A Comparative Assessment of Performance between HT and Cartridge IVT Expression Arrays

March 2007

I. Summary

This study provides data and analyses describing the performance of HT IVT plate arrays with emphasis on the magnitude and effects of potential sources of systematic variation as well as the effects of image anomalies.

The analyses conclude that although systematic variation is detectable (e.g., plate, row and column effects), the magnitude of these effects is minimal as compared to biological effects and does not affect the ability of HT IVT plates to meet their performance specifications.

In addition, the presence of small anomalies in array images, probably from bubbles in the stain and/or hybridization stages of array processing and debris in solutions during scanning, does not affect data quality. A reliable quality measure of array expression data reproducibility is described and used in these studies, “Mean Absolute Relative Log Expression” (MA(RLE)). A workflow is proposed for HT IVT array processing and analysis based on the use of this metric.

II. Introduction

The Affymetrix GeneChip® HT Array Plate System facilitates large-scale studies, simplifying the management of processing multiple samples and microarrays in parallel. The HT System consists of the GeneChip® Array Station (GCAS), used for target preparation and array processing, and the HT Array Plate Scanner and its associated software, used for processing and scanning the plates as well as for downstream data analysis. The plates are constructed to be spatially compatible with conventional 96-well plate formats and liquid-handling equipment.

The GeneChip® HT Human Genome U133 Array Plate Set, the GeneChip® HT Mouse Genome 430 Array Plate Set and the GeneChip® HT Rat Focus Array Plates have content derived from the respective cartridge arrays in a plate format. Both HT System and IVT cartridge arrays exceed the performance specifications established for the Affymetrix IVT expression assay. Although the probe sets on the HT plate arrays are identical in content to their counterparts from the appropriate cartridge arrays, there are potential differences between the array formats that need to be considered.

One major difference between cartridge arrays and plates is that the cartridge array flow cell is a closed system, while the array plates are open to the environment. The cartridge

arrays are processed on the GeneChip® Fluidics Station, with buffers and stains injected into the flow cell through septa. The HT array plates are manipulated on the GCAS and transferred to a series of trays for the hybridization, wash and stain steps. These differences in array processing result in different sources of variability for cartridge and plate arrays. Cartridge arrays are processed four at a time on a GeneChip Fluidics Station. While each set of four is washed with the same buffers, they are processed on independent fluidics modules with separate temperature controllers and tubes of stain. Each plate array is batch processed in one of two different formats, either 24 or 96 samples at a time, in the same wash and stain trays on one GCAS and scanned on one HT Array Plate Scanner. The batch effects for cartridge and HT arrays are different due to the differences in the respective systems. Sources of variability unique to the HT platform include plate, format (e.g., 24-array plate versus 96-array plate) and row-column effects.

Additionally, because of the differences in array processing, users may occasionally observe artifacts on the array images. Most of the artifacts result from array processing and are not manufacturing defects. The array designs and default analysis algorithms for both cartridge and plate arrays add robustness to the entire system. Probes from individual probe sets are distributed across the entire array surface to minimize the impact from image artifacts on any one probe set. Similarly, the default multi-chip analysis algorithm (RMA) adds robustness during data analysis. The effects of typical image artifacts are evaluated in this study.

This study examines the issues described above and answers some of the common questions regarding HT IVT plate performance. Do plates perform as well as cartridges? Are there systematic effects and, if so, will they confound results? What is the impact of major and minor image anomalies? What quality control (QC) metric can be used with the arrays?

In addressing these questions, we will introduce a standard statistical variability metric, the mean absolute relative log expression, or MA(RLE), to characterize the magnitude of system variation.

III. Materials and Methods

The following series of experiments were designed to characterize performance with respect to systematic sources of variability. In essence, common pools of HeLa cell-derived target with additional spikes of transcripts at known concentrations were hybridized to cartridge and plate arrays. This provided a data set to measure systematic variation as well as absolute performance. By pooling and then dividing the target, target preparation variation was removed as a variable.

The following factors as sources of variation were measured: plate, format, row and column. While running a single, pooled hybridization sample does not directly reflect the typical experimental designs that users of the system would employ, it is highly sensitive to measuring the targeted sources of system variability.

Target Preparation: The Automated Target Preparation protocol (TP_0001) was used to prepare labeled cRNA hybridization target according to the *GeneChip® Expression Analysis Technical Manual for Cartridge Arrays Using the GeneChip® Array Station* (P/N 702064) from total RNA from the HeLa human cell line. The remaining target preparation steps were carried out manually. cRNA was pooled, fragmented, then split to two aliquots to create hybridization cocktail for plate or cartridge arrays. Twenty-eight pre-labeled spike transcripts derived from human cDNA clones were added in a mini-Latin square configuration at the following concentrations: 0, 0.75, 1.5 or 3.0 pM (See Table A1). Four spike pools containing spikes at various concentrations were assigned to arrays at different well positions to create the Latin square configuration. Table A1 describes the spike pool constitutions and Table A2 shows the pool to plate well assignment to create the Latin square configuration on the 24-array and 96-array plates.

Probeset_ID	Pool 1	Pool 2	Pool 3	Pool 4
203508_at	0	3.0	1.5	0.75
200665_s_at	0	3.0	1.5	0.75
204563_at	0	3.0	1.5	0.75
207641_at	0	3.0	1.5	0.75
204513_s_at	0	3.0	1.5	0.75
207540_s_at	0	3.0	1.5	0.75
204959_at	0	3.0	1.5	0.75
207655_s_at	0.75	0	3.0	1.5
205291_at	0.75	0	3.0	1.5
203471_s_at	0.75	0	3.0	1.5
203921_at	0.75	0	3.0	1.5
209795_at	0.75	0	3.0	1.5
207777_s_at	0.75	0	3.0	1.5
204912_at	0.75	0	3.0	1.5
205569_at	1.5	0.75	0	3.0
203927_at	1.5	0.75	0	3.0
207160_at	1.5	0.75	0	3.0
205692_s_at	1.5	0.75	0	3.0
212827_at	1.5	0.75	0	3.0
212886_at	1.5	0.75	0	3.0
204951_at	1.5	0.75	0	3.0
209606_at	3.0	1.5	0.75	0
205267_at	3.0	1.5	0.75	0
207968_s_at	3.0	1.5	0.75	0
210895_s_at	3.0	1.5	0.75	0
205903_s_at	3.0	1.5	0.75	0
206060_s_at	3.0	1.5	0.75	0
205790_at	3.0	1.5	0.75	0

Table A1: Spike pool description. Concentration in pM. Spikes at 1.5 pM are approximately equivalent to 1 transcript per 100,000.

24-array plate

	5	7	9
A	1	3	4
B	4	1	2
C	2	4	1
D	1	4	3
E	3	2	4
F	2	3	1
G	3	1	2
H	4	2	3

96-array plate

	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	4	2	1	3	4	2	1	3	4	2
B	4	1	2	3	4	1	2	3	4	1	2	3
C	2	4	1	3	2	4	1	3	2	4	1	3
D	1	4	3	2	1	4	3	2	1	4	3	2
E	3	2	4	1	3	2	4	1	3	2	4	1
F	2	3	1	4	2	3	1	4	2	3	1	4
G	3	1	2	4	3	1	2	4	3	1	2	4
H	4	2	3	1	4	2	3	1	4	2	3	1

550001_4025828_241

	5	7	9
A	1	4	1
B	4	2	4
C	2	1	2
D	1	3	1
E	3	4	3
F	2	1	2
G	3	2	3
H	4	3	4

Table A2: Spike pool to well assignment for 24-array and 96-array plates. Plate 550001_4025828_241 had a different spike pool to well assignment, shown at bottom.

