Evaluating the Influence of Quality Control Decisions and Software Algorithms on SNP Calling for the Affymetrix 6.0 SNP Array Platform

Mariza de Andrade¹  Elizabeth J. Atkinson¹  William R. Bamlet¹
Martha E. Matsumoto¹  Sooraj Maharjan¹  Susan L. Slager¹  Celine M. Vachon²
Julie M. Cunningham³  Sharon L.R. Kardia⁴

Divisions of ¹Biomedical Statistics and Informatics, and ²Epidemiology, Department of Health Sciences Research, and ³Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minn., and ⁴Department of Epidemiology, University of Michigan, Ann Arbor, Mich., USA

Abstract

Objective: Our goal was to evaluate the influence of quality control (QC) decisions using two genotype calling algorithms, CRLMM and Birdseed, designed for the Affymetrix SNP Array 6.0. Methods: Various QC options were tried using the two algorithms and comparisons were made on subject and call rate and on association results using two data sets. Results: For Birdseed, we recommend using the contrast QC instead of QC call rate for sample QC. For CRLMM, we recommend using the signal-to-noise rate $\geq 4$ for sample QC and a posterior probability of 90% for genotype accuracy. For both algorithms, we recommend calling the genotype separately for each plate, and dropping SNPs with a lower call rate ($<95\%$) before evaluating samples with lower call rates. To investigate whether the genotype calls from the two algorithms impacted the genome-wide association results, we performed association analysis using data from the GENOA cohort; we observed that the number of significant SNPs were similar using either CRLMM or Birdseed. Conclusions: Using our suggested workflow both algorithms performed similarly; however, fewer samples were removed and CRLMM took half the time to run our 854 study samples (4.2 h) compared to Birdseed (8.4 h).

Background

With the advances of technology, several DNA analysis arrays have been developed to measure variation in single nucleotide polymorphisms (SNPs) and copy number variation (CNV). In this paper, we focused on the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy SNP 6.0) that features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 additional probes for the detection of CNV. This chip was an updated version of the previous two-chip design based on the NspI and StyI restriction enzymes.

There are two genotype calling algorithms (GCAs) broadly used for the Affy SNP 6.0: Birdseed and CRLMM. Birdseed is an evolution of the Bayesian Robust Linear
Model with Mahalanobis distance classifier (BRLMM) GCA \[1, 2\] and is the default algorithm used by Affymetrix \[3\]. It performs a multiple-chip analysis to estimate signal intensity for each allele of each SNP, fitting probe-specific effects to increase precision (similar to the BRLMM-P algorithm) \[2\]. The Birdseed algorithm then makes genotype calls by fitting a Gaussian mixture model in the two-dimensional A allele intensity signal versus B allele intensity signal space, using SNP-specific models to increase performance. Carvalho et al. \[4\] observed that this algorithm did not perform well for data from different laboratories and developed a new algorithm named CRLMM (corrected RLMM) to correct for various batch-related effects. Similarly, Lin et al. \[5\] observed, using HapMap samples, that Birdseed had similar limitations to BRLMM in cross-laboratory stability of their calls. They also found CRLMM provided confidence measures that better correlated with observed accuracy.

As with many new technologies, the Affymetrix analysis workflow is still a work in progress \[6–10\]. In the laboratory, a sequence of time-consuming steps must be performed taking one lab technician five days to process a sample (fig. 1). In our experience, this platform is very sensitive to sample DNA quality such as concentration, buffer solution, and DNA degradation, furthering the challenges within the laboratory. In addition to sample quality problems, we experienced issues with hardware such as filter plate failures during the PCR step and alignment problems during the scan process, further delaying the sample processing. Besides challenges in the laboratory, the recommended analysis software has undergone several revisions. Initially the Birdseed-v1 algorithm was used and one year later this was updated to Birdseed-v2 \[9, 10\] to fix several bugs. Additionally, the quality control (QC) step used to identify poor quality samples was modified in February 2008 by Affymetrix \[8\]. The QC call rate (QCR) was used first, identifying samples with values less than 0.86 as ‘failures’ that needed to be repeated \[7\]. However, QCR was shown to be of limited value for predicting genotyping performance in problematic samples \[11\]. Thus, the current recommended algorithm is the contrast QC (CQC), where values less than 0.40 are considered failures \[8, 11\]. Although the QCR is no longer the recommended QC step, it is still being used by some genotyping facilities.

