

Quality Control Assessment in Genotyping Console™

Introduction

Prior to the release of Genotyping Console™ (GTC) 2.1, quality control (QC) assessment of the SNP Array 6.0 assay was performed using the Dynamic Model QC (DM) call rate analysis of a subset of 3,022 SNPs following chip scanning. In certain problematic data sets, DM call rates are of limited value for predicting genotyping performance. Such problematic data sets may occur when sample DNA is of less than optimal quality or when assay performance deviates from Affymetrix-recommended procedures.

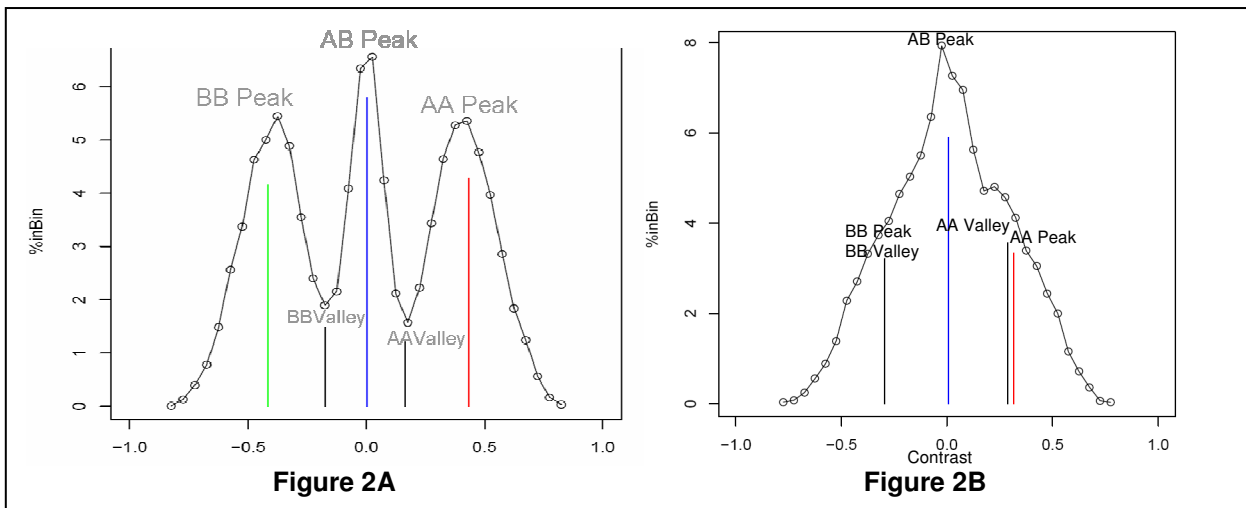
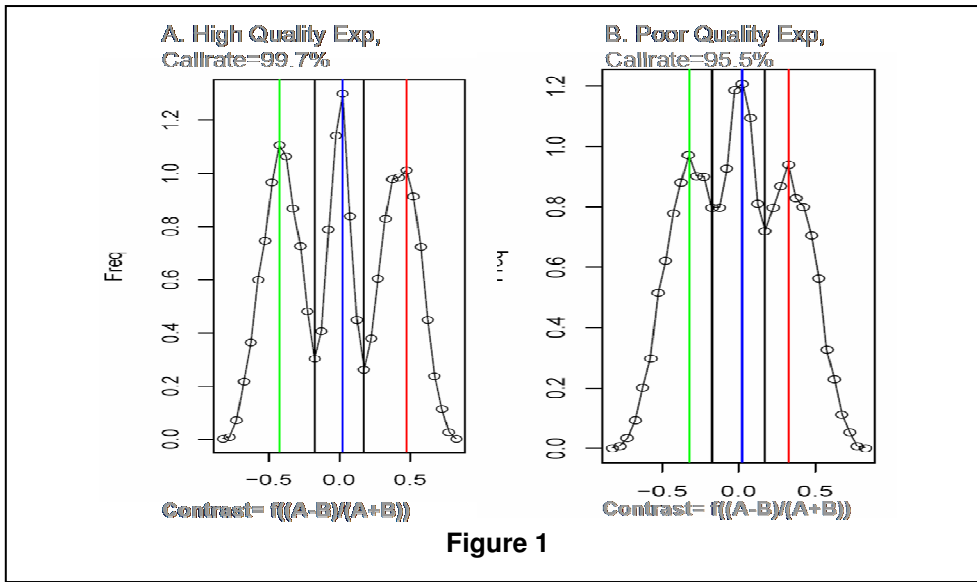
DM call rates measure the consistency of PM and MM intensities within each SNP, with four possible genotyping states (Null, AA, AB and BB). The DM algorithm does not measure the degree to which the PM intensities of all SNPs cluster by genotype. In high-quality samples, the A and B allele probe intensities will display three clusters for the AA, AB and BB genotypes of the sample. In poor-quality samples, these clusters will merge. This property of cluster resolution is a better predictor of a sample's genotyping performance (call rate and concordance) resulting from cluster-based genotype calling algorithms.

