Genome analysis

GEL: a novel genotype calling algorithm using empirical likelihood

Dan L. Nicolae\textsuperscript{1,*}, Xiaolin Wu\textsuperscript{2}, Kazuaki Miyake\textsuperscript{2} and Nancy J. Cox\textsuperscript{2}
\textsuperscript{1}Department of Statistics and \textsuperscript{2}Department of Medicine, The University of Chicago

Received on April 22, 2006; revised on June 16, 2006; accepted on June 20, 2006
Advance Access publication June 29, 2006
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Preliminary results on the data produced using the Affymetrix large-scale genotyping platforms show that it is necessary to construct improved genotype calling algorithms. There is evidence that some of the existing algorithms lead to an increased error rate in heterozygous genotypes, and a disproportionately large rate of heterozygotes with missing genotypes. Non-random errors and missing data can lead to an increase in the number of false discoveries in genetic association studies. Therefore, the factors that need to be evaluated in assessing the performance of an algorithm are the missing data (call) and error rates, but also the heterozygous proportions in missing data and errors.

Results: We introduce a novel genotype calling algorithm (GEL) for the Affymetrix GeneChip arrays. The algorithm uses likelihood calculations that are based on distributions inferred from the observed data. A key ingredient in accurate genotype calling is weighting the information that comes from each probe quartet according to the quality/reliability of the data in the quartet, and prior information on the performance of the quartet.

Availability: The GEL software is implemented in R and is available by request from the corresponding author at nicolae@galton.uchicago.edu

Contact: nicolae@galton.uchicago.edu

1 INTRODUCTION

Large-scale genotyping platforms are used in genetic association studies that aim to identify the DNA variation affecting risk to disease, or modifying the distribution of phenotypic variables such as quantitative traits and response to drug. The Affymetrix genotyping sets (Matsuzaki et al., 2004), GeneChip\textsuperscript{TM} Mapping 10K, 100K and 500K Arrays, were the first widely available platforms, and are used in many of current studies. They provide a large number of genotypes at low (per genotype) costs, but the technology used in producing the data can lead to high variation in the genotype calling rates. The high variability in calling rates can be owing to many factors including quality of DNA and reagents, and small variation in the experimental protocols. We have observed that the decrease in the calling rate is usually associated with a disproportionate increase (relative to other genotypes) in the number of heterozygous genotypes assigned a ‘NoCall’ label (see also, section 4). We think that a significant percentage of the ‘NoCall’ heterozygotes are due to the difference in binding efficiency of the probes corresponding to the two alleles, a difference that is not properly modeled by the algorithm implemented in the GeneChip DNA Analysis Software. We propose an algorithm that is based on a weighted likelihood, using weights that incorporate information on the properties of each quartet (e.g. the interrogating base), on the sample specific characteristics (e.g. presence of outliers), and prior information on the performance of the quartet that is obtained from data for which the genotypes are known. Calling genotypes is, for biallelic markers, a three-label classification problem so it is possible that many well-developed supervised learning algorithms can be modified for this problem. We think that an off-the-shelf approach might not be suitable for these data because the arrays are generally not exchangeable. The proposed likelihood-based approach can be used even when there is only one array available (and no training data), and it offers a framework for simple generalizations to methods for detecting chromosomal abnormalities such as copy-number polymorphisms.

2 DATA STRUCTURE AND GENOTYPE CALLING ALGORITHMS

The Affymetrix GeneChip Mapping Arrays consist of a large number of probes containing 25-base oligonucleotides designed to test a given sequence. The probes are grouped in quartets: perfect match for allele A (PA), mismatch for allele A (MA), perfect match for allele B (PB) and mismatch for allele B (MB). The mismatch probes are used to measure the magnitude of cross-hybridization. For the ‘AA’ genotype, it is expected that the intensity of PA is higher than that of MA, PB and MB, and for the ‘AB’ genotype it is expected that the intensities of PA and PB are higher than those of MA and MB. For each SNP there are 10 quartets used on the Affymetrix GeneChip 100K Mapping Arrays (Matsuzaki et al., 2004), and six quartets on the Affymetrix GeneChip 500K Mapping Arrays. The information available in the data files that are part of the output of the imaging software consist, for each probe, of the number of pixels (\(n_i\)) corresponding to the probe, the mean (\(\mu\)) and the standard deviation (\(\sigma\)) of the pixel intensity. These summaries of the pixel-level information are used by the Affymetrix GeneChip DNA Analysis Software to make the genotype calls. Figure 1 shows summaries of the probe data for the same SNP in two reference samples, one with a ‘NoCall’ assigned by the Dynamic Model (DM) Mapping Algorithm (Di et al., 2005) implemented in the Affymetrix GeneChip DNA Analysis Software, and one with a correctly assigned ‘AB’ genotype.

