Evaluating the Reproducibility of Microarray Technology

Glenda Delenstarr, Scott Vacha, Mark Hartnett, Condie Carmack, William Love, and Manoj Nair
Agilent Technologies, Inc.

OVERVIEW

This paper serves as a guide for evaluating the reproducibility of microarray technology* and shows:

- **How to design the experiments needed to evaluate reproducibility**
- **Why omission of outlier features and low-signal microarrays increases reproducibility**
- **Why the use of biological replicates increases reproducibility**
- **How high reproducibility of Agilent microarray technology increases confidence in differential expression results.**

High reproducibility is essential to correctly assess the extent of differential gene expression. Each step in the preparation of a microarray contributes noise that can affect the reproducibility of microarray signal data. To assess reproducibility, metrics are needed that can be compared across microarray platforms.

This paper uses data from the Agilent microarray system, in particular, to demonstrate metrics for evaluating quality and reproducibility. These metrics show that the Agilent microarray system, along with Rosetta Resolver algorithms, produces data that provides high confidence for gene expression interpretations.

These metrics can be used to calculate the reproducibility of microarray technology:

- Metrics to evaluate the microarray preparation process
- Signal statistics (%CV) for each gene on each array
- Signal statistics (%CV) for each gene across microarrays spotted with the same sample
- Log ratio statistics (SD) for all genes on each self-self microarray (Figure 1a)
- Log ratio statistics (SD) for each gene on each differential expression microarray (Figure 1b)
- Log ratio statistics (SD) for each gene across all differential expression microarrays
- Biological replicate metrics

This paper presents comparisons of these metrics with and without features identified as anomalous by the Agilent Feature Extraction software:

- Reproducibility with and without data points flagged as outliers
- Reproducibility with and without data points subjected to a test that evaluates if a signal is well above background
- Reproducibility with and without data points from microarrays containing an abnormal number of flagged outliers or low signals

Because an early version of the Agilent Feature Extraction software (v. 4.0.45) was used to produce the results, researchers can expect even better reproducibility with current versions of the Feature Extraction software than that presented in this document.

* Microarray technology refers to the entire microarray process, including the steps of array synthesis/deposition, target labelling, hybridization, wash, scan, and feature extraction/data processing.

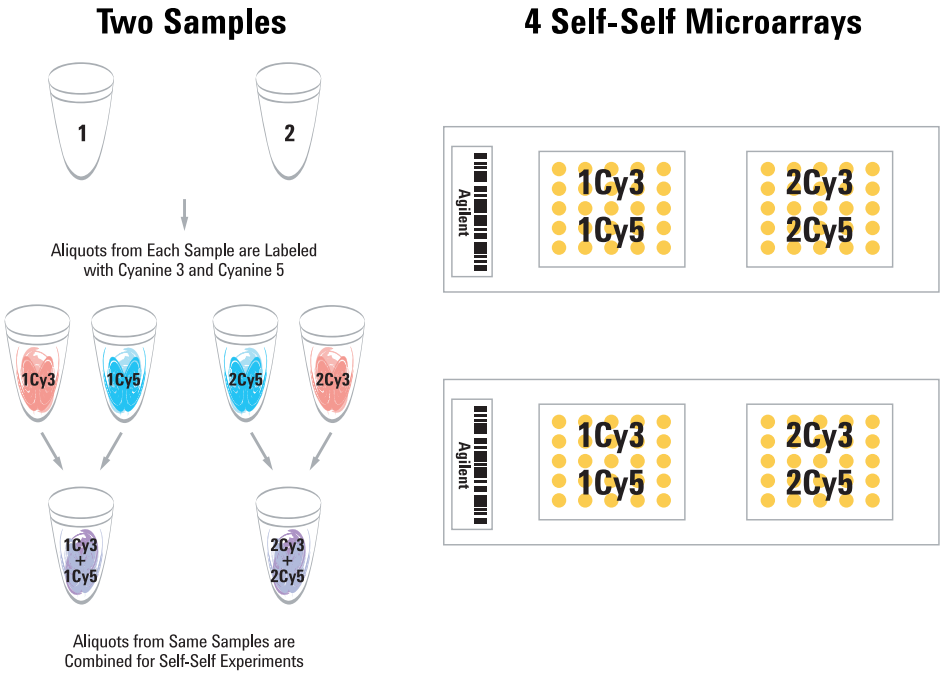


Figure 1a: Experiment type 1 with 4 self-self microarrays

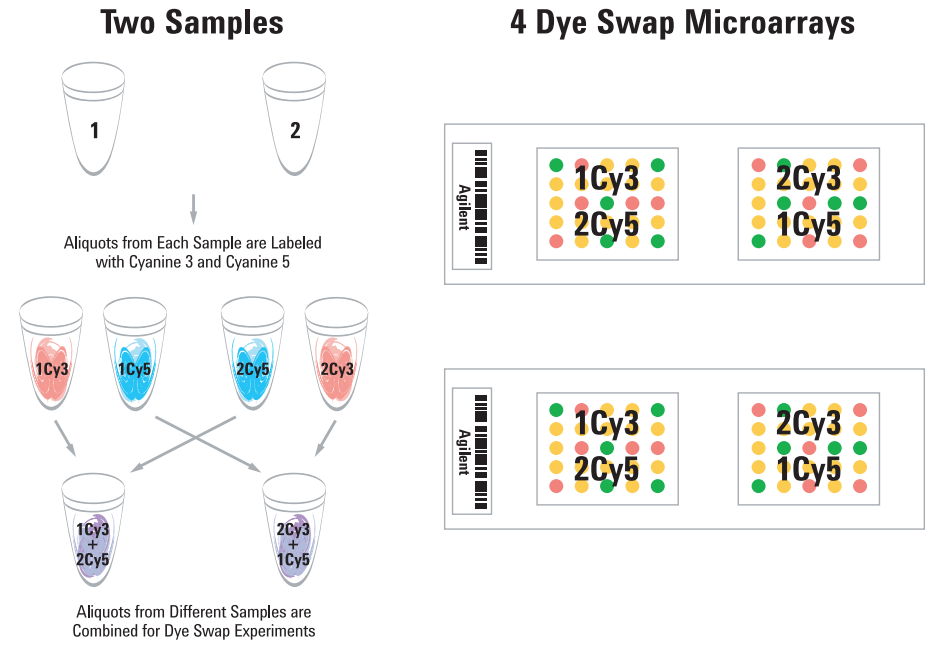


Figure 1b: Experiment type 1 with 4 dye-swap differential expression microarrays

EXPERIMENTAL DESIGN

Theoretical design

Researchers need to design only two experiments in order to produce sufficient data to reliably calculate the metrics described in the introduction.

- Experiment type 1—
  Four self-self microarrays (Figure 1a), two of which contain one sample labeled with both dyes and two of which contain a second sample labeled with both dyes. Four dye-swap differential expression microarrays (Figure 1b), two of which contain both samples, each labeled with a different dye and the other two of which contain both samples, each with the labels switched from the order in the first two microarrays.

- Experiment type 2—
  Same microarray design and sample types as in Experiment type 1. To be able to measure biological replicate metrics, the RNA sample preparations for Experiment type 2 must be different that those in Experiment type 1.

Both types of experiments should use replicate probe sequences per gene per microarray, which enables the calculation of intra-array statistics. Preferably, each microarray should contain at least 10 genes to evaluate intra-array reproducibility. These genes should span the full signal range, and each gene should have at least 15 feature replicates per gene.

Agilent designed the microarrays used for this paper under constraints for the needs of a customer. These constraints led to a design with 29 genes, each with 5 feature replicates.

Actual design and conditions

Five experiments (A-E, see below) were carried out to assess the performance of Agilent microarray technology.

Experiment A	Sample 1 vs. 2, RNA prep A (Experiment type 2)
Experiment B	Sample 1 vs. 2, RNA prep B (Experiment type 2)
Experiment C	Sample 3 vs. 4 (Experiment type 1)
Experiment D	Sample 5 vs. 6 (Experiment type 1)
Experiment E	Sample 7 vs. 8 (4 Self-self microarrays; 4 Dye-swap differential expression microarrays)

- Only experiments A and B were necessary for a complete performance evaluation.
- The customer requested that experiments C-E be carried out for biological interest.
- Each experiment used eight custom in-situ microarrays (60-mers; 8,455 features) to measure reproducibility and differential expression of a biological sample pair (40 microarrays total).
- The customer requested that Agilent design the microarray probe content to a eukaryotic organism of interest.
- Agilent used probe selection algorithms developed by Agilent Technologies and Rosetta Biosoftware.
- The probe design of the arrays was the same for all experiments and contained 29 genes of primary interest to the customer. These 29 genes had five replicate probes per array, thus enabling intra-array probe statistics.
- Agilent microarray technology used in these experiments included the Agilent 2100 Bioanalyzer; Agilent labeling kits, Agilent in-situ microarrays, Agilent hybridization and wash protocols, Agilent scanner, Agilent Feature Extraction software (v.4.0.45); and Rosetta Resolver software (v.2.0).
- To reproduce the experimental variability that microarray users may experience, each experiment used Agilent microarrays from different lots manufactured on different days from different printers.
- Also, multiple scientists performed sample labeling, hybridization and washing of the microarrays.

EVALUATION OF THE PREPARATION PROCESS

Before summary statistics are calculated on individual genes, the microarray preparation process is evaluated. If there is a major flaw in the process—either with microarray printing or sample labeling, hybridization or washing—the number of anomalous features across the microarray will be greater than average.

The first step is to check for slide and background anomalies by viewing the images, with different color scales, in the Image Analysis portion of the Agilent Feature Extraction software.

The second step is to review the quantitative metrics for each microarray [See Box 1 for definitions of terms in Italics and Box 2 for the Agilent microarray data analysis (feature extraction) process.]:

- Percent of features flagged as non-uniform outliers in either the green or red channel (control type = 0)
- Percent of features flagged as saturated in both the green and red channels (control type = 0)
- Average of net green and net red signals, averaged over entire array (control type = 0)
- Average of net local background green and net red signals, averaged over entire array (inliers)
- The SD of net local background green and net red signals, over entire array (inliers).