With this in mind, an effort has been undertaken to develop a new QC algorithm that better tracks with genotyping call rates, produced by Birdseed, the standard cluster-based genotyping algorithm for the SNP Array 6.0. This new algorithm is referred to as Contrast QC (CQC). In this paper, a description of the algorithm will be followed by advanced interpretative information.

CQC measures the separation of allele intensities into three clusters in “contrast space.” Contrast space is a projection of the two-dimensional allele intensity space into an informative single dimension. Specifically, a SNP's contrast value is a function of *AllelicContrast* (Equation 1) whose values range from -1 for the ideal BB genotype to +1 for the ideal AA genotype with the ideal AB genotype at zero. A and B are the medians of the replicate probes for A and B alleles.

$$\text{Equation 1} \qquad \text{AllelicContrast} = \frac{(A - B)}{(A + B)}$$

A high-quality sample produces a histogram of contrast values that resolves into three clusters for the three genotypes in the sample. An example is shown in Figure 2 (left). In poor-quality samples, there is a reduction in homozygous peak resolution as data quality begins to degrade (Figure 1, right). Poorly resolved clusters mean that genotyping is unlikely to work well and such samples will have lower call rates and accuracy. CQC is a metric that captures the property of cluster resolution in contrast space. A high-quality sample produces three genotype peaks, separated by two valleys, in the histogram of contrast values (Figure 2A) for a representative sample of SNPs.



The difference between each homozygous peak and its associated valley is calculated, and the least of these two values becomes the CQC value (Equation 4, discussed below). In poor-quality samples, the increase of signal within the AB cluster results in significant loss of peak/valley resolution and a concomitant decrease in the CQC value. The CQC algorithm will always fit three peaks, even when only one cluster is present (Figure 2B).

In such cases, the CQC value will be zero or negative, which means that the sample produces no cluster resolution. CQC is one of several approaches to measuring clustering resolution. It has few assumptions, and among various alternatives tested, was found to work best as a predictor of genotyping performance.

CQC values are generated for three populations of autosomal SNPs. One is a random, representative sampling of the total SNP population (*CQC (random)*), one is a sampling of

SNPs that reside only on Nsp fragments (*CQC (Nsp Only)*) and one is a sampling of SNPs that reside only on Sty fragments (*CQC (Sty Only)*). The CQC values for the random, representative SNPs are the most predictive of the sample's genotyping performance. The enzyme-only CQC values are not as predictive of the genotyping performance of the combined targets; however, a large difference between them can flag a single-enzyme target prep failure. Therefore, the final CQC value shown in the CQC column of the GTC 2.1 QC report view relies on one additional test of the data to screen for a gross enzyme failure. This will be described in the Algorithm Basics section.

In summary, CQC measures the separation of allele intensities into three clusters in *contrast space*. This property of cluster resolution is a better predictor of a sample's genotyping performance than is the DM call rate analysis, especially in certain problematic data sets (discussed below).

Summary of Best Practices for Quality Control

The following lists the current best practices steps for pre-clustering QC of samples and data sets. The details will be explained in this paper. Note that best practices QC recommendations are continuing to evolve in the field, and these are a subset of best practices steps for study design, and filtering data for use in association studies.

- Perform CQC analysis for all samples within a batch.
- Re-hybridize, reprocess or fail those samples that have a CQC < 0.4.
- Calculate the mean of the CQC for those samples that pass the 0.4 threshold.
- Apply the following rules for flagging potentially problematic data sets:
 - When more than 10 percent of samples in a batch do not pass the CQC cutoff of 0.4.
 - After plotting CQC Nsp Only vs. CQC Sty Only, samples that do not fall within the major cluster and are significantly separated should be suspected of decreased enzyme performance.
 - Batches where the mean passing CQC is < 1.7.