\textsuperscript{*}To whom correspondence should be addressed.
The DM algorithm is calculating log-likelihood statistics for four models (NULL, AA, AB and BB) for the 10 probe quartets using a Gaussian model that assumes independent pixel intensities and equal mean intensities for probes in the same group (e.g. same mean intensity for PB, MA and MB if the genotype is AA, same mean intensity for PA and PB if the genotype is AB). The data for the first quartet in Figure 1 show that the latter assumption is in general not valid [see also Zhan and Kulp (2005)]. A one-sided Wilcoxon signed-rank test is then used on differences in log-likelihood values to calculate reliability rank scores (p-values corresponding to the signed-rank test). The smallest reliability score indicates the most probable genotype. In the ‘NoCall’ example shown in Figure 1, the fifth quartet favors the model ‘AA’, and the sixth and seventh quartet favor the model ‘BB’; the Wilcoxon rank test is not powerful enough to overcome this, and therefore the DM algorithm is unable to call this genotype. The (lack of) power of the Wilcoxon signed-rank test is even more problematic for the Affymetrix GeneChip 500K Mapping Arrays where only six quartets are available for each SNP; one outlier quartet is sufficient to yield reliability rank scores as high as 0.4375.

The DM algorithm performs remarkably well and it has several strengths. The main strength is that calling genotypes does not require prior training that is difficult for rare alleles, or information from other variants on the array. As a consequence, the accuracy of the call is not affected by the SNP allele frequency. Also, the DM algorithm efficiently uses the low signal (small differences in PM and MM), and the ‘no-signal’ quartets (i.e. quartets where the PMs means are smaller than the MMs means), and it aggregates robustly the quartet information. Some of the deficiencies of the approach are in the modeling data: it assumes that the PM means of the two alleles are equal for heterozygous genotypes, and independence of pixel intensities in calculating the likelihoods. Also there is no obvious way to incorporate the Quality Scores from the output of the software in the analytic methods for association studies.

Several data characteristics are important for design of a more effective allele calling algorithm. The standard deviations of the probe intensities are highly correlated with the mean intensities (for our arrays, the correlation between the mean and the standard deviation of the mis-matches are between 0.891 and 0.925), so the pixel standard deviations offer mostly redundant information. This has been discussed also in more general contexts of microarray-based hybridization (Zhan and Kulp, 2005). Also, the standard deviations of the statistics used in the algorithm depend on the location of the interrogation nucleotide, with lower variation if the offset (distance between the SNP location and the interrogation position) is 0 or 1. The offset is used explicitly in other algorithms (LaFramboise et al., 2005), but we use it only in designing weights for the quartets, with higher weights on the quartets with smaller standard deviations [see also Hekstra et al. (2003)].

Finally, the mean intensities for the two alleles are not equal for heterozygous genotypes. To investigate this we looked at the heterozygous genotype in the reference samples (where the genotypes are known), and we calculated the number of times the mean PA is bigger than the mean PB (an integer between 0 and 10). The distribution of this deviates substantially (data not shown) from what is expected under exchangeable mean intensities (i.e. a binomial 0.5 distribution).

The DM algorithm and the method proposed in this paper can be applied to individual arrays, and do not require training data or a large number of arrays to be available. Rabbee and Speed (2006) introduced an algorithm, RLMM, based on a multi-chip model that takes advantage of similarities of the intensities for the same probes across arrays. RLMM is using a supervised learning algorithm, and the construction of the classifier requires that the genotypes are known for some of the arrays. We compare the performance of the three algorithms in the Results section.