A summary of these metrics is shown for all five experiments in Table 1. There are four arrays that have a high number of features flagged as non-uniform (> 5 %). These appear in bold in the table.

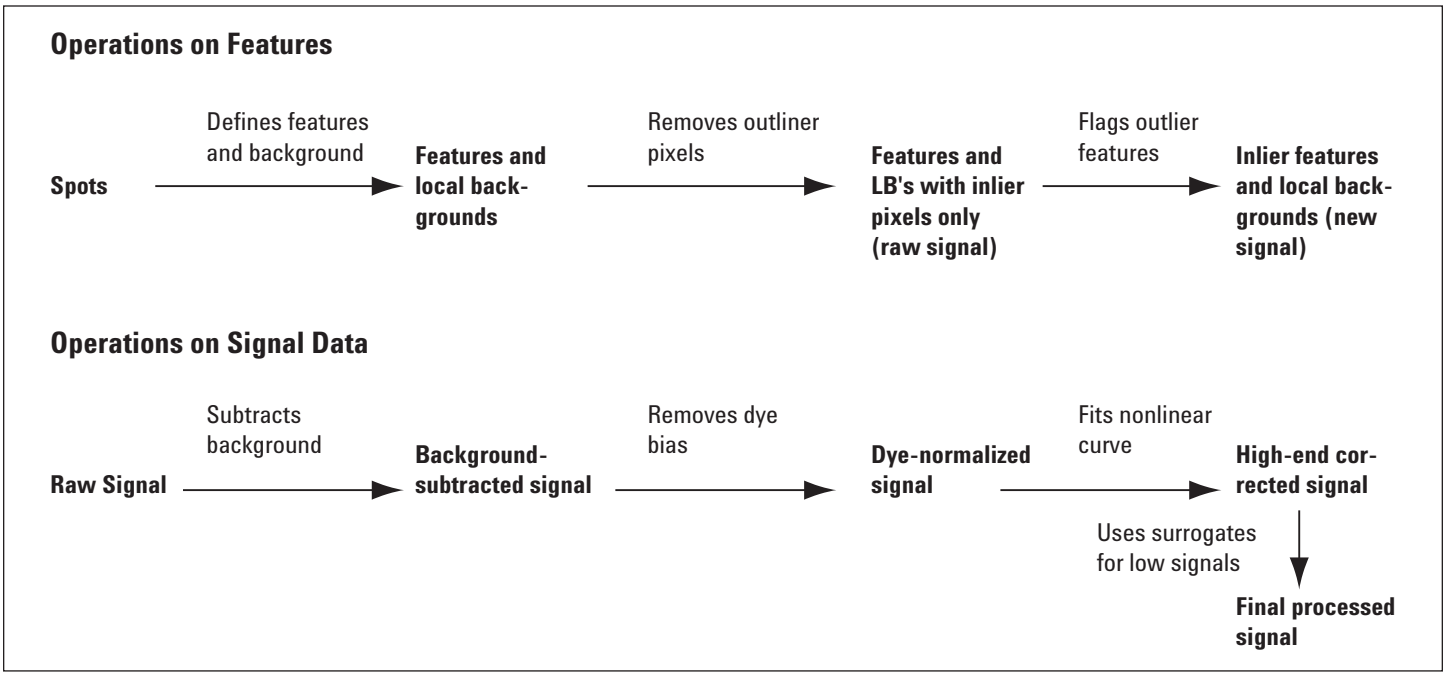
Non-uniform outliers are features flagged by the Agilent Feature Extraction software if the population of pixels within a feature spot used to calculate the signal average has a high signal standard deviation (SD). These features often lead to incorrect log ratio results. Thus, they are flagged and not used in statistical and Rosetta Resolver analyses.

All features on Agilent microarrays are assigned a *control type:* 0, 1, and -1:
Type 0 = all experimental features
Type 1 = positive control features
Type -1 = negative control features

Net signal is the raw signal with scanner constant offset subtracted.

Statistics for local backgrounds ring around each feature) are calculated using local backgrounds from all control types of features. In addition, global statistics are calculated on "*inlier*" local backgrounds. That is, they pass both non-uniform outlier and population outlier flags.

Box 1: Terminology



Box 2: Agilent microarray data analysis

Table 1: Microarray Quality Metrics Summary

				Experimental Features			Local Backgrounds (Inliers)			
				%Flags		Net Signal[4] Avg. G. R	Signal Averages		Signal Std. Dev.	
Expt.	Array Type[5]	Sample # (red)	Array Name[6]	Non-uniform	Resolver[7]		Green	Red	Green	Red
A	Diff	1A	08_A01	1.4	1.4	592	65	8	1.9	1.3
A	Diff	1A	37_A02	3.2	3.2	949	55	5	1.4	0.9
A	Diff	2A	08_A02	1.1	1.1	511	70	10	2.4	1.3
A	Diff	2A	37_A01	2.6	2.6	946	53	4	1.3	0.8
A	Self	1A	07_A02	2.5	2.7	994	59	6	1.3	1.0
A	Self	1A	38_A01	1.3	1.5	1083	53	4	0.9	0.8
A	Self	1A	07_A01	2.6	2.7	802	59	5	1.8	1.1
A	Self	1A	38_A02	1.7	1.7	784	56	5	1.1	0.9
B	Diff	1B	10_A01	2.1	2.1	1152	61	4	1.2	1.0
B	Diff	1B	35_A02	1.1	1.1	1054	52	5	1.6	1.0
B	Diff	2B	10_A02	3.9	3.9	967	64	5	2.2	1.4
B	Diff	2B	35_A01	1.2	1.2	683	50	4	1.4	0.9
B	Self	1B	09_A02	2.3	2.6	1198	58	5	1.3	1.0
B	Self	1B	36_A01	0.5	0.7	1092	50	4	1.0	1.1
B	Self	2B	09_A01	1.3	1.4	943	57	4	1.3	1.0
B	Self	2B	36_A02	1.0	1.1	797	51	5	0.9	1.2
C	Diff	3	12_A01	2.9	3.0	1131	71	11	2.4	2.0
C	Diff	3	33_A02	3.7	3.8	1552	59	5	1.3	1.4
C	Diff	4	12_A02	1.7	1.8	1104	70	11	3.4	2.0
C	Diff	4	33_A01	10.1	10.1	1425	58	4	1.2	1.2
C	Self	3	11_A02	3.5	3.6	1379	74	12	3.5	2.8
C	Self	3	29_A01	16.1	16.2	1861	64	8	1.6	1.9
C	Self	4	11_A01	2.6	2.6	835	63	7	5.3	3.8
C	Self	4	29_A02	1.0	1.1	1418	66	9	2.8	2.6
D	Diff	5	20_A01	5.0	5.2	1147	55	7	1.3	1.4
D	Diff	5	34_A02	1.5	1.7	1398	58	5	1.2	1.1
D	Diff	6	20_A02	3.8	3.9	1035	58	8	1.4	1.2
D	Diff	6	34_A01	3.5	3.7	1124	57	4	1.4	1.1
D	Self	5	19_A02	4.3	4.5	1131	59	10	1.5	1.6
D	Self	5	30_A01	62.8	62.8	1660	61	9	2.4	1.7
D	Self	6	19_A01	1.3	1.6	970	57	8	1.3	1.6
D	Self	6	30_A02	1.4	1.8	1462	63	7	2.4	1.5
E	Diff	7	22_A02	49.0	49.0	3316	52	4	0.9	0.8
E	Diff	7	27_A01	1.0	1.1	1330	60	5	1.2	1.0
E	Diff	8	22_A01	3.6	3.6	994	53	3	0.8	0.8
E	Diff	8	27_A02	1.6	1.7	1098	60	6	0.8	1.0
E	Self	7	21_A01	0.8	0.8	1101	54	4	1.9	1.3
E	Self	7	28_A02	0.5	0.6	987	59	7	1.4	1.1
E	Self	8	21_A02	1.0	1.2	1473	56	5	1.3	1.0
E	Self	8	28_A01	0.6	0.8	1343	58	6	1.1	0.9

[1] Net signal is the raw signal with scanner constant offset subtracted.

[2] The Array Type is either Self (self-self hybridization microarray) or Diff (differential expression dye-swap microarray). The Sample # (red) is listed in many tables, along with the Array Type. From the Sample # (red) and the Array Type, one also knows the type of green sample used on that array.

[3] The naming convention for microarrays is NN_A0s, where NN are the two right-most numbers of the microarray barcode and where s is the side number (A01 is the left side and A02 is the right side).

[4] Agilent Feature Extraction software uses two criteria to flag Rosetta Resolver not to use Agilent data in its calculations: 1) if features are flagged as non-uniform in either the green OR red channel, or 2) if features are flagged as saturated in both the green AND red channels.

Outlier microarrays

A plot of the % features flagged as non-uniform for each microarray in each experiment is shown in Figure 2. The average of the % features flagged as non-uniform for all microarrays is 5.3%, and the average for the % features flagged as either Non-uniform or saturated, as flagged for Rosetta Resolver, is 5.4%. The four microarrays with > 10% features flagged as Non-uniform are shown as stars in the figure. When these outlier microarrays are omitted from the calculation, the averages decrease to 2.1% flagged as Non-uniform and 2.2% flagged for Rosetta Resolver.

Low signal microarrays

In Table 1, Experiment A, microrrays 8_A01 and 8_A02 both showed significantlyreduced net signal compared with their respective duplicate microarrays, 37_A02 and 37_A01.

This paper discusses the impact of omitting these potential outlier and low signal microarrays on composite signal and log ratio statistics and analyses.