Algorithm Basics

The CQC algorithm transforms A and B values into a contrast value with a given subset of SNPs. A and B are the median intensities of replicate probes for the A and B alleles of a SNP. The full contrast equation is shown below, where k is a stretch parameter.

$$\text{Equation 2} \quad \text{Contrast} = \sinh \left(\frac{\left(k * \frac{(A - B)}{(A + B)} \right)}{\sinh(k)} \right)$$

The transformation in equation 2 gives the same functional form as equation 1 but changes the spread between heterozygous and homozygous genotypes, based on the parameter k , which we chose as $k = 2$.

Ten thousand autosomal SNPs were randomly selected without any bias for performance or occurrence on enzyme fragments. The selection was performed during algorithm development and now constitutes a constant set applied each time the algorithm is performed. The actual SNPs are located in the SNP Array 6.0 .QCC file found in the library folder for GTC 2.1.

The magnitude of the peaks and valleys in the histogram of contrast values for the 10,000 random SNPs are used to compute the CQC (random) value for a sample. The *AA_Peak*, *AB_Peak* and *BB_Peak* values, such as those shown in the Figure 1A example, are identified with the following procedure.

First, the algorithm locates the centers of the AA, AB and BB clusters in contrast space, using a Gaussian Mixture Model, fit by EM (expectation maximization). The Gaussian Mixture model always finds the best fit for three clusters. This explains why homozygous peaks are found in very poor-quality data where only a central heterozygous peak is visible (see Figure 1B). Second, the *AA_Peak*, *AB_Peak* and *BB_Peak* values are set to the *%Bin* histogram values (y axis) at the contrast locations, identified by the EM fit for each genotype cluster. The contrast histogram is constructed by dividing the contrast values, produced into bins of size 0.02. Therefore, the center bin would be from -0.01 to 0.01, with bins extending in both directions. There are 100 bins from -1 to +1. The *%Bin* values are determined by calculating the percent of values in each bin relative to the total number of SNPs tested as given by Equation 3.

$$\text{Equation 3} \quad \text{Percent} = \frac{(\text{SNPs_in_bin})}{(\text{Total_Number_SNPs})}$$

Finally, the AA and BB valleys are located by finding the bins with minimal *%Bin* values between the AA/AB and AB/BB peaks, respectively.

The difference between each homozygous peak and its associated valley is calculated. The least of these two values becomes the CQC (random) value (Equation 4) when the contrast input is calculated for the 10,000 random SNPs.

$$\text{Equation 4} \quad \text{ContrastQC}(\text{random}) = \min \left(\begin{array}{l} BB_Peak - BB_Valley, \\ AA_Peak - AA_Valley \end{array} \right)$$

The final CQC value shown in the *Contrast QC* column of the GTC 2.1 quality control report view relies on one additional test of the data to screen for a gross failure of one of the target preps (Nsp or Sty). Instead of using a random set of 10,000 SNPs, a carefully selected panel of 20,000 SNPs for each enzyme (Nsp, Sty) was selected. These SNPs represent those markers that only occur on single fragments (Nsp, Sty). Enzyme-specific CQC values are calculated in the same manner described in the previous section, and are referred to as *Contrast QC Nsp Only*, and *Contrast QC Sty Only* in the GTC 2.1 QC report view. Note that for completeness, CQC is computed for a subset of 20K SNPs that occur on both Nsp and Sty markers and is referred to as *Contrast QC Nsp/Sty*. The selection of these three SNP classes was performed during algorithm development and now constitutes a constant set applied each time the algorithm is performed. The actual SNPs are located in the SNP Array 6.0 .QCC file found in the library folder for GTC 2.1.

A single-enzyme target prep failure is automatically flagged by asking whether the values of the Nsp-only and Sty-only CQC values differ by greater than two (Equation 5). If so, the algorithm will output a value of zero. When an enzyme failure occurs, the CQC column will have a value of zero and the CQC (random) column will have a different numerical value.

$$\text{Equation 5} \quad \text{If } |(CQC_StyOnly - CQC_NspOnly)| > 2, \text{ then } CQC = 0$$

Interpreting the Contrast QC Report in GTC 2.1

In the GTC 2.1 QC report view, five columns are displayed by default. The final CQC value is shown in the *Contrast QC* column of the GTC 2.1 QC report view, and is the value generated by Equation 5. Values generated by the procedure described in the Algorithms Basics section and Equation 6 are found in the *Contrast QC (random)*, *Contrast QC Nsp Only*, *Contrast QC Nsp/Sty* and *Contrast QC Sty Only* columns for the 10K random, 20K Nsp-only, 20K Nsp/Sty and 20K Sty-only SNPs, respectively.