Thus, our objective was to evaluate the influence of QC decisions for the Affymetrix SNP Array 6.0 using two GCAs, CRLMM and Birdseed v2, and determine whether these decisions had an effect on association analysis results.

**Fig. 1.** Description of the Affymetrix SNP Array 6.0 processing. This workflow details the steps required for a single individual performing the process. The QC steps are essential: verifying restriction enzyme activity and DNA concentration (by pico green); verifying PCR by gel analysis, and lastly verifying optimal fragmentation. Any sample that does not pass these steps is not applied to the Affymetrix array. The two starred points (filter plates and Birdseed 2) indicate the unreliability of filter plates recommended by Affymetrix and Birdseed 2 that cannot be recalled.

---

- **Pre-PCR**
  - DNA concentration
  - Test restriction enzymes

- **PCR**
  - Pool PCR
  - Specific for concentration

- **Fragment DNA**

- **Hybridization and wash to GeneChip**

- **Scan**
  - Can fail alignment
  - DTT pulls all data
  - GC analysis for QC cutoff

- **Notes**
  - DTT is real time, so GC QC is not calculated before pulling – need to add in pass/fail step before DTT

*Fig. 1: Description of the Affymetrix SNP Array 6.0 processing. This workflow details the steps required for a single individual performing the process. The QC steps are essential: verifying restriction enzyme activity and DNA concentration (by pico green); verifying PCR by gel analysis, and lastly verifying optimal fragmentation. Any sample that does not pass these steps is not applied to the Affymetrix array. The two starred points (filter plates and Birdseed 2) indicate the unreliability of filter plates recommended by Affymetrix and Birdseed 2 that cannot be recalled.
Methods

Data
Two data sets were used in our evaluation. The first data set included samples from 854 non-Hispanic white adults enrolled in the Genetic Epidemiology Network of Arteriopathy (GENOA) study of the Family Blood Pressure Program (FBPP), a study designed to identify genetic determinants of hypertension in multiple ethnic groups [12]. The second data set included samples from 746 non-Hispanic white adults enrolled in the Genetic Epidemiology study of Chronic Lymphocytic Leukemia (CLL) [13]. These samples were genotyped at the same time at the Genotyping Shared Resource of the Mayo Clinic Technology Center using Affymetrix SNP Array 6.0.

Affy SNP 6.0
There are 934,968 SNPs on the 6.0 chip. Of these, 3,022 were duplicates and were included purely for QC purposes (2,906 of these have rsIDs). An additional 25,346 SNPs have been excluded from the officially released SNPs due to poor performance. The default extraction of the SNPs using most software packages yielded only 909,622 SNPs. Once the duplicates were excluded, 906,600 SNPs remained for analysis (table 1).

Birdseed Algorithm
Birdseed-v2 uses the expectation-maximization (EM) algorithm to derive a maximum likelihood fit of a 2-dimensional Gaussian mixture model in the A versus B allele space. A key difference between Birdseed-v1 and Birdseed-v2 is that the former uses SNP-specific models or priors only as an initial condition from which the EM fit is free to wander; on rare occasions this allows for mislabeling of the clusters. For Birdseed-v2 the SNP-specific priors are used not only as initial conditions for EM, but are incorporated into the likelihood as Bayesian priors. This constrains the extent to which the EM fit can wander off. Correctly labeling SNP clusters, whose centers have shifted relative to the priors, is problematic for both Birdseed versions. However, given the additional constraint on the EM fit, Birdseed-v2 is more likely than Birdseed-v1 to either correctly label the clusters or set genotypes to no-calls [8]. The algorithm by default uses minor allele frequency (MAF) values from the Caucasian HapMap data as ‘seeds’. It is also possible to specify your own ‘seeds’ based on other data sources. Each call is given a confidence score that varies from 0 (best) to 1 (worst). Confidence is calculated as Confidence = 80% × E_i + 20% × E_2, where E_i is the posterior probability (PP) ratio between the 2nd closest peak to the closest peak, and E_2 is the deviation penalty from the closest peak. From the confidence we can stipulate a confidence score threshold, which is the maximum score at which the Birdseed algorithm will make a genotype call. Larger values of the confidence threshold indicate less certain calls. Calls with confidence scores above the threshold are assigned a no-call. The confidence score threshold default is 0.5.

