

MONITORING PERFORMANCE LEVELS IN MICROARRAY CORE FACILITIES



Wendell Jones and Laura Reid
 Expression Analysis, Inc., Durham, NC

Special Thanks to
 the Anonymous Laboratories
 Participating in this Program

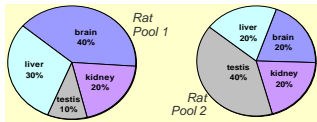
INTRODUCTION

The quality of microarray results has improved over the last few years. Recent publications have demonstrated good laboratory-to-laboratory reproducibility using the same RNA samples. Additional quality improvements are ongoing at several levels, including chip manufacturing, reagent manufacturing, and laboratory processes. However, due to the complexities of microarray-related experiments and the hyper-dimensionality of microarray measurements, it is very difficult to quantify laboratory performance or to even understand what indicates excellent versus good results.

We have conducted a multi-laboratory study over time of rat mixed tissue reference RNA materials using the Affymetrix platform. During each of three rounds of testing, approximately 16 participating laboratories prepared targets and hybridized three replicates of two biologically different RNA sources to six Rat230.2 GeneChips. During our analysis of the entire experiment, we observed, along with our study partners, characteristics of lab performance that translate into larger, higher-quality lists of differentially expressed genes. These relatively simple lab performance metrics and characteristics will be described in this poster as well as their association with differential expression sensitivity and specificity.

DATA SET

Participating laboratories received six Rat230.2 Affymetrix GeneChips and three replicates of two different RNA sources. The samples were derived from pools of rat total RNA derived from four rat tissues (Thompson, 2005 *Nucl. Acids Res.* 33:e187).



The labs prepared and hybridized targets from the six RNA samples using their own protocols.

Number of Labs	Reverse Transcriptase	RT RNA Polymerase	CDNA/RNA Purification
3	Invitrogen	Enzo	PLG/Qiagen
1	Invitrogen	Affymetrix	PLG/Qiagen
6	Invitrogen	Enzo	Affymetrix
5	Affymetrix	Affymetrix	Affymetrix
3	Ambion	Ambion	Ambion

The data files from each laboratory were collected for analysis at a central location. Signal values were calculated using Microarray Suite version 5 (MAS5) algorithms. In order to monitor variation over time, the testing was repeated every three months.

Round	Number of Labs	Number of Hybridizations
Round 1 June 2004	13	15
Round 2 Sept. 2004	16	15
Round 3 Dec. 2004	15	15

Data set of 264 hybridizations

DATA ANALYSIS

Good microarray assays require both sensitivity and specificity in order to reliably detect a certain threshold of differential expression. In our study, we used the following definitions (which include a somewhat arbitrary 2-fold threshold):

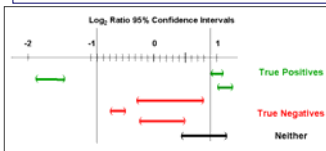
Sensitivity = the ability to detect a ≥ 2 -fold change in RNA abundance between sample groups
Specificity = the accuracy of a ≥ 2 -fold assessment relative to transcripts where there is strong evidence of less than a 2-fold change

These assessments of lab performance require prior knowledge of expression levels for a large number of transcripts, which we can define based on consensus values from the overall proficiency testing data.

Definition of Expected Results

Using MAS5-based results, we defined a large subset of the 31,000 probe sets on the Rat230.2 GeneChip as either a "True Positive" (~2000 total) or "True Negative" (~27,000 total) for ≥ 2 -fold change. These distinctions were based on confidence intervals from individual ANOVA analysis (gene-by-gene) of $\text{Log}_2(\text{Signal})$ results of 168 hybridizations. The model examined $\text{Log}_2(\text{Signal})$ intensity as a function of RNA pool, protocol, lab, round, chip lot, etc. To avoid bias, the hybridizations were stratified and a subset randomly chosen from a given class yielding a similar number from each of the predominant protocols.

True Positives = probe sets with strong evidence of ≥ 2 -fold change in abundance
 The 95% confidence interval of the $\text{Log}_2(\text{Ratio})$ RNA pool effect parameter does not overlap the interval [0.9,0.9]. A threshold of 0.9 was chosen rather than 1 due to the mixture used for the rat MTRRM samples and to the general compression of signal observed in microarray assays.
True Negatives = probe sets with little evidence of ≥ 2 -fold change in abundance
 The 95% confidence interval for $\text{Log}_2(\text{Ratio})$ completely within [-0.9, -0.9].



Calculation of Performance Scores

A list of differentially expressed (DE) genes was calculated for each of the lab-round data sets in the proficiency testing program. Probe sets with ≥ 2 -change and $p < .001$ using a two-sample t-statistic were included on the DE gene lists. Sensitivity and Specificity Scores were then calculated based on the composition of the DE List to the previously defined lists of True Positives and True Negatives.

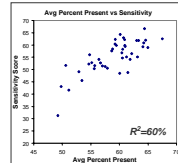
$$\text{Specificity Score} = 1 - \frac{\# \text{TrueNegatives on DE List}}{\text{Size of DE List}}$$

$$\text{Sensitivity Score} = \frac{\# \text{TruePositives on DE List}}{\text{Size of TruePositives List}}$$

SENSITIVITY

We identified three primary factors that were predictive of the Sensitivity Score and which are consistent with our current understanding of the microarray assay.

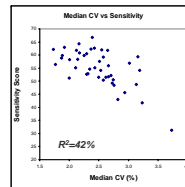
1- Percent Present



Average Percent Present:
 • GCOS generates a Percent Present for each hybridization.
 • The Average Percent Present is calculated by averaging the Percent Present value from each of six hybridizations from the same laboratory within a given round of testing.

