathbf{U}_j$ is estimated with the sample variance-covariance of $\hat{\boldsymbol{\lambda}}_i \equiv (\hat{\lambda}_{iAA}, \hat{\lambda}_{iAB}, \hat{\lambda}_{iBB})'$, $i = 1, \ldots, I$. With $\hat{\mathbf{U}}_j$, the estimate of $\mathbf{U}_j$, in place, we shrink the $\hat{\boldsymbol{\lambda}}_{ig}$ as done in (2):

$$\tilde{\boldsymbol{\lambda}}_i = (\hat{\mathbf{U}}_j^{-1} + \mathbf{W}_i^{-1})^{-1} \mathbf{W}_i^{-1} \hat{\boldsymbol{\lambda}}_i \quad (4)$$

with $\mathbf{W}_i$ as above.

3.5 Producing posteriors

Using the current CRLMM method, posterior calls were particularly over-confident. This is consistent with the fact that the estimated $\tilde{\boldsymbol{\mu}}_i$ are assumed to be known. We developed a procedure that permits us to account for the uncertainty associated with estimating the SNP-specific and batch-specific shifts. In this Section, we illustrate the idea by demonstrating the approach when there are no batch-specific shifts and the ϵ's are normally distributed. In the Appendix, we describe the details needed for the full model, including the batch-specific shifts and the t-distribution assumption.

Consider the simplified model with no batch effect (thus j is omitted)

$$[M_{ik}|Z_{ik} = g, \mu_{ik} = \hat{\mu}_{ik}] = f_{kg}(S_{ik}) + \hat{\mu}_{ik} + \epsilon_{ikg}, \tag{5}$$

with ϵ normally distributed with mean 0 and variance σ^2_{ig}. In our approach, we estimate with a shrunken version of the sample average, but for simplicity we will assume we used the sample average. In this case, the estimated SNP-specific shifts, $\hat{\mu}_{ig}$, are normally distributed with mean zero and variance σ^2_{ig}/N_{ig}, with N_{ig} the number of points available in the training data to estimate μ_{ig} as in (4). We can then show that

$$\begin{aligned}
\mathbb{E}\left[M_{ik}|Z_{ik} = g\right] &= \mathbb{E}_{\mu_{ig}}\left[\mathbb{E}\left(M_{ik}|Z_{ik} = g, \mu_{ig}\right)\right] \\
&= \mathbb{E}_{\mu_{ig}}\left[f_{kg}(S_{ik}) + \mu_{ig}\right] \\
&= f_{k,g}(S_{ik}) \qquad (6) \\
\mathbb{V}\left[M_{ik}|Z_{ik} = g\right] &= \mathbb{V}\left[\mathbb{E}\left(M_{ik}|Z_{ik} = g, \mu_{ig}\right)\right] + \mathbb{E}\left[\mathbb{V}\left(M_{ik}|Z_{ik} = g, \mu_{ig}\right)\right] \\
&= \mathbb{V}\left[f_{kg}(S_{ik}) + \mu_{ig}\right] + \mathbb{E}\left(\sigma^2_{ig}\right) \\
&= \frac{\sigma^2_{ig}}{N_{ig}} + \sigma^2_{ig} \\
&= \left(1 + \frac{1}{N_{ig}}\right)\sigma^2_{ig}. \qquad (7)
\end{aligned}$$

The posterior probabilities are produced by normalizing the joint densities of the log-ratios M and genotypes g:

$$Pr(Z_{ik} = g|M_{ik} = m) = \frac{P(Z_{ik} = g)\phi_{M_{ik}|Z_{ik}=g}(m)}{\sum_{g=1}^{3} P(Z_{ik} = g)\phi_{M_{ik}|Z_{ik}=g}(m)}. \tag{8}$$

with $\phi_{M_{ik}|Z_{ik}=g}(m)$ representing a normal density with mean and variance shown in equations (6) and (7) respectively. A similar calculation, delineated in the Appendix, provides posteriors for the full model.

3.6 Quality scores

Carvalho et al. (2007) present a powerful procedure for detecting problematic arrays based on the estimated f. Here we present a quality assessment procedure for SNPs and hybridization batches. The quality of batch j can be quantified by the diagonal entries of $\hat{\mathbf{U}}_j$. We demonstrate the utility of this approach in the results section. For SNPs, we can quantify quality by assigning a posterior probability of being an outlier to each shift, i.e. $\boldsymbol{\mu}_i$ or $\boldsymbol{\lambda}_{ij}$. Using the fitted prior distributions for $\boldsymbol{\mu}_i$ and $\boldsymbol{\lambda}_i$ we introduce a density function h_0 for outlying $\boldsymbol{\mu}$ and compute the posterior probability:

$$\Pr(\text{Shift } i \text{ is outlier}|\boldsymbol{\mu}_i) = \frac{h_0(\boldsymbol{\mu}_i)}{h_0(\boldsymbol{\mu}_i) + \phi(\boldsymbol{\mu}_i)}$$

with $\phi(\boldsymbol{\mu}) = (2\pi)^{-3/2}|\mathbf{V}|^{-1/2}\exp(\boldsymbol{\mu}'\mathbf{V}^{-1}\boldsymbol{\mu})$. A practical choice for h is the three dimensional uniform distribution covering all possible values of $\boldsymbol{\mu}$. We perform a similar computation for $\boldsymbol{\lambda}_{ij}$ for each batch j. To illustrate the advantage of the empirical Bayes approach, we plotted the λ_{ijAA} versus λ_{ijAB} and λ_{ijAA} versus λ_{ijBB} (Figure 4). The large number of SNPs permitted us to borrow strength across SNPs. The non-zero correlations permitted us to borrow strength across genotypes.