There are two QC assessments available for use prior to executing the Birdseed calling algorithm: CQC and QCR. The QCR uses a dynamic model QCR analysis of a subset of 3,022 duplicate SNPs following chip scanning. However, the QCR was shown to be dependent on sample quality, and it did not measure the degree to which the perfect match intensities of all SNPs cluster by genotype. In high quality samples, the A and B allele probe intensities display three clusters for AA, AB, and BB genotypes. For poor quality samples, these clusters will merge. Since the genotype calling is based on cluster resolution, this QC algorithm was not ideal. The CQC algorithm was developed to better track the genotyping call rates yielded by the Birdseed algorithm. This algorithm measures the separation of allele intensities into three clusters in a single dimension generated by the projection of the two-dimensional allele intensity space. A detailed description of these two QC algorithms is described in the Affymetrix White Paper [10]. According to the initial Affymetrix SNP 6.0 workflow, re-processing was recommended for samples with a CQC < 0.86 [6]. The current Affymetrix recommendation is to re-process samples with CQC < 0.4 [9, 10].

CRLMM Algorithm
CRLMM uses a supervised learning approach based on a two-stage hierarchical model [4]. Unlike other algorithms, CRLMM models M = \log_2 (I_A) – \log_2 (I_B) instead of the intensity pair, where I_A and I_B are the intensity of alleles A and B, respectively. This choice makes CRLMM more robust to probe effects [Carvalho, unpubl. data]. To account for the dependence of M on the overall intensity S, defined as S = (\log_2 (I_A) + \log_2 (I_B))/2, Carvalho et al. [4] fitted splines using a mixture model and corrected the bias using 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 has a different mean and standard deviation which are estimated from HapMap samples. Because of its design, HapMap provides few rare genotypes in some SNPs, so Carvalho et al. [4] used an empirical Bayes approach in which the means, conditioned on genotype, follows a multivariate normal distribution and the variances have a scale-inverse chi² distribution. This approach permits CRLMM to borrow strength from other SNPs. To make calls, CRLMM treats the estimated parameters as known and computes PP_s for each genotype given the observed log-ratio M. The posteriors are then used as a confidence measure. Lin et al. [5] found that the confidence measures provided by CRLMM were not optimal and proposed an ad-hoc adjustment based on a training approach. These are the confidence measures currently implemented in CRLMM. There are other measures such as SNPQC and SNR. SNPQC is a quality score for each SNP, and a SNPQC value > 0.25 indicates a SNP with very accurate genotype call. SNR is a signal-to-noise ratio

<table>
<thead>
<tr>
<th>Table 1. Distribution of the 906,600 SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
</tbody>
</table>

Evaluating Calling Algorithms for Affy 6.0

Hum Hered 2011;71:221–233
quantity, and when it is close to zero it is impossible to distinguish between homozygous and heterozygous clusters [4]. By default, CRLMM considers SNPs with SNR $\geq 5.0$ to be of excellent quality.

The CRLMM workflow requires the user to make several decisions about which samples and genotype calls are going to be trusted. The impact of choosing different cut-off values for the sample quality scores, SNR, and for the PP should be investigated for each study. In our investigation we tried the following:

- SNR cutoffs of 3, 4, and 5
- PPs of 75, 80, 85, 90, and 95%.

Additional Decisions
Our recommended workflow for Birdseed and CRLMM algorithms is shown in table 2. However, there are additional decisions that need to be made by the investigator which can have a large influence on the final results. After applying the appropriate QC algorithm, samples can be called separately for each plate (individual) or using all samples together (pooled). The current workflows leave this decision up to the investigator. Additionally, order matters when determining call rates for SNPs and subjects. We investigated the impact of dropping SNPs with call rates $< 95%$.

