BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array

Introduction

Highly accurate and reliable genotype calling is an essential component of any high-throughput SNP genotyping technology. BRLMM, the method of choice for the Mapping 500K product, is effective, but requires the presence of mismatched probes (MM) probes on the array to create "seed" genotypes. We present here a method that only uses perfect-match probes, BRLMM-P. The primary difference between BRLMM-P and BRLMM is that BRLMM-P derives seed genotypes directly from the clustering properties of the data (as opposed to BRLMM's reliance on initial genotype seeds from DM). Several secondary differences exist, such as using only the most informative dimension for clustering and some modifications to the exact choices for likelihood function.

As an extension of the RLMM concept [1,2,3], BRLMM-P (like BRLMM) performs a multiple chip analysis, enabling the simultaneous estimation of probe effects and allele signals for each SNP. Just as it has in the now reasonably mature field of probe-level expression analysis, accounting for probe specific effects results in lower variance on allele signal estimates. This step is retained even in arrays employing multiple copies of the same probe, even though the probe specific effects are only minimally different between copies. The distribution of summary values across arrays is then used to evaluate the likely genotypes.

Figure 1 presents an overview of the BRLMM-P approach. The first step is to normalize the probe intensities and estimate allele signal estimates for each SNP in each experiment. The allele signal estimates are then transformed to a 2-dimensional space in which the underlying genotype clusters are 'well behaved' in terms of having similar variance for each of the clusters. As the primary discriminator of genotype is the "contrast" dimension, the "size" dimension is discarded. In the resulting 1-dimensional space, for each SNP, we evaluate the posterior likelihood of all plausible divisions of the observed data into three (or fewer) seed genotypes using a Gaussian likelihood model combined with prior information. The highest likelihood divisions of the data into plausible genotypes are retained, and combined to form a final estimate of seed genotype assignments. These final seeds are combined with the data to form a posterior distribution summarizing the best current estimate of genotype cluster center and variance for the SNP. Finally, a genotype and confidence score are assigned for each observation according to the relative distance to the cluster centers.

The remainder of this manuscript steps through each of the above steps in detail and then presents a detailed assessment of BRLMM-P performance.

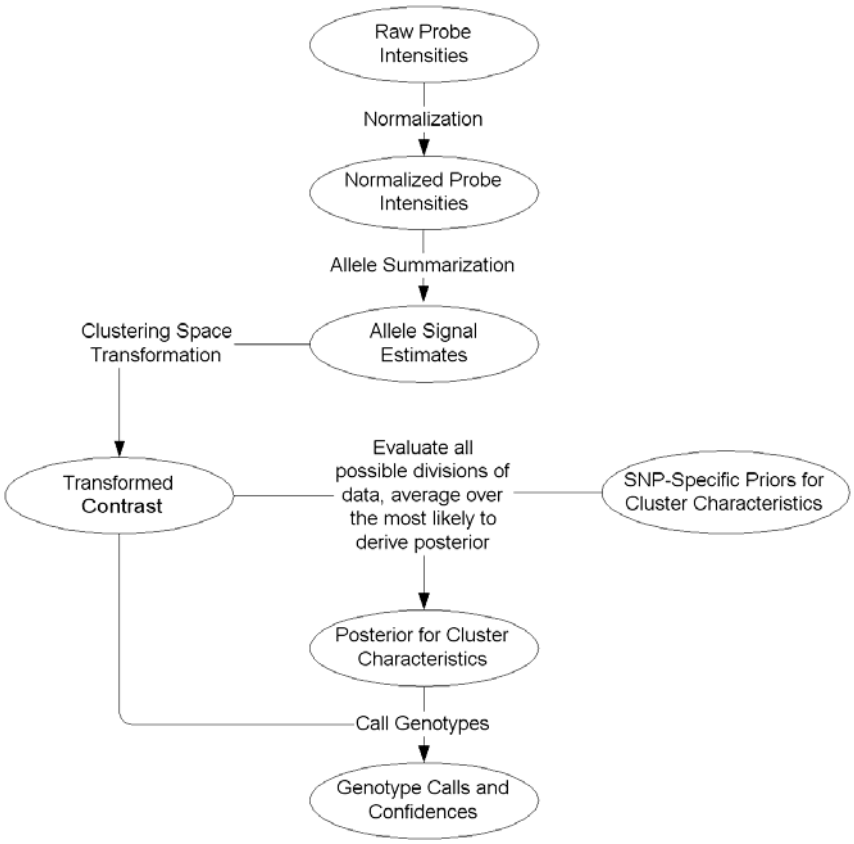

Figure 1: BRLMM-P algorithm workflow

Normalization and Allele Summarization

The normalization and allele summarization steps of the BRLMM-P method consist of producing a summary value for each allele of a SNP in each experiment. (They are identical to the steps used in BRLMM.) The "A" allele summary value increases and decreases with the quantity of the "A" allele in the target genome, and similarly the "B" allele summary value increases and decreases with the quantity of the "B" allele in the target genome. These summary values are calculated to remove extraneous effects – chip-chip variation, background, and the relative brightness of different probes on the array. This section explains the technical details of this summarization process, which is similar to that used on expression arrays.