### 3 THE PROPOSED ALGORITHM

The ideas behind the algorithm we introduce are straightforward. For each quartet, we calculate two statistics: $t_A$ that measure the difference between PA and the mis-matches and $t_B$ that measure the difference between PB and the mismatches. The statistics are used to quantify the support for the presence of allele A and/or B. Optimal
statistics can be inferred from assumptions on the distribution of the probe intensities, and our choice is described in the next subsection. We denote with $D_A^0$, $D_A^1$ and $D_A^2$ the sampling distributions of $t_A$ when the genotype is $AA$, $AB$ and $BB$ respectively. Note that these distributions generally depend on the array, on the SNP, and on the allele under investigation. Also note that we assume that the distributions of the statistics are different when allele $A$ is present in a homozygous or heterozygous genotype (LaFramboise et al., 2005). If these distributions are known, the likelihood of the data under each genotype can be inferred, e.g.

$$P(\text{data} \mid AA) = \prod_{q=1}^{10} P(t_q^A \mid AA)P(t_q^B \mid AA),$$

where $t_q^A$ is the observed value for the $q$-th quartet (we implicitly assume that $t_A$ and $t_B$ are sufficient statistics). The (posterior) probabilities for each genotype can be calculated using Bayes rule as

$$P(AA \mid \text{data}) = \frac{P(\text{data} \mid AA)P(AA)}{\sum_G P(\text{data} \mid G)P(G)},$$

where $G$ indexes the genotypes ($AA$, $AB$ or $BB$), and $P(G)$ are the prior probabilities on the genotypes. Note that we use a non-informative prior $P(AA) = P(AB) = P(BB)$, and that the performance of the algorithm can be improved by using more accurate information (e.g. genotype frequencies estimated from HapMap). The called genotype is the mode of the posterior distribution.

There are several issues to deal with: (1) what are the optimal statistics for classification? (2) how to calculate/estimate the distributions of the statistics used in the likelihood calculations? (3) how to deal with outliers, data quality etc.? Note that optimal statistics are such that their distributions given the presence and given the absence of the allele should be well separated, similar for all levels of hybridization, and similar across arrays (so we can combine information from multiple arrays).

### 3.1 Empirical likelihood classification

We noted before that the data on the arrays we analyzed show that the standard deviation of the probe intensities is proportional to their mean, suggesting a Gamma model for the intensities. It is well-known that optimal tests for difference in the means of two Gammas are based on the ratio of the sample means [e.g. Shuie and Bain (1983)], and therefore we propose to use

$$t_A = \frac{x_{PA}}{x_{PA} + x_{MA}/2 + x_{MB}/2}$$

and

$$t_B = \frac{x_{PB}}{x_{PB} + x_{MA}/2 + x_{MB}/2}.$$

Note that we are not using the standard deviations, and that these statistics are bounded by 0 and 1. Using the same mis-matches in both statistics creates dependence among them, and we ignore this dependence in the probability calculations. Note that the mismatches can be ignored if we use data across arrays because it is possible to obtain empirical distributions for $x_{PA}$ and $x_{PB}$ based on data on a large number of normalized arrays. We plan to do this in future work.

It is easy to see that the distributions of the $t_A$ and $t_B$ statistics can be approximated by Beta distributions, and one solution for calculating the likelihoods is to estimate the parameters of the distributions from the data (an empirical Bayes approach). The disadvantage of this method is that it relies on a good fit of the Beta distribution to the observed data. We therefore propose using empirical approximations to the distributions of $t_A$ and $t_B$, and the corresponding algorithm is called GEL: genotype calling using empirical likelihood. Based on the results of a preliminary genotype calling algorithm (e.g. the DM algorithm), we can assign a genotype call to each SNP, and we can use the observed values of the statistics under a given genotype to build the empirical estimates for the distributions. Ideally, we would like to build empirical distributions for each SNP separately based on the data on a large number of arrays. This can be done if it would be possible to normalize the arrays such that the data on them are exchangeable. A simple alternative is to assume that the SNPs on the same array are exchangeable (the statistics have the same marginal distribution) and estimate the three distributions for each array separately.