[FIGURE 4 ABOUT HERE]

3.7 Software

The methodology described here is available via the `crlmm` R/BioConductor package. To demonstrate its performance, we compared CRLMM to Birdseed, the standard genotyping tool for SNP 6.0 arrays, on the 270 HapMap samples. On this set, the maximum amount of memory used by CRLMM, during preprocessing, was 3.2 GB. After preprocessing, the memory usage was reduced to 2 GB. CRLMM needed 52 minutes to complete the task. Birdseed used 845MB for most of the process, increasing slowly to 900 MB and took 150 minutes. The comparisons were executed on a four-processors system (3GHz Dual-Core AMD Opteron Processor 2222) with 32GB RAM.

4. RESULTS

We assessed the performance of the enhanced CRLMM with a comparison to CRLMM and Birdseed, the default algorithm provided by the manufacturer. We use three datasets:

A) 143 Hapmap samples hybridized by Affymetrix.

B) 55 HapMap samples hybridized at Johns Hopkins.

C) 3,050 samples from the GoKinD dataset (Mueller et al. 2006) made available through the Genetic Association Information Network (GAIN).

We used HapMap samples because knowing the "truth" permitted us to effectively assess our methodology. Note that, although the same samples, the hybridizations used here were not the same as the set used to train our algorithm. The Dataset C provided a large set with 34 different batches defined by the 96-well plate in which the samples were stored. To assess performance with this dataset we computed the concordance between calls obtained by running the algorithm on all samples to calls obtained by running the algorithms by batch. We obtained calls for each dataset with the birdseed, CRLMM and the enhanced CRLMM.

4.1 Overall accuracy

We first compared over-all accuracy using Datasets A and B. We calculated accuracy, i.e. proportion of correct calls, for calls with confidence scores above a given cutoff. Various cutoffs were considered. We then plotted accuracy against the proportion of calls below the confidence cutoffs. The accuracy versus drop rate (ADR) plots demonstrated that, overall, the the enhanced CRLMM outperformed the other two algorithms (Figure 5).

[FIGURE 5 ABOUT HERE]

4.2 Posteriors

To assess the validity of the posteriors, we compared observed accuracy to reported posteriors. Specifically, we stratified calls by their associated posterior and each strata we computed the proportion of correct calls. We then plotted these against each other with the expectation that they fall on the identity line. Although the current CRLMM does not use posteriors as

a confidence measure, we obtained the posteriors by modification to the CRLMM code. The enhanced CRLMM improved the posteriors provided by the current CRLMM which clearly were optimistic (Figure 6).

[FIGURE 6 ABOUT HERE]

4.3 SNP quality metrics

For Datasets A and B we computed SNP quality scores as described in the Methods Section. Namely, for each SNP, we computed the posterior probability of the estimated λ not being an outlier. To demonstrate the utility of this metric we stratified SNPs by the quality score reported for Datasets A and B, and created ADR plots for each strata (Figure 7). Note, that by restricting attention to SNPs with QC scores above 0.25, we obtained near perfect results. For Datasets A and B, 98.63% and 99.18% of the SNPs surpassed this cutoff.

[FIGURE 7 ABOUT HERE]

4.4 Batch quality metrics

For Dataset C we generated calls in two ways: by using and ignoring batch information. We then computed the concordance between these two sets of calls for each batch. We considered batches with lower concordance to be problematic. The percentage of samples with signal to noise ratios, as defined by Carvalho et al. (2007), below 5 was the best predictor of low quality batches (Figure 8A). The quality score, based on the distribution of the $\boldsymbol{\lambda}_j$s and summarized by $\mathbf{U}_j$, predicted low quality as well (Figure 8B).

[FIGURE 8 ABOUT HERE]

Our batch quality score also effectively predicted the differences in accuracy observed in Figure 5. Note that Datasets A and B had batch quality scores of 0.0337 and 0.0745 respectively.

5. DISCUSSION

We have presented a multi-level enhancement to the CRLMM model described by Carvalho et al. (2007). Our model accounts for three levels of variability in SNP array data: 1) SNP specific

shifts, 2) hybridization batch shifts to each SNP and 3) heavy tailed measurement error. By explicitly modeling these sources of uncertainty, the estimated posterior probabilities are much improved as compared to those offered by the current version of CRLMM. We also incorporate the variability associated with estimating model parameters with training data. Our approach produced priors with superior properties to those produced by the current CRLMM model. The refinements will improve the accuracy of downstream results obtained from probability based association tests such as the one described by Ruczinski et al. (2009).

We have also described methodology useful for detecting problematic SNPs and hybridization batches. We find the latter contribution particularly important. Adapting analysis tools to deal with hybridization batch effects should be a priority of analysis groups working with genome-wide association study (GWAS) data. Due to experimental logistics, GWAS rarely control or randomize for well-plate, for example when using external controls. Therefore, an undetected problematic batch could make it difficult, if not impossible, to distinguish reported associations from artifactual ones such driven by hybridization batches. We have presented a powerful solution that predicts problematic batches and can be easily incorporated into any analysis pipeline.