For each SNP of interest, the array contains multiple probes designed to hybridize to each allele of the SNP. The intensities of these features typically vary together in systematic ways for each genotype of the SNP. We therefore summarize these intensities in a single value for the features corresponding to each allele, the "signal" for that allele. (Note: due to cross-hybridization with the alternate allele, this signal does not directly correspond to the concentration of the perfectly matched allele.) The intensities of the probes matched to the "A" allele are expected to decrease with decreasing quantities of the "A" allele, and similarly

for the "B" allele probes. Since these change in opposite directions, we summarize the probes for each allele as independent signals. Therefore, for each SNP in each experiment, we obtain two values – an "A" signal and a "B" signal, which summarize the probes.

From the field of expression analysis on arrays, we know how to summarize several probes to a single signal value effectively. We need to account for extraneous effects on the probe intensity that vary from experiment to experiment (normalization), account for potential differences in background from chip to chip (background adjustment), and account for the systematic differences in feature intensity due to probe composition (feature effects). For the SNP5.0 array the multiple features used to interrogate each allele have identical probe sequences but even so we sill use an approach that allows for systematic differences between probes from sources other than probe composition. While there are many options available for each of these effects, we have chosen to use off-the-shelf options: quantile normalization at the feature level, no background adjustment, a log-scale transformation for the perfect match intensities, and a median polish to fit feature effects to the data obtaining a signal. This is exactly the same methodology that can be applied to summarize an expression array and produce a signal for a probe-set.

Quantile normalization is performed as in the literature – the intensities on each chip are ranked, and then the average intensity across experiments for each rank of intensity is substituted within each experiment for the given rank. [If R(I) is the rank of intensity within a chip, and Q(R) is the average intensity for a given rank, the quantile normalized intensity within a chip is Q(R(I))]. Because the quantile function is slowly varying and smooth, we approximate the Q(R) function for each chip with a linear interpolation for processing speed ["sketch" normalization]. This allows us to normalize millions of data points per chip rapidly with compact summaries of the data.

Several background adjustments were explored during development, and we settled on using no adjustment for background. Unlike expression arrays, the target concentrations are well above background for the majority of the fragments containing SNPs. For this assay and genotype clustering algorithm background adjustment was not useful, and therefore the (normalized) perfect match intensities are used without adjustment for background.

To account for systematic differences in relative brightness between features, we fit the standard log-scale additive model to the probes for each allele separately: $\log(I_{i,j}) = f_i + t_j + \varepsilon_{i,j}$, where f_i is the effect due to feature i across experiments, t_j is the effect with experiment j responding to the genotype of the SNP and the relative quantity of the fragment on which it is located (because of cross-hybridization to the other allele it cannot be interpreted as simply the effect due to the concentration of target for allele A), and $\varepsilon_{i,j}$ is the multiplicative error for the observation. We fit this model using the standard median polish procedure for f and t, and for each experiment output the fitted value for t as the signal for that allele. For identifiability, we require sum(f) = 0. The output signal value is retransformed to lie on the original linear intensity scale: signal = exp(t).

These stages constitute the normalization and allele summarization portion of the algorithm. At the end of these steps, we have for each SNP in each experiment two signal values: one for the "A" allele probe set, and one for the "B" allele probe set. Each SNP therefore has a 2xN matrix of values output – 2 signals for each of N experiments. This output matrix is then used to evaluate each SNP for the genotype present in each experiment.

Clustering Space Transformation

Now that we have signals for the two alleles of the SNP across all experiments, we will be evaluating distances between a prototype (cluster center) for a given genotype (AA, AB, BB) and the actual data seen in any one experiment. However, raw "signal" value, while very useful for expression analysis, is not perfectly suited for genotype cluster analysis (figure 2a). We transform each pair of signals for each experiment into a space with properties more suitable for evaluating genotypes.

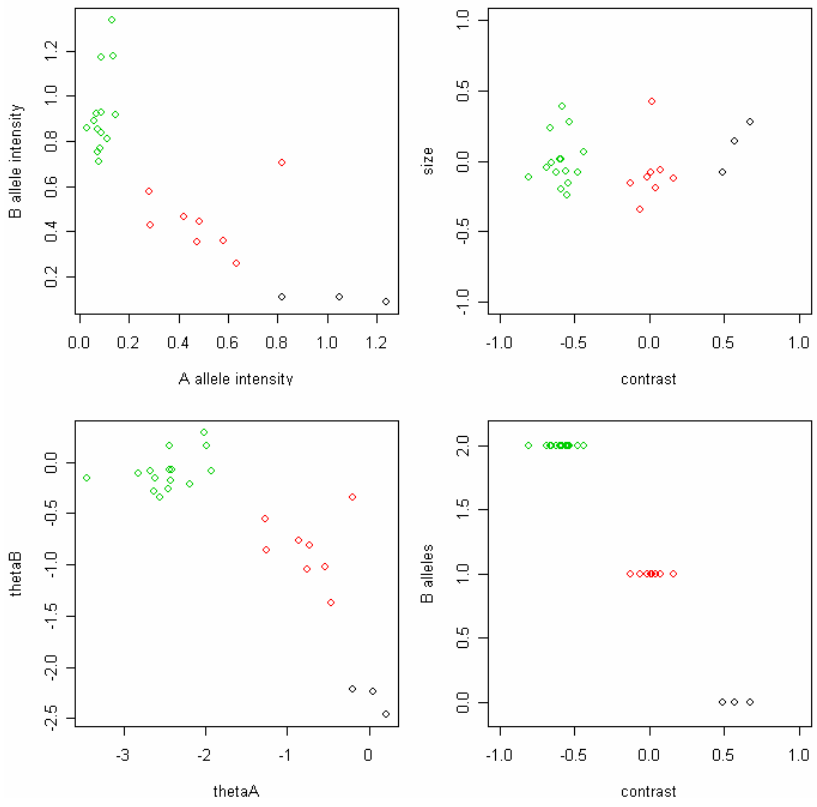

Figure 2: Clustering Space Transformations. Here a simulated SNP is taken through the transformations used in BRLMM-P. The upper left shows summarized allele intensities, the lower left shows the log-transformed intensities, the upper right shows the transformation to contrast (sinh(K*(A-B)/(A+B)/sinh(K)), and size (log(A+B)), and the lower right shows the assignment of the number of B alleles to each data point for a potential seeding. In all cases, BB points are green, AB are red, AA points are black with genotypes assigned by the design reference.