The design used in the 24-array plates was replicated on three sets of 24 cartridges selected from three different array manufacturing lots. To make the spike-in assessment comparable between 24-array and 96-array plates, we selected a subset of three columns on the 96-array plates in which the spike-in configuration from the 24-array plates was replicated.

The three lots of cartridges were processed in one experiment with a single hybridization target preparation. The plates were run in a series of three experiments spanning six months. A pool of hybridization target was prepared separately for each of the three plate experiments. Figure 1 shows the response of the 28 individual spikes over the concentration range described. One probe set, 207641_at, showed poor response in both cartridge and plate experiments and was left out of the subsequent analyses. The presence of this probe set would not have substantially changed the results.

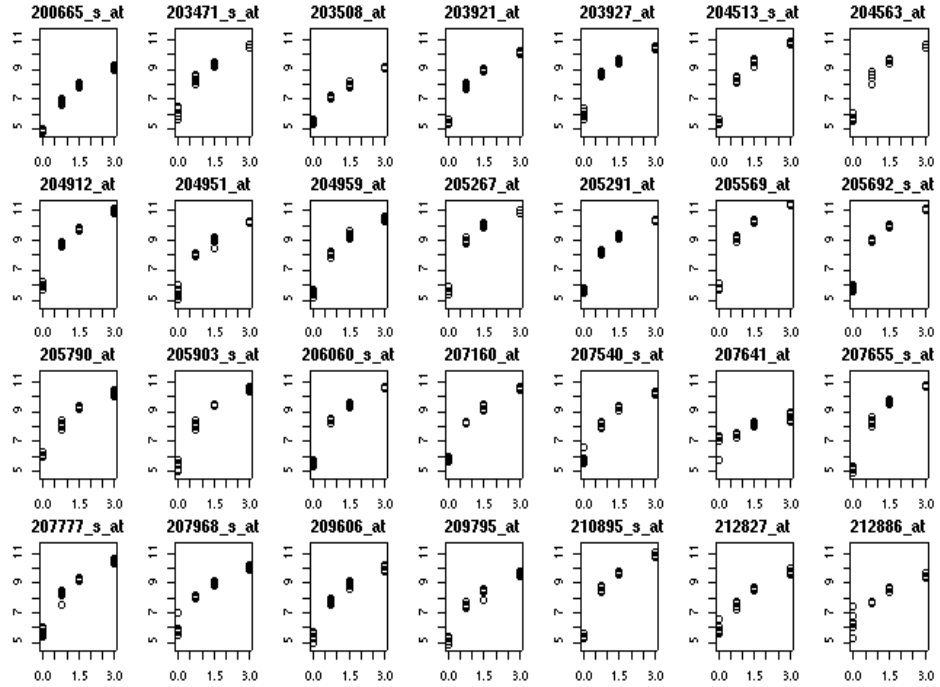


Figure 1: Plots of RMA probe set summaries (log base 2) on the y axis versus spike-in concentration on the x axis for probe sets corresponding to the 28 pre-labeled spike transcripts that were added at various concentrations to the hybridization cocktails. Because of the poor response to changes in concentration on both cartridge and plate arrays for the spike corresponding to the target of probe set 207641_at, we excluded this spike from the performance assessment figures and summaries. These plots were created from data pooled from arrays from multiple 24-array plates. N = 6 for each probe set versus concentration graph.

Plate Processing: 24-array and 96-array HT HG-U133A plates were hybridized, washed and stained using protocol WS_0001 according to the *GeneChip® Expression Analysis Technical Manual for HT Array Plates Using the GeneChip® Array Station* (P/N 702063). The GCAS instrument was upgraded to Maestro Software (v. 4.0). Samples were randomized across the wells. Plates from three separate experiments were used for this study. The number and formats of plates run in the three experiments are as follows: 1) six 24-array and six 96-array plates; 2) four 24-array plates; 3) two 24-array and two 96- array plates, for a total of twelve 24-array and eight 96-array plates.

Data Analysis: All plates were scanned on the HT Scanner using HT Image Reader 1.0 software to acquire ".dat" file images and generate ".cel" files. Quantile normalization and probe set signal summarization were computed according to the RMA procedure as implemented in the Affymetrix Power Tools (APT) software. In addition to APT, Affymetrix Expression Console™ software produces equivalent RMA probe set signal summarizations, and reports MA(RLE) for each array. 'R' was used for routine statistical analysis and visualization.

IV. Data Quality and Array Performance Assessment

MA Plots

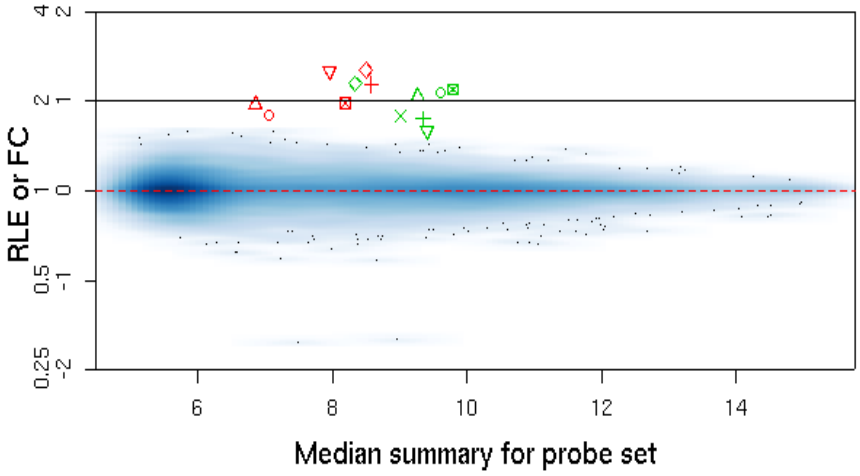
The MA plot provides a comprehensive visual assessment of array expression data quality in the context of assessing the platform's ability to detect a few differentially expressed transcripts among the vast majority of transcripts present at invariant concentrations (1). The MA plot is used to compare expression values from two sources, typically two arrays. On the y axis, the difference in estimated expression for each probe set, $\log_2(E_1)-\log_2(E_2)$, is plotted against an estimate of the expected expression (e.g., the average of the log estimated expression, $(\log_2(E_1)+\log_2(E_2))/2$). Log transformation of the expression values brings their distribution closer to normality where the statistics are simpler. The values plotted on the y axis can also be thought of as a log ratio since $\log_2(E_1/E_2) = \log_2(E_1)-\log_2(E_2)$. We refer to this quantity as a Relative Log Expression (RLE). We use summaries of the RLE distribution for probe sets targeting transcripts that do not change concentration across arrays to assess data reproducibility as explained below. The RLE values for probe sets corresponding to transcripts spiked at different concentrations across arrays provide an indication of the array's ability to detect differentially expressed transcripts, or sensitivity at the spike-in concentration fold changes.

The MA plot as described above is frequently used to assess reproducibility and sensitivity based on a pair of arrays. To assess the performance of a larger set of arrays, one can compare each array to expected expression derived from a common reference, or baseline, array. The reference array could be an actual array, or as in the approach adopted in this evaluation, a virtual array with expression values for each probe set computed from a summary of the entire set of arrays. For example, each probe set could be plotted in the MA plot with the Y-value as the relative expression value for a particular array $RLE = \log_2(E_c)-B$, and with the X-value B, where the reference or baseline log expression, B, is computed as the median $\log_2(E)$ value over a set of arrays. E_c is the estimated expression value for a specific array.