A. INCORPORATING BATCH EFFECTS AND USING A MORE ROBUST DISTRIBUTION FOR RESIDUALS

Here, we describe further details on how to derive posterior probabilities that account not only for the uncertainty induced by the estimation of the SNP-specific shifts, but also for batch effects and heavier tails for the density of the residuals in model (1), which we recapitulate:

$$[M_{ijk}|Z_{ij} = g, \mu_{ig}, \lambda_{ijg}] = f_{jkg}(S_{ijk}) + \mu_{ig} + \lambda_{ijg} + \sigma_{ig}\epsilon_{ijkg},$$

where:

$$\begin{aligned}
[\boldsymbol{\mu}_i|Z_{ij} = g] &\quad iid \quad N_3(\mathbf{0}, \mathbf{V}) \\
[\boldsymbol{\lambda}_{ij}|\boldsymbol{\mu}_i, Z_{ij} = g] &\quad iid \quad N_3(\mathbf{0}, \mathbf{U}_j) \\
[\epsilon_{ijkg}|\boldsymbol{\mu}_i, \boldsymbol{\lambda}_{ij}] &\quad iid \quad t_\nu \\
\sigma^2_{ig} &\quad iid \quad d_g s^2_g \frac{1}{\chi^2_{d_g}}
\end{aligned}$$

As described earlier, $i = 1, \ldots, I$ represents SNP, $j = 1, \ldots, J$ represents the batch, $k = 1, \ldots, K$ represents the sample, and $g = AA, AB$ or BB is the genotype. Note that, $f_{jkg}(S_{ijk})$ is a deterministic function estimated during preprocessing.

In order to obtain the posteriors $Pr(Z_{ij} = g|M_{ijk} = m)$, we initially derive the joint distribution of the log-ratio and genotype. This is easily achieved by integrating the complete joint density over μ_{ig}, λ_{ijg} and σ_{ig}.

$$\begin{aligned} & f\left(M_{ijk} = m, Z_{ij} = g\right) = \\ & = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty f\left(M_{ijk} = m, Z_{ij} = g, \mu_{ig}, \lambda_{ijg}, \sigma_{ig}\right)\, d\mu_{ig}\, d\lambda_{ijg}\, d\sigma_{ig} \\ & = \iiint f\left(M_{ijk} = m | Z_{ij} = g, \mu_{ig}, \lambda_{ijg}, \sigma_{ig}\right) f\left(\mu_{ig} | Z_{ij} = g, \lambda_{ijg}, \sigma_{ig}\right) \times \\ & \quad \times f\left(\lambda_{ijg} | Z_{ij} = g, \sigma_{ig}\right) f\left(\sigma_{ig} | Z_{ij} = g\right) Pr(Z_{ij} = g)\, d\mu_{ig}\, d\lambda_{ijg}\, d\sigma_{ig} \end{aligned} \tag{A.1}$$

The conditional densities of $[M_{ijk}|Z_{ij} = g, \mu_{ig}, \lambda_{ijg}, \sigma_{ig}]$ and $[\sigma_{ig}|Z_{ij} = g]$ can be derived with simple variable transformations:

$$\begin{aligned} \mathbb{F}_{M_{ijk}|Z_{ij}=g,\mu_{ig},\lambda_{ijg},\sigma_{ig}}(m) &= \mathbb{F}_{\epsilon_{ijkg}}\left(\frac{m - f_{jkg}(S_{ijk}) - \mu_{ig} - \lambda_{ijg}}{\sigma_{ig}}\right) \\ f_{M_{ijk}|Z_{ij}=g,\mu_{ig},\lambda_{ijg},\sigma_{ig}}(m) &= \frac{1}{\sigma_{ig}} f_{\epsilon_{ijkg}}\left(\frac{m - f_{jkg}(S_{ijk}) - \mu_{ig} - \lambda_{ijg}}{\sigma_{ig}}\right) \end{aligned} \tag{A.2}$$

$$\begin{aligned} \mathbb{F}_{\sigma_{ig}|Z_{ij}=g}(s) &= \mathbb{F}_{\sigma^2_{ig}|Z_{ij}=g}(s^2) \\ f_{\sigma_{ig}|Z_{ij}=g}(s) &= 2s f_{\sigma^2_{ig}|Z_{ij}=g}\left(s^2\right) \end{aligned} \tag{A.3}$$

Using results (A.2) and (A.3), the joint density (A.1) is rewritten below, on Equation (A.4) and later simplified on Equation (A.5).