The desirable qualities for such a space include approximate independence of the difference between genotypes and the magnitude of signal, and controlling the variation within the various clusters to be comparable. For example, the standard "MvA" or "MA" transformation used to plot expression analysis could be applied to the two signals, resulting in $M = \log(S_A)-\log(S_B)$ and $A = (\log(S_A)+\log(S_B))/2$. This isolates most of the difference between genotypes into the M axis, leaving a mostly irrelevant "brightness" component in the A axis.

The MvA transformation is useful, but does not allow any fine tuning of cluster properties. One hazard is that the spread of homozygous clusters (where one allele is completely absent) can be very large if, after background adjustment, the resulting signal for that allele is near zero. Signals near zero can be extremely variable after taking logarithms, and the MvA transformation inherits this variability.

We therefore wish to use a space in which the spread of homozygous clusters can be controlled, even when a signal estimate is near zero, and where the typical variation can be adjusted to be similar between heterozygous and homozygous genotype clusters. Let us define two axes: $\text{Contrast} = (S_A-S_B)/(S_A+S_B)$ and $\text{Strength} = \log(S_A+S_B)$. Strength of course measures the overall brightness, which is mostly independent of genotype, and Contrast is a quantity that will depend most strongly on genotype ranging from -1 for the ideal BB genotype to +1 for the ideal AA genotype. However, while this transformation limits the range of the resulting value, and so limits the variation, there is no guarantee that the result of this transformation will have similar variation between the heterozygous cluster and the homozygous clusters. We further generalize the Contrast axis to define a Transformed $\text{Contrast} = \text{asinh}(K(S_A-S_B)/(S_A+S_B))/\text{asinh}(K)$, where K is a tuning constant. Figure 3 shows the functional form of this transformation for different values of K. The effect of varying K is to change the amount of "stretch" of the difference between A and B signals when the difference is small (i.e. likely to be heterozygous), vs. the difference between A and B signals when the difference is large (i.e. likely to be homozygous), thus K can be used to balance the variability in homozygous and heterozygous genotypes and remove any heterozygous dropout. By experimentation across several data sets, we ascertained that the value K=2 worked well to balance the variation of genotype clusters (figure 2d).

While many other transformations of the data could be used, this space worked well for clustering genotypes while avoiding heterozygous dropout. We therefore implemented this as "Contrast Center Stretch" (CCS) option within the software, and cluster in this transformed signal space. We have also noticed that the largest quantity of information about the genotype is contained within the contrast dimension, with minimal information about the genotype in the "size" dimension. For BRLMM-P, we only retain the contrast information for each SNP, and cluster in the resulting 1-dimensional space.

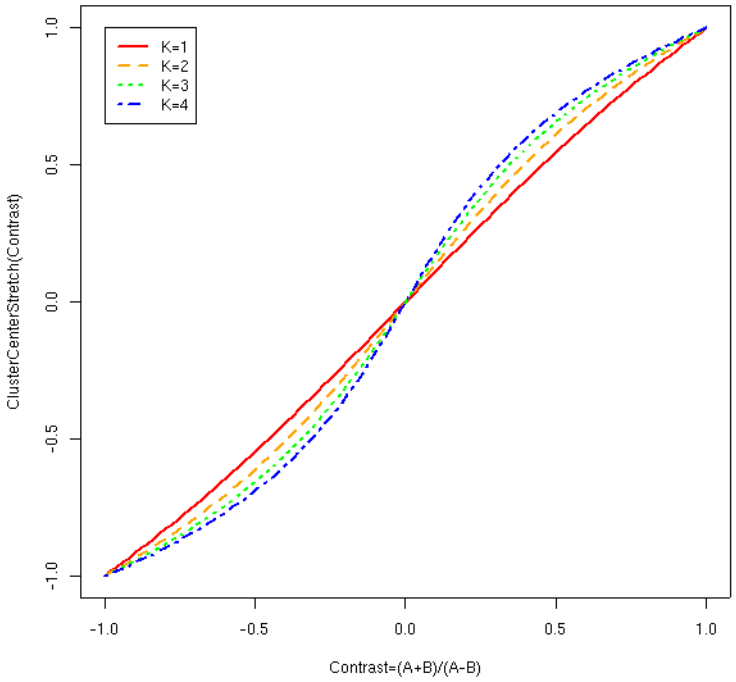

Figure 3: Examples of the Cluster Center Stretch (CCS) transformation. The CCS transformation is defined as asinh(K*Contrast)/asinh(K) where Contrast is defined as $(S_A-S_B)/(S_A+S_B)$. The effect of the transformation is to stretch contrast values near zero (corresponding to heterozygous genotypes) and to compress contrast values near -1 and +1 (corresponding to homozygous genotypes). Higher values of K apply a more extreme transformation, setting K to 1 yields effectively an identity transformation. The value of K can thus be tuned to alter the balance between performance on homozygotes and heterozygotes, with higher K values making het calls more likely.