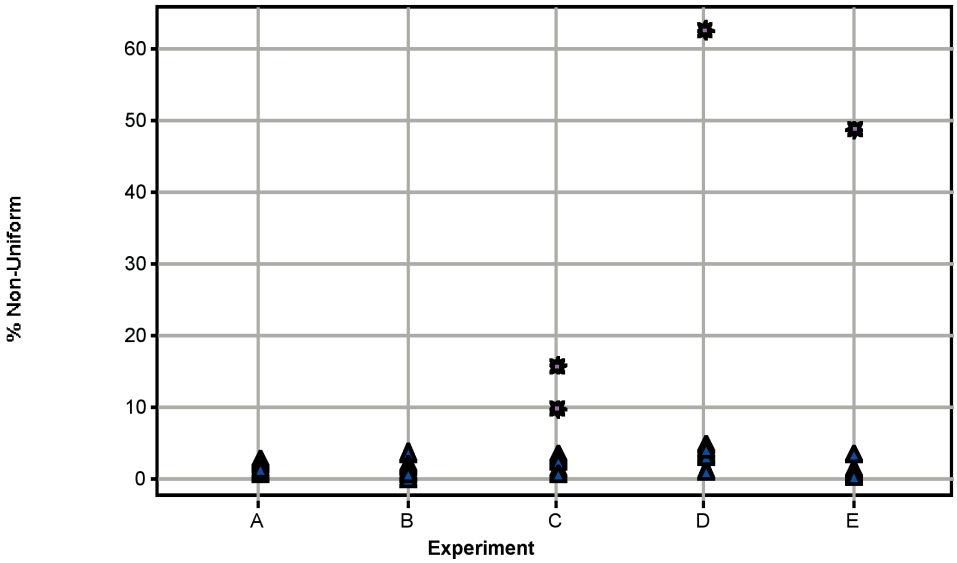
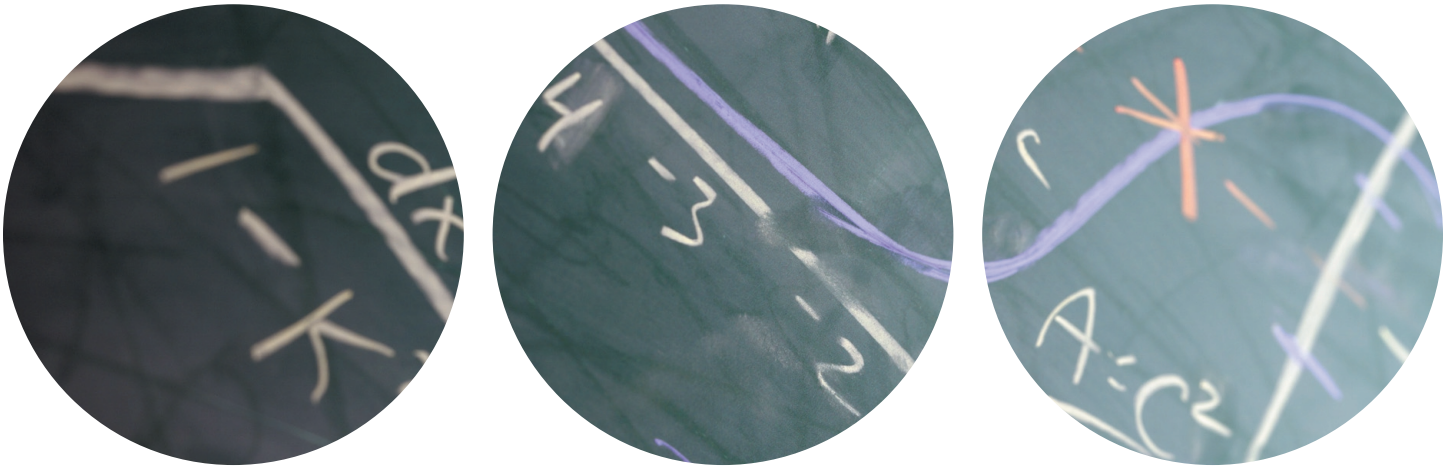


Figure 2: % of experimental features flagged as non-uniform for Experiments A-E



EVALUATION OF REPRODUCIBILITY WITH SIGNAL STATISTICS (%CV)

Calculation of signal statistics (intra-array, intra-gene)

The first step of this analysis is to calculate the average and standard deviation (SD) of the signals from the green or red channel for each gene (intra-gene) of the 29-gene set. These 29 genes have five replicate features (same probe sequences) randomly spread across the microarray (Figure 3), allowing probe statistics (intra-array). The reproducibility metric is the % coefficient of variation, or %CV, where %CV = 100% * (SD/average). For Experiment B, the average, SD and %CV of the 5 replicate feature signals within a microarray were calculated for each gene using the green final processed signals.

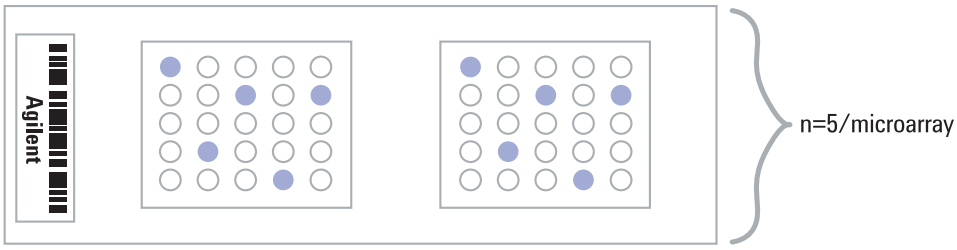


Figure 3: Signal % CV calculated for five replicate probes (intra-gene)

The probes in Figure 3 are randomly located on each microarray (intra-array) and represent only one set of the 29 genes on the microarray.

Table 2: Experiment B — Summary of intra-array, intra-gene green signal statistics

Experimental Features	N	Avg. of %CV	SD of %CV
All features	232	26.3%	40%
All features except those flagged for Rosetta Resolver	226	25.7%	40%
Inlier features [all features except those flagged for Rosetta Resolver or not Well Above Background (WABk)]	212	17.3%	12%

Calculation of summary signal statistics

The second step is to calculate the summary statistics for all the genes on all the microarrays for one experiment. From the signal statistics (%CV) of Experiment B, the average and SD of the %CV's were calculated for all features and are presented in Table 2, Row 1, along with the number of data points. Because experiment B has 8 microarrays and the analysis used 29 genes, the maximum number of data summary points used for experiment B = 8*29 = 232. There are fewer data points if flagged features are omitted.

[5] The "*Well Above Background*" test is a more stringent test than the t-test that determines if a feature is positive and significant against background. This test requires that a feature's background-subtracted signal be 2.6 times greater than the background SD. This requirement approximates the requirement that the feature's raw signal be greater than 99% of the background population's signal.

Terminology

Effect of omitting flagged outlier features on summary signal statistics

Signal statistics should be calculated with a subset of data that is of the highest quality. Only feature data that has passed three criteria are used. Features flagged as non-uniform are omitted from calculations of signal statistics. Saturated features are also omitted from statistics, as their signals have hit a ceiling. Finally, features that are not "Well Above Background" (WABk) are omitted from signal statistics, as their signal is very low. However, the features that are not well above background are still used for calculating log ratio statistics.

Single experiment

The impact of omitting these flagged features on summary signal statistics is shown in Table 2 and Figure 4a and Figure 4b for experiment B.

- If a gene has features that have been flagged as non-uniform or saturated (flagged for Rosetta Resolver), the %CV of the signals from feature replicates calculated for that gene will be higher than if the flagged features had been omitted (Table 2, rows 1 and 2, respectively). Genes that have features flagged for Rosetta Resolver are shown as dark pink symbols in Figure 4a.

- If genes have features that fail the WABk test, they generally also have higher %CV's than if these features are omitted (Table 2, rows 1 and 3, respectively). Genes that fail the WABk test are shown as triangles in Figure 4a.

- The inlier data (Table 2, row 3), shown as light blue circles in Figure 4a, are also shown in a zoom view (note change in Y-scale) in Figure 4b. The average %CV using only inlier features (not flagged for Rosetta Resolver and passing WABk) is 17.3% (Table 2, Row 3), which is less than the 26.3 %CV (Table 2, Row 1) using all features, as expected.

- The distribution of the %CV's of the inlier features is also tighter than when all features are used. That is, the SD of the %CV's is 12% vs. 40%, respectively (Table 2, rows 3 and 1, respectively).

- Figure 4b also demonstrates the expected result that signal reproducibility is better (lower %CV) at higher average signals.

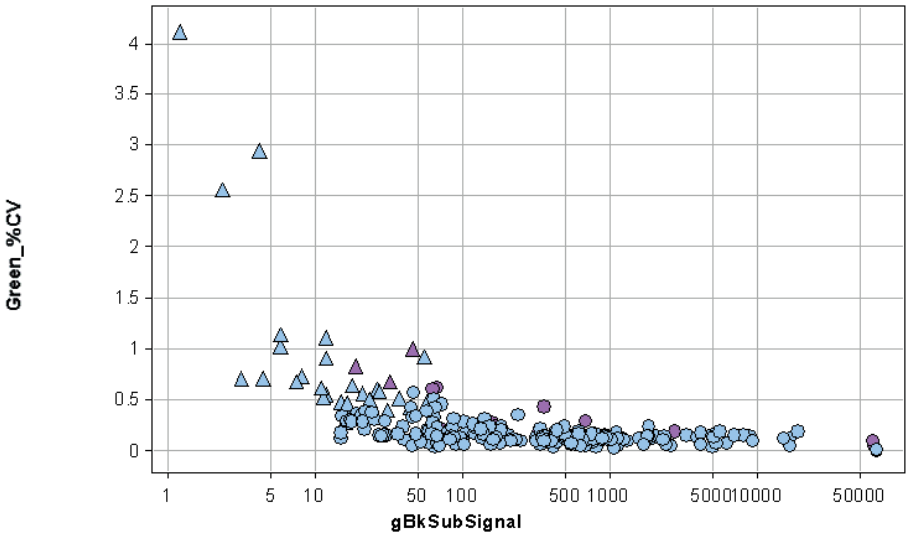


Figure 4a: Expt. B—Intra-array signal statistics vs gBkSubSignal for ALL features, including flagged features

Figure 4a shows the intra-array, intra-gene %CV's of experiment B for all features (same as Table 2, row 1) plotted against the green background-subtracted signal. All features, including flagged data, are shown. The features that are flagged as non-uniform or saturated are shown in dark pink; inlier features are light blue. The shapes show the features that pass the Well-Above Background (WABk) flag in both the red and green channels (circles), one channel (triangle), or neither channel (squares). The average of the %CVs for all features equals 26.3%.