Finally, we need to adjust the inference for the presence of outliers and other data quality issues. We propose to do this by constructing weights for each quartet that incorporate several aspects of the data. The weights are used in combining the information across quartets,

$$P(\text{data} \mid AA) = \prod_{q=1}^{10} [P(t_q^A \mid AA)P(t_q^B \mid AA)]^w_q,$$

where $w_q$ is the weight of the $q$-th quartet. Obviously, quartets determined to be of high quality will be given higher weights.

The algorithm has the following steps (the current implementation is run for each algorithm separately): (1) run a ‘preliminary algorithm’ (described next) to get calls for each SNP; (2) use the calls from (1) to construct empirical estimates for $D_A^0$, $D_A^1$ and $D_A^2$; (3) use the empirical distributions to calculate the likelihoods for each quartet; (4) calculate the weight of each quartet; (5) combine the data on each quartet and make the genotype call.

### 3.2 The preliminary algorithm

This step can be skipped if calls are available from a different algorithm such as DM. These initial calls are used to build empirical distributions for the statistics $t_A$ and $t_B$.

1. For each quartet, calculate statistics that indicate presence or absence of the two alleles,

$$s_A = \frac{x_{PA} - x_{MA}}{\sqrt{s_{PA} + s_{MA}}},$$

$$s_B = \frac{x_{PB} - x_{MB}}{\sqrt{s_{PB} + s_{MB}}},$$

$$s_{AB} = \frac{x_{PA} - x_{PB}}{\sqrt{s_{PA} + s_{PB}}},$$

$$s_M = \frac{x_{MA} - x_{MB}}{\sqrt{s_{MA} + s_{MB}}}.$$

2. For each SNP, average the $s$ statistics from Step 1. Ignore the quartets for which both PM means are smaller than
their corresponding MM means. The corresponding statistics are:

\[ t_A = \frac{1}{n} \sum_{i=1}^{n} x_i^A, \]

where \( i \) indexes the quartets that are used, and \( n \) is their number. Similarly, get \( t_B, t_{AB}, t_M \).

(3) For each value of \( t_A \) calculate \( q_A \), the value of the distribution function of \( t_M \) evaluated at \( t_A \), i.e. \( q_A \) is the proportion of \( t_M \) smaller than \( t_A \). Similarly for \( t_B, t_{AB}, t_M \).

(4) Call the genotypes:
- if \( q_A > 0.99 \) and \( q_{AB} > 0.95 \) call ‘AA’;
- if \( q_A > 0.99 \) and \( q_{AB} < 0.05 \) call ‘BB’;
- if \( q_A > 0.975 \) and \( q_B > 0.975 \) call ‘AB’;
- otherwise, call ‘NoCall’.

3.3 The main algorithm

The main part of the algorithm is described in this subsection. Preliminary genotype calls (e.g. using the above module) and specification of the weights assigned to individual quartets (next module) are necessary to run this part of the algorithm. The user needs to specify the threshold used in calling the genotypes, \( q \in [0,1] \). For example, if \( q = 0.95 \), the SNPs for which the maximum posterior probability is \(< 0.95 \) are assigned a ‘NoCall’.

(1) Calculate statistics that indicate presence or absence of the two alleles,

\[ t_A = \frac{x_{PA}}{x_{PA} + x_{MA}/2 + x_{MB}/2}, \]

and

\[ t_B = \frac{x_{PB}}{x_{PB} + x_{MA}/2 + x_{MB}/2}. \]

(2) Define empirical distributions for the statistics. We will define three distributions: the distribution of the statistic when the allele is present in a homozygous genotype \( (D^2) \), the distribution of the statistic when the allele is present in a heterozygous genotype \( (D^1) \), and the distribution of the statistic when the allele is absent \( (D^0) \). We use the calls from the ‘Preliminary Algorithm’ (PrAl) to do this:

- \( D^2 \) is the union of the set of \( t_A \)s when the Preliminary Algorithm call is AA and the set of \( t_B \)s when the PrAl call is BB;
- \( D^1 \) is the set of \( t_A \)s and \( t_B \)s when the PrAl call is BB;
- \( D^0 \) is the union of the set of \( t_A \)s when the PrAl call is BB and the set of \( t_B \)s when the PrAl call is AA.