Calling Genotypes

We call genotypes by a template-matching procedure comparing the transformed allele signal values observed in an experiment to the typical values (prototype) we expect for each genotype. The genotype that is estimated to have the highest probability of having produced the data point is reported as the call. The approximate confidence we report for that call is the estimated probability that the data point belongs to one of the other clusters. This allows us to rank the genotype assignments by quality, and hence make the decision not to call in cases of ambiguity.

Every SNP is expected to have three genotypes, “AA”, “AB”, and “BB” (There is an exception for X chromosomes in males, in which case we only have two genotypes “A” and “B”). For each genotype for a given SNP, we expect to have a prototype (typical observed values for that genotype, or cluster center), with some scatter of values around the prototype. We approximate the scatter by a normal distribution (and the careful choice of the CCS transformation ensures this is a good approximation). For clusters of this type, the relative

probability of belonging to a cluster is computed as a function of the distance from the cluster center and the variation within the cluster. The standard settings for BRLMM-P (which may be altered by advanced users) compute a common variance for all clusters.

Within any experiment, we derive transformed contrast values x for a SNP and compare to the SNP-specific prior on cluster characteristics, the derivation of which is outlined in the following section. The SNP-specific prior includes three cluster centers μ_{AA}, μ_{AB}, and μ_{BB} with covariance matrices Σ_{AA}, Σ_{AB} and Σ_{BB}, from which we obtain relative probabilities p(AA), p(AB),p(BB). (Note that we retain where possible similar notation to that used in the description of the BRLMM algorithm, though for BRLMM-P the covariance matrices are just scalar values since the clustering is performed in a one-dimensional space). We call the genotype of the SNP as the genotype with the highest probability, X, where P(X)>P(Y)>P(Z).

The confidence we assign to this call is P(Y)+P(Z) , where P(X) is the estimated probability for the called cluster. This confidence is always between zero and 1 (in fact, it is difficult for it to be above 0.66). It is a rough measure of the quality of the call (but is not a "p-value"). We set a threshold for quality of 0.05 for a call/no-call decision, based on the performance on several test data sets. This can be adjusted by the user to tune the tradeoff between call rate and accuracy – see the results section for a comparison of performance at various thresholds. Users accustomed to BRLMM should carefully note these changes in the nature of the confidence score, and the scale upon which it operates (BRLMM has a standard cutoff of 0.5 for call/no-call).

The next section describes how we learn the prototypes and their variation for each SNP from the data.

Estimating Cluster Centers and Variances

The above section dealt with how to call genotypes and ascribe confidence values to those calls given an appropriate prototype. This section deals with how to derive these prototypes.

This is achieved using a Bayesian procedure, in which we visit every SNP and combine a prior for that SNP with the data observed to obtain a posterior estimate of cluster centers and variances. The prior used may be a generic prior common to all SNPs, or a specific prior computed for that SNP from a set of training data. A prior has entries for the expected center of each genotype, the expected variance of each genotype, the uncertainty in those estimates (measured in 'pseudo-observations'), and covariances between those genotype centers. The posterior estimate has the same structure (mean, variance, uncertainty, and covariances). The posterior estimate is what is then used to call genotypes.

The trick here is that we do not know the actual division of the data into genotypes when combining the prior and the data. BRLMM solved this problem by using an external source of seed genotypes (DM) giving reliable data for a subset of data points. BRLMM-P solves

this problem by evaluating all plausible assignments of 'seed' genotypes to the full set of data points with respect to their likelihood, and then averaging over the most likely seeds. That is, we repeatedly make a "hard" (every data point is assigned to exactly one genotype cluster) assignment of data points to "seed" genotypes, and evaluate the likelihood of the data under a Gaussian cluster model to evaluate the quality of this 'hard' assignment of genotypes to data points (this is similar to a K-means procedure). We combine the most likely 'hard' assignments into a "soft" (allowing a data point to be partially assigned to more than one genotype cluster) assignment of seed genotypes, which we treat as a reliable seed. Once we have a reliable assignment of seed genotypes, we can compute the posterior distribution of the locations of the three clusters.

We observe that plausible "hard" assignments of genotypes to data points have the following structure: sweeping from left to right in contrast space, we will always see some number (possibly zero) of BB genotypes, followed by some number (possibly zero) of AB genotypes, followed by some number (possibly zero) of AA genotypes. That is, the more copies of the B allele we have, the higher the relative intensity of the B probes relative to the A probes. Given the data, plausible genotypes are assigned as though there were two dividing contrast values (corresponding to vertical lines in contrast/size space) that determine the transitions between genotypes (BB to AB, and AB to AA).