$$\text{Equation 6} \quad \text{Contrast QC } \langle \text{SNP subset} \rangle = \min \left(\begin{array}{l} BB_Peak - BB_Valley, \\ AA_Peak - AA_Valley \end{array} \right)$$

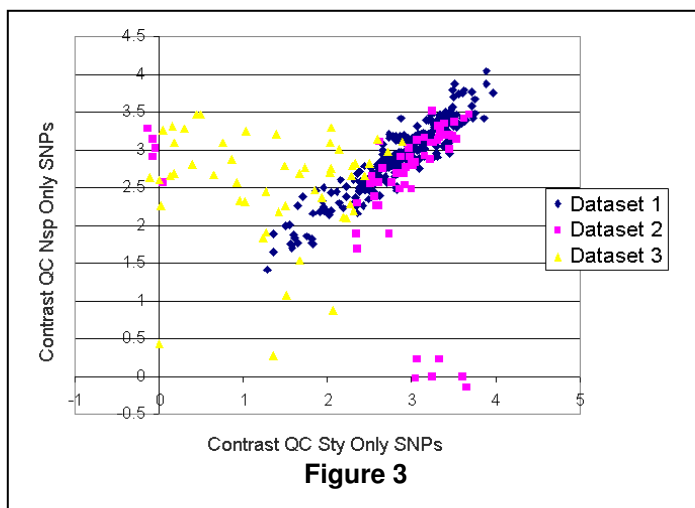
As will be discussed later in this paper, CQC values track well with Birdseed call rate and concordance. Low values are associated with a drop in these performance metrics. A CQC value of 0.4 has been chosen as the cutoff for performance (discussed below). Samples with a CQC ≥ 0.4 are considered passing. Usually, the CQC value will equal the CQC (random) value. The exception occurs when the CQC value is set to zero, when a large difference between the CQC Nsp-Only and CQC Sty-Only values causes the rule in Equation 5 to be applied. A large difference between these enzyme-specific CQC values is

likely due to a single-enzyme target prep failure. The difference between the enzyme-specific CQC values is more sensitive to this problem than the CQC (random) value alone. The rule in Equation 5 helps ensure that a likely single-enzyme target failure will be automatically flagged.

The CQC Nsp Only and CQC Sty Only are the CQC values for markers that only occur on single fragments (Nsp, Sty). Accordingly, these SNPs represent the most extreme in terms of performance. The majority of SNPs on the SNP Array 6.0 are present on multiple fragments and therefore would not be expected to completely fail when there is decreased performance of a single enzyme. With this in mind, altered performance of the assay due to reagent degradation, sample quality or experimental technique will be immediately visible when viewing the enzyme-specific CQC values. However, this has led to some confusion when interpreting the CQC report.

In some situations, there may be a passing CQC score; however, one enzyme may show a zero or negative value. This is not reason to fail the sample in question; rather, it simply indicates that one enzyme has functioned poorly. The genotyping algorithm may still be able to generate calls at high levels; however, this situation may have a slight effect on concordance. As part of every experiment, it is critical that laboratories maintain adequate tracking of potentially problematic samples for troubleshooting downstream. A simple flag on samples with low single-enzyme values may be useful when filtering data for use in association studies. Performance of the association analysis with and without the flagged sample may be helpful in identifying contributors to false-positive associations.

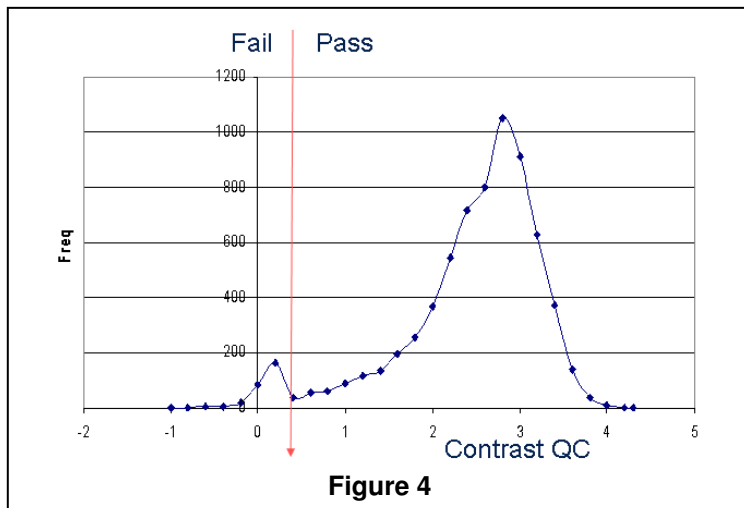
An additional QC step that is very useful in identifying single-enzyme performance issues requires the end user to plot the single-enzyme CQC Nsp versus CQC Sty values. For gross failures, the additional step in the CQC algorithm (Equation 5) easily flags samples by returning an overall CQC value of zero. However, subtle degradation of a single enzyme will result in movement of the plotted Nsp vs. Sty value away from the central cluster of data. In a perfectly performing data set, the Nsp vs. Sty plot will show a linear relationship of data points, along the diagonal.