In this study, the set of arrays used to compute the baseline expression for each spike-in probe set is the subset where the transcript was spiked at half of the concentration. For example, for any of the 28 probe sets corresponding to a transcript spiked at a concentration of 3.0 pM or 1.5 pM, the arrays where the corresponding transcript was spiked at 1.5 pM or 0.75 pM, respectively are used to obtain the baseline value. In this manner, we would expect to see an RLE of 1 (log base 2) for the spiked target and an RLE of 0 for all other unspiked probe sets (assuming no variation or bias). Except for a few rare cases where we have incomplete data (seven out of 1,056 arrays had scanning failures), six values go into the computation of the baseline values. For comparability purposes, we also use six arrays selected at random from the array set under investigation to get the baseline expression values for the non-spike probe sets. One could also use all of the arrays in the array set under analysis as is done in APT software and in the RLE summaries discussed below. Differences are insignificant and in the latter case we lose the clean interpretation of the MA plot as giving us a sense of the reproducibility (looking

at non-changing probe sets) and sensitivity (looking at RLE for two-fold change spikes) of probe set summaries when comparing one array to a pooled summary computed from six arrays.

Figure 2 shows an example MA plot. Fold Change values ($FC=2^{RLE}$) are also indicated on the y axis. The blue density cloud corresponds to the probe sets that, in an ideal experiment, would not change since they are measuring the same molar amount of transcripts in all microarrays. We'd also see a straight horizontal line about the zero fold change line. In practice, however, variation results in the observed blue density cloud. Some individual points that occur on the outskirts of the cloud are plotted. Points corresponding to transcripts spiked at two-fold differences in concentrations are indicated in the colored symbols. See Figure 3 for spike-in key. Some summaries of the absolute value of RLE distribution are listed as a table at the foot of the Figure 2.



Stat:	mean	50%	99%	99.9%	99.99%
|RLE|:	0.1	0.08	0.42	0.58	0.79
FC:	1.07	1.05	1.34	1.49	1.72

Figure 2: Example MA plot on which RLE = $\log_2(E_c)$- B is plotted on the y axis and the baseline value on the x axis. Fold Change values = 2^{RLE} are also indicated on the y axis. The blue density cloud represents all points corresponding to the non-changing probe sets. Some individual points that occur on the outskirts of the cloud are plotted. Points corresponding to transcripts spiked at two-fold differences in concentrations are indicated in the colored symbols. See Figure 3 for key. MA plots were generated using the smoothScatter function of the geneplotter R package, version 1.6.1.

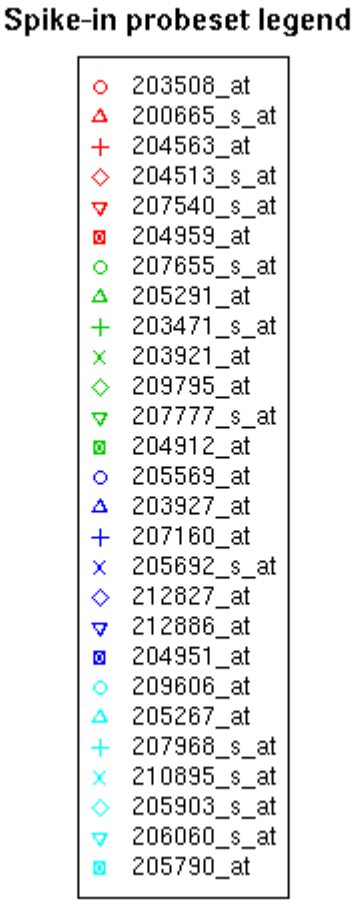


Figure 3: Legend for 27 spike-in probe sets used in this evaluation.

As described above, the RLE provides a metric for the performance of a given probe set on a given array relative to a set of arrays. The MA(RLE) for a given array is the mean of all of the absolute RLE values on that array. This is a performance metric which summarizes the reproducibility of signal of an entire array and includes both bias and variation. It should be noted the RLE and MA(RLE) values are dependent on the specific experimental data included in the analysis. Although CV values have frequently been used to assess array performance, the MA(RLE) metric described above has the added advantage of indicating performance for each specific array, allowing for its use as a quality metric in an experimental workflow.

Relative Log Expression Summaries

Although we stress expression data reproducibility in our assessment of array performance, it should be obvious that reproducibility is directly related to sensitivity – given that the expression values are sensitive to changes in target concentration, the greater the reproducibility of estimated expression values for probe sets targeting RNA transcripts that do not change concentration level between arrays, the greater will be the array's ability to detect the transcripts that do change in concentration. The sensitivity of the various platforms to changes in target transcript concentration was verified with a set of spike-in transcripts. Referring back to Figure 1 showing the relationship between

expression values and spike concentration, we see that after excluding probe set 207641_at, the average $\log_2$ expression value increases with spike concentration level in all probe sets. This figure is typical of all of the platforms being evaluated here – cartridges, 24-array plates and 96-array plates.

We use the MA(RLE) value to summarize the RLE distribution. This is only one of many possible ways to summarize the distribution. One may use tail quantiles of the RLE distribution expressed on the linear or fold change scale to estimate levels of reliable detectability. For example, the 99.9 percentile of the fold change distribution of values computed from technical replicates can be interpreted as the fold change that could be detected with a 1 in 1,000 false positive rate. Figure 4 displays the relationship between MA(RLE) values and detectable fold change at a 1 in 1,000 false positive rate for a set of arrays. From this graph, we see that on an array with a MA(RLE) value of 0.10, one could detect a 1.5-fold change at 1 in 1,000 false positive rate. Note that this figure is only meant as an illustration of the relationship between the MA(RLE) values and tail quantiles. The values for the x and y scale in this plot are characteristic of the MA(RLE) distribution that we observe when hybridizing aliquots of prepared HeLa targets to a set of arrays. We emphasize that the distribution of RLE values will be dependent on the source of RNA and amount of biological variability present in the samples being analyzed. One should also note that although tail quantiles may be appealing for interpretability, they are more statistically variable than estimates of the center of the distribution.

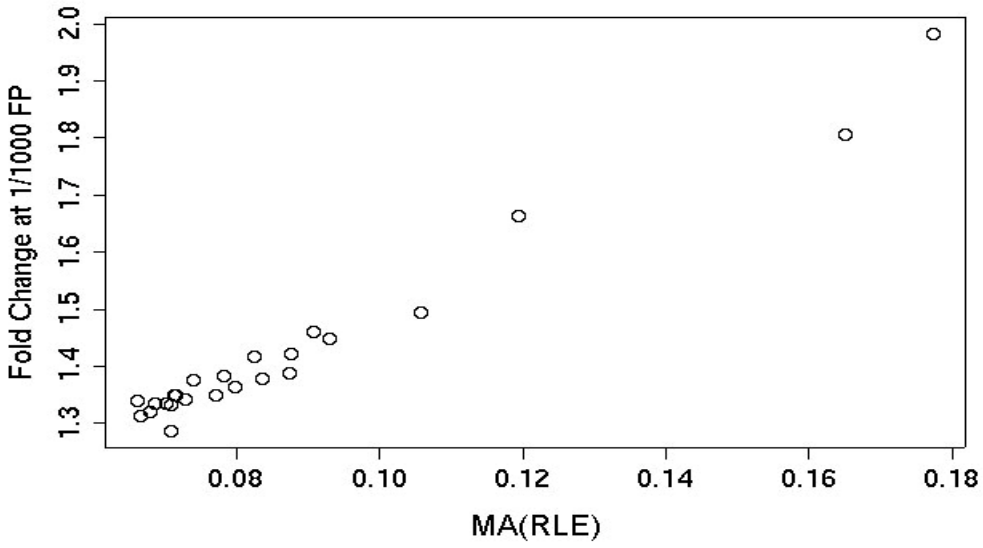


Figure 4: Relationship between MA(RLE) and Fold Change at a 1 in 1,000 false positive rate. MA(RLE) values are shown for a single 24-array plate, 550001_4023845_245. MA(RLE) is represented on the x axis and Fold Change at 1/1000 false positive rate is represented on the y axis. These values are specific to technical replicates of HeLa sample hybridized to HT HG-U133A Arrays.