<table>
<thead>
<tr>
<th>Action</th>
<th>CRLMM</th>
<th>Birdseed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove low quality samples</td>
<td>SNR $\geq 4.0$</td>
<td>QC contrast $\geq 0.4$</td>
</tr>
<tr>
<td>Set unreliable genotype calls to missing</td>
<td>PP $&lt; 0.90$</td>
<td>Confidence measure $&lt; 5$</td>
</tr>
<tr>
<td>Remove SNPs with a low call rate</td>
<td>SNP call rate $\geq 95%$</td>
<td>SNP call rate $\geq 95%$</td>
</tr>
<tr>
<td>Remove samples with a low call rate</td>
<td>Sample call rate $\geq 95%$</td>
<td>Sample call rate $\geq 95%$</td>
</tr>
</tbody>
</table>

Table 2. Suggested workflow for Birdseed and CRLMM

**Fig. 2.** Comparison between the two Affymetrix QC measures: QCR and CQC for the GENOA and CLL samples. The x-axis displays the QCR with its cut-off value of 0.86. The y-axis displays the CQC with its cut-off value of 0.4. For GENOA, 48 samples failed QCR and 102 samples failed CQC. For CLL, 30 samples failed QCR and 53 samples failed CQC.
then dropping subjects with call rates $< 95\%$ or reversing the order and dropping subjects first.

**Statistical Analysis**

For the association analysis we used the quantitative trait, estimated glomerular filtration rate (eGFR), adjusting for age and sex, available only in the GENOA data set, and ran a mixed effects linear model taking into account the familial relationship among the sibships using the R library multic (http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm) for the association analysis. We used PLINK for data management, for calculating identify-by-state analysis to confirm familial relationships and to identify hidden familial relationships, and for performing population stratification analysis [14].

### Results

**Comparison of the Birdseed QC Measures**

The purpose of QCR and CQC is to identify samples that are of poor quality and might negatively influence the Birdseed calling of other subjects. Figure 2 compares the QCR and CQC values for each sample for the GENOA and CLL cohorts, where 94 and 93% of the samples passed QCR and 83 and 96% passed CQC, respectively. Additionally, the most recent Affy SNP 6.0 user guide [10] suggests that plates should be flagged as potentially problematic if less than 90% of a given plate passes the CQC threshold or if the average CQC of those samples that passed is less than 1.7. For the GENOA samples, the mean and median CQC over all the plates were 1.67 and 1.72, respectively, and for the CLL samples, the mean and median CQC over all the plates were 1.81 and 1.91, respectively. As shown in figure 3, these values can vary significantly by plate. Based on this recommendation, 60% of the GENOA plates were flagged as opposed to 11% of the CLL plates. Despite this recommendation we did not exclude these plates from our evaluation because we were interested in assessing how the performance of these algorithms impacts the performance of the Birdseed calling algorithm. As was observed by Affymetrix White Paper [10], we also observed that although more samples passed the initial screening using QCR, more samples ultimately passed Birdseed using CQC suggesting that the ‘bad’ samples perhaps contaminated the results. Of the GENOA and CLL samples that passed QCR (n = 806 and 693, fig. 2A, B), 703 (87.2%, fig. 4A) and 693 (93%, fig. 4B) passed Birdseed (sample call rate $\geq 0.95$). On the other hand, of the 709 GENOA samples that passed CQC in figure 2B, 716 passed Birdseed (100%, fig. 4B). Thus, the samples that passed QCR but failed CQC generally failed the Birdseed algorithm. We also observed that these ‘bad’ samples influenced the Birdseed calling rate of the ‘good’ samples if the prior filtering (CQC $\geq 0.4$) was not applied (fig. 5: plates 8 and 9, GENOA; plate 7, CLL).

**Pooled versus Individual Plate Calling Using Birdseed**

The impact of pooling samples was evident in the GENOA samples but not in the CLL samples, regardless of which options were tried for the GENOA samples, and we ultimately passed more GENOA samples running the
Birdseed algorithm per plate rather than pooled. The influence of the QC algorithms was also evident when comparing calling using pooled or individual plates. The increased failure rate for the pooled samples was considerably higher after using the CQC in the GENOA samples compared to the CLL samples. More failures were observed from samples using QCR compared to CQC for pooled plates (31 vs. 12% for GENOA; 1.5 vs. 0.7% for CLL) and for individual plates (13 vs. 0.4% for GENOA; 0 and 0% for CLL). Based on these results, we recommend running the Birdseed algorithm separately for each plate if there are underperforming plates as we observed in GENOA (fig. 4A).