(3) Discretize the empirical distributions for efficient computations, i.e. we use histogram approximations to their distribution. The first step is to define a partition of the range of the statistics. We use the intervals: (0,0.02], [0.02, 0.04], . . . , (0.98,1]. Because some of the intervals might contain no points, we add a count of 10 to each bin, and this will lead to shrinkage estimators for the bin probabilities. For each interval \( I_j \), \( j = 1,\ldots,10 \) calculate \( p_i^j \) as the fraction of points in \( D^2 \) that are in \( I_j \) (and similarly for \( D^1 \) and \( D^0 \)).

(4) For each quartet calculate the likelihood of AA, AB, BB, NoCall in the following way:
- Find \( j_A \) the index for the interval where \( t_A \) is, and similarly for \( j_B \).
- Calculate

\[ P(\text{AA} \mid t_A, t_B) = p_i^1 p_i^0, \]

\[ P(\text{AB} \mid t_A, t_B) = p_i^1 p_i^1, \]

\[ P(\text{BB} \mid t_A, t_B) = p_i^0 p_i^2, \]

\[ P(\text{NoCall} \mid t_A, t_B) = p_i^1 p_i^0. \]

(5) The contribution of individual quartets to the final call depends only on the relative probabilities; e.g. (AA, AB, BB)-genotype probabilities equal to (0.001, 0.001, 0.01) have the same information as the probabilities (0.01, 0.01, 0.1), i.e. BB is 10 times more likely than AA and AB. To make more robust calls, we shrink them by adding \( 1/2^2 \) to each of them. We denote the resulting quantities by \( p_{AA}, p_{AB}, p_{BB}, p_{\text{NoCall}} \).

(6) Calculate weights for each quartet, \( w_i, i = 1, 2, 3, \ldots, 10 \), using the module described next.

(7) For each SNP, calculate the probabilities for AA, AB, BB, NoCall, by averaging the corresponding probabilities from the ten quartets with the weights from Step 6. That is,

\[ \prod_{i=1}^{10} p_{AA,i} \]

and similarly for AB and BB. Calculate \( P(\text{AA}), P(\text{AB}) \) and \( P(\text{BB}) \) from these probabilities such that their sum is equal to 1.

(8) Make the call as the genotype that has the highest probability.

(9) Calculate the quality measure, \( Q \), as the probability of the called genotype.

(10) All the genotype calls with quality measures lower than \( q \) are changed to ‘NoCall’.

3.4 The weights for each quartet

As discussed above, the 10 quartets that are used to call the genotype for one SNP do not yield data of the same quality. The differences are because of design (location of the interrogation probe), variable hybridization kinetics owing to different oligonucleotide sequence composition, and local array characteristics. Initially all the weights, \( w_i \), are equal to 1, and the weight changes are done in two modules: ‘standard’ (applied all the time), and ‘multi-chip’ (uses information from many arrays to quantify the quartet efficiency and reliability). Standard weighting:

- Down-weight the quartets with a high probability of NoCall. That is we multiply the weight with \( 1 - p_{\text{NoCall}} \).
- Up-weight the probes with the polymorphic nucleotide close to the center of the oligo. We multiply the weights with 1.43 when the offset is zero, and with 1.07 when the offset is 1 (the
The weights are modified as follows: quartet probabilities are generally consistent with the made calls. Given a quartet we calculate the set of probabilities \( p_j \) such that \( p_j \) is the probability of the called genotype for the \( j \)-th array in the given quartet. If this set consist mostly of high probabilities, the quartet probabilities are generally consistent with the made calls. The weights are modified as follows:

- For each quartet we calculate the median of \( (p_j)^{1\leq j\leq n} \), and the weights are multiplied by these reliability scores.