Receiver Operator Characteristic (ROC) Curves
ROC curves are used to compare assay performance and illustrate the reciprocal relationship between sensitivity and specificity. To evaluate the sensitivity to detect signal differences for two-fold changes in spike probe sets, the accuracy (y axis) is plotted against the false change rate for all probe sets with unchanged concentration (x axis). Two comparisons were generated for this study: spike concentrations at 3.0 pM were compared to spikes at 1.5 pM, and spike concentrations at 1.5 pM were compared to spikes at 0.75 pM. The area under the curve (auROC) is calculated to indicate the measured sensitivity and specificity. A value of 100 percent for auROC indicates 100 percent sensitivity and 100 percent specificity (i.e., 0 percent false positive).

V. Results

Variation Study
Within-lot and Plate Variation
The first question that we address is how the reproducibility of expression values extracted from arrays processed on a single 24-array or 96-array HT HG-U133A plate compares with the reproducibility of expression values extracted from arrays processed on HG-U133A 2.0 cartridges selected from a single array manufacturing lot. This question addresses the comparability of platforms on small scale studies. To address this question we analyzed data for the following array sets: three sets of 24 cartridge arrays selected from three HG-U133A 2.0 cartridge lots, 12 sets of 24 arrays from 24-array HT HG-U133A plates, and eight sets of 96 arrays from 96-array HT HG-U133A plates. The MA(RLE) values for each array set are summarized by box plots in Figure 5. The majority of the MA(RLE) values are below 0.l2 (mean reproducibility to within 8.7 percent of baseline value), with a median per cartridge lot or plate typically less than or equal to 0.08 (mean reproducibility to within 5.7 percent of baseline value). The figure illustrates that the within-plate reproducibility is comparable to within-lot reproducibility for cartridge arrays. We expect that while reproducibility across platforms will remain comparable when other sources of RNA are used, the level of reproducibility will depend on both the RNA source and biological variability in the samples being analyzed.

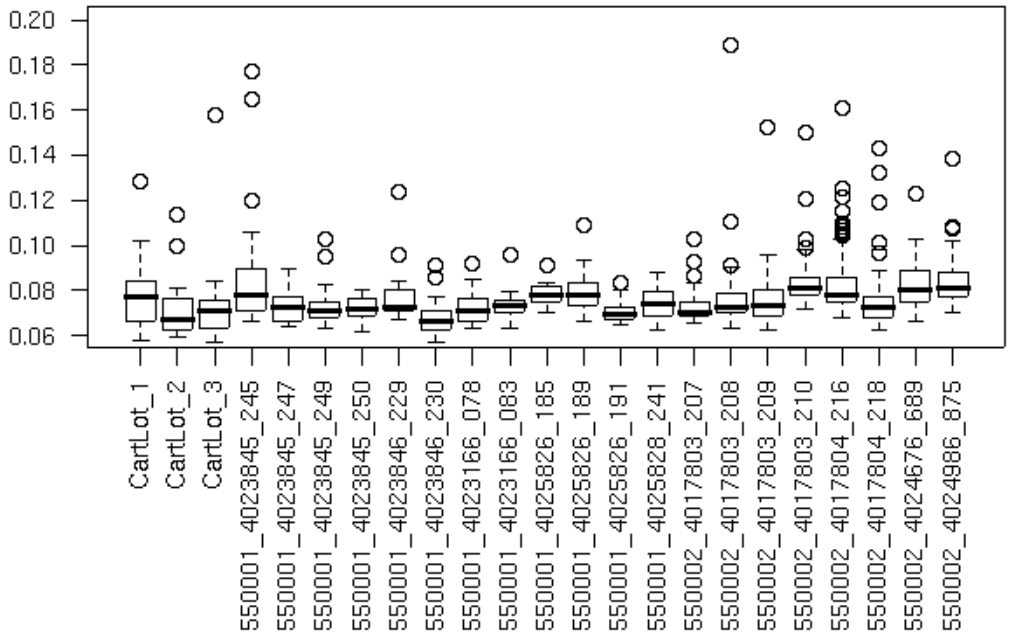


Figure 5: Box plot of array MA(RLE) values corresponding to different chip sets. Each box represents the MA(RLE) values for chips from a cartridge lot (CartLot_1, CartLot_2 and CartLot_3) or plate. The upper and lower boundaries of the box represent the 25th and 75th percentile and the line in the box represents the median value. Plates beginning with the numbers 550001 and 550002 are 24-array and 96-array plates, respectively.

Between-lot and Plate Variation

The next question that we address is how the reproducibility of expression values extracted from arrays processed on multiple 24-array or 96-array HT HG-U133A plates compares with the reproducibility of expression values extracted from arrays processed on HG-U133A 2.0 cartridges selected from multiple array manufacturing lots. This question addresses the comparability of platforms on large-scale studies. Because the cartridges were run in one experiment with one pooled hybridization sample, for this comparison we selected six 24-array plates and six 96-array plates that were also run in one experiment with one pooled hybridization target for comparison. To address the question, we re-analyzed the array data, pooling all cartridge arrays coming from the three different lots into one set, all of the arrays from the six 24-array plates into a second set, all of the arrays from the six 96-array plates into a third set, and all of the arrays coming from either 24-array or 96-array plates (six of each) into a fourth array set. The MA(RLE) values for each array set are summarized by box plots in Figure 6. While the reproducibility of expression values for both platforms is still high, when expression values are compared across plates, higher MA(RLE) values show that reproducibility is slightly lower on the HT System than with cartridge arrays.

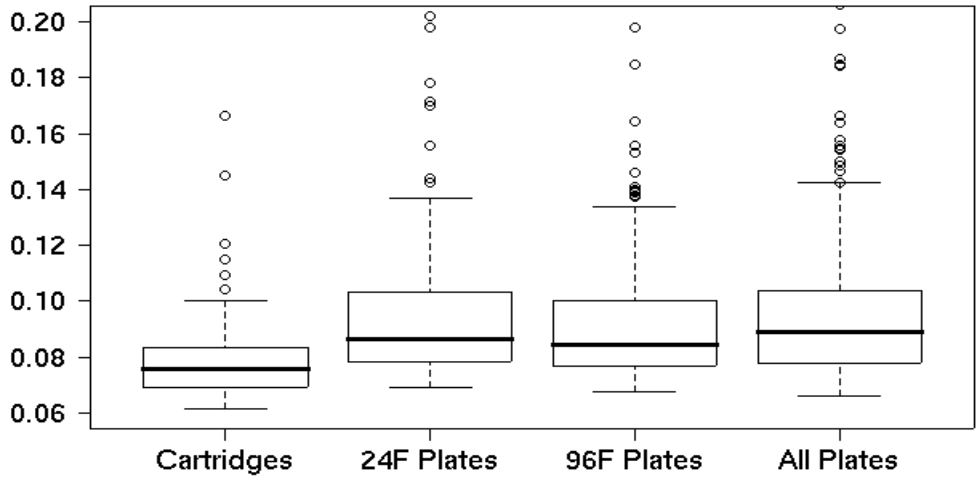


Figure 6: Box plot of array MA(RLE) values for multiple cartridge lots or multiple plates. Because the cartridges were run in one experiment with one hybridization target preparation, we used only plates from one of the three plate experiments with one hybridization target preparation as a comparison. Each box represents the MA(RLE) values for all cartridge lots (Cartridges), six 24-array plates (24F Plates), six 96-array plates (96F Plates) and the 24-array and 96-array plates analyzed together. The upper and lower boundaries of the box represent the 25th and 75th percentile and the line in the box represents the median value. As seen in Figure 7, even arrays with MA(RLE) values of 0.18 have near perfect auROC of greater than 99.00, indicating a very high level of sensitivity and specificity to detect two-fold changes at low concentrations.