$$\begin{aligned} & f\left(M_{ijk} = m, Z_{ij} = g\right) \\ & = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty f\left(M_{ijk} = m, Z_{ij} = g, \mu_{ig}, \lambda_{ijg}, \sigma_{ig}\right)\, d\mu_{ig}\, d\lambda_{ijg}\, d\sigma_{ig} \\ & = Pr(Z_{ij} = g) \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{\left(d_0 s_0^2/2\right)^{d_0/2} \Gamma\left(\frac{\nu+1}{2}\right)}{\pi^{3/2}\sqrt{u_{gg} v_{j,gg} \nu}\, \Gamma\left(\frac{d_0}{2}\right) \Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sigma_{ig}^{2+\nu}} \times \\ & \quad \times \left[1 + \frac{1}{\nu}\left(\frac{m - f_{jkg}(S_{ijk}) - \mu_{ig} - \lambda_{ijg}}{\sigma_{ig}}\right)^2\right]^{-\frac{\nu+1}{2}} \times \\ & \quad \times \exp\left\{-\frac{1}{2}\left(\frac{\lambda_{ijg}^2}{u_{j,gg}} + \frac{\mu_{ig}}{v_{gg}} + \frac{d_0 s_0^2}{\sigma_{ig}^2}\right)\right\}\, d\mu_{ig}\, d\lambda_{ijg}\, d\sigma_{ig} \end{aligned} \tag{A.4}$$

$$
\begin{aligned}
& f\left(M_{ijk}=m, Z_{ij}=g\right) \\
& \propto \quad Pr(Z_{ij}=g)\frac{1}{\sqrt{u_{gg}v_{j,gg}}}\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\sigma_{ig}^{2+\nu}} \times \\
& \qquad \times\left[1+\frac{1}{\nu}\left(\frac{m-f_{jkg}(S_{ijk})-\mu_{ig}-\lambda_{ijg}}{\sigma_{ig}}\right)^2\right]^{-\frac{\nu+1}{2}} \times \\
& \qquad \times \exp\left\{-\frac{1}{2}\left(\frac{\lambda_{ijg}^2}{u_{j,gg}}+\frac{\mu_{ig}}{v_{gg}}+\frac{d_0 s_0^2}{\sigma_{ig}^2}\right)\right\}\, d\mu_{ig}\, d\lambda_{ijg}\, d\sigma_{ig}. \qquad \text{(A.5)}
\end{aligned}
$$

The posterior probabilities can be found by using the Bayes theorem and the Law of Total Probabilities (A.6).

$$
Pr(Z_{ij}=g|M_{ijk}=m) \quad = \quad \frac{f\left(M_{ijk}=m, Z_{ij}=g\right)}{\sum_{g=1}^3 f\left(M_{ijk}=m, Z_{ij}=g\right)} \qquad \text{(A.6)}
$$

Because the integral on Equation (A.5) does not have a closed-form, either the integrand should be approximated or Monte-Carlo methods be used. In particular, a saddle-point approximation for the kernel of the Student-t density will be effective.

REFERENCES

Affymetrix (2006), "BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set," *White Paper*, pp. 1–18.

Affymetrix (2007), "BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array," *White Paper*, pp. 1–16.

Bash, L., Erlinger, T., Coresh, J., Marsh-Manzi, J., Folsom, A., and Astor, B. (2008), "Inflammation, Hemostasis, and the Risk of Kidney Function Decline in the Atherosclerosis Risk in Communities (ARIC) Study," *Am J Kidney Dis*, p. Ahead of print.

Carlin, B. P., and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, 3rd. edn, : Chapman & Hall, CRC Press.

Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. (2007), "Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data," *Biostatistics*, 8(2), 485–99.

Di, X., Matsuzaki, H., Webster, T. A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., mei Shen, M., Kulp, D., Kennedy, G. C., Mei, R., Jones, K. W., and Cawley, S. (2005), "Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays," *Bioinformatics*, 21(9), 1958–63.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data.," *Biostatistics*, 4(2), 249–264.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. (2008), "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nat Genet*, 40(10), 1253–60.

Li, C., and Wong, W. H. (2001*a*), "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proc Natl Acad Sci USA*, 98(1), 31–6.

Li, C., and Wong, W. H. (2001*b*), "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biol*, 2(8), RESEARCH0032.

Lin, S., Carvalho, B., Cutler, D., Arking, D., Chakravarti, A., and Irizarry, R. A. (2008), "Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays," *Genome Biol*, 9(4), R63.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007), "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nat Genet*, 39(7), 906–13.

Mueller, P. W., Rogus, J. J., Cleary, P. A., Zhao, Y., Smiles, A. M., Steffes, M. W., Bucksa, J., Gibson, T. B., Cordovado, S. K., Krolewski, A. S., Nierras, C. R., and Warram, J. H. (2006), "Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes," *J Am Soc Nephrol*, 17(7), 1782–90.

Naef, F., and Magnasco, M. O. (2003), "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," *Phys. Rev. E*, 68(1), 011906.

Plagnol, V., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2007), "A method to address differential bias in genotyping in large-scale association studies," *PLoS Genet*, 3(5), e74.

Rabbee, N., and Speed, T. P. (2006), "A genotype calling algorithm for affymetrix SNP arrays," *Bioinformatics*, 22(1), 7–12.

Ruczinski, I., Li, Q., Carvalho, B., Fallin, M. D., Irizarry, R. A., and Louis, T. A. (2009), "Association Tests that Accommodate Genotyping Errors," *Submitted to Journal of the American Statistical Association*, .

The International HapMap Consortium (2003), "The International HapMap Project.," *Nature*, 426(6968), 789–796.

Wellcome Trust Case Control Consortium (2007), "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447(7145), 661–78.

Wu, Z., Irizarry, R. A., Gentleman, R. C., Martinez-Murillo, F., and Spencer, F. (2004), "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association*, 99(468), 909–917.