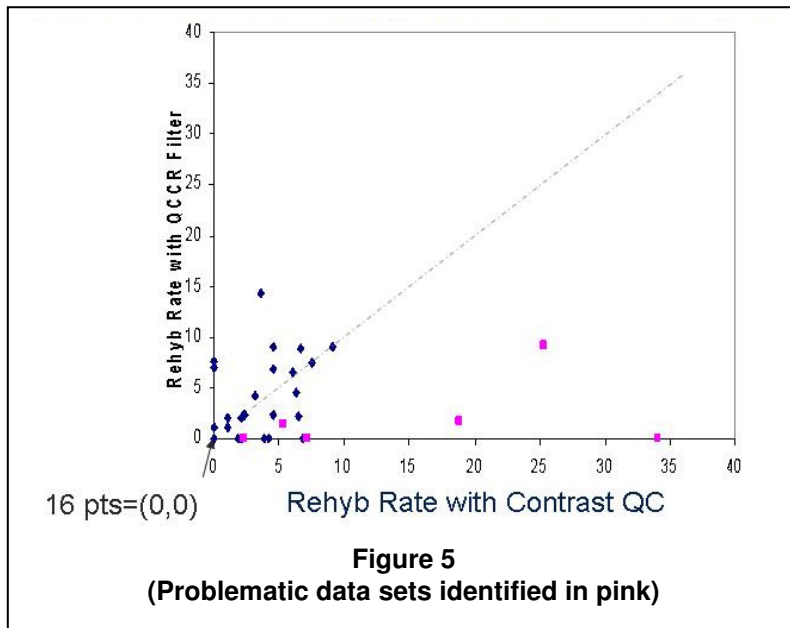
In Figure 3, an Nsp vs. Sty graph has been generated for three data sets. Data set 1 (blue) shows the relationship of a perfectly performing data set. Points that lie close to either axis (data sets 2 and 3) are, for the most part, identified by the algorithmic correction from Equation 5; however, there are clearly a number of points that fall away from the central cluster of high-quality data that indicate potential problematic samples. Again, this does not require an immediate failing of the sample, but rather a cautionary flag that there might be some slight degradation in overall Birdseed call rate or concordance due to the decrease in enzyme performance.

Contrast QC Passing Threshold and Re-hybridization Rates

A passing threshold for CQC has been set at 0.4. The evidence for this setting is based on a careful examination of more than 6,000 samples (discussed in the next section). This setting was selected to ensure the detection of samples that would potentially experience lower genotyping call rates using Birdseed. As seen in Figure 4, the majority of samples falls within a broad distribution centered near a CQC value of 3. A smaller peak just below 0.4 captures the majority of poorly performing samples. Although not perfect, the CQC algorithm detects a significant proportion of samples that will perform poorly when genotyped.