While the reproducibility of expression values when pooling arrays across plates is lower than within-plate reproducibility, the platform still very cleanly detects the transcripts spiked at two-fold changes in concentration. To demonstrate this point, for each of the 24-array plate well locations, we selected arrays at random from all of the 24-array plate arrays at the corresponding well location to create a virtual plate which mimics a 24-array plate in well composition and therefore replicates the Latin square design for the spike-in transcripts described in Tables A1 and A2. Figure 7 shows MA plots for arrays in this virtual 24-array plate. The expression values were computed from the pooled RMA model fitted to all 24-array plate arrays but MA plots were drawn using data from the specific 24 arrays making up the virtual plate, as described in the **Data Quality and Array Performance Assessment** section above. The title line for each MA plot is Well (position in plate by row-column number) / MA(RLE) / auROC, where auROC is the area under the ROC constructed by thresholding on fold change or RLE to discriminate between the spike-ins and the non-spikes in each MA plot. The area under the ROC exceeds 99.9 (out of a perfect 100) in 22 of the 24 arrays analyzed in this set. Results

from an analysis of the 96-array plate arrays are similar (not shown here). To contrast the effects of low and high MA(RLE), Figure 8 displays MA plots for two arrays from the pooled 24-array array set representing arrays at opposite ends of the reproducibility range.

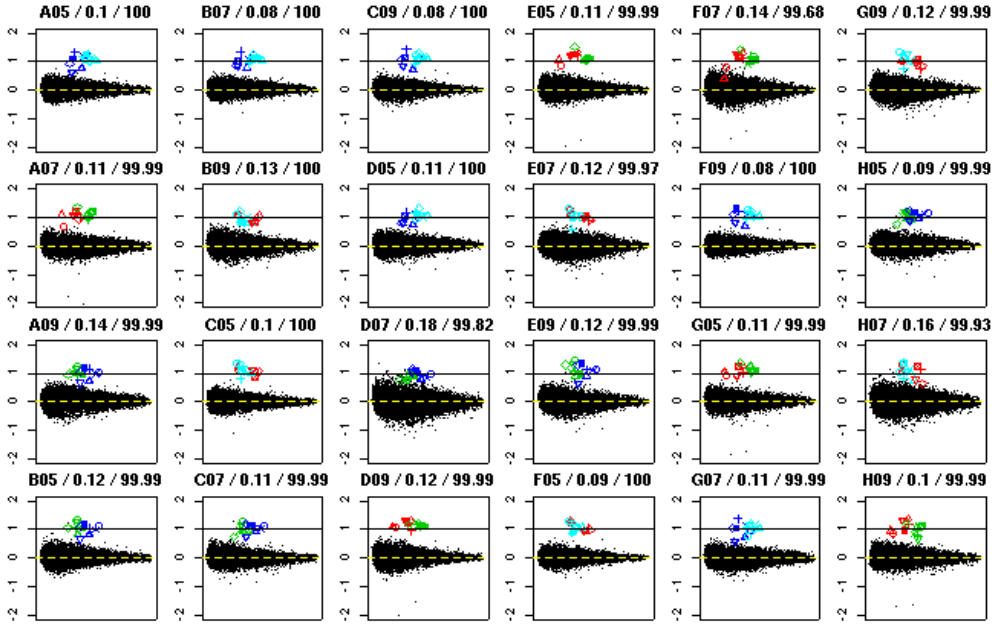
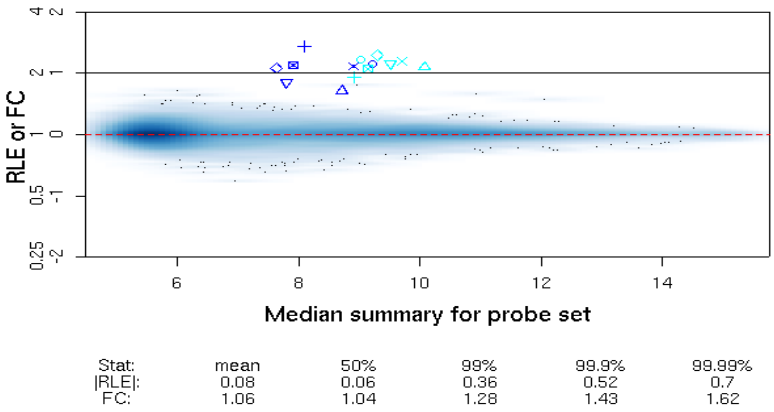


Figure 7: MA plots for chips from the pooled set of 11 of the 12 24-format plates analyzed in this study. The sample layout in one 24-array plate (550001_4025828_241) was different from all of the others, preventing a direct comparison of samples in the same row and column position for that plate. For each well location, a chip was selected at random from all chips from the corresponding well location. Expression values were computed from the pooled model fitted to all 24-format plate chips. MA plots were drawn using data from the 24 chips making up the virtual plates. Title for each MA plot is Well / MA(RLE) / auROC.



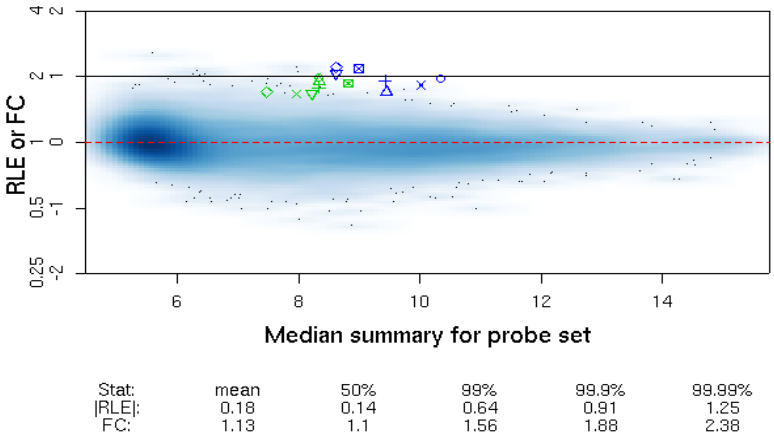


Figure 8: MA plots for two chips from the pooled 24-format chip set representing chips at opposite ends of the reproducibility range. Top panel is for array C09 from a pooled 24-array virtual plate. Bottom panel is array D07.

Figure 9 shows box plots of RLE values computed from probe set values corresponding to two-fold changes in spike concentrations. Recall that ideally we expect RLE of two-fold change would be around $\log_2(2)=1$. The first three box plots summarize the RLE of spike-ins from cartridges analyzed within array lot. The “550001_” box plots summarize RLE values for spike-in probe sets on 24-array plates. The “550002_” box plots summarize RLE values for spike-in probe sets from columns 03, 06 and 09 of 96-array plates. These three columns were selected to replicate the Latin square design in the 24-array plates. The 24F_Pooled and the 96F_Pooled box plots summarize RLE values computed from values for spike-in probe sets from the virtual 24-array and 96-array

plates, respectively. The majority of RLE values for true two-fold changes as measured by cartridges and plates lies between 1.0 and 1.2 (fold change of 2.0-2.3).