Product	Release Date	SNPs
10K Array	July/2003	10,000
100K Array Set	April/2004	116,000
500K Array Set	September/2005	500,000
SNP 5.0 Array	February/2007	500,000
SNP 6.0 Array	May/2007	906,600

Table 1: General information about Affymetrix SNP arrays platforms. The coverage has increased significantly over the past five years. The 100K and 500K array sets are comprised of two arrays each, Xba+Hind and Nsp+Sty, respectively. The SNP counts are approximate.

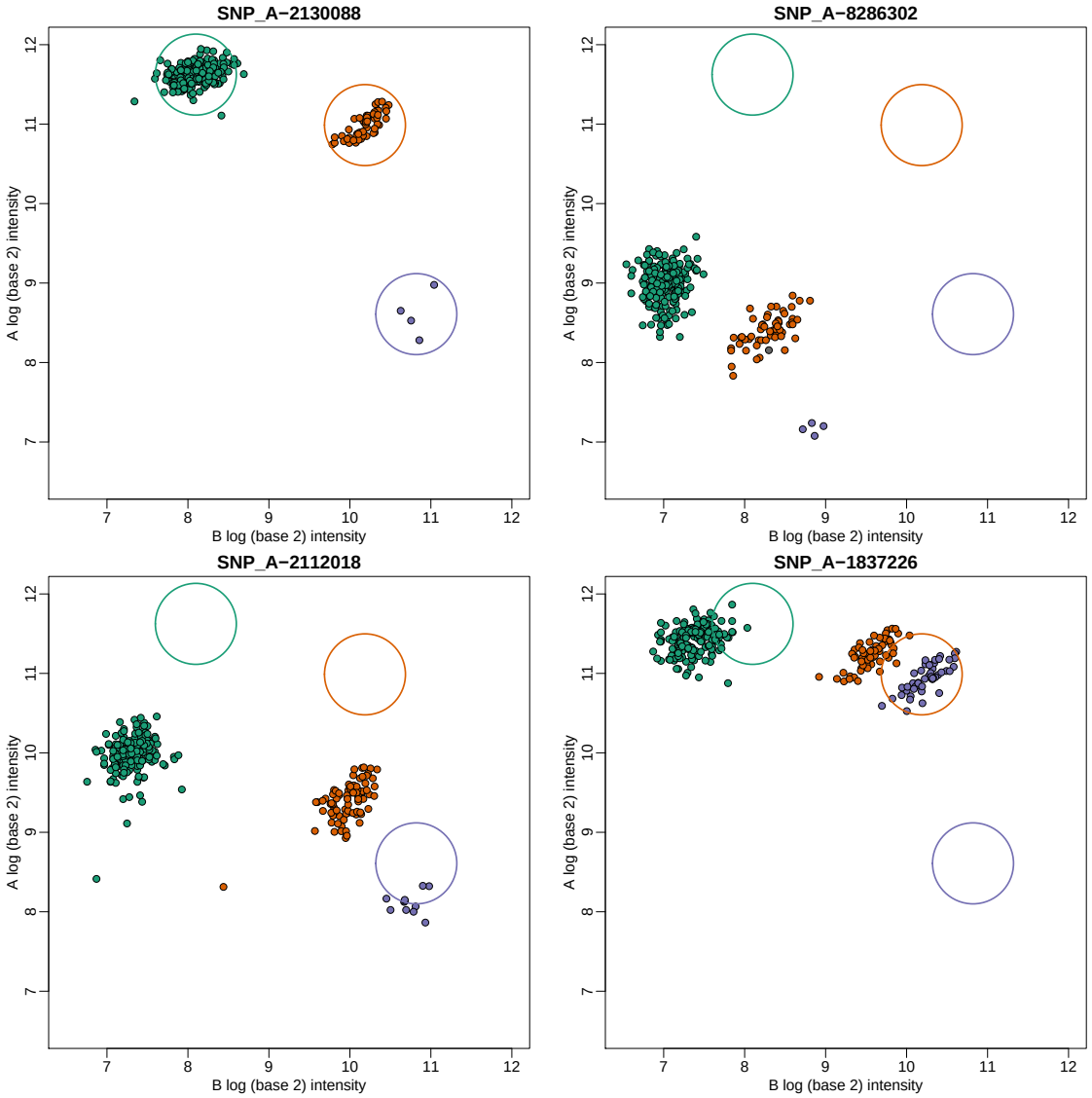

Figure 1: The intensity of both alleles is plotted against each other, i.e. I_A versus I_B, for four randomly selected SNP's. The three circles illustrate the distribution of the data for each genotype (AA: green; AB: orange; BB: violet) for the first SNP. Note that these regions are incompatible with the data for the three other SNPs. This figure illustrates that the SNP to SNP variability is much larger than the within SNP variability and that naive genotyping algorithms that define global thresholds are not appropriate.