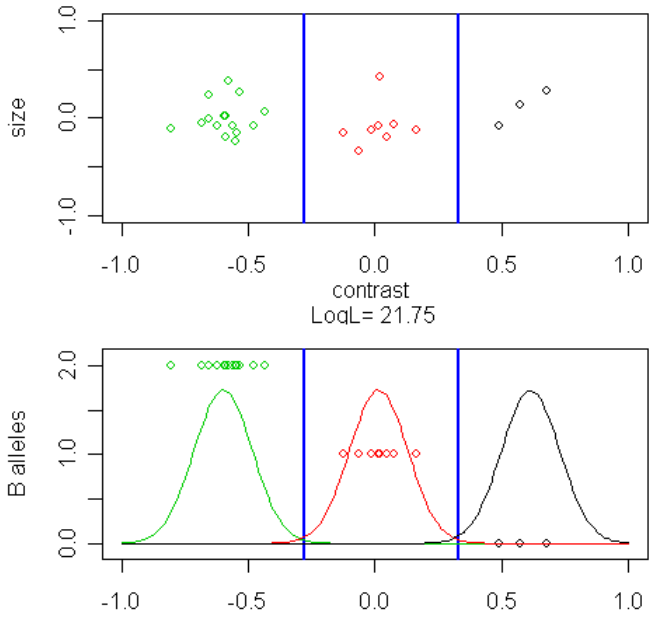

Figure 4. An example division of (simulated) data is shown in this figure. Two dividing lines divide the data into three assigned genotypes, BB in green, AB in red, and AA in black. Within each genotype, we compute a mean and variance, and combine with the prior to obtain a posterior estimate of mean and variance for each cluster. The log-likelihood of the data is computed given these distributions and the hard assignment of seed genotypes to clusters.

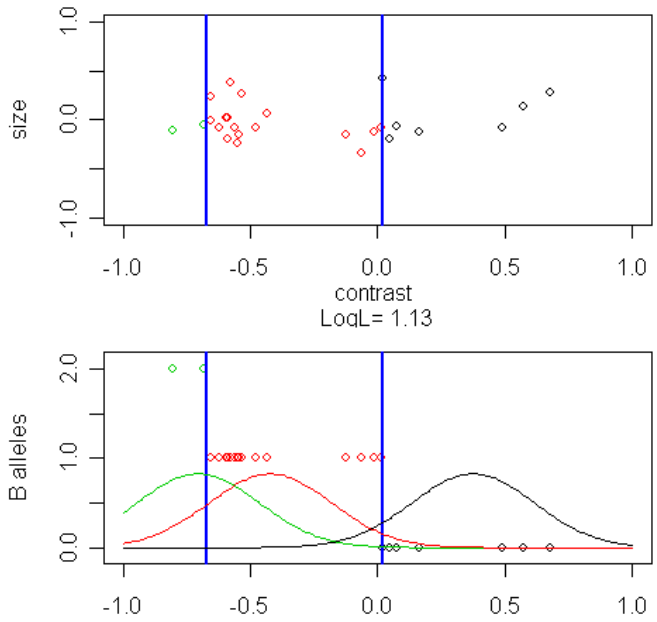

Figure 5. This shows another example of dividing the data, with a lower likelihood. While a human eye can clearly see that dense clusters in the data are split, the computer must evaluate the likelihood of the data given this clustering to find that this is a suboptimal assignment of genotypes to data.

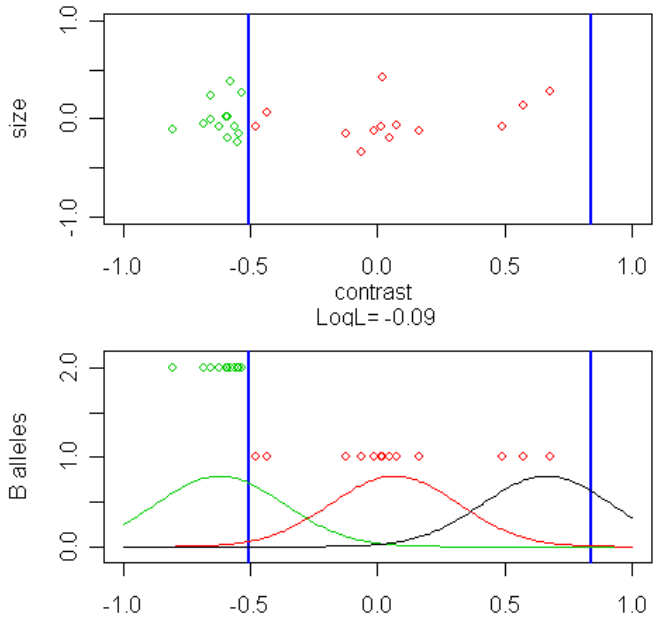

Figure: 6 This shows a division of the data that includes no AA genotypes (black, 0 B alleles). Note that there is still a computed mean and variance for the AA cluster, despite no seeds designated AA being present. This mean and variance is computed via the prior and is based on the prior center, as well as the covariances between cluster centers.

This implies that for N data points, there are only (N+1)*(N+2)/2 plausible ways of dividing the data into genotypes. Rather than following an iterative procedure (such as K-means or EM), we simply enumerate them all and thereby avoid any problems of being trapped in local maxima of the likelihood when looking at the fit to the data. This is not particularly time consuming because the use of Gaussian likelihoods allows us to evaluate each plausible assignment in O(1) time by means of running sums. (We have employed additional computational tricks that allow for linear scaling to large numbers of data points.)

For BRLMM-P, we use the normal-inverse-gamma prior for the distribution of the mean and variance of each cluster center. (This differs from the semi-conjugate distribution used in BRLMM) If we denote:

K = covariance matrix between cluster centers (expressed in pseudo-observations)
S = variance of observations
u = prior means
m = observed means
N = diagonal – number of observations in each cluster

Then for the conjugate prior, the variance of observations factors out of the updating formula for the cluster centers and results in the update formula :(K^-1 + N)^-1 * (K^-1*u + N*m). This is different from BRLMM in that we compute the shift in means without first constructing a SNP-specific variance matrix. The conjugate prior links the mean and variance of each cluster more tightly than the semi-conjugate prior used in BRLMM, and therefore simplifies the computation. This lightens the computational load of computing posteriors and is an advantage of using the conjugate prior rather than the semi-conjugate prior used in BRLMM.