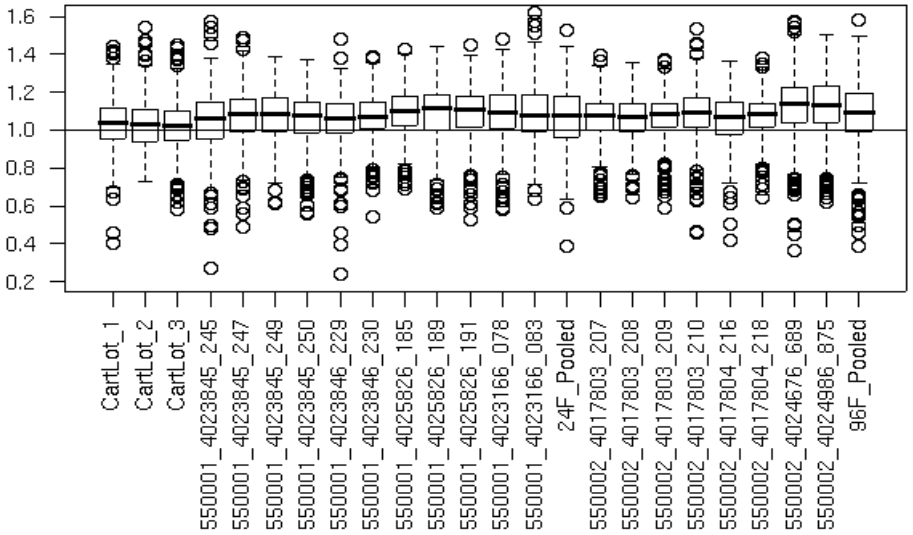


Figure 9: Box plots of RLE values computed from probe set values corresponding to two-fold changes in spike concentrations (including comparisons between 3.0 and 1.5 pM and comparisons between 1.5 and 0.75 pM), stratified by array set. The first three box plots summarize the RLE of spike-ins from cartridges analyzed within manufacturing lot. The "550001_" box plots summarize RLE values for spike-in probe sets on 24-array plates. The "550002_" box plots summarize RLE values for spike-in probe sets from columns 03, 06 and 09 of 96-array plates. The 24F_Pooled and the 96F_Pooled box plots summarize RLE values computed from values for spike-in probe sets from the virtual 24-array and 96-array plates, respectively.

The final question that we want to address is the effect of well location, row and column on data reproducibility, and how this effect compares to plate effect. To address this question we assess how the MA(RLE) values for arrays from the 96-array plates vary by plate, and column and row within plate. We use the MA(RLE) values computed from the plate-specific models in this analysis. Figure 10 displays box plots of the MA(RLE) values for the 96-array plates, stratified by plate, columns and rows. From this picture we can see that that the plate effects are largest, row effects (A to H) are smallest and column effects somewhere in the middle, with column effects largest around the edges, in columns 1, 2, 11 and 12. To quantify these effects we estimate the following analysis of variance model to the array summaries: MA(RLE) = intercept + Plate + Row + Column + error. Note that the estimated plate effects are relative to the first plate, the row effects are relative to the first row and column effects are relative to the first column. None of the row effects are significant either statistically (see column titled p-value in Table B) or

practically (see column titled “estimate”). Although some of the column effects are statistically significant, they are small in magnitude. While plate effects are larger than row or column effects, they still remain small when evaluated in terms of our ability to detect two-fold changes in spiked transcripts concentrations. To see this, note that reproducibility, as measured by MA(RLE), is 0.075 (see intercept estimate in Table B), and the largest plate effect on reproducibility is 0.011. Figure 7 shows that the MA(RLE) has to be 0.14 or greater before our ability to detect the spikes as captured by the auROC statistic is affected. The difference between the MA(RLE) where auROC first drops significantly below 100 percent (0.14) and the overall MA(RLE) (0.075) is about six times greater than the most significant plate effects.

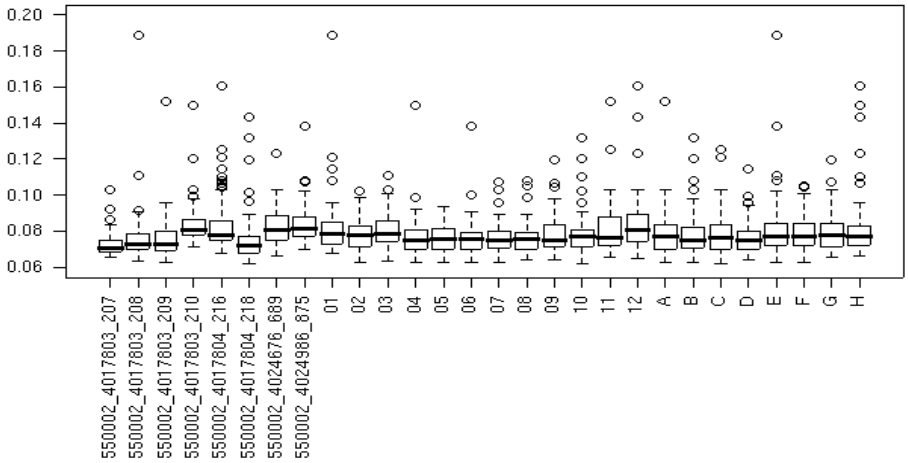


Figure 10: Box plot of MA(RLE) values for 96-array plate arrays, stratified by plate, columns and rows.

	Estimate	Std. Error	t value	p-value
(Intercept)	0.075	0.002	36.94	0.0000
Plate550002_4017803_208	0.004	0.002	2.27	0.0237
Plate550002_4017803_209	0.004	0.002	2.19	0.0287
Plate550002_4017803_210	0.011	0.002	6.88	0.0000
Plate550002_4017804_216	0.011	0.002	6.71	0.0000
Plate550002_4017804_218	0.003	0.002	1.72	0.0858
Plate550002_4024676_689	0.010	0.002	6.15	0.0000
Plate550002_4024986_875	0.011	0.002	7.10	0.0000
RowB	0.000	0.002	-0.01	0.9890
RowC	0.000	0.002	-0.20	0.8388
RowD	-0.002	0.002	-1.31	0.1891
RowE	0.002	0.002	1.26	0.2090
RowF	0.000	0.002	0.24	0.8109
RowG	0.001	0.002	0.33	0.7401
RowH	0.002	0.002	1.42	0.1552
Col02	-0.003	0.002	-1.67	0.0962
Col03	-0.002	0.002	-1.02	0.3067
Col04	-0.005	0.002	-2.67	0.0077
Col05	-0.006	0.002	-3.18	0.0015
Col06	-0.005	0.002	-2.61	0.0093
Col07	-0.006	0.002	-2.86	0.0043
Col08	-0.006	0.002	-2.97	0.0031
Col09	-0.004	0.002	-1.97	0.0491
Col10	-0.004	0.002	-1.91	0.0566
Col11	-0.001	0.002	-0.72	0.4706
Col12	0.002	0.002	0.91	0.3636

Table B: Plate, row and column effects for 96-array ANOVA model: MA(RLE) = intercept + Plate + Row + Column + error. Note that the estimated plate effects are relative to the first plate, the row effects are relative to the first row and column effects are relative to the first column.