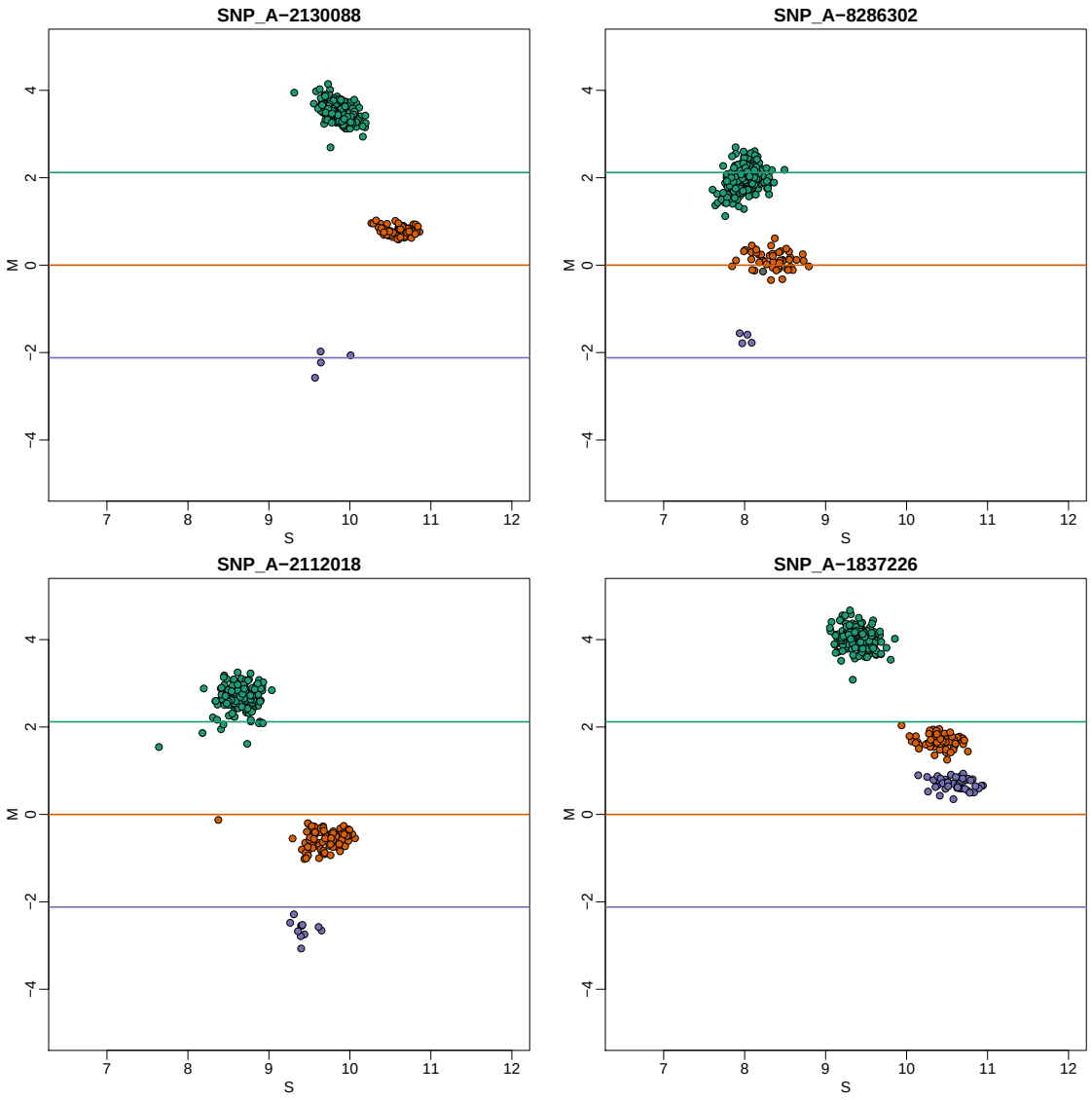

Figure 2: The advantage of modeling M instead of (I_A, I_B): Here we plot M versus S for the same data shown in Figure 1. The across SNP variability is smaller for M than for S. However, the probe effect is not completely removed as seen in the SNP in the bottom right pane. Note that for this SNP the cluster centers are substantially shifted.

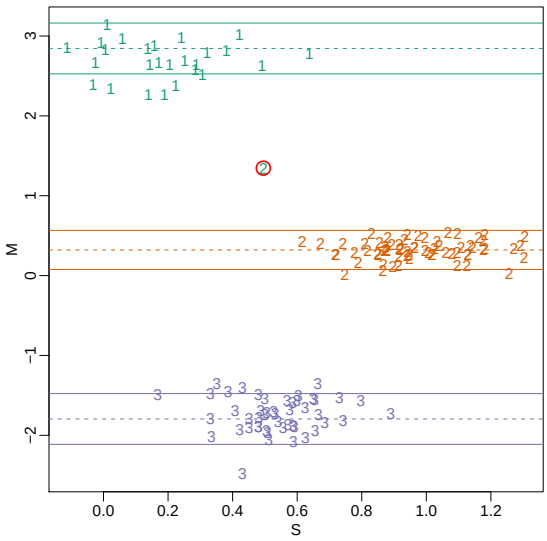

Figure 3: An example of a SNP with three clear clusters: The calls derived from the algorithm are represented by the colors (AA: green; AB: orange and BB: violet). The observation with the red circle around it was incorrectly called AA and, under the normal assumption for the residuals, the posterior was greater than 0.999. With the assumption that the residuals follow a t-distribution, the posterior was penalized and reduced to 0.500.