**Comparison of the CRLMM Options**

CRLMM also provides a QC measure for samples (SNR) to assess sample quality. We tried multiple SNR thresholds, and observed that there were overall more samples failing under the default SNR value of 5 (table 3). Thus, we concluded that a SNR threshold of 5, though recommended, may be too conservative and in our data a SNR threshold of 4 appeared to be more appropriate (table 3). Using this SNR threshold of 4, we investigated the appropriate threshold for the PP that determined an acceptable sample call rate and SNP call rate. We observed that a PP of 90% yielded an acceptable sample call rate and SNP call rate (table 4). Defining what is acceptable for sample and SNP call rates is left to the investigators. In our case, this decision was made after balancing the number of samples and SNPs we would keep and how comparable these numbers were with the ones obtained with Birdseed. We also observed similar results using the CLL samples (results not shown).

**Pooled versus Individual Plate Calling Using CRLMM**

Since the impact of pooling was evident in the GENOA samples using Birdseed, we decided to perform a similar
Fig. 4. Description of the Affymetrix workflow for the two QC measures: QCR and CQC. B CLL samples.

Table 3. CRLMM evaluation for SNR thresholds under pooled and individual plate calling for 854 samples and 906,600 SNPs

<table>
<thead>
<tr>
<th>SNR threshold</th>
<th>Plate calling</th>
<th>Samples passed after SNR, n (%)</th>
<th>Samples passed after CRLMM*, n (%)</th>
<th>SNP passed, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>individual</td>
<td>836 (98)</td>
<td>812 (95)</td>
<td>679,968 (75)</td>
</tr>
<tr>
<td>3</td>
<td>pooled</td>
<td>836 (98)</td>
<td>808 (95)</td>
<td>679,985 (75)</td>
</tr>
<tr>
<td>4</td>
<td>individual</td>
<td>805 (94)</td>
<td>799 (94)</td>
<td>692,011 (76)</td>
</tr>
<tr>
<td>4</td>
<td>pooled</td>
<td>805 (94)</td>
<td>798 (93)</td>
<td>693,008 (76)</td>
</tr>
<tr>
<td>5</td>
<td>individual</td>
<td>703 (82)</td>
<td>703 (82)</td>
<td>712,076 (79)</td>
</tr>
<tr>
<td>5</td>
<td>pooled</td>
<td>703 (82)</td>
<td>703 (82)</td>
<td>711,985 (79)</td>
</tr>
</tbody>
</table>

* Using the PP threshold of 0.90 and ‘drop SNP first’ filter.
Fig. 5. Influence of a CQC filter in the Birdseed genotype calling for the GENOA and CLL samples. A higher sample missingness rate is observed when no CQC filter is applied, particularly for plates 8 and 9 for GENOA samples and plate 7 for CLL samples. P = Passed; F = failed.
Evaluation of CRLMM algorithms for Affy 6.0

To see whether we observed a similar effect, the number of samples with an acceptable call rate was similar using either pooled or individual plate calling. Using our suggested threshold of 90% for PP and 4 for SNR, 94% of the samples and 76% of the SNPs passed under individual plate calling compared to 93% of the samples and 76% of the SNPs that passed under pooled plate calling (Table 3, Fig. 6A). Using the same thresholds for PP and SNR in the CLL samples, we observed that 98% of the samples and 80% of the SNPs passed with either pooled or individual plate calling (Fig. 6B). Based on these results, we observed that pooled plate calling was more susceptible to underperforming plates than individual plate calling. Thus we recommend running the CRLMM algorithm separately for each plate.

**Table 4.** CRLMM evaluation for PP using a SNR threshold of 4 under individual plate calling and ‘drop SNP first’ filter for 805 samples and 906,600 SNPs

<table>
<thead>
<tr>
<th>PP threshold</th>
<th>Samples passed, n (%)</th>
<th>SNP passed, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>794 (93)</td>
<td>639,107 (70)</td>
</tr>
<tr>
<td>0.90</td>
<td>799 (94)</td>
<td>692,011 (76)</td>
</tr>
<tr>
<td>0.85</td>
<td>805 (94)</td>
<td>728,910 (80)</td>
</tr>
<tr>
<td>0.80</td>
<td>805 (94)</td>
<td>754,930 (83)</td>
</tr>
<tr>
<td>0.75</td>
<td>805 (94)</td>
<td>773,633 (85)</td>
</tr>
</tbody>
</table>

**Fig. 6.** Comparison between CRLMM and Birdseed algorithms dropping SNP first. A GENOA samples.
since it is less susceptible to underperforming plates and does not appear to negatively influence the genotype calling.