Impact of Image Artifacts

The open system used to process the array plates has led to occasional artifacts caused by bubbles during the hybridization, wash or stain steps or very small debris out of the focus plane during scanning. We have determined that dark circles and some dark regions are due to bubbles that form predominantly during the wash/stain procedure, but also during hybridization. Bright-ringed dark circles, some bright circles and bright regions are due to bubbles or debris floating in the holding buffer during scanning. Figure 11 shows examples of some of the artifacts observed in these studies and their associated MA(RLE) values. The artifacts most often observed include dark circles, bright circles and bright regions. Haze (typically uneven background on array surface) and high background occur rarely in both cartridge and plate arrays and have a more deleterious impact. Images A and H in Figure 11 are representative of the most extreme artifacts observed, affecting a large percentage of the array surface area. The number of arrays with greater than 10 percent of the surface area impacted is low, only 14 of the 1,056 arrays run in this study (~1 percent). The two affected arrays in Figure 11 have elevated MA(RLE) values of 0.19. Although they would be identified as outliers in the MA(RLE) box plots in Figure 5, these arrays are still capable of detecting two-fold changes. The other six examples shown are more typical, with MA(RLE) values ranging from 0.07 to

0.13, indicating minimal effect on probe set signal estimates. Most of the image artifacts observed tend to be small and not at the extremes of the range of intensities, either completely dark or completely saturated. For cases where a small percentage of the surface area is affected, the impact to data quality is minimal. This is demonstrated by evaluating MA(RLE) box plots in Figure 12. The first box represents the 835 arrays with no visible image artifacts, the second box represents the 200 arrays with image artifacts affecting less than 10 percent of array surface area and the last box represents the 14 arrays with more than 10 percent of the array surface area impacted by image artifacts. The MA(RLE) distribution shifts higher when more than 10 percent of the array surface is impacted, but the distributions are very similar for arrays with and without small image artifacts. The array design, which includes 11 probe pairs per probe set distributed spatially over the array surface, and the default analysis algorithms (RMA), makes the system fairly robust to areas of corrupted data (see Statistical Algorithms Description Document, located at: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf). The MA(RLE) values illustrate variability at the array level. MA plots can be used to assess the impact at the probe set level. Figure 13 shows MA plots from the images shown in Figure 11.

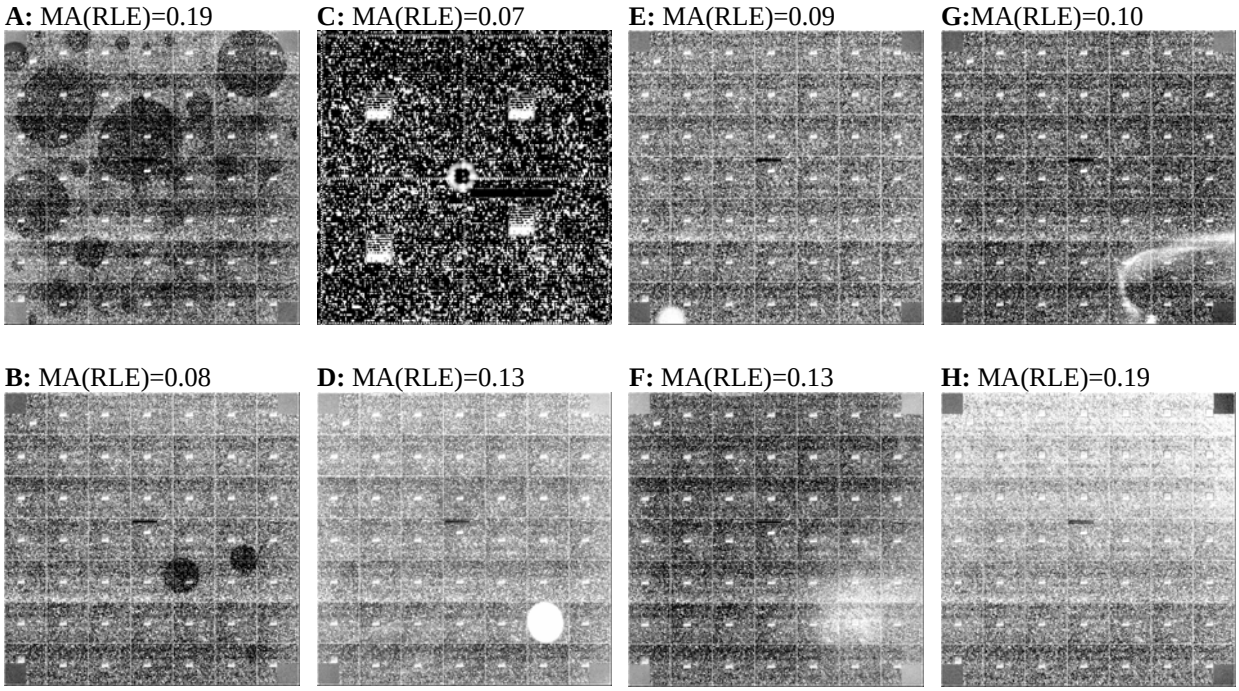


Figure 11: Screen shots of arrays with image artifacts, with MA(RLE) value for each. A and B, dark circles. C, bright-ringed dark circle (portion of original image magnified). D, bright circle. E, bright circle. F and G, bright regions. H, haze.

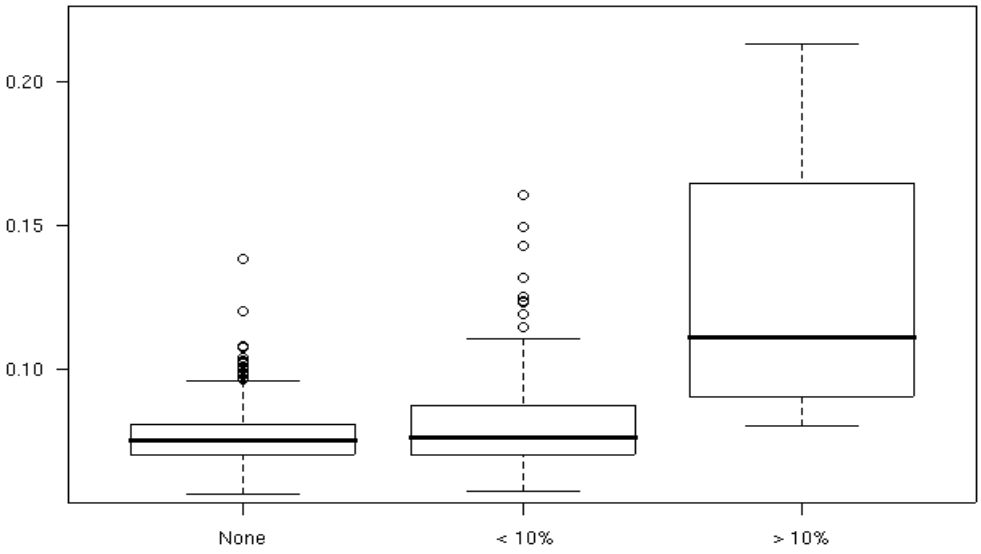


Figure 12: Box plot of MA(RLE) for arrays with no image artifacts (N = 835), arrays with small image artifacts (less than 10 percent of surface area, N = 200) and arrays with large image artifacts (N = 14). MA(RLE) values were calculated from signal estimates generated from within plate analyses with RMA.

