

# COMPARING MICROARRAY DATA GENERATED IN DIFFERENT LABORATORIES

Laura Reid<sup>1</sup>, Wendell Jones<sup>1</sup>, Donald Cox<sup>1</sup>, Thomas Goralski<sup>1</sup> and Andrew Brooks<sup>2</sup>

<sup>1</sup>Expression Analysis, Inc., Durham, NC; <sup>2</sup>University of Rochester Medical Center, Rochester, NY

Winner of Oral Poster Competition  
2004 meeting of the Association of  
Biomolecular Resource Facilities

## INTRODUCTION

Although microarray facilities use similar RNA labeling and hybridization procedures, minor differences in lab protocols can affect the expression results and confound data comparisons. At Expression Analysis (a microarray service provider), we are collecting and analyzing microarray results generated in different laboratories using the same RNA samples. This program allows us to determine the level of variability between laboratories. It also helps develop standardization methods for confirming the proficiency of individual laboratories and for evaluating the quality of specific data sets.

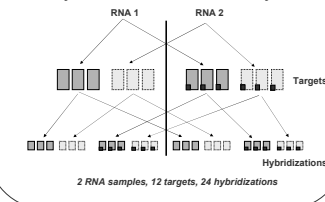
Our initial research compared data generated in two microarray laboratories using a pair of mouse RNA samples (experiment 1). A later study analyzed a pair of human RNA samples and results from a third lab (experiment 2). All labs demonstrated good reproducibility, based on the consistency of data between the three replicates of each RNA sample.

Interestingly, a subset of the 25-mer probes that contain a particular sequence element had consistently higher hybridization intensities in one laboratory. The comparability of the data between labs was evaluated by comparing lists of differentially expressed genes between the two RNA sources. The comparability of the microarray results was apparently related to the nature of the RNA samples.

We are now extending this research to other microarray facilities. As before, replicate targets from two RNA samples are processed and hybridized in multiple laboratories. The program can be used to evaluate the performance of proposed RNA standards and to develop distribution curves for quality control metrics and expression results for a subset of transcripts.

## EXPERIMENTAL DESIGN

Each experiment compared data generated at two microarray facilities using two RNA samples. Three biotin-labeled cRNA targets were prepared for each sample at both facilities. After the first use, the hybridization cocktails were exchanged for a second hybridization in the other laboratory.



## ANALYSIS METHODS

Reproducibility was examined by comparing hybridization intensities of individual probes (derived from CEL files) and by looking for agreement in the Signal and Detection Call data (derived from CHP files).

**Correlation.** The intensity of each probe cell was graphed against its corresponding intensity in a replicate hybridization. The Pearson Product moment correlation was computed after curvature (bias) was removed using a loess smooth on an M vs. A plot.

**False Positives.** This measure is based on the "Detection Call" and "Signal" values calculated by MAS 5.0 (Affymetrix). False positives are defined here as probe sets with a greater than 2-fold change in signal between replicates. Affymetrix control probe sets and transcripts called "Absent" in both replicates are excluded. The signals were censored at the low end (i.e. values below 2<sup>0</sup> were treated as 2<sup>0</sup>).

Comparability was evaluated by comparing the fold change differences for individual probe sets and by looking for agreement in the lists of differentially expressed genes.

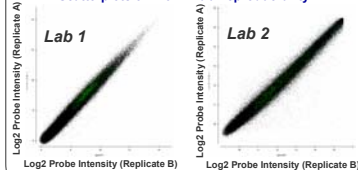
**Log Ratio Comparison Graph.** Signal for each replicate was computed using a modified PDNN method (Zhang et al. Nature Biotech. 21:818-821, 2003) and averaged for each RNA sample. The log<sub>2</sub> ratio of the average signal for RNA 1 to the average signal for RNA 2 was computed for each transcript. The results from one laboratory are plotted against corresponding values in the other lab. A log<sub>2</sub> ratio of 1 is equivalent to a fold change of 2.

