

COMPARABILITY OF AFFYMETRIX DATA GENERATED IN MULTIPLE LABORATORIES

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



Special Thanks to
 the Anonymous Laboratories
 Participating in this Program

ABSTRACT

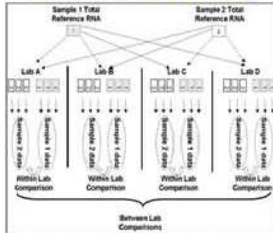
Much of the promise of genomic research relies on the comparison of microarray data generated in multiple facilities. However, the comparability of gene expression results is affected by a variety of technical, biological and data analysis factors. In order to quantify and reduce technical variation, we established a year-long testing program to compare microarray data generated in seventeen different laboratories using the Affymetrix platform and identical RNA samples.

The data sets were evaluated with a number of quality, sensitivity, reproducibility and comparability metrics. All of the participating laboratories produced high quality results that repeatedly demonstrated good agreement between the facilities and within the same laboratory over time. These findings are encouraging for data comparisons involving multi-site consortiums or central databases. They also help establish quality review procedures that may be required in establishing microarray standards or submitting data for regulatory review.

EXPERIMENTAL DESIGN

Participating laboratories received six rat 230.2 Affymetrix GeneChips and three replicates of two different RNA sources. The samples were derived from pools of rat total RNA and included unique combinations of bacterial polyadenylated transcripts. The labs prepared targets from the six RNA samples using their own protocols. After hybridization, the image files from each laboratory were collected for data analysis at a central location.

Example of One Round of Testing



In order to monitor variation over time, the testing is repeated every three months. Results from round 1 (June 2004) and round 2 (September 2004) are presented in this poster. Thirteen to sixteen labs participated in each round of testing.

Participating Labs



QC METRICS

The hybridization quality metrics calculated by the MAS 5.0 software were generally consistent between laboratories and over time.

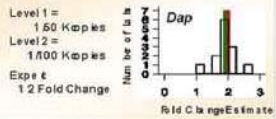
Summary of Quality Control Metrics

	Rd. 1 Median (IQR) n = 78 samples	Rd. 2 Median (IQR) n = 8 samples
Hybridization QC Metric	2.2 (1.4-3.3)	3.0 (1.5-7.0)
Skilling Factor	1.8 (0.86-2.0)	1.9 (0.7-3.3)
Background	7.2 (4.6-2.3)	9.5 (4.8-8.8)
Percent Present	60 (47-67)	58 (47-70)
Average Signal (R)	100.9 (91-22)	103.9 (89-11.8)
3.5 GA PDI	1.1 (0.9-1.1)	1.1 (0.9-1.1)
3.5 Actin	1.9 (1.1-1.7)	1.8 (1.1-1.3)

SENSITIVITY

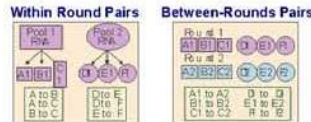
Three polyadenylated transcripts of bacterial genes were added to each sample at known concentrations. Using multiple probe sets, the sensitivity of each laboratory was evaluated by comparing the expected to the observed fold changes for these spikes.

Example of Lab Distribution of Sensitivity Results



REPRODUCIBILITY

Reproducibility was evaluated by comparing the Signal values for the same probe set in pairs of technical replicates. Measurements for each laboratory were averaged from six pairs of replicate samples.



Two metrics were used to measure reproducibility.

Signal Correlation For each probe set the log₂ Signal from one hybridization was compared against its corresponding intensity in a replicate hybridization. The Pearson product-moment correlation coefficient (R) was calculated. A Pearson product-moment correlation coefficient of 0.95 or greater was considered a high level of reproducibility.

Probe Position A set of ~2000 transcripts that were detected in >20% of the hybridizations was selected. A probe position was defined as a probe set within a gene that had a fold change in Signal between technical replicates. The signals were ordered at the low end (i.e. values below 2) and at the high end (i.e. values above 2).

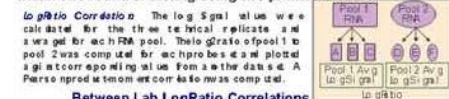
Summary of Reproducibility Measures using MAS 5.0

Reproducibility Measure	Within Round		Between Rounds
	Rd. 1 vs Rd. 2 Median (IQR) n = 78 samples	Rd. 1 vs Rd. 2 Median (IQR) n = 8 samples	Rd. 1 vs Rd. 2 Median (IQR) n = 8 samples
Signal Correlation	0.96 (0.93-0.97)	0.96 (0.93-0.98)	0.94 (0.89-0.97)
Probe Position	1.4% (0.5-6.6)	1.9% (0.4-3.9)	3.2% (0.7-11.8)

The reproducibility measures were dramatically improved (correlation coefficient ~ 0.95; false positives < 0.01%) when Signals were generated by the PDNN method

COMPARABILITY

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. LogRatio data from each lab was compared to every other lab within one round of testing using two primary metrics:

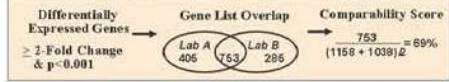


Between Lab LogRatio Correlations

13 data sets = 78 comparisons
 12 green values from Lab A

Comparability Score Differentially expressed genes were identified as having > 3 fold changes in a two-sample t-test of gene lists. For each pair of data sets, a percent score was calculated to indicate the commonality of the gene lists.