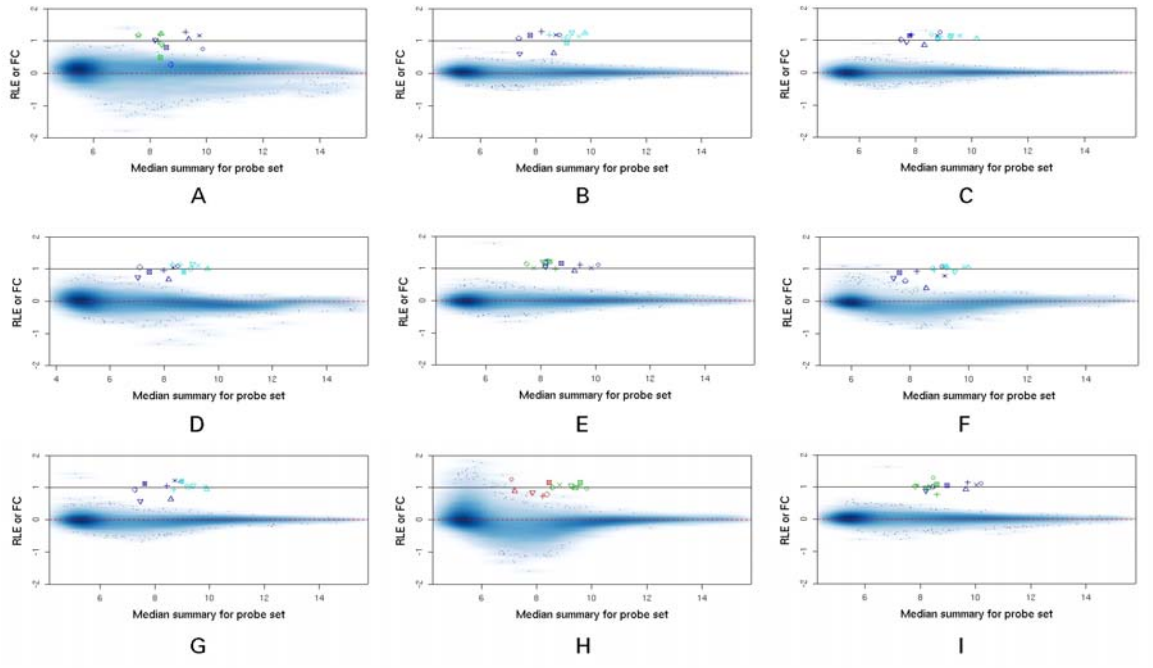


Figure 13: A-H: MA plots of arrays shown in Figure 11. I: MA plot from array with no image artifacts (550002_4017804_218.D10, MA(RLE) 0.08).

Consistent with the MA(RLE) values, the MA plots show that small image artifacts, such as those in images B, C, E and G of Figure 11, have virtually no impact on the probe set signal estimates. In addition to being small, the artifacts in these images are not at the extremes of the intensity range. Arrays with slightly larger regions impacted and/or brighter intensities (D and F) have a slightly higher MA(RLE) value (0.13) and a small, but noticeable increase in the spread in the probe set signal.

The MA(RLE) values will vary depending on samples and array design used. For HeLa on HT HG-U133A Arrays, typical MA(RLE) values range from 0.05-0.12. Arrays with MA(RLE) values much higher than 0.12 were identified as outliers in this study. Each user will have to evaluate their results with respect to the sample types and arrays used to define what an outlier array is in their system. To minimize the occurrence of image artifacts, we recommend keeping the GCAS in a clean area of the laboratory, keeping the doors closed during array processing, filtering the holding buffer and making sure that scan trays are clean. In addition, we recommend the following workflow for users to identify outlier arrays with their samples:

1. Run the experiment, including target preparation, hybridization, wash, stain and scanning
2. Generate probe set signal estimates using the RMA (or similar) algorithm
3. Calculate the MA(RLE) for each array in the experiment
4. Evaluate the distribution of MA(RLE) values to identify outlier arrays
5. Exclude outliers and regenerate probe set signal estimates with RMA

VI. Discussion

The GeneChip® HT System simplifies batch processing of many arrays in parallel, although there are some trade-offs that users must consider when determining whether to use cartridge or plate arrays. To evaluate array performance and quantitate these trade-offs, we have described a global metric, Mean Absolute Relative Log Expression (MA(RLE)) that we recommend using to assess array quality as well as system variability for both cartridge and plate arrays. MA(RLE) distribution summaries have certain advantages over traditional reproducibility measures such as the coefficient of variation, which both lacks a per-array reproducibility metric and is highly influenced by the worst performing arrays in a batch. For samples that are expected to have substantial similarity in expression profile, MA(RLE) can be used to provide quality assessments at the array level, and to provide an overall array set quality assessment. While the expected level of the MA(RLE) values will vary with the context of the experimental design, the relative size of the MA(RLE) distribution spread estimates should remain a valuable quality assessment metric in most experimental settings. For example, one context of an experimental design is the type of tissue used. Array MA(RLE) distribution will change as a response to differing tissues. Even within a common tissue type, MA(RLE) values may change due to the magnitude of experimentally induced perturbations to gene expression profiles. When multiple tissues are analyzed in a single experiment, MA(RLE) values should be computed within each tissue separately to maximize the metric's sensitivity to processing artifacts.

In the introduction, we posed four questions about the performance of plate arrays: Do plates perform as well as cartridges? Are there systematic effects and, if so, will they confound results? What is the impact of major and minor image anomalies? What QC metric can be used with the arrays?

To address these questions, we evaluated performance of both cartridge and plate arrays with respect to variability, which impacts the sensitivity and specificity of the results. We used MA(RLE) as a quality metric to identify outlier behavior in a data set. The MA(RLE) for each array is an output in both Expression Console software and the Affymetrix Power Tools (APT). Using MA(RLE), we found that within-plate variability for HT arrays was comparable to cartridge arrays, but when comparing arrays between multiple plates, the variability for HT arrays was slightly higher than cartridge arrays. Second, we identified three sources of quantifiable systematic variation: plate, column and row. Of these, plate and some column variation were found to be statistically significant, but small in magnitude. The largest plate effect was six times smaller than the increase in MA(RLE) required to lower the auROC for two-fold changes from 100 percent to 99.68 percent (see array F07 in Figure 7). What this means in practical terms is that systematic effects do not impact the ability of the HT System to meet our performance specifications with respect to detecting two-fold changes at low concentrations. The slight increase in variability compared to cartridge arrays means that in some cases it may be more difficult to detect smaller than two-fold changes.

We identified image artifacts that are predominantly a result of plate processing on the GCAS. The artifacts are typically small and have minimal impact on the probe set signal. Of these, only those covering greater than 10 percent of the array surface could, but not always, increase probe set signal variation. Severe haze and background, while rarely seen in this data set, contribute to increased variability and are easily detected as outliers by MA(RLE).

The most difficult question to answer about any analysis of array performance is where to set thresholds that determine outlier behavior. The data set we have described is very sensitive to detection of systematic variation, since we used technical replicates of a single, pooled hybridization target. The MA(RLE) values that we have used to identify outliers only apply for HeLa on the HT-HG-U133A array plates. In practice, users must determine an appropriate threshold based upon familiarity with the experimental design and the actual behavior of the data in their system.

VII. References

[1] Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12:111-139 (2002).
[2] White Paper, Statistical Algorithms Description Document: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf
[3] White Paper, Quality Assessment of Exon Arrays v. 1.0: http://www.affymetrix.com/support/technical/whitepapers/exon_arrays_qa_whitepaper.pdf
[4] Affymetrix Power Tools: http://www.affymetrix.com/support/developer/powertools/index.affx