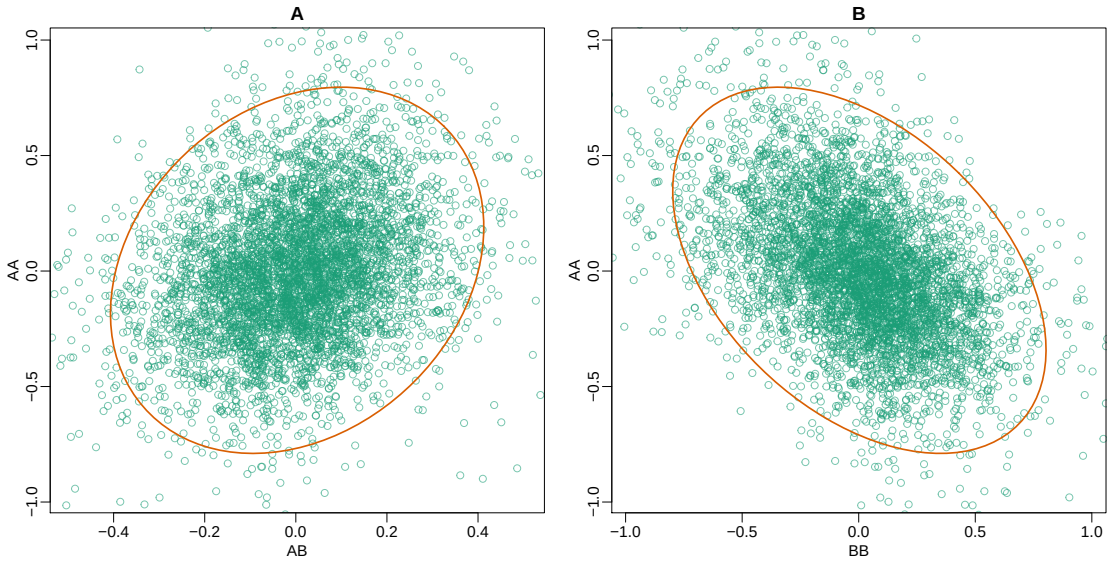

Figure 4: Plots of $\hat{\boldsymbol{\lambda}}$ for a given batch. Note that they are correlated. We take advantage of this correlation to predict or improve precision of shifts when not enough training data is available. The ellipses delimit the 95% confidence regions of the estimated distribution. SNPs with Points outside these regions are associated with abnormal movements and are flagged as possible outliers. Panel A shows $\hat{\lambda}_{AA}$ versus $\hat{\lambda}_{AB}$. Panel B $\hat{\lambda}_{AA}$ versus $\hat{\lambda}_{BB}$. The plot for $\hat{\lambda}_{AA}$ versus $\hat{\lambda}_{AB}$ is similar to that shown Panel A.

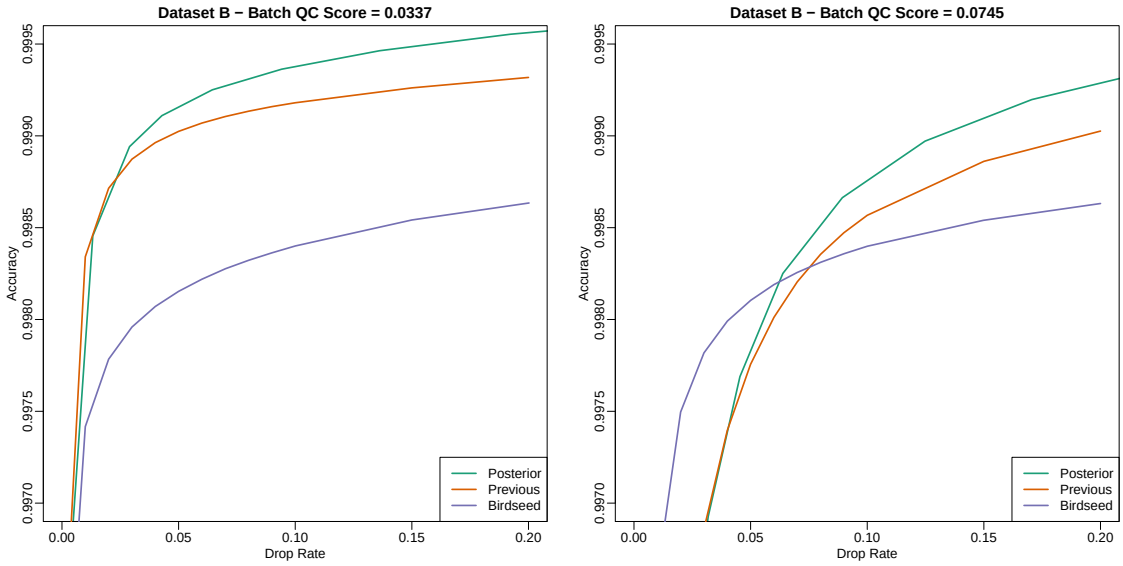

Figure 5: Accuracy versus drop rate (ADR) plots for Datasets A and B. For the first set, the enhanced CRLMM outperforms both Birdseed and the previous CRLMM implementation. For the second set, it outperforms the other two methods roughly at a drop rate of 6%. Also note that the accuracy on the second dataset is lower when compared to the first one, indicating significant variation on the quality of the two sets.