**Identifying Poor Quality SNPs and Samples**

Regardless of the calling algorithm, the order of investigating call rates of samples and SNPs needs serious consideration. For the GENOA samples, when we dropped samples first using our suggested workflow for Birdseed, 17% of the samples and 33% of the SNPs were dropped (online suppl. fig. 1, for all online suppl. material, see www.karger.com/doi/10.1159/000328843); using our suggested workflow for CRLMM, 11% of the samples and 22% of the SNPs were dropped (online suppl. fig. 1). When we dropped SNPs first using our suggested workflow for Birdseed, 17% of the samples and 33% of the SNPs were dropped (fig. 6A); using our suggested workflow for CRLMM, 6% of the samples and 24% of the SNPs were dropped (fig. 6A). These 6% samples were among the 11% of samples dropped using CRLMM when samples were dropped first, indicating that by removing underperforming SNPs first, we can retain more samples for analysis. For the CLL samples, when we dropped samples first using our suggested workflow for Birdseed, 4% of the samples and 19% of the SNPs were dropped (online suppl. fig. 2); using our suggested workflow for CRLMM, 3% of the samples and 20% of the SNPs were dropped (online suppl. fig. 2). When we dropped SNPs first, using our suggested workflow for Birdseed, 4% of the samples and 19% of the SNPs were dropped (fig. 6B); using our suggested workflow for CRLMM, 2% of the

---

**Fig. 6.** Comparison between CRLMM and Birdseed algorithms dropping SNP first. **B** CLL samples.
samples and 20% of the SNPs were dropped (fig. 6B). Based on our results, we recommend dropping SNPs first rather than samples, since fewer samples are dropped and dropped SNPs can be imputed [15]. The same recommendation was also suggested recently by another genome-wide association QC study [16].

Comparison of Birdseed and CRLMM in Practice

To perform this comparison, we ran a genome-wide association analysis using the GENOA cohort and the continuous trait eGFR as the outcome variable. Samples were analyzed using the genotyping algorithms by plate using our recommended workflow (fig. 6A). Using the samples and SNPs that passed both algorithms, we also observed similar genotypes (99.8% concordance rate).

The final data sets using CRLMM and Birdseed consisted of 799 (93.6%) and 709 (83.0%) samples, respectively. From those samples, 709 (83.0%) passed both algorithms, 55 (6.4%) failed both, and 90 (10.5%) failed Birdseed. We observed that the genomic inflation factor \( \lambda \) was the same for both algorithms (1.019), and the number of significant SNPs using either calling algorithm was similar for p values more significant than \( 10^{-3} \); however, for more stringent significant levels, more SNPs were observed using Birdseed (fig. 7A, B, table 5, online suppl. tables 1–5). Furthermore, the most significant SNP identified by one algorithm was not always revealed by the other (fig. 7A, B, online suppl. fig. 3–5). For instance, the most significant SNP (rs100488947 on chromosome 4) identified by CRLMM (p value = \( 2.35 \times 10^{-7} \)) was seen by Birdseed at a slightly lower level of significance (p value = \( 7.42 \times 10^{-6} \)). One sample did not have information

![Fig. 7. Q-Q plots for the association analysis p values using GENOA samples with genomic inflation factor \( \lambda \). A Birdseed GCA. B CRLMM GCA.](image)

<table>
<thead>
<tr>
<th>p value cutoff</th>
<th>Birdseed only*, n (%)</th>
<th>CRLMM only*, n (%)</th>
<th>Both, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>19,692 (3.38)</td>
<td>19,833 (3.40)</td>
<td>40,408 (6.93)</td>
</tr>
<tr>
<td>0.01</td>
<td>2,886 (0.49)</td>
<td>2,819 (0.48)</td>
<td>3,476 (0.59)</td>
</tr>
<tr>
<td>0.001</td>
<td>328 (0.06)</td>
<td>389 (0.07)</td>
<td>323 (0.06)</td>
</tr>
<tr>
<td>1.00–4</td>
<td>56 (0.01)</td>
<td>27 (0.005)</td>
<td>25 (0.004)</td>
</tr>
<tr>
<td>1.00–5</td>
<td>12 (0.002)</td>
<td>2 (0.0003)</td>
<td>1 (0.0002)</td>
</tr>
<tr>
<td>1.00–6</td>
<td>0 (0)</td>
<td>1 (0.0002)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