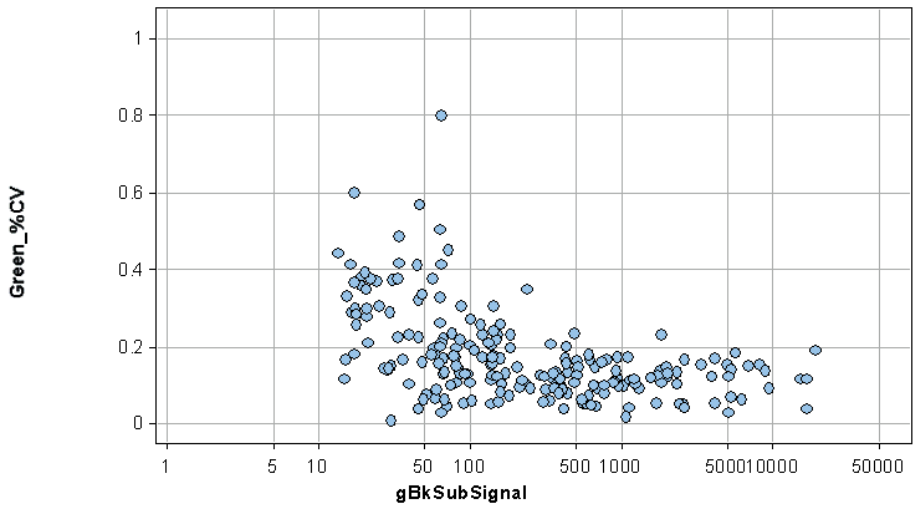


Figure 4b: Expt. B—Intra-array signal statistics of INLIER features versus gBkSubSignal

Fgure 4b shows the intra-array, intra-gene %CV's of experiment B for only inlier features that pass the WABk test (same as Table 2, row 3) plotted against the green background-subtracted signal. The average of the %CVs for inlier features equals 17.3%.

All experiments

Table 3 shows the intra-array statistics for all 5 experiments, using the inlier set of features (e.g., Table 2, row 3 for Experiment B). The analysis was performed on the final processed signals for both the green and red channels. Table 3 shows the average %CV and the median %CV. Medians better reflect the central tendency of the data because they are much more robust to outliers than are averages. In Table 3, the median %CV in every experiment is less than the average %CV.

Table 3: All Experiments—Summary of intra-array, intra-gene inlier signal statistics

		Green %CV Statistics		Red %CV Statistics	
Expt.	N	Median	Avg.	Median	Avg.
A	214	15.7%	17.8%	13.8%	17.4%
B	212	14.2%	17.3%	11.7%	13.2%
C	216	18.6%	14.4%	12.7%	16.9%
D	176	12.5%	15.3%	11.1%	12.6%
E	198	7.6%	11.5%	6.4%	8.5%

Calculation of composite signal statistics (intra-gene, intra-sample, inter-feature)

Another metric evaluates a composite of both intra-array and inter-array reproducibility. This metric calculates the signal statistics for the replicate features for each gene (intra-gene) within an array (intra-array) and among microarrays (inter-array) that have a given sample (intra-sample) for that signal channel. For each of the experiments, there were 4 microarrays (inter-array) with one type of sample in the green channel and 4 microarrays with the second type of sample in the green channel. There was a maximum of (5 features/gene/microarray)*(4 microarrays/sample-type) = 20 features summarized per gene (inter-feature). Inter-feature statistics can only be calculated on a gene if at least 2 of the 20 features/gene remain, after flagged features are omitted.

The average, SD, and %CV of these 20 signals were calculated for each gene, for the two sample types. There were 29 genes in the set analyzed and two types of samples used per channel. Therefore, 29*2= 58 data points are possible, if all features are used. After flagged or low signal features are omitted, there may be fewer than 58 data points (see Table 4a, Experiments D and E).

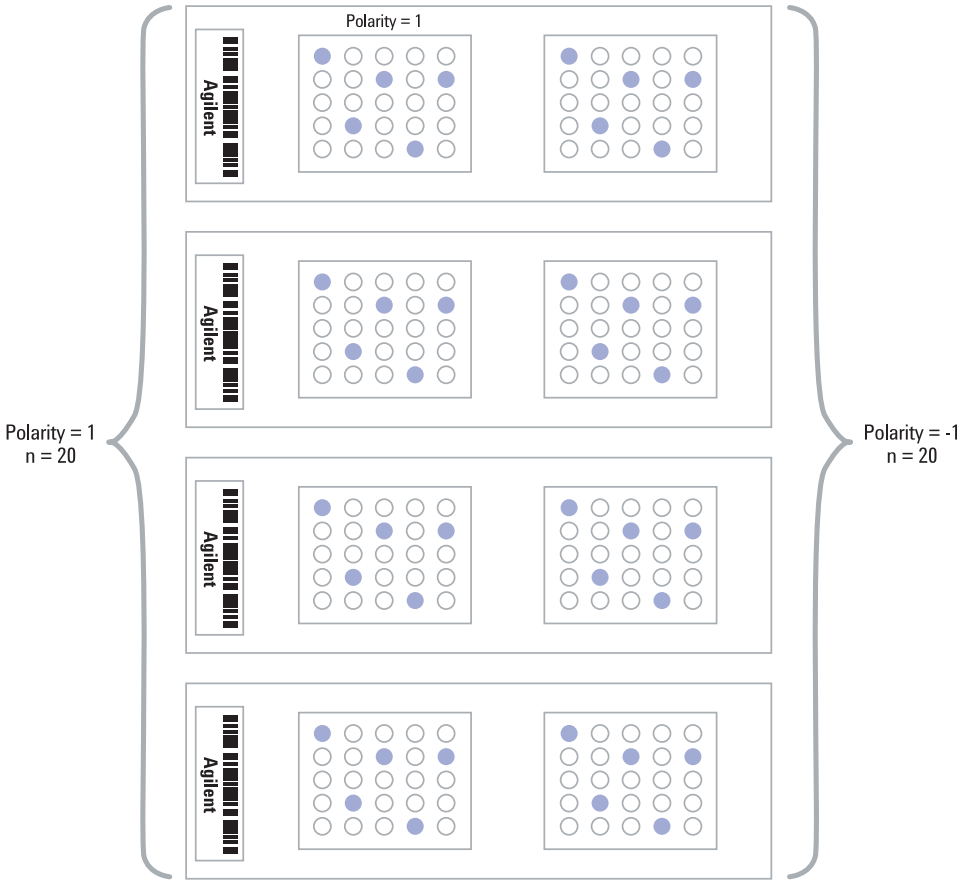


Figure 5: Signal %CV calculated for twenty features (inter-feature)

The features (blue spots) for each gene (intra-gene) for each labeled sample (intra-sample) are randomly located across four microarrays (inter-array) in Figure 5. Microarrays with "+1" polarity have sample_1 labeled with the red dye and sample_2 labeled with the green. The opposite is true for microarrays with "-1" polarity (e.g. red = sample_2; green = sample_1). The features in the figure represent just one set of the 29 genes on the microarray. See Figure 1 for a description of the dye-swap microarrays used in the experiment. See Figure 3 for a representation of intra-array calculations.

Effect of omitting flagged outlier features on composite signal statistics

The features were filtered for two flags: passing the non-uniform flag and not being saturated. The summary average of the signal %CV's and median of the signal %CV's were calculated from the remaining data points. The filtered features were then tested to see if their signals were well above the background (WABk). A second set of summary statistics (avg. %CV and median %CV) were calculated for features that were inliers and that passed the WABk test. Results of these calculations appear in Table 4a. Note that N for features not subject to the WABk test is less than 58 for Experiments D and E, whereas N for features subject to the WABk test is less than 58 for all five experiments. See the section, "Calculation of composite signal statistics" on the previous page for an explanation.
In Table 4a the calculations for the filtered signals that were not screened for the "Well Above Bk" test (WABk) are shown in the rows where "Well Above Bk Test?" = "No". The statistics for the filtered signals that passed the WABk test are shown in the rows where "Well Above Bk Test?" = "Yes". The calculations on features that passed the WABk test yielded %CV's that were predictably lower than those for the feature sets that included non-WABk features.

Table 4a: All Experiments—Summary of composite signal statistics

	Well Above_Bk	Green (%CV)			Red (%CV)		
Expt.	Test?	N	Median	Avg.	N	Median	Avg.
A	No	58	35.5%	40.7%	57	36.0%	47.0%
	Yes	46	33.9%	35.5%	45	34.9%	40.5%
B	No	58	27.9%	32.1%	57	25.2%	30.8%
	Yes	45	24.7%	27.9%	40	25.1%	31.4%
C	No	58	27.1%	31.5%	58	26.9%	30.4%
	Yes	39	23.3%	27.6%	34	23.3%	26.9%
D	No	57	35.1%	39.1%	57	35.1%	37.9%
	Yes	34	31.9%	36.8%	30	31.9%	37.4%
E	No	55	36.2%	38.6%	55	32.1%	36.3%
	Yes	43	26.9%	32.3%	41	23.3%	30.5%

For example, in experiment A, "sample 1" is in the red channel for two self-self microarrays (7_A02 and 38_A01, see section of Table 1 on the next page) and two dye-swap microarrays (8_A01 and 37_A02, see section of Table 1 below). The inter-feature statistics (average, SD, and %CV across the inliers of the 20 features) are calculated for each of the 29 genes for this sample. The 4 microarrays with "sample 2" in the red channel provide 29 more gene-sample data points, for a total of 58 data points, before filtering data for flagged features. After the features are screened for non-uniform outliers and for saturation, 58 data points still remain.

When the "Well Above Bk" test is applied to the features, only 46 data points remain (See Table 4a, Row 2) and the %CV's are lower than when the test was not applied (Table 4a, Row 1).

The %CV's for this composite analysis (Table 4a) are higher for each experiment than the corresponding intra-array %CV's (Table 3), since the composite %CV's are encompassing the additional noise of combining features across 4 different microarrays (inter-array) and across microarrays that are self-self or dye-swap types (inter-type).