**Venn Diagram.** The number of probe sets identified as having a 2-fold difference in average signal between the RNA 1 and RNA 2 samples. Average signal is calculated as in the log ratio graph.

## PROTOCOL DIFFERENCES

The two labs in experiment 1 had different stain and scan procedures. In Lab 1, the GeneChips were scanned once after signal amplification (two rounds of SAPE stain plus a biotin-conjugated antibody). The scanner PMT setting was ~1,800 CPSM and data analysis was generally performed with a target intensity of 500. In Lab 2, the scanner was set at a higher PMT setting (~18,000 CPSM). GeneChips were scanned two times, once before and once after signal amplification. Only data from the second scan were used in this study. Both labs used GeneArray 2500 scanners. The protocol and equipment differences are reflected in the within lab reproducibility graphs and in the hybridization quality parameters.

### Scatterplots of Within Lab Reproducibility



## EXPERIMENT 1

In experiment 1, the two RNA samples were isolated from the brains of wild type and knockout mice. Targets were hybridized to Mouse U74Av2 GeneChips (Affymetrix).

### Hybridization Quality Parameters

Lab 1, Experiment 1	Noise	Scaling Factor	Back-ground	Percent Present	3'/5' Actin	3'/5' GAPDH
Minimum Value	2.57	5.35	64.00	42.60	1.81	1.48
Maximum Value	3.56	6.45	75.00	48.50	4.80	2.37
Average Value	2.96	5.86	70.33	45.20	3.63	1.82
Coefficient of Variation	11.3%	7.0%	5.7%	4.8%	31.3%	18.9%

Lab 2, Experiment 1	Noise*	Scaling Factor*	Back-ground*	Percent Present	3'/5' Actin	3'/5' GAPDH
Minimum Value	37.20	0.20	412.00	45.90	2.20	0.68
Maximum Value	42.10	0.89	779.00	51.10	3.25	1.84
Average Value	39.37	0.34	677.50	48.12	2.69	1.04
Coefficient of Variation	5.7%	78.8%	20.1%	4.1%	16.3%	45.7%

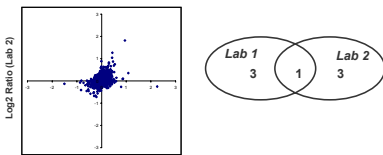
\*The hybridization quality parameters were computed at a target intensity of 500. This setting generated atypical Noise, Scaling Factor and Background measures for Lab 2, which uses a high PMT setting during scanning.

### Reproducibility Measures

Experiment 1	Within Lab 1		Within Lab 2		Between Labs	
	RNA 1	RNA 2	RNA 1	RNA 2	RNA 1	RNA 2
Average Correlation Coefficient	0.976	0.973	0.988	0.986	0.911	0.895
Average False Positives*	1.7%	1.7%	1.2%	1.5%	9.5%	8.6%

\*Signals generated by Microarray Suite 5.0 algorithms (Affymetrix).

### Comparability



Signals generated by PDNN method (Zhang et al. Nature Biotech. 21:818-821, 2003).

## EXPERIMENT 2

In Experiment 2, the two RNA samples were isolated from human lymphoblasts before and after LPS stimulation. Targets were hybridized to human U95Av2 GeneChips (Affymetrix).

### Hybridization Quality Parameters

Lab 1, Experiment 2	Noise	Scaling Factor	Back-ground	Percent Present	3'/5' Actin	3'/5' GAPDH
Minimum Value	2.08	2.16	54.13	41.07	1.15	1.20
Maximum Value	2.63	3.46	75.62	44.60	1.31	1.35
Average Value	2.38	2.74	64.61	42.98	1.23	1.26
Coefficient of Variation	6.8%	13.8%	8.5%	2.7%	3.9%	3.9%

Lab 3, Experiment 2	Noise	Scaling Factor	Back-ground	Percent Present	3'/5' Actin	3'/5' GAPDH
Minimum Value	1.74	1.65	45.91	40.34	1.21	1.08
Maximum Value	2.39	3.07	66.94	45.39	1.75	1.67
Average Value	2.07	2.09	58.19	43.25	1.33	1.24
Coefficient of Variation	8.8%	17.6%	10.7%	4.3%	9.5%	10.7%

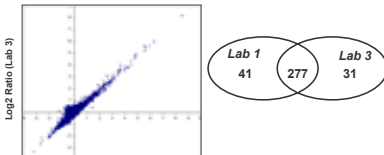
\*The hybridization quality parameters were computed at a target intensity of 300.