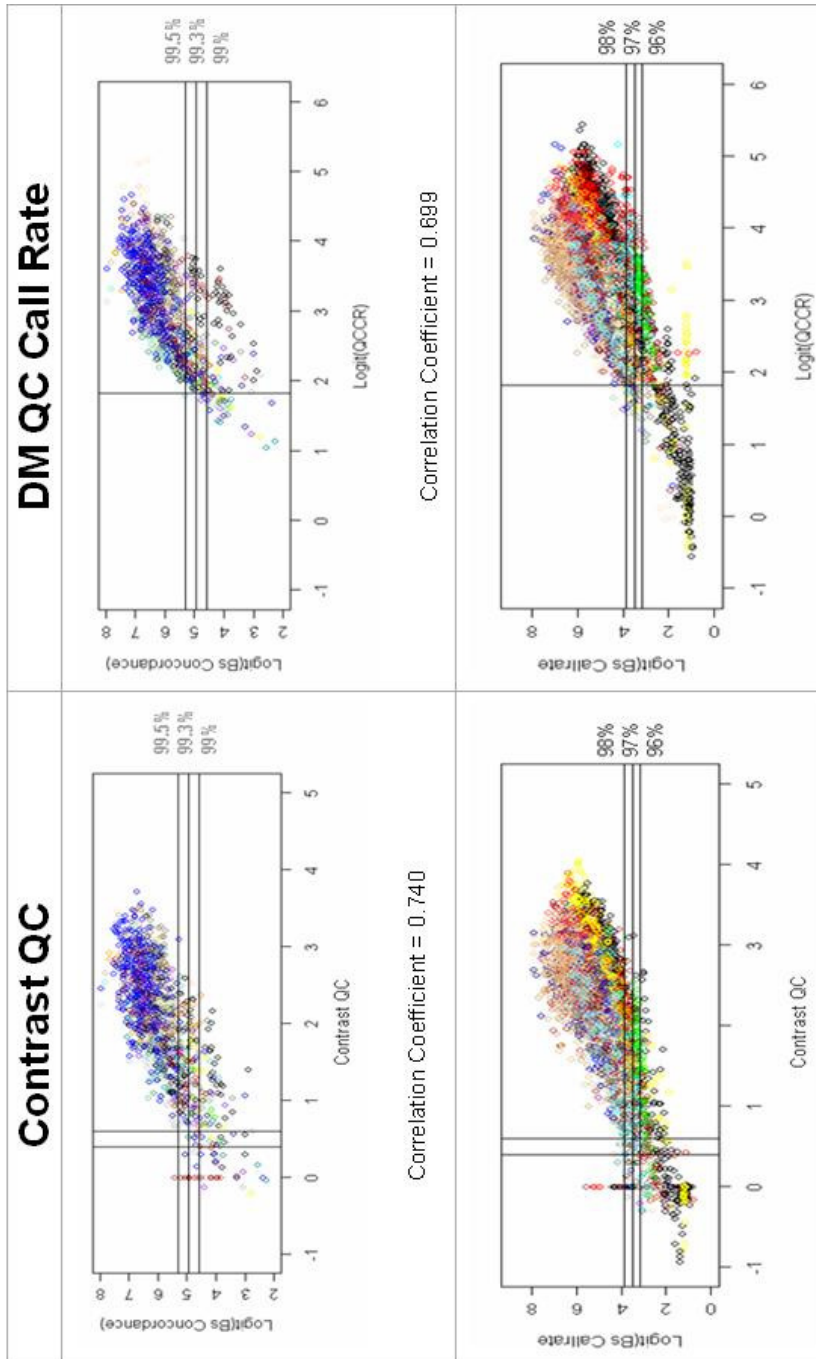
A careful study of “normal” data sets (samples exhibiting a typical range of target quality that are not associated with a systematic assay performance issue) has demonstrated a virtually identical re-hybridization rate (the percentage of samples failing the QC test) between DM QC call rate and Contrast QC. This data supports the ability of both algorithms to identify poorly performing samples at roughly the same level. Graphing of the CQC and DM QC re-hybridization rates for 50 data sets clearly demonstrates that both metrics perform similarly for most samples in identifying problems; however, CQC clearly differentiates a number of problem data sets (pink) that would not have been identified with DM QC call rate (Figure 5).



Correlation of QC Metrics with Birdseed Call Rate and Concordance

An effective QC algorithm must be capable of accurately predicting chip performance during subsequent genotyping steps. To better understand the ability of CQC and DM QC call rate to predict Birdseed call rate, more than 6,000 samples representing more than 50 data sets were analyzed. These data sets utilized both high- and poor-quality assay runs as well as HapMap and customer sample-generated data. Each data set was clustered separately using Birdseed v2 with a DM QC call rate cutoff > 86 percent. For each data set, samples failing the > 86 percent DM QC call rate filter were subsequently clustered with their entire data set in a separate run. The data was plotted to visualize the correlation between each QC metric and the subsequent data call rate and concordance. In Figure 6, overlaid graphs using different colors for each data set are shown along with the correlation coefficient over all samples in all data sets. In both cases of concordance and call rate, CQC was more tightly correlated (0.74 and 0.76) than that observed with DM QC call rate (0.70 and 0.68). In routine analysis of customer-generated data, the correlation coefficient between CQC and Birdseed v2 call rate in single experiments frequently exceeds 0.90.