### 4 RESULTS

The methods described above, the DM and RLMM algorithms, were applied to data from 15 XbaI arrays that were assayed at The University of Chicago. The hybridizations for these arrays were done on reference genomic DNA. The reference DNA was used for quality control and troubleshooting, and consensus genotypes for this sample were provided by Affymetrix based on nine independent replicates. Note that it is possible that some of the consensus genotypes are different than the underlying genotypes because they reflect only the majority call from the DM algorithm. The RLMM and GEL algorithms require the probe intensities for the genotype calling, and these were extracted from the .CEL files using the Gtype software provided by Affymetrix. The R code for RLMM was downloaded from Bioconductor, and the necessary internal files were downloaded from http://www.stat.berkeley.edu/nrabbee/RLMM/.

The RLMM algorithm relies on information from files containing information necessary for classification/genotype calling, files that are provided by the developer. Although we used the default of calling all genotypes, 4743 SNPs out of the total of 58960 were not called. We believe that this is a consequence of the small training sample (the 90 CEU individuals in HapMap). For example, the SNP SNP A-1641749 (rs16936866) is monomorphic in HapMap and no training data are available for two of the possible genotypes, and as a consequence it is not called by RLMM. Because \( \sim 8\% \) of the genotypes are missing even when the default is used, there is no meaningful comparison of call rates between RLMM, DM and GEL. The RLMM classification is based on determining a cutoff for the quality scores, and it was suggested (Rabbee and Speed, 2006) to use the quantiles of \( X^2 \). The quality scores for these arrays do not follow a \( X^2 \) distribution (median scores range from 2.06 to 24.76), so we could not decide on threshold based on this distribution. To allow a direct comparison of error rate for RLMM and GEL, we use only the most accurate RLMM genotypes that give the same call rate as GEL. For example, for the first array in Table 1, we used the top (lowest distances to clusters) 88.7\% genotypes out of the 54217 RLMM-called genotypes.

Table 1 shows the results of the three algorithms applied to the reference samples. GEL was applied in its basic form (multi-chip weights were not used). Note that there is a trade-off between accuracy and call rates, i.e. decreasing the threshold \( q \) will increase both the call and error rates. We choose a threshold for GEL \( (q = 0.85) \) that gives similar error rates to those given by the DM algorithm. Note that we could have chosen a higher threshold that would lead to call rates that are similar to those obtained using DM, but much smaller error rates.

The arrays show variable call rates, e.g. DM call rates range from 88.7 to 99.6\%. The variability in calling rate is surprising because hybridization was done on DNA of similar quality coming from a single subject. We think that this variability is owing to many factors including differences in the reagents used, sample preparation and variation in experimental protocol. Note that the call rates obtained using GEL are uniformly higher than those given by the DM algorithm, and the improvement is larger for the samples with low-throughput. For example, there is an additional \( \approx 4\% \) genotypes called for the two arrays with DM call rates <90.

<table>
<thead>
<tr>
<th>Call rate</th>
<th>Error (%)</th>
<th>AB in NoCalls</th>
<th>AB in errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>RLMM</td>
<td>GEL</td>
<td>DM</td>
</tr>
<tr>
<td>88.70</td>
<td>92.66</td>
<td>0.90</td>
<td>3.17</td>
</tr>
<tr>
<td>89.08</td>
<td>93.01</td>
<td>0.58</td>
<td>1.76</td>
</tr>
<tr>
<td>91.27</td>
<td>93.55</td>
<td>0.84</td>
<td>3.03</td>
</tr>
<tr>
<td>94.63</td>
<td>96.88</td>
<td>0.33</td>
<td>1.50</td>
</tr>
<tr>
<td>96.09</td>
<td>98.06</td>
<td>0.21</td>
<td>0.86</td>
</tr>
<tr>
<td>96.91</td>
<td>98.63</td>
<td>0.11</td>
<td>0.69</td>
</tr>
<tr>
<td>97.51</td>
<td>98.35</td>
<td>0.12</td>
<td>0.44</td>
</tr>
<tr>
<td>97.62</td>
<td>98.84</td>
<td>0.16</td>
<td>1.22</td>
</tr>
<tr>
<td>97.98</td>
<td>99.03</td>
<td>0.08</td>
<td>0.71</td>
</tr>
<tr>
<td>98.12</td>
<td>99.93</td>
<td>0.09</td>
<td>0.48</td>
</tr>
<tr>
<td>98.33</td>
<td>99.25</td>
<td>0.06</td>
<td>0.53</td>
</tr>
<tr>
<td>98.99</td>
<td>99.38</td>
<td>0.04</td>
<td>0.56</td>
</tr>
<tr>
<td>99.36</td>
<td>99.76</td>
<td>0.03</td>
<td>0.41</td>
</tr>
<tr>
<td>99.47</td>
<td>99.55</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>99.61</td>
<td>99.73</td>
<td>0.02</td>
<td>0.31</td>
</tr>
</tbody>
</table>