### Reproducibility Measures

Experiment 2	Within Lab 1		Within Lab 3		Between Labs	
	RNA 1	RNA 2	RNA 1	RNA 2	RNA 1	RNA 2
Average Correlation Coefficient	0.991	0.991	0.993	0.994	0.888	0.916
Average False Positives*	0.8%	0.9%	0.3%	0.3%	2.0%	3.0%

\*Signals generated by Microarray Suite 5.0 algorithms (Affymetrix).

### Comparability

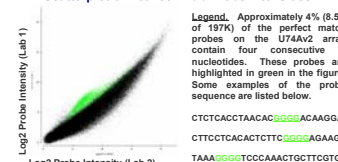


Signals generated by PDNN method (Zhang et al. Nature Biotech. 21:818-821, 2003).

## PROBE-SPECIFIC BIAS

Scatterplots of the probe cell intensities in experiment 1 identified a subset of 25-mer probes that had consistently higher intensities in lab 1 than lab 2 (highlighted in green below). Sequence analysis revealed that the affected probes all contained four consecutive G nucleotides (which hybridize to four consecutive biotin-labeled C ribonucleotides in the targets). The position of this common sequence element varied within the probes. The "multiple-G" phenomenon was not detected in experiment 2, but has been observed in other data comparability studies.

### Scatterplot of Between-Lab Probe Intensities



Legend. Approximately 4% (8.5K of 197K) of the perfect match probes on the U74Av2 array contain four consecutive G nucleotides. These probes are highlighted in green in the figure. Some examples of the probe sequence are listed below.

The "multiple-G" effect was detected in data sets hybridized in both lab 1 and lab 2. These results suggest that the hybridization changes are related to a protocol difference during target preparation.

### Comparisons that Demonstrated Multiple-G Effect

Data Set 1 versus Data Set 2		Multiple-G Detected?
Lab 1 Targets/Lab 1 Hybs	Lab 2 Targets/Lab 2 Hybs	Yes
Lab 1 Targets/Lab 2 Hybs	Lab 2 Targets/Lab 2 Hybs	Yes
Lab 1 Targets/Lab 1 Hybs	Lab 1 Targets/Lab 2 Hybs	No

## CONCLUSIONS

- Each lab generated reproducible data with high quality hybridization parameters.
- Probe cell intensities were highly correlated and false positive rates were relatively low between labs, even when different protocols and scanner settings were used in experiment 1.
- The differential gene lists in experiment 2 had almost 90% agreement (277 of ~310) between laboratories.
- Data comparability was easier to quantify when the input RNA samples were known to contain a large number of differentially expressed genes at high magnitude fold changes.
- In some comparisons, probes with common sequence elements had consistent differences in the hybridization intensities between labs. This finding suggests that differences in target preparation can result in both lab-specific and probe-specific bias.

## DISTRIBUTION OF EXPRESSION CHANGES

Only one probe set was identified as differentially expressed in both labs in experiment 1, while more than 200 probe sets were in agreement in experiment 2. This difference in comparability is most likely related to the protocol differences and the nature of the RNA samples, rather than the quality of the data generated. The two mouse brain RNA samples used in experiment 1 had relatively few differentially expressed genes, with generally less than 2-fold changes. In contrast, the two human lymphocyte RNA samples used in experiment 2 had dozens of differentially expressed genes, many with more than 10-fold changes.