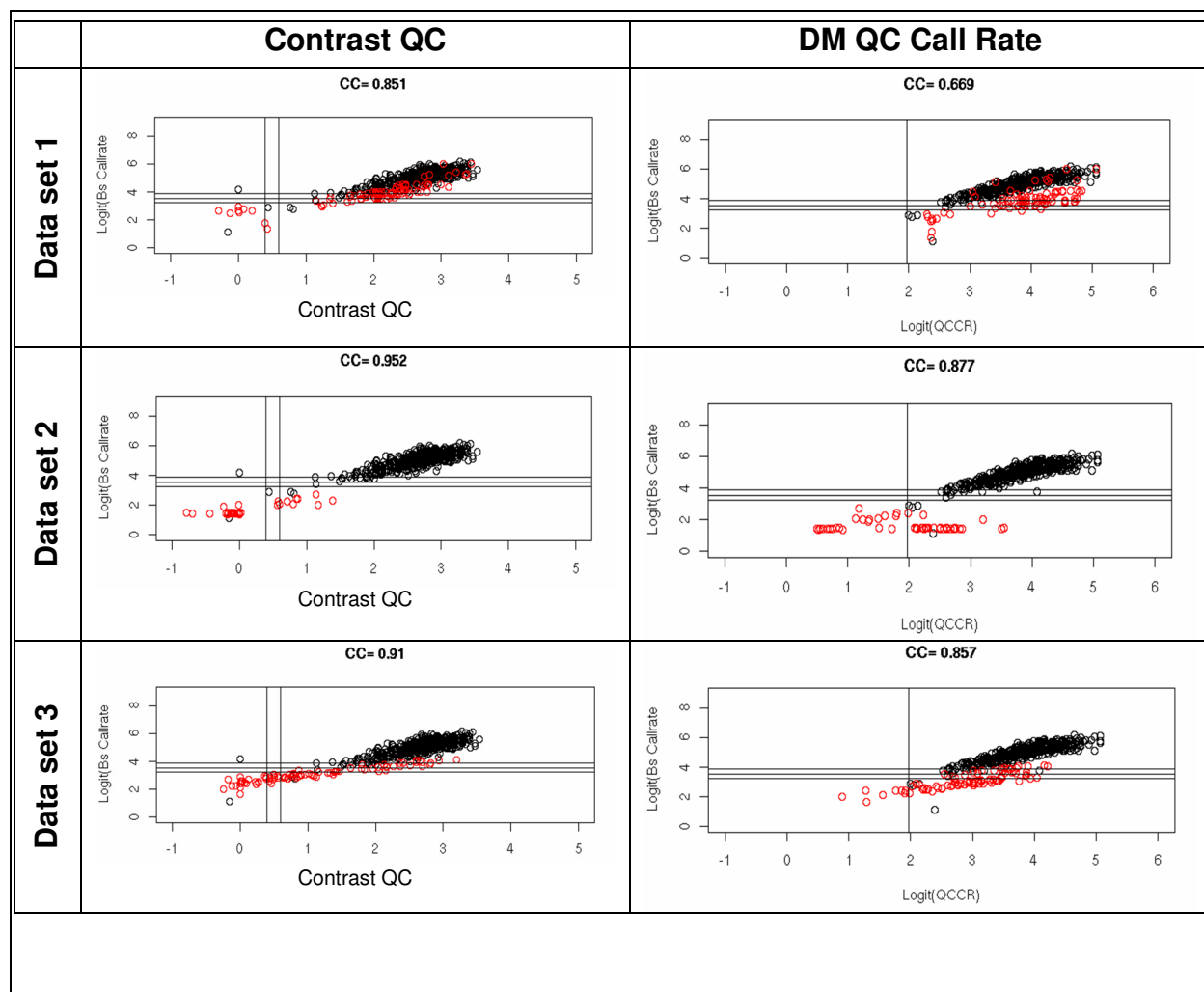
Figure 6: Call rates (DM QC and Birdseed) are plotted on the logit scale where $\text{Logit}(x) = \ln(x/(100-x))$



How does CQC Improve the Detection of Problem Data Sets?

One of the shortcomings associated with the DM QC call rate metric was its occasional failure to pick up certain problematic data sets at the QC step. Once Birdseed genotyping was performed, these problem data sets were easily identified. How then, does CQC perform in its ability to identify these problem data sets?