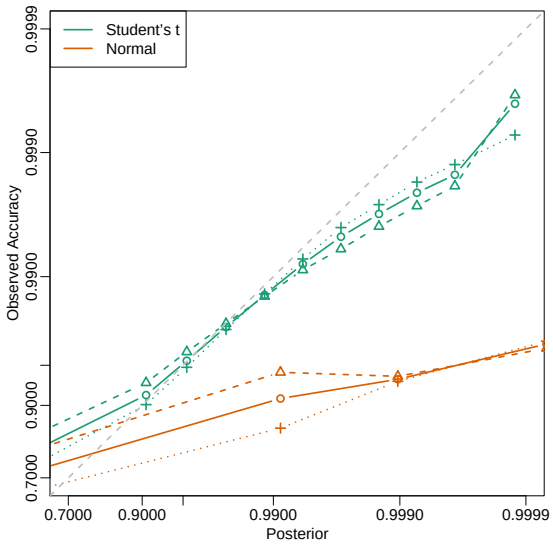

Figure 6: For Dataset A, calls were stratified by their associated posterior. For each strata the observed accuracy was computed by comparing to HapMap gold standard calls. The new CRLMM is compared to the current CRLMM which is clearly to optimistic.

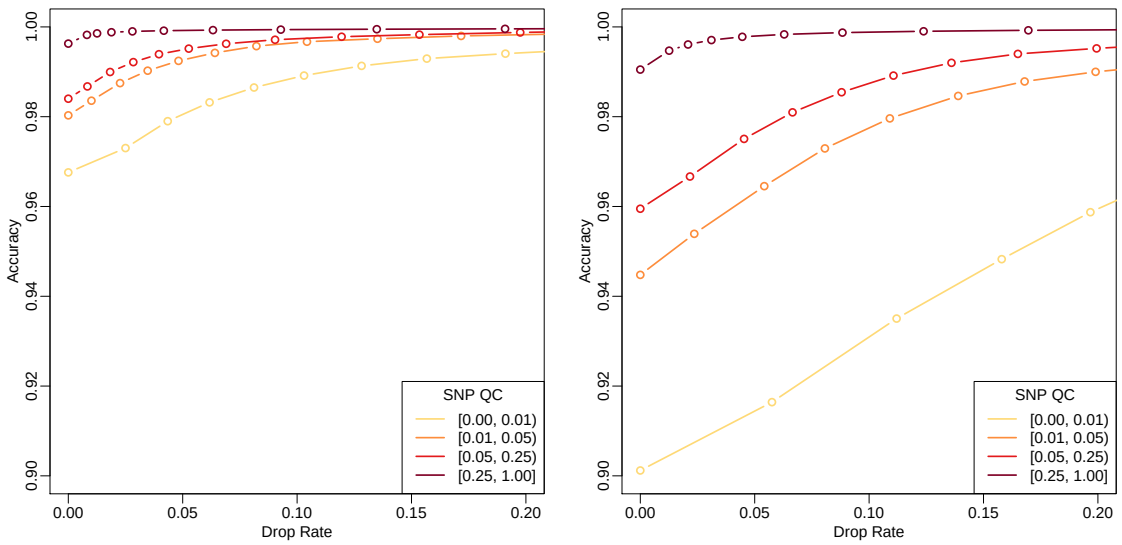

Figure 7: ADR plots for Datasets A and B. SNP's were stratified by their quality scores and accuracy versus drop rate curves were produced for each stratum. The scores are shown to successfully identify SNP's with lower accuracies.

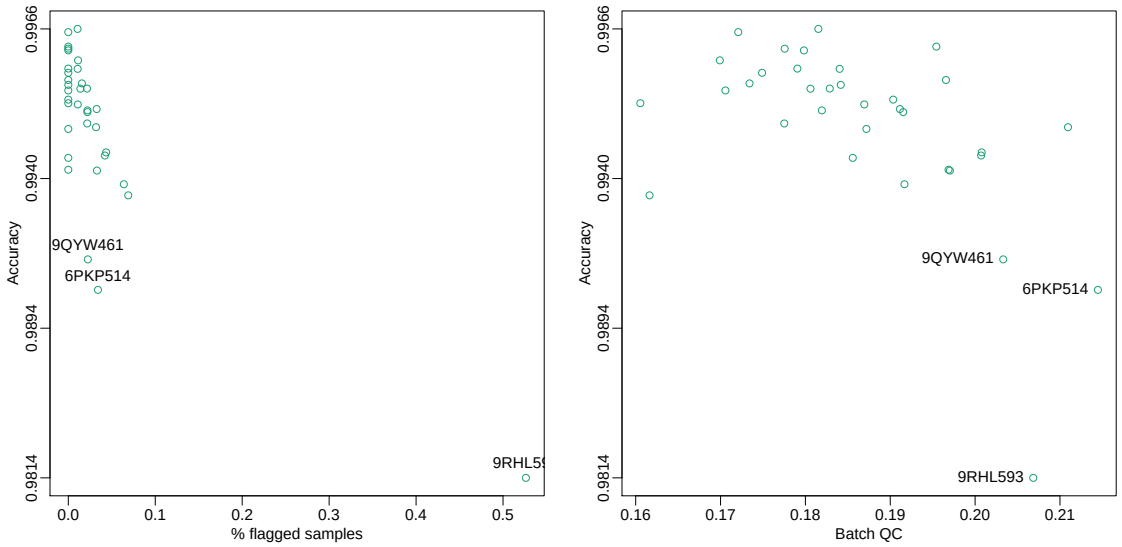

Figure 8: Batch quality plots. A) The accuracy with a 5% drop rate is plotted against the percentage of sample flagged by the SNR score. B) The accuracy with a 5% drop rate is plotted against our batch quality score.