Example of Between Lab Comparability Scores



Even with the arbitrary gene list thresholds, a relatively few technical replicates and MAS 5.0 algorithms, we observed approximately two-thirds of the gene lists are identical between labs.

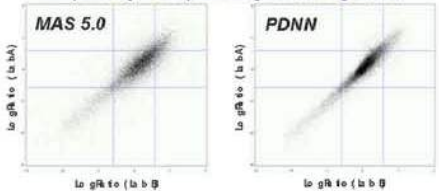
Summary of Comparability Measures

Comparability Measure	Between Labs		Within Lab
	Rd. 1 vs Rd. 2 Median (IQR) n = 78 samples	Rd. 1 vs Rd. 2 Median (IQR) n = 8 samples	Rd. 1 vs Rd. 2 Median (IQR) n = 8 samples
LogRatio Correlation	0.8 (0.6-0.9)	0.82 (0.6-0.9)	0.82 (0.7-0.9)
MA Score	66% (59.7)	64% (48.7)	70% (61.7)
PDNN Score	78% (66-8)	77% (4.6-8)	82% (64-8)

DATA ANALYSIS DIFFERENCES

Although lab protocols differed, data analysis methods did not vary between laboratories. For most metrics, we relied on Signal and Detection Call values generated with the Affymetrix Microarray Suite version 5.0 software (MAS 5.0). Interestingly, the reproducibility and comparability measures were dramatically improved when Signals were generated by the Positional Nearest Neighbor (PDNN) method**.

Comparability Scatterplots using Different Algorithms

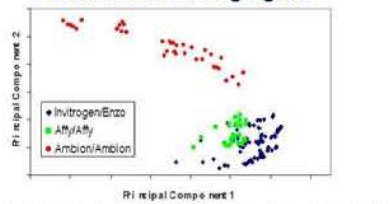


Legend: For each transcript, the log₂ Signal (a vs a) Signal pool 1 to a vs a Signal pool 2) was plotted against the log₂ Signal (a vs a) Signal pool 1 to a vs a Signal pool 2) with the MAS 5.0 or PDNN algorithms. A set of ~20,000 transcripts that were detected in >20% of the hybridizations are presented.

PROTOCOL DIFFERENCES

In order to capture a realistic range of results, we did not attempt to standardize the participating laboratories to specific procedures. Labs used the same protocol in all rounds of testing, but different labs used different protocols and reagent sources. Operators, enzyme lots and GeneChip lots also varied during the testing program. These sources of variability are reflected in the principal component analysis (PCA).

PCA with Protocols Highlighted

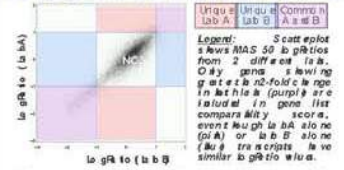


Legend: Each point represents one hybridization from the round of testing. Points are highlighted by color to indicate the reagent source and associated protocols used during the hybridization. The x-axis is the first principal component and the y-axis is the second principal component.

GENE LIST LIMITATIONS

We used Gene List overlap as one measure of comparability between laboratories and over time. However, this measure typically underestimates agreement in data sets in part due to the arbitrary fold change and statistical significance thresholds required to create gene lists. As shown in the scatterplot below, many transcripts with similar estimated LogRatio values do not meet the gene list criteria.

Fold Change Thresholds Underestimate Agreement



Close examination of the gene lists revealed that most of the "unique" genes (i.e. present on one but not both gene lists for a pair of data sets) showed concordant expression patterns.

Many "Unique" Genes show Concordant Changes



CONCLUSIONS

Our results demonstrate that similar biological results can be identified when comparing microarray data generated in different laboratories over time.

- Multiple labs repeatedly generated similar QC metrics, sensitivity, reproducibility and comparability measures.
- Between-round metrics indicate consistent data over time.
- Using MAS 5.0 algorithms, two-thirds of the differentially-expressed genes identified in one facility were also present on a similar list from another facility. The comparability scores increased to three-quarters of the gene lists when using PDNN.
- Comparability scores were higher when comparing gene lists generated at different times in the same laboratory, than when comparing data sets from different labs.
- Reproducibility measures also improved when using PDNN methods. And were only slightly impacted in between-round comparisons.
- Simple comparisons of differentially expressed gene lists underestimate the comparability of data sets.
- There are measurable, protocol-related impacts to a subset of probe sets.
- The data sets showed good agreement, even though the participating labs used different protocols, instruments and reagent sources.