A series of data sets generated both internally and externally were examined for their performance using both CQC and DM QC call rate in association with their Birdseed v2 call rate. For the majority of samples and data sets, CQC and DM QC call rate track Birdseed v2 call rate with equivalent sensitivity. But for a few problem samples and data sets, DM QC call rate has decreased sensitivity to the magnitude of the drop in Birdseed call rate. In Figure 7, three of such data sets, each composed of multiple sample plates, are presented as red points in graphs of Birdseed v2 call rate versus QC metric.



The black points are for samples in a normal data set, where CQC and DM QC call rate track Birdseed v2 call rate with equivalent good sensitivity. In data set 1, all samples passed the 86% DM QC call rate cutoff and were subsequently genotyped. However, data set 1 points (red) clearly track separately from the normal sample points (black), reducing the correlation coefficient for DM QC call rate to 0.669. When CQC is applied, the correlation coefficient between CQC and Birdseed v2 is very high (0.851).

Furthermore, the 0.4 cutoff eliminates almost all samples with Birdseed call rates less than 96 percent. A careful examination of the remaining two data sets reveals high correlation coefficients between Birdseed v2 and CQC (all > 0.91), as well as significant filtering of samples with poor Birdseed v2 performance.

Flagging Problematic Data Sets and Best Practices Recommendations

After completion of a single batch using the SNP Array 6.0 assay, all samples should be screened using the CQC algorithm to assess individual performance and to flag those samples that should not be forwarded to Birdseed genotyping. A cutoff of 0.4 is recommended because the minimum performance for a sample to pass QC. Samples with a CQC less than 0.4 should NOT be genotyped using the Birdseed algorithm.

An additional QC measure is recommended to identify batches with more subtle systematic problems. This additional step requires the user to calculate the mean of the samples passing the CQC cutoff of 0.4. In optimally performing data sets, this mean should be greater than 1.7, and the percentage of samples passing the CQC 0.4 threshold should be ≥ 90 percent. If either condition is not met, the batch should be flagged as potentially problematic. Internal analysis of problem data sets with low Birdseed performance supports the use of the 1.7 mean as an effective method of identifying those data sets with less than desired performance when genotyped.

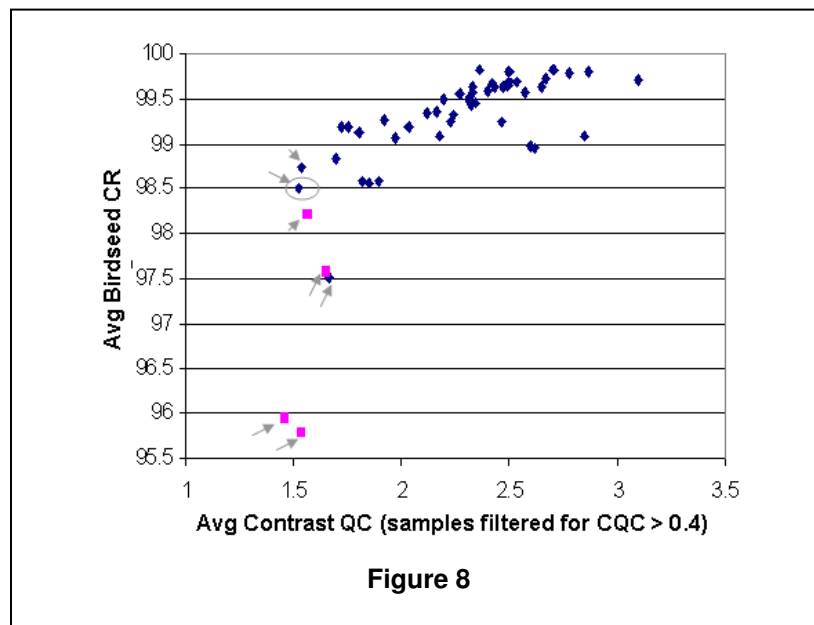


Figure 8

As shown in Figure 8, known problematic data sets (in pink) all have CQC means, after 0.4 cutoff filtering, of less than 1.7. It is clear that their subsequent performance in Birdseed genotyping is significantly affected. In addition, three additional data sets were identified as problematic (blue diamonds identified by arrowheads) using this filtering method. This batch mean-based filtering should serve as a method to flag problematic data sets. This is not a reason to immediately exclude the samples from further analysis, but rather, should serve as a warning that if spurious results are obtained during association analysis, the user may wish to exclude the problematic plate and rerun the association study.