Section of Table 1

				Experimental Features		
				%Flags		Net Signal
Expt.	Array Type	Sample # (red)	Array Name	Non-uniform	Resolver	Avg. G. R
A	Diff	**1A**	**08_A01**	1.4	1.4	**592**
A	Diff	**1A**	**37_A02**	3.2	3.2	949
A	Self	**1A**	**07_A02**	2.5	2.7	994
A	Self	**1A**	**38_A01**	1.3	1.5	1083
A	Diff	2A	08_A02	1.1	1.1	**511**
A	Diff	2A	37_A01	2.6	2.6	946
A	Self	2A	07_A01	2.6	2.7	802
A	Self	2A	38_A02	1.7	1.7	784

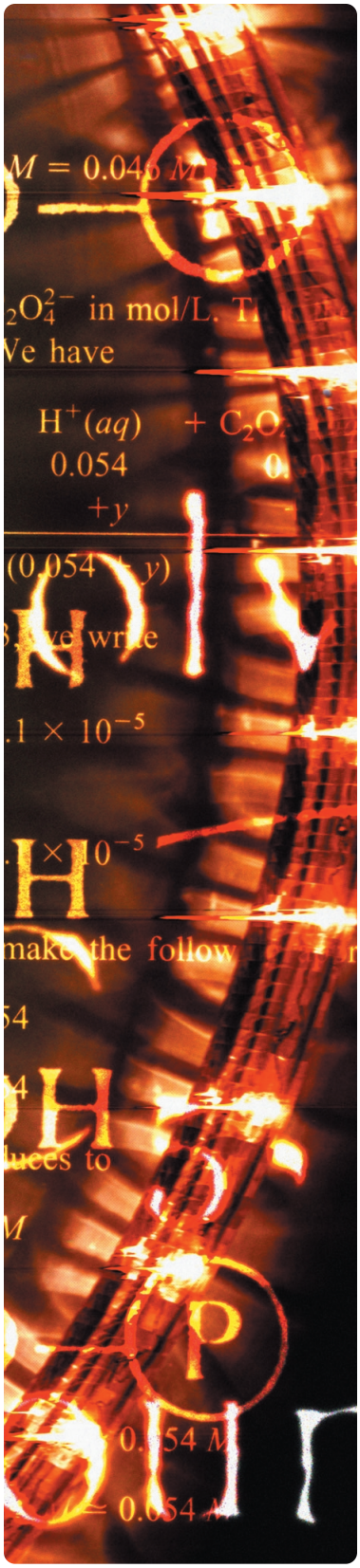
Effect of omitting low signal microarrays on composite signal statistics

Experiment A demonstrates the power of evaluating microarrays for quality before including them in results. Microarrays 8_A01 and 8_A02 (see above) had much lower average net signals than the other microarrays of experiment A. The intra-gene, intra-sample, inter-feature signal statistics were repeated with experiment A, omitting these two microarrays (Table 4b). The %CV's for the green and red channels decreased, when compared with Experiment A using all microarrays (Table 4a). For example, for Experiment A with all arrays and features that had not passed the WABk test, the median %CV was 35.5% (Table 4a), and without the two low signal arrays, the median %CV for the same feature set was 25.1%.

Table 4b: Experiment A—Composite signal statistics after omitting low signal microarrays

Well Above_Bk Test?	Green (%CV)			Red (%CV)		
	N	Median	Avg.	N	Median	Avg.
No	57	25.1%	35.4%	57	27.8%	34.7%
Yes	45	23.3%	29.2%	45	26.4%	31.7%

The above analysis demonstrates the impact of microarray variation upon signal statistics. However, since differential expression experiments are performed as ratio experiments, variations in red or green channel signals within or across microarrays are attenuated when calculating the log ratios, as discussed in the next section.



EVALUATION OF REPRODUCIBILITY WITH LOG RATIO STATISTICS (SD)

Calculation of log ratio statistics for a self-self microarray (intra-array, inter-gene)

A very important metric for microarray reproducibility is the standard deviation (SD) of the log ratios of self-self arrays. A log ratio is the log of the ratio of the final processed signal in the red channel to the final processed signal in the green channel. This metric evaluates the level of noise produced by the preparation process for each microarray. Therefore, microarray technology that produces lower SD's of the self-self log ratios also has higher sensitivity for detecting significant differentially expressed genes.

All non-control features that are inliers can be used for this metric, since the log ratios of all genes should be approximately equal to 0 for a self-self array. A global average and SD of log ratios for all the genes (inter-gene) are calculated for each self-self array (intra-array). For accuracy, we expect the average of the ratios of self-self arrays to be equal to 1. Therefore, the average of the log ratios is expected to be equal to 0. For reproducibility, we expect that the global SD of all the log ratios to be close to 0. As the SD decreases, the "noise envelope" decreases, and the sensitivity of detection increases.

Figure 6a shows all the data (inlier and outlier) from a self-self array (expt. B, microarray 9_A01) as log of red vs. log of green background-subtracted signal. This type of plot is instructive, as it shows the underlying signal data, which is subsequently normalized, converted to ratio of (red/green) and finally converted to log ratio (red/green). The tightness of the points about the diagonal relates to the SD of the log ratios. Log ratios of all experimental features (control type = 0) were analyzed for this metric, not just the set of 29 genes used with other reproducibility metrics. All features, including flagged data, are shown. The features that are flagged as non-uniform or saturated are shown in dark pink; inlier features are light blue. The shapes show the features that pass the Well-Above Background (WABk) flag in both the red and green channels (circles), one channel (triangle), or neither channel (squares).

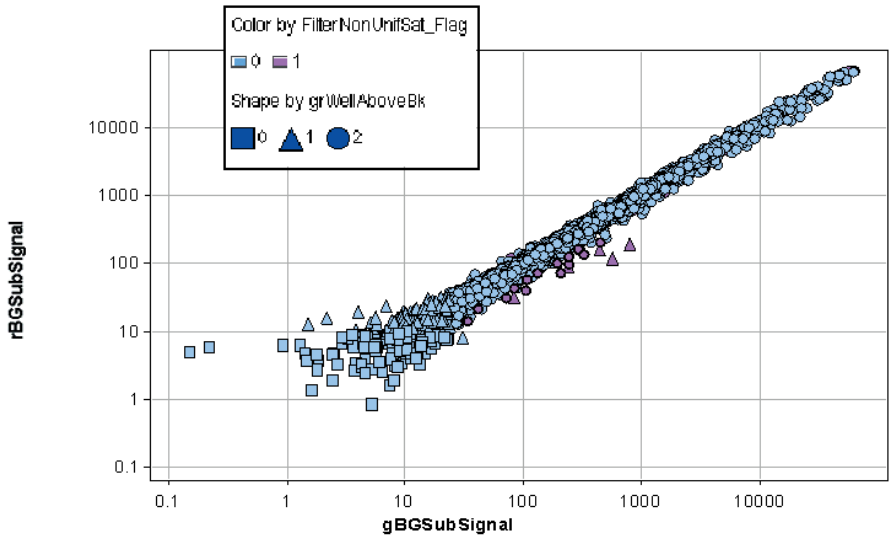


Figure 6a: Experiment B, microarray 9_A01 (Self)—Experimental features including flags (n=7,986)

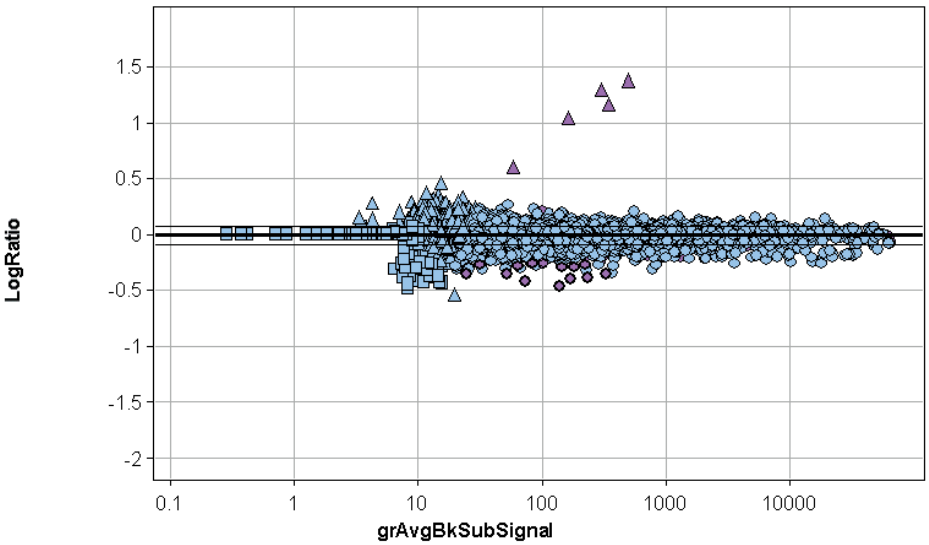


Figure 6b: Expt. B, microarray 9_A01 (Self)—Log ratio versus the log of the Avg (green and red) BkSubSignal

Effect of omitting flagged outlier features on log ratio statistics for self-self microarrays

Single experiment

The same data shown in Figure 6a is presented in a more typical log ratio vs. log signal plot in Figure 6b. If the flagged data are not omitted, the SD of the log ratios is 0.078. The average log ratio (approximately 0) and SD bars are shown in Figure 6b. When the features that are flagged for either non-uniform or saturation (pink symbols) are removed (as is done during the data import into Rosetta Resolver), the SD of the log ratios decreases to 0.071.

[11] A log ratio is the log of the ratio of the final processed signal in the red channel to the final processed signal in the green channel.

All experiments

The analysis with flagged features omitted was performed on each of the 4 self-self arrays in a given experiment, yielding 4 SD's of the log ratios. The average of the 4 SD's of the log ratios for each experiment is shown in Table 5. The averages of the SD of the log ratios are very similar across the 5 experiments. The average number of inlier features that were used for each microarray calculation is also shown in Table 5. Experiment D has a lower average number of features used because of the high number of non-uniform features omitted, especially from microarray 30_A01 (see Table 1).