Interpreting the formula as English, the cluster centers move to the average between the mean of the data, weighted by the number of observations of each genotype, and the prior location, weighted by the effective number of pseudo-observations provided in the prior. The variance is then computed as a weighted average between the observed variation, the prior variance, and the distance by which the centers have moved from the prior location.
This Bayesian update has the sensible property that when there is little or no data available the estimate of cluster centers will be driven mainly by the prior estimate u, and when there is a lot of data available for a given genotype the estimate will be driven by the observed means m.

The complete computation loop looks like the following: for each plausible assignment of seed genotypes, evaluate the likelihood of the assignment given the posterior likelihood of the clusters. Given the likelihoods of all assignments, compute a relative probability for each data point to be each genotype (i.e. a “soft” assignment obtained by a weighted average over all plausible “hard” assignments). Use this resulting “soft” assignment to seed the final computation of the posterior distribution of centers and spread for each genotype. With these posterior estimates of center and spread for each cluster, genotypes and confidences are then determined as outlined in the previous section.

Special Cases

The preceding algorithm assumes that the observations for each SNP are well described by prototypes for each genotype. However, for SNPs on the X chromosome, there are distinct clusters for each gender due to males having one fewer copy of the X chromosome. This not only changes the location of the cluster centers for XY individuals, but the SNPs located on chrX may end up being called as heterozygote. We therefore treat the chrX SNPs differently for XX individuals than for XY individuals. Note that the special treatment of chrX SNPs described here is only applied to SNPs on chrX in the nono-pseudo-autosomal region, and for the rest of this section when we talk about chrX it is to be interpreted as chrX excluding the pseudo-autosomal region

We detect the difference between XY and XX individuals by the distribution of observations in contrast space for all chrX snps. XY individuals are estimated as those where the distribution of all chrX SNP contrast values within the sample divides into three clusters by EM with fewer than 10% of chrX SNP values in the middle cluster. This decision rule allows for some frequency of misclassification of chrX SNPs when treating them uniformly, while robustly discriminating males from females in natural populations. The remaining individuals are classified as XX. For each chrX SNP, we treat XX individuals and XY individuals as separate data sets.

XX individuals are handled using the standard BRLMM-P methodology for all chrX SNPs, that is, three cluster centers are learned from the data along with within-cluster spread and used to classify observations. However, no data from XY individuals is used in this calculation.

XY individuals are handled using a modification of the BRLMM-P methodology for all chrX SNPs. Only two cluster centers can be learned from the data (AA and BB), and only the data for the XY individuals are used. Therefore the following modifications are performed First, we only evaluate assignments of AA and BB genotypes as a “seed”, and ignore the computed AB cluster completely for both likelihood and making genotype calls from the posterior. Thus, for XY individuals, only “AA” and “BB” genotypes are fit, and for any observed data, “AB” will never be called.

Fitting of XX and XY individuals separately improves the genotyping performance within each group. Modifying the prior for XY individuals to avoid heterozygous calls improves the genotyping performance for XY individuals. This is the justification for having a special purpose modification for chrX SNPs within BRLMM-P.

Another special case is that of a SNP with unusual behavior, such as a SNP with probes for the A allele having a different sequence than probes for the B allele (for example, the A allele probe could be from one strand and the B allele probe from the other strand). Such a SNP may have a very unusual location for the cluster centers when compared to the typical SNP on the array. This may lead to erroneous assignment of cluster identity if, for example, the AB cluster is located where the BB cluster is on a typical SNP. With sufficient data to show

examples of all three clusters, such a mis-assignment will usually be corrected, however, for those SNPs with rare minor alleles, this may require a large number of samples.

To handle exceptional cases such as this and to improve the performance on more conventional SNPs, we allow for the provision of a SNP-specific prior for each SNP. This takes data (labeled and/or unlabeled) from a training set and provides information on where the training genotypes are located. This is very similar in effect to clustering the observed data with the training data, and requires that lab procedures be sufficiently similar between training and observed data so that they may be clustered together.

Pre-screening samples

In the typical workflow it is very useful to have a simple metric that can be computed based on a single experiment to determine if the experiment is of high enough quality to be considered as completed and ready for future multiple-sample analysis, or if it should be repeated. BRLMM-P yields very high quality genotype calls but it is inherently a multiple sample method, and the exact results for any given sample will depend in part on the batch in which the sample is analyzed, so it is not ideally suited for in-lab single-chip quality determination.

With this application in mind, we have taken advantage of the DM genotype calling algorithm [4] – it is a single chip analysis method and call rates with DM are very strongly correlated with call rates and concordance when experiments are ultimately re-called with BRLMM-P or other multiple-sample genotype calling methods. The SNP 5.0 array has a set of 3,022 SNPs tiled with both PM and MM probes so that each chip can be analyzed with DM (at a confidence threshold of 0.33) to produce a call rate. We call this metric the QC call rate.

The 3,022 SNPs tiled for calling with DM are a subset of the 500,568 SNPs on the Mapping500K product (1,511 from each of the Nsp and Sty arrays), but they are not a random sample – the pool is intentionally enriched for SNPs that were more challenging to call in the Mapping500K to yield a more sensitive metric of quality.