* The genomic inflation factor \( \lambda \) was 1.019 for both calling algorithms.
Discussion and Conclusion

According to the most recent Affy SNP 6.0 user guide [10], data sets should be identified as being potentially problematic if less than 90% of the samples pass the CQC threshold or if the average CQC of the remaining samples is less than 1.7. The GENOA data did not fulfill the first condition and was close to the cut-off value for the second condition due to an abysmal CQC of more than 40% in two plates. Even with these potential problems, we ultimately were able to identify significant genome-wide association study results. We also observed that CQC is a more reliable measure than QCR, which is more sensitive to problematic samples as were observed in the GENOA data. In the GENOA data, 99.5% of the samples that passed CQC also passed Birdseed compared to 87.2% using QCR, whereas 100% of the CLL samples passed using either CQC or QCR. Our results have also shown, using both cohorts, that more samples passed if Birdseed was performed by individual plate instead of pooling the plates. Finally, run time for each algorithm may be another consideration; it took CRLMM 4.2 h (3 h) and Birdseed 8.4 h (8.4 h) to run 854 GENOA samples by plate (run time for pooled data).

We created guidelines based on two data sets; however, given the sensitivity of these algorithms we suggest examining each project carefully. Recently there have been several articles published discussing similar problems, focusing either on statistical issues of genome-wide association [17] or the usefulness of QC measures when using high-throughput genotyping from existing public databases [18]. Several factors may influence call rates such as mixtures of populations within a plate or unbalanced numbers of cases and controls in a plate that can result in bad genotype clustering [17, 18]. These factors did not apply in either of our two data sets since the phenotypes for GENOA were continuous, and the primary phenotype for the CLL study, although discrete, was controlled by randomization of the samples to each plate. Additionally, our samples were all white non-Hispanics which was also confirmed by population stratification analysis implemented in PLINK [14]. Furthermore, each of our plates had a minimum of 15 females as required by Affymetrix – Best Practices SNP 6.0 Workflow [7–9]. We have performed our evaluation on two data sets that were genotyped at the same time in the same platform and using the same genotyping center. By comparing these two studies we observed that QC measures are important to detect problematic samples. If the samples are not appropriately filtered, it can influence the sample and SNP call rates, regardless of which algorithm is used. We believe that evaluating these two GCAs in a real set-up can provide a useful workflow for other researchers. The factors that influence genotyping are difficult to capture in a realistic simulation of the intensity files and are outside the scope of this paper. Additionally, real data has issues not seen by simply using the HapMap samples, and previous validation of GCAs has used the intensity files of HapMap data and not simulated data [1, 4, 5].

In summary, we have shown that the overall performance of the Birdseed genotyping process varies by choice of the two pre-cluster sample QC checks (QCR and CQC) and the number of samples used to call the genotypes (by plate versus using all samples). We have summarized study results using the two GCAs Birdseed and CRLMM, investigating the impact of several options within the CRLMM function and comparing advantages and disadvantages of the two algorithms. We have performed a genome-wide association analysis in the GENOA samples using the genotype calls from those two algorithms and observed similar association results when the level of significance was more significant than 10^{-3}, but not when the level of significance was below 10^{-3} where more significant SNPs were picked up by Birdseed; we could not evaluate the number of false positives since truth was unknown. Finally, we conclude that it is useful to validate findings using either another platform or GCA if available, and to look at SNP genotype clustering plots if intensity data are available.

Acknowledgments

We thank DeLaine Anderson for her technical assistance with the manuscript and Benilton Carvalho for his helpful comments about CRLMM. The Genotyping Shared Resource of the Mayo Clinic Technology Center did all the genome-wide association genotyping described in this paper. This paper was supported by the National Institutes of Health research grant R01 HL87660 (to M.d.A. and S.L.R.K.) and National Cancer Institute grant U01 CA118444 (to S.L.S.).
References

14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.