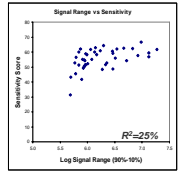
2- Median CV

Median CV Score:
 • Calculate a coefficient of variation (CV) of the log_2 MASS Signal values for each probe set across the three replicates for both pools of RNA.
 • Determine median CV value across the 31,000 probe sets representing the replicates in each pool.
 • Average the median CV score for the two pools.



3- Range of Signal

Average Range of Signal:
 • From the distribution of the log_2 MASS Signal values, for each hybridization, compute the 90th and 10th percentile values.
 • Determine the difference between the 90th percentile and the 10th percentile Signal for each hybridization.
 • Calculate the mean difference for the six hybridizations in a group.



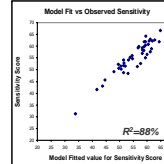
Combination of Predictive Factors

An important aspect of these three measures is that they are not redundant and have much more predictive power together than they do individually, as evidenced by having a multiple R^2 of 88% in a multiple regression model.

Regression Model
 Dependent Variable: Sensitivity Score

Ind. Var.	Est. Coeff	t-test	p-value
Constant	5.35	0.73	0.472
Med. CV%	-19.57	-2.81	<0.0001
Range of Signal	8.95	7.53	<0.0001
Percent Present (%)	0.347	2.74	0.0093

$R^2 = 88.1\%$
 R (adjusted) = 87.1%
 $s = 2.462$

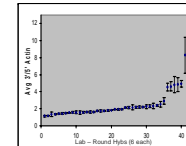
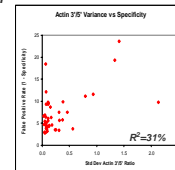
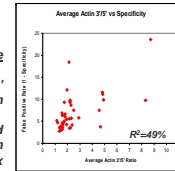


SPECIFICITY

We identified a primary factor that was predictive of the Specificity Score. This measure was also consistent with our current understanding of the microarray assay.

1- Actin 3/5'

Average and Variance in the Actin 3/5' Ratio:
 • GCOS generates Actin 3/5' ratio metric for each hybridization.
 • Compute the average and standard deviation of the Actin 3/5' ratios across all six hybridizations in a given round.



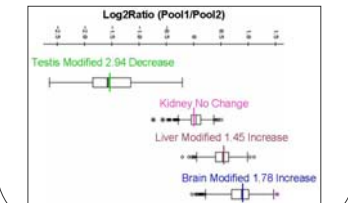
Average Value or Variance

We examined both the average value and the standard deviation, or variance, of the Actin 3/5' ratios for each set of six lab-round hybridizations. In our experience, these two values are greatly confounded, so that the correlation to Specificity Score may be related to the variation, rather than the average value. Variation in the transcription reactions may induce false differences in signal intensity for certain probe sets that are more 5' oriented and induce a false indication of differential expression between sample groups. This can happen even in normal circumstances but may be more frequent when the reactions are more variable.

WHAT IS TRUTH?

Due to the variable nature of RNA samples, it is difficult to define the expected or "true" results in microarray assays. In this analysis, we relied on the consensus estimates which may include unknown biases.

Thompson et al. (2005, *Nucl. Acids Res.* 33:e187) have developed an alternative method of defining expected results using a set of approximately 200 "tissue-selective" genes that are predominantly expressed in one of the four tissues in the rat MTRRM source pools. The investigators initially expected 4-fold decrease, 2-fold increase, 1.5-fold increase and no change in expression in the testis, brain, liver and kidney tissue-selective genes, respectively, based on the compositions of the two source pools. However, these ratios were later modified to correct for low level expression in the other three tissues. The data in this study agrees well with the modified ratios for the tissue selective genes. Future analyses may rely on these tissue-selective genes to distinguish true positives and true negatives.



CONCLUSIONS

- Given one characterization of sensitivity and specificity (based on an ANOVA analysis of 168 hybridizations equally representing different protocols across multiple laboratories, chip lots, and hybridization conditions), we have detected four primary factors that indicate good processing of the rat MTRRM samples when using MAS5 signal algorithms.
- Three metrics, average percent present, range of (log) Signal values and median CV, are predictive of sensitivity (i.e., ability to detect true positives) and account for more than 88% of its variation
- High average (and variable) 3/5' Actin Signal values are associated with lower specificity (i.e., ability to avoid false positives).
- The metrics are intuitively related to the microarray assay, as good quality would be expected from labs detecting more transcripts with highly repeatable Signal results spanning a larger dynamic range.
- Together, these four metrics can be used to monitor performance of microarray core facilities.

OTHER QC METRICS

Several other common Affymetrix Quality Control (QC) metrics were examined in terms of their ability to predict Sensitivity and Specificity Scores. Although some of these QC metrics appear correlated, their utility relative to those measures already identified is not clear except in special circumstances.

Correlation of QC Metrics to Sensitivity and Specificity Scores

QC Metric	Sensitivity Score r & (R^2)	Specificity Score r & (R^2)
Average Background	-.32 (10%)	-.21 (4%)
Variance in Background	-.22 (5%)	.03 (1%)
Average Scaling Factor	-.28 (8%)	-.12 (1%)
Variance in Scaling Factor	-.35 (12%)	-.15 (2%)
Average Percent Present for External Controls (i.e., Spike-ins)	.56 (31%)	-.16 (3%)
Variance in Percent Present for External Controls (i.e., Spike-ins)	-.39 (15%)	-.04 (<1%)
Average Noise	-.25 (6%)	-.18 (3%)
Variance in Noise	-.16 (3%)	.31 (11%)
Average Pairwise Correlation of Replicates	.57 (32%)	-.07 (1%)