The recommended protocol is that experiments with a QC call rate of 86% or better should be considered as complete and are expected to result in a call rate of 97% or better when re-called with BRLMM-P. Note that this threshold of 86% is specific to the SNP 5.0 array and the particular set of 3,022 SNPs tiled on it.

Results

The ideal way to assess performance would be to evaluate the tradeoff between accuracy and call rate in data generated from a collection of samples for which the true reference genotypes are available for all SNPs on the SNP5.0 array. Fortunately something closely

approximating this has been made possible by the International HapMap Consortium – the phase 2 release provides reference calls on a collection of 270 samples for approximately 70% of the SNPs on array. If submissions to HapMap by Affymetrix are included this rises to 97% of the SNPs, however for the sake of computing concordance we try to avoid overestimating concordance by including only the non-Affymetrix submissions to HapMap. This constitutes an excellent resource for the performance evaluation; though it is important to bear in mind the caveat that the genotype calls in HapMap themselves do have some small but non-zero error rate. Additionally, the HapMap samples consist of some trios, enabling the evaluation of Mendelian inheritance error rates. Finally, we also look at reproducibility of genotype calls on sample replicates.

For evaluation of call rates, accuracy and Mendelian inheritance error rate we use four datasets consisting of HapMap samples. The first dataset consists of all 270 HapMap samples, processed jointly by Affymetrix and the Broad Institute. The remaining three sets use a collection of 44 HapMap samples comprising 30 unique DNAs (10 trios) with five of the samples run multiple times to evaluate reproducibility. Sets two and three were run at customer sites and set four was run by a group within Affymetrix that was newly-trained on the assay. Performance is assessed by looking at the 440,794 SNPs that meet rigorous performance standards with the BRLMM-P algorithm, future analysis improvements will move more of the 500,568 SNPs tiled on the chip into this class.

To account for the fact that one can adjust the confidence threshold to trade off between call rate and accuracy we look at performance at all possible thresholds and plot the relationship between HapMap concordance and no-call rate, as shown in Figure 7. Table 1 presents performance for BRLMM-P at its default confidence threshold.

Dataset	Number of samples	Overall Call Rate	Hom Call Rate	Het Call Rate	Overall Conc-ordance	Hom Conc-ordance	Het Conc-ordance	Mendelian Consistency	Reprod-ucibility
Set1	270	**99.71%**	99.76%	99.62%	**99.69%**	99.70%	99.67%	99.96%	NA
Set2	44	**99.55%**	99.83%	98.86%	**99.67%**	99.70%	99.56%	99.95%	99.90%
Set3	44	**99.37%**	99.71%	98.51%	**99.56%**	99.60%	99.44%	99.94%	99.85%
Set4	44	**99.63%**	99.73%	99.40%	**99.69%**	99.69%	99.70%	99.96%	99.92%

Table 1: Performance on HapMap dataset for BRLMM-P at various fixed thresholds. Results are based on 440,794 SNPs.

One caveat about evaluating concordance with HapMap is that to some extent it provides only a lower bound estimate for accuracy, since HapMap itself does have a certain error rate. With this in mind, it is useful to look at additional measures of performance. All four datasets summarized here contain (father,mother,child) trios of samples which can be assessed for Mendelian consistency. The Mendelian consistency is estimated looking only at informative trios (those in which we have a call for all three samples where the parents are not both called heterozygous), call this number T. If the number of such trios which exhibit a Mendelian inconsistency is E then the Mendelian consistency is estimated as (T-E)/3T,

which is based on the assumption that when there is an inconsistency in a trio that only one of the three calls is erroneous.

The final metric of performance we evaluate is reproducibility on sample replicates. Arguably this metric is less useful than those above since it only reports on the consistency of calls made but not on whether or not those calls are actually correct. Nevertheless, other things being equal a reproducible method will generally be preferable to one that isn't. The first dataset does not contain replicates, for the other three datasets the pairwise reproducibility of calls is on average 99.9%.