Calculation of log ratio statistics for differential expression microarrays (intra-array, intra-gene)

The previous analysis could be performed across all genes (inter-gene) on a microarray because the microarray set studied included only self-self microarrays. However, metrics for differential expression microarrays are also necessary. The microarray design should therefore include a set of genes with multiple feature replicates per microarray when microarray performance is being evaluated. This microarray design lends itself to calculating "intra-gene" statistics of log ratios for a given microarray (intra-array). See Figure 3. Again, all log ratio statistics are calculated using features that are inliers to both the non-uniform and saturated flags.

Single experiment

The SD of the log ratios for each gene was calculated (max N = 5 inlier features/gene) for each microarray. The figures below use data from one experiment (B) and one array (10_A01) to show that the log ratio SD's and the signal %CV's are greater when the log ratio is near 0 and/or when the overall signal is low. The figures also show the attenuating effect of using log ratios.

Table 5: Summary of log ratio standard deviations (SD's)

Expt.	Avg. # of Features/Array	Average SD_Log Ratio
A	7844	0.082
B	7899	0.090
C	7547	0.074
D	6603	0.088
E	7954	0.076

- **In Figure 7a** the log ratios are extremely tightly clustered for the replicates at moderate to high signals but are less tightly clustered at lower signals. This noise versus signal pattern appears for most of the reproducibility metrics for both signals and log ratios.
- The 2 genes with the highest SD's have log ratios near 0 (**Figure 7b**) and have 2 of the lowest average signals (**Figure 7c**, dark pink symbols). The average SD of the log ratios is 0.037 and the median SD is 0.017.
- The signal %CV's (average of green and red signal %CV's) were also highest for these same 2 genes (**Figure 7d**, **7e,** dark pink symbols).
- **Figure 7e** shows the attenuating effect of using ratios. There are 27 genes with average signal %CV's ranging from 2% to 18%. Yet the majority of these genes have log ratio SD's < 0.03, with no relation between the log ratio SD and %CV. Thus, even though there may be a wide range of signal variability, there is a narrow range of log ratios.

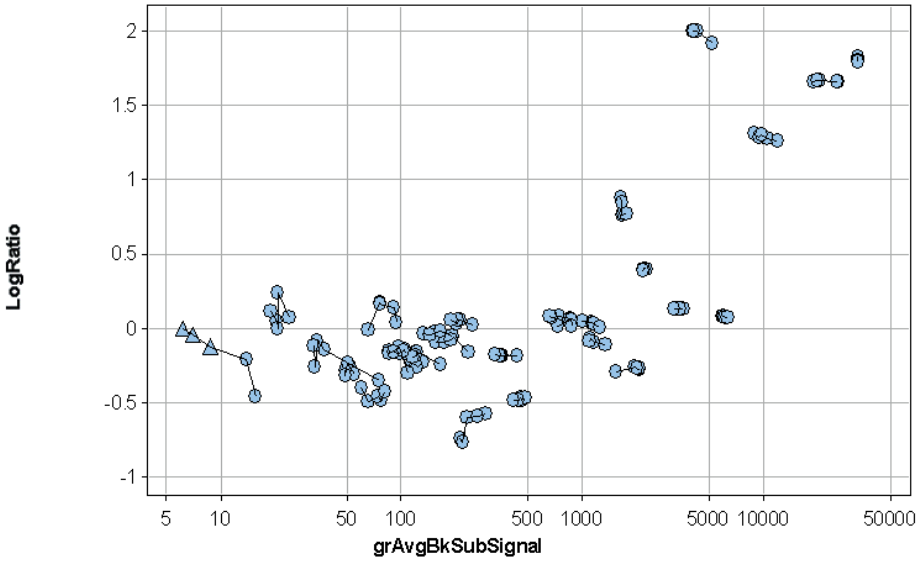


Figure 7a: Experiment B—Intra-array log ratio reproducibility (max N = 5 features/ gene; 29 genes/microarray)

Figure 7a shows log ratios for 29 genes from experiment B, microarray 10_A01 (Diff array), plotted versus the average background-subtracted signals for the red and green channels. The 5 feature replicates for each gene are connected by lines.

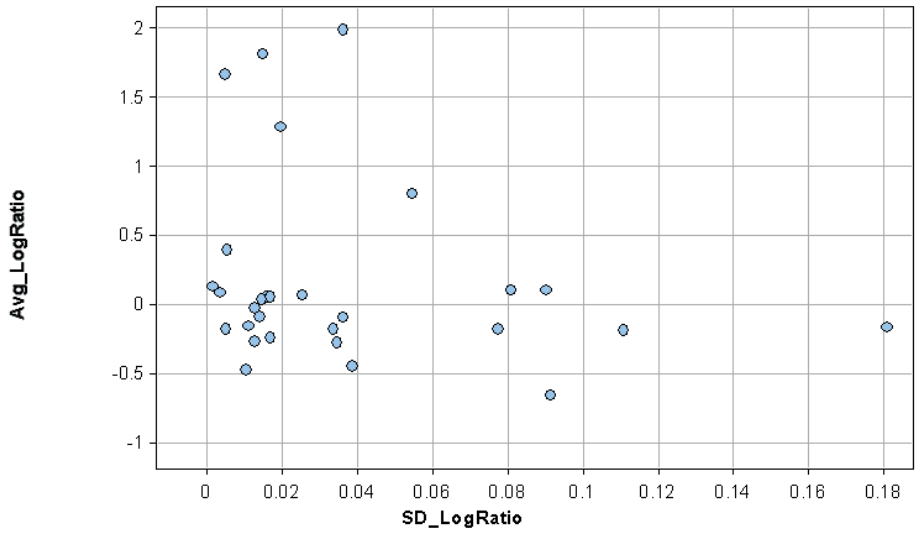


Figure 7b: Experiment B—Intra-array log ratio averages and SD's (median SD of log ratio = 0.017)

Figure 7b shows a plot of the intra-array, intra-gene average log ratios and log ratio SD's for the 29 genes on microarray_10_A01 from experiment B.

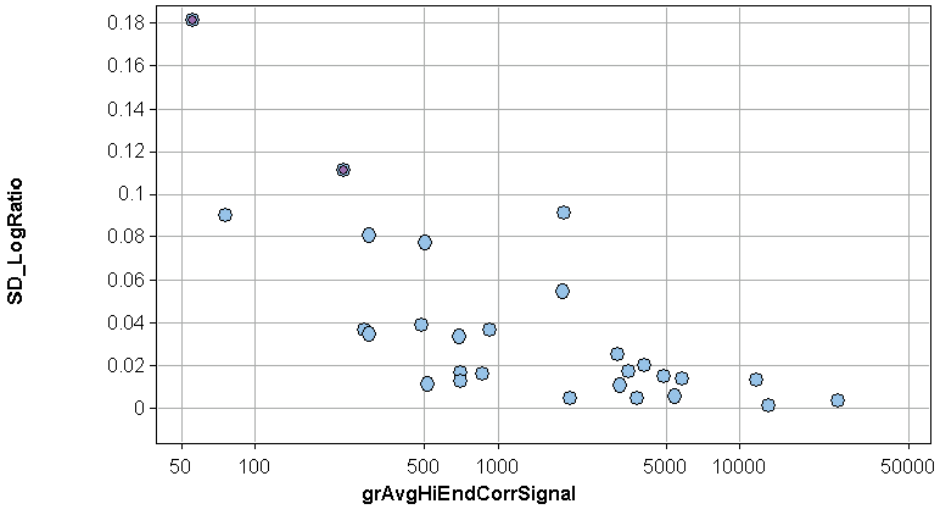


Figure 7c: Experiment B—Relation between log ratio SD's and high-end corrected signal.

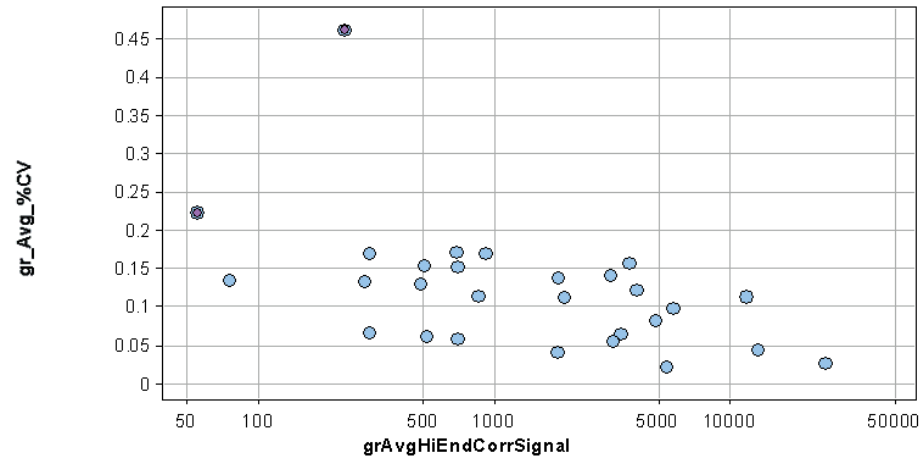


Figure 7d: Experiment B—Relation between signal %CV and high-end corrected signals

In Figures 7d and 7e signal %CV's are the average of the green %CV and the red %CV for each gene.