The methods are compared on: the percentage of present genotypes (call rate), the percentage of discordant genotypes (error %), the percentage of heterozygotes in the set of missing genotypes (AB in NoCalls), and the percentage of heterozygotes in the set of discordant genotypes (AB in errors).
Genotyping errors were calculated with respect to the consensus genotypes. Note that for both DM and GEL, the proportion of errors are <1%, and they are negatively correlated with the call rates. The proportion of errors when using RLMM was much higher. To investigate whether this is a consequence of incorrect consensus genotypes, we also looked at the replicability of the genotype calls. For RLMM and GEL we calculated the number of SNPs for which at least two genotypes were called at least three times each in the 15 reference arrays. For GEL, there were 111 SNPs with inconsistent genotypes, and for RLMM there were 659.

Strikingly, when using the DM and RLMM algorithms, the percentage of heterozygous genotypes in the set of missing genotypes or in the set of discordant genotypes is very large (29% of the reference genotypes are heterozygous). This implies that there is a higher chance for a heterozygous genotype to be missing or miss-called than a homozygous genotype. When using GEL, there is a significant decrease in the proportion of heterozygous in the missing and miscalled genotypes, but there is still an excess of heterozygous genotypes in the missing dataset.

We also applied the GEL algorithm using multi-chip weights. The reliability of the quartets was inferred from 30 independent arrays for which there was no prior information on the genotypes. We noticed an improvement in call rates (data not shown); as before the improvement is larger in the low-throughput samples—almost 1% increase over basic GEL for the two arrays with DM call rates under 90%. There was also a substantial improvement in the heterozygous rates.

5 DISCUSSION
The algorithm described in this paper provides results for the Affymetrix 100K platform that are more accurate than the DM results for similar call rates, or more abundant (higher call rates) for similar error rates. The algorithm can be applied to individual arrays, but can also incorporate information from other arrays by defining reliability measures for each quartet that are used in averaging the information across quartets. The biggest improvement when using GEL is that there seems to be no inflation in the percentage of heterozygous genotypes in the missing data or in the errors.

There are several directions for improvement of the algorithm. First, the error rates are not constant across arrays, and this complicates the adjustment for errors in the genetic association studies. Array-specific genotype calling thresholds can be defined such that the error rates are constant at a pre-specified level. Another improvement to the current implementation can be made by calibrating the quality scores (posterior probabilities), i.e. in the long run, \( p \times 100\% \) of the genotypes with maximum posterior probability equal to \( p \) are called correctly. Also, we assume the independence of the \( A \) and \( B \) statistics. An algorithm that would look at their joint distribution (i.e. derive the bivariate empirical distributions for \( (x_i, y_i) \) and use them to calculate posterior probabilities) can provide more accurate posterior probabilities. Finally, the method we introduce in this paper is not designed to handle structural polymorphisms such as copy number variation.

It is probable that an optimal genotyping calling algorithm would be based on a combination of single-chip and multi-chip approaches. Approaches such as RLMM that require training data and exchangeable arrays are likely to be optimal when the training data are informative (e.g. when sufficient data on all three genotypes are available). Methods that can be applied for individual arrays with no prior information, such as DM and GEL, work well for rare variants and 'unusual' arrays. We plan to investigate hybrid algorithms in future work.

ACKNOWLEDGEMENTS
The authors would like to thank Graeme Bell, Xinmin Li, Laura Martinolich and Raluca Nicolae for their contribution to producing the data used in this paper. The research was supported in part by NIH grants DK-55889, DK-20595, and DK-47486. K.M. was supported in part by a fellowship from the Manpei Suzuki Diabetes Foundation.

Conflict of Interest. none declared.

REFERENCES