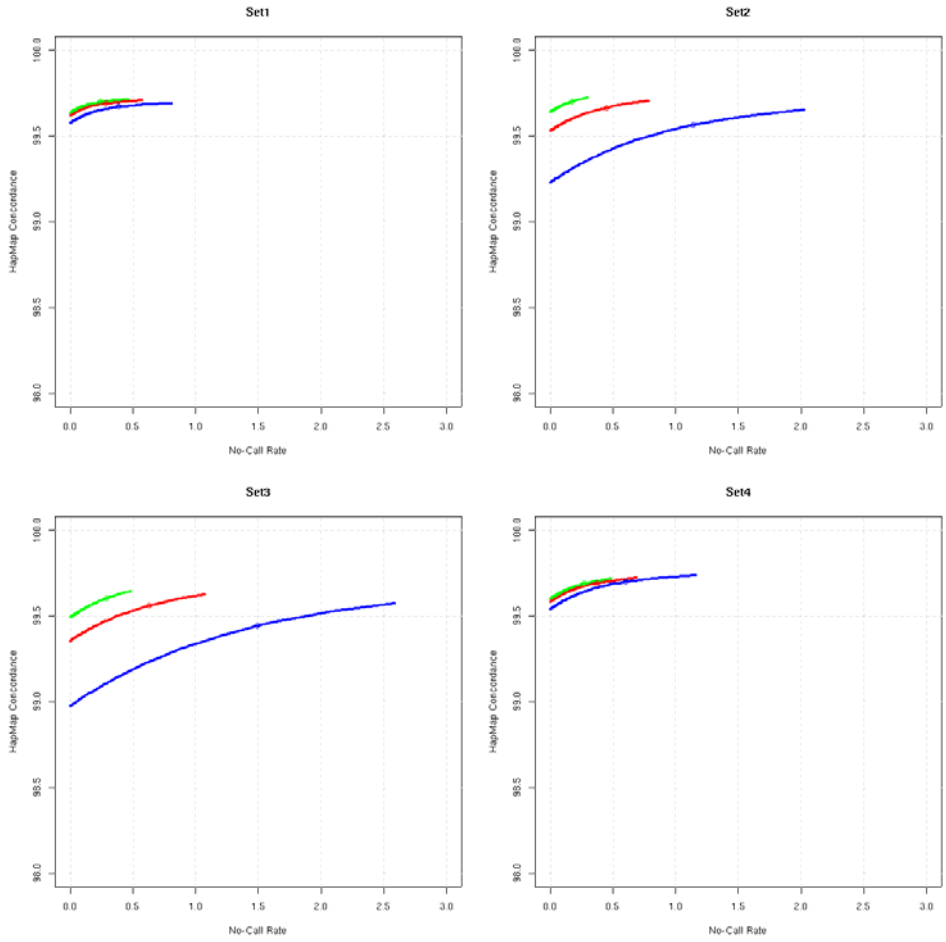

Figure 7: Performance of BRLMM-P on HapMap samples. Concordance with HapMap and genotype call rate is determined at all possible thresholds, the plots depict the tradeoff between the two performance metrics. The default confidence threshold of 0.5 is indicated as a point on each curve. The red curves present the tradeoff on all genotypes combined, the blue and green curves summarize performance looking only a genotypes that HapMap indicates are heterozygous and homozygous respectively.

Discussion

BRLMM-P enables accurate calling of genotypes using only PM probes. This allows for more SNPs on an array of a given size. The performance is comparable to the performance of BRLMM on Mapping500K arrays. As a multiple chip method it has some extra considerations which need to be taken into account in practice.

One matter to consider is the batch size in which to apply BRLMM-P. More samples will generally lead to better performance, however we have found that very high performance can be attained with 44 or even fewer samples. BRLMM-P performs slightly better on SNPs for which there are observations of all three genotypes. As a result the addition of more samples is expected to be mainly of benefit to SNPs of lower minor allele frequencies, which will be more likely to have only one or two observed genotypes for low number of samples. Another observation is that reliability of calls improves as the the number of observations in the genotype cluster increases. Thus the addition of more samples will tend to be of most benefit to rare genotypes. Since the main benefit is to rarer genotypes, addition of more samples may appear to provide marginal benefit when one focuses on overall performance.

Another important consideration is the extent to which datasets can be combined. On the one hand, as discussed above, more samples should improve performance, particularly for rare genotypes. On the other hand the validity of combination of datasets will depend on the degree to which the combined datasets have the same underlying probe intensity distribution and SNP cluster properties. A good way to check the appropriateness of combining datasets is to inspect SNP cluster centers for each dataset separately and to check the degree to which the cluster centers and variances are consistent both with each other and with the SNP-specific prior distributions being supplied to BRLMM-P.

Finally, while BRLMM-P represents a significant step forward in genotype calling by removing the reliance on MM probes, it is only one of a variety of genotype calling algorithms that either exist already or are in development. Currently the list of alternatives includes CRLMM [5], GEL [6], Birdseed [7] and Chiamo++ [8], and if the field of expression analysis is anything to judge by the number of alternatives will continue to increase. One or more of these alternatives and/or future updates to BRLMM-P should lead to even better performance than presented here.

References

1. **Nusrat Rabbee and Terence P. Speed**, *"A genotype calling algorithm for Affymetrix SNP arrays"* UC Berkeley Statistics Online Tech Reports, August 2005. http://www.stat.berkeley.edu/users/nrabbee/693.pdf
2. **Nusrat Rabbee and Terence P. Speed**. *"A genotype calling algorithm for Affymetrix SNP arrays"* Bioinformatics Advance Access published online on November 2, 2005 http://bioinformatics.oxfordjournals.org/cgi/content/abstract/bti741v1
3. **RLMM**: http://www.stat.berkeley.edu/users/nrabbee/RLMM/

4. **Xiaojun Di, Hajime Matsuzaki , Teresa A. Webster , Earl Hubbell , Guoying Liu , Shoulian Dong , Dan Bartell , Jing Huang , Richard Chiles , Geoffrey Yang , Mei-Mei Shen , David Kulp , Giulia C. Kennedy , Rui Mei , Keith W. Jones and Simon Cawley**. *"Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays"*. Bioinformatics 2005 21(9):1958-1963
5. **Benilton Carvalho, Terence P. Speed, and Rafael A. Irizarry**, *"Exploration, normalization and genotype calls of high density oligonucleotide SNP array data"* (July 2006). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 111 http://www.bepress.com/jhubiostat/paper111
6. **Dan L. Nicolae, Xiaolin Wu, Kazuaki Miyake and Nancy J. Cox** "*GEL: a novel genotype calling algorithm using empirical likelihood*" Bioinformatics 22(16) , 1942-1947
7. **Finny Kuruvilla, Josh Korn, Alex Wysoker**, *Birdseed*, personal communication
8. **Jonathan Marchini**, *Chiamo++*, personal communication