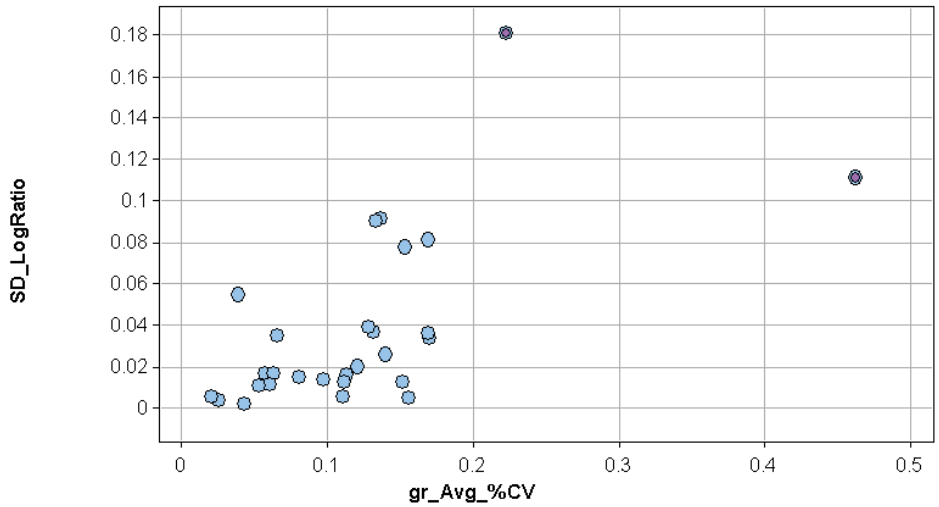


Figure 7e: Experiment B—Relation between log ratio SD's and signal %CV

All experiments

This standard deviation analysis was repeated for all microarrays. A summary of the intra-gene, intra-array log ratio SD's is shown in Table 6. The median SD of the log ratios for each experiment and array type is summarized, as well as the total number of genes used in the calculations. The small variation in the standard deviations of the log ratios reflected in Table 6 demonstrates the power of using log ratios for differential expression analyses. That is, even though the feature replicates may have fluctuating signals (Table 3), the replicate log ratios of the red/green signals are less variable.
The median of the SD's of the log ratios is presented by experiment and by microarray type (i.e. Self or Diff). The intra-array SD of the log ratios is calculated for each gene (of the set of 29 genes) for each array (n= 5 features/gene per array). Each experiment has 4 microarrays of a given type (i.e. Self or Diff). Thus, the total number of genes summarized is (29 genes/microarray)*(4 microarrays/type) = 116 genes/microarray type. The median SD is calculated from these 116 SD's of the log ratios. Microarrays with a high number of non-uniform outliers may have only 1 replicate (of max 5 feature replicates) of a gene as an inlier. Intra-array SD's are only calculated if there are at least 2 feature inliers for a given gene on a given microarray. Thus, those experiments with fewer than 116 genes per array type had genes with too few inliers.

Table 6: Summary of the SD's of log ratio statistics (intra-array, intra-gene)

	Self		Dye-swap	
Expt.	Genes / Array Type	Median SD_Log Ratio per Gene	Genes / Array Type	Median SD_Log Ratio per Gene
A	116	0.021	116	0.030
B	116	0.022	116	0.026
C	116	0.018	116	0.024
D	102	0.022	116	0.020
E	110	0.014	110	0.013

Calculation of composite log ratio statistics for differential expression microarrays (inter-feature, intra-gene, inter-array)

The final metric used to evaluate log ratio reproducibility is to evaluate all replicates (inter-feature) of a given gene (intra-gene), across all dye-swap microarrays (inter-array). The log ratios calculated from the Agilent Feature Extraction software are used for those microarrays with "+1" polarity (e.g. red = sample_1; green = sample_2). A "polarity-corrected" log ratio, calculated as (log ratio)*(-1), is used for those microarrays with "-1" polarity (e.g. red = sample_2; green = sample_1). This allows summary statistics to be performed on the "polarity" corrected log ratios across both polarities of dye-swap microarrays. For a given gene, in a given experiment, the average and the SD of the polarity-corrected log ratios is calculated. (See Figure 8) The maximum number of features used is (5 features/gene/microarray)*(4 microarrays) = 20 features/gene/experiment. If any features are flagged, there will be fewer than 20 features used/gene. As discussed above, all log ratio statistics are calculated using features that have not been flagged as non-uniform and saturated.

Single experiment

Figure 9 shows the average and the SD (n = 20 features) of the polarity-corrected log ratios ("PolLogRatio") for each of the 29 genes, from experiment B. As seen in Figure 7b, log ratios near zero typically have higher SD's. The median of these 29 SD's is 0.11.

The gene shown highlighted (blue star) in Figure 9 is discussed in the next section, "Producing high confidence in differential expression results".

All experiments

This analysis is summarized in Table 7 for each experiment, that is, across 29 genes and 4 dye-swap microarrays for each experiment, A through E. As discussed above and shown in Figure 8, each gene has 20 feature replicates that are used in this calculation. There will be fewer than 20 if outliers are omitted. The actual average number of features used for each SD calculation is shown in Table 7. The median SD is calculated from the 29 gene SD's for each experiment. The resulting inter-feature, inter-array SD's of the log ratios shown in Table 7 are higher than the intra-array SD's (Table 6), because additional noise from multiple microarrays (inter-array) and from dye-swap arrays (inter-polarity) contributes to the overall noise.

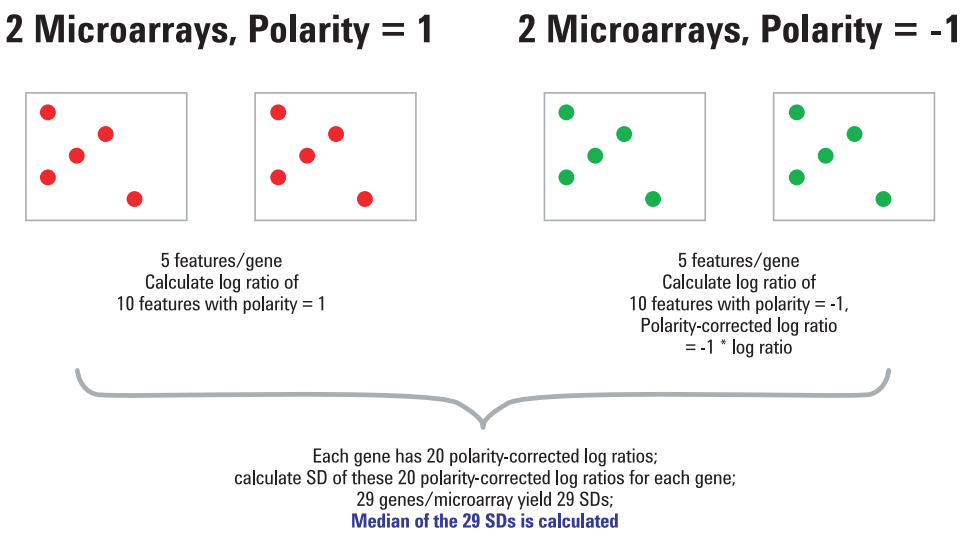


Figure 8: Summary statistics on "polarity" corrected log ratios

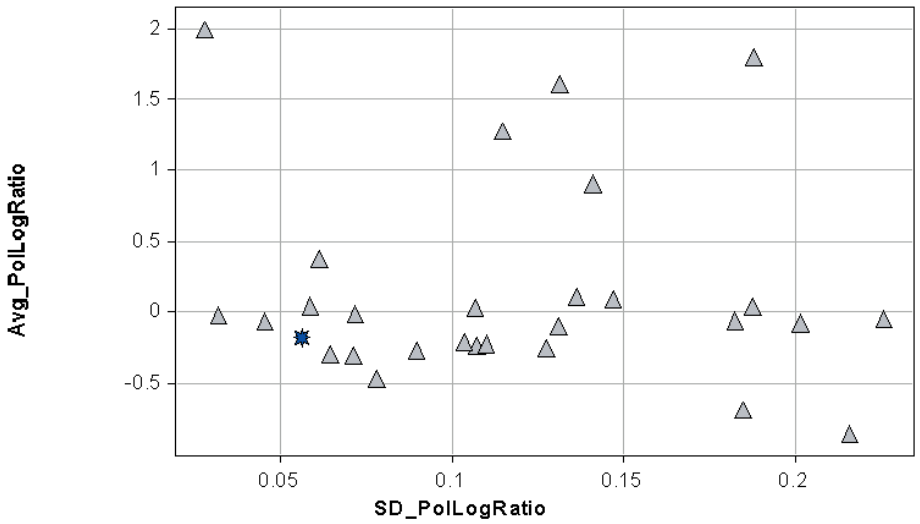


Figure 9: Experiment B—Composite log ratio statistics of dye-swap microarrays (intra-gene, inter-array, inter-polarity)

Table 7: Summary of composite log ratio statistics of dye-swap (diff) microarrays

Expt.	Avg. Features/ Gene	Median of SD_Log Ratio
A	19.8	0.0120
B	19.7	0.110
AB	39.4	0.127
C	19.3	0.081
D	19.4	0.056
E	17.8	0.126
*E	14.7	0.077

{*E= E, omitting microarray 22_A02} :

Effect of using biological replicates

Experiments A and B used different biological preparations of two samples for the {sample 1, sample 2} dye-swaps. Thus, an "inter-biological replicate" metric can be calculated. The SD of the log ratios was calculated for each gene by grouping all inlier features across the 8 dye-swap microarrays combined from experiments A and B (i.e. "inter-experiment" and "inter-biological replicate"; thus, max N = 8 * 5 = 40 features/gene). The median SD of the polarity-corrected log ratios = 0.127, which was very similar to the median SD's from the individual experiments A (median SD = 0.12) or B (median SD = 0.11).

Effect of omitting outlier and low signal microarrays on composite log ratio statistics

Experiment E demonstrates the power of evaluating the preparation process for each microarray before including it in the results. Microarray 22_A02 had an extremely high number of non-uniform flagged features (Table 1). Even though the flagged features are omitted from analyses, the high number of outliers may indicate that the overall microarray quality is low and that even the inlier features may be problematic. This type of problem can result from hybridization or wash artifacts. The inter-feature, intra-gene, inter-array statistics of the log ratios was repeated for experiment E, after omitting microarray 22_A02. The median of the SD of the log ratios decreased from 0.126 to 0.077 (Table 7).

Experiment A had two microarrays (8_A01, 8_A02) with lower signal than the other dye-swap microarrays. Analyses in Tables 4a,b demonstrated that omitting these 2 microarrays would decrease the inter-array signal %CV, as expected. However, the omission of these same 2 microarrays had much less of an impact on the inter-array SD's of the log ratios. The median SD of the log ratio = 0.105 when the 2 microarrays were omitted, compared with median SD = 0.120 with all microarrays of experiment A. This demonstrates the attenuating power of using ratios for differential expression analysis; that is, even though there may be large variations in the green or red signals of gene replicates, the variation of the resulting ratios will be much reduced (see also the discussion of Figure 7e).

Confidence in differential expression results

High reproducibility of the final calculated results is absolutely essential for the correct interpretation of differential gene expression. This paper uses custom calculations to show the reproducibility of Agilent microarray technology across multiple microarrays. Rosetta Resolver also uses the final results produced by the Agilent Feature Extraction software to evaluate reproducibility across multiple microarrays and ultimately the extent of gene expression. The Agilent results for single microarrays that are used by Rosetta Resolver include the log ratios and their associated log ratio errors. Below, a common threshold test using unweighted data is compared with significance tests (p-value) using unweighted data or Resolver weighted data to show the advantage of the latter for interpreting gene expression results.

Disadvantage of constant threshold tests

All the analyses in this paper used statistics of a population (average, SD, etc) without weighting the data points. With unweighted data points, a common method to determine if a log ratio is significantly different from zero is to use a constant threshold. For example, a constant threshold of significance may be to require the change in differential expression to be > 2-fold, which is the equivalent of requiring that the ratio be greater than 2 or less than 0.5, or requiring that the log ratio be greater than 0.30 or less than –0.30. This constant threshold method of determining significance of ratios can result in a high number of false calls at low signal ranges (false positives) and a high number of missed calls (false negatives) at moderate to high signal ranges (Figure 10).

Advantage of significance tests using Agilent error propagation and Rosetta Resolver weighted data

Rather than relying upon simple constant thresholds, the Agilent and Rosetta systems determine differential expression based upon signal/noise. Agilent Technology Feature Extraction software extends the accuracy of the data by propagating errors for each signal and background measured. Calculations using this propagated error and Rosetta's Universal Error Model yield a log ratio error and p-value for the log ratio of each feature. A log ratio error can be thought of as a SD about the log ratio. A p-value is a metric that indicates if the log ratio is significantly different from zero. A signal to noise (S/N) metric can be calculated for a single feature by dividing the absolute(log ratio) by the log ratio error. Similarly, an un-weighted, population-based S/N can be calculated for replicate features by dividing the absolute(average log ratio) by the SD of the log ratios. The error around log ratios is generally signal-dependent (see Figures 6b and 7c). That is, a log ratio of a certain value may not be significant, using log ratio errors or p-values, if the features have low or noisy signals (i.e. low S/N), or the same log ratio value may be significant if the log ratio was obtained with features at higher signals, or with features of lower signal and low noise (i.e. high S/N).

Rosetta Biosoftware's Resolver software improves the accuracy of the S/N analysis by weighting each log ratio, using Agilent Feature Extraction log ratio errors. The weighted log ratios, errors and p-values can be calculated for a single microarray or across multiple arrays, using Rosetta Resolver's weighted "combine" algorithms. The "combine" algorithms use the individual errors associated with each log ratio as well as reproducibility metrics of the log ratios (i.e. using intra-probe or intra-gene and/or inter-array replicates). The result of the combine analysis yields a single log ratio and p-value obtained from the replicate log ratios. Users can set the p-value threshold that determines significance for each experiment (e.g. typically choosing a p-value 0.01 or 0.001).

Example of the effect of using significance testing with or without weighting data on the interpretation of differential expression results

Comparing the results of these different types of significance testing, when used on a gene with a low signal and high reproducibility, demonstrates the power of the Agilent and Rosetta method to evaluate differential expression. These tests were used on a gene that can be seen in Figure 9 (see highlighted gene; blue star symbol). This gene has an unweighted average polarity-corrected log ratio = - 0.18 (i.e. a ratio of 0.66) and a SD of log ratios = 0.056 (N= 20; experiment B). A significance test using a constant threshold of 2-fold change (ratio > 2 or < 0.5, or log ratio > 0.30 or < –0.30) would not label this gene as significantly differentially expressed because the ratio of 0.66 is not less than 0.5 (or, the log ratio of –0.18 is not less than –0.30). The ratio of the unweighted log ratio to its error (i.e. the S/N) is equal to {absolute(-0.18)/0.056 } ~ 3.2. A p-value calculated from this data would be just < 0.01; thus, showing that this gene is borderline down-regulated.

However, when this gene was analyzed with Rosetta Resolver, its weighted log ratio was calculated to be -0.21, and its error was calculated to be 0.03. These values lead to a ratio of the weighted log ratio to its error (i.e. S/N) of 7. Therefore, the gene was determined to be very significantly down-regulated (p-value < 0.001), since this feature had good signal/noise and reproducibility across microarrays. Rosetta's weighted combine algorithms thus improve the sensitivity of detection (S/N = 7) compared to significance testing of unweighted data (S/N = 3.2). In contrast, the unweighted constant threshold method was not able to detect this gene as significantly differentially expressed.

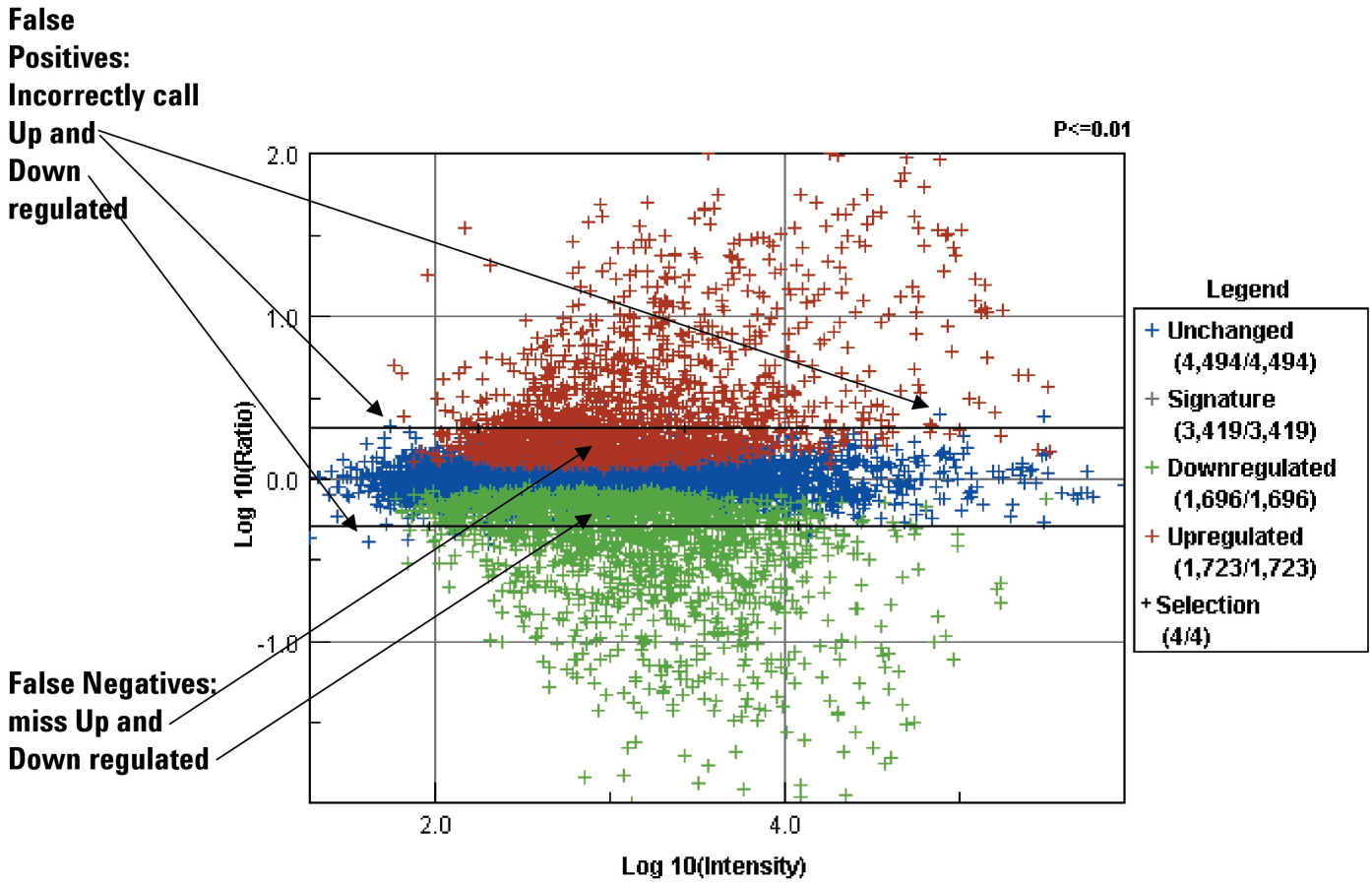


Figure 10: Disadvantage of a constant threshold to evaluate differential expression

Figure 10 shows a plot of the log ratio versus the log of the intensity for all the features from the four differential expression microarrays used in Experiment B. This is a Rosetta Resolver "combine" plot. The constant threshold is represented on the plot by the two lines at log ratios of +0.30 and –0.30. The genes whose p-values were calculated as less than or equal to 0.01 were labeled up-regulated (red) or down-regulated (green).

CONCLUSION

This paper has presented a series of custom metrics to evaluate the reproducibility of microarray technology. In particular, experiments were performed to illustrate the reproducibility of Agilent microarray technology. This paper has also shown that the combination of the reproducibility of Agilent microarray technology, Agilent error propagation, and Rosetta Resolver algorithms provide high confidence in differential expression results.

www.agilent.com
800 123-4567

Sales, Service and Support

If you do not have access to the Internet, one of these centers can direct you to your nearest representative:

United States
1 800 123 4567

Canada
1 877 123 4567
(905) 123 4567 (FAX)

Europe
(31 20) 123 4567
(31 20) 123 4567 (FAX)

Japan
(81) 123 45 6789
(81) 123 45 67 89 (FAX)

Latin America
(305) 123 4567
(305) 123 4567 (FAX)

Australia
1 800 123 4567
(613) 123 4567 (FAX)

New Zealand
0800 123 4567
64 4 123 4567 (FAX)

Asia-Pacific
(852) 123 4567
(852) 123 4567 (FAX)

Print Date: May 1, 2003
Publication Number: 5988-8611EN

Agilent Technologies