AFFYMETRIX®

Technical Note

New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays

Affymetrix has designed new algorithms for monitoring GeneChip® expression data. These *statistical algorithms*, created in response to customer input, were designed to accommodate the typical distribution of data found in microarray experiments. The new *statistical algorithms* employ standard statistical techniques and are optimized to accommodate advancements in array and probe selection technology. They provide accurate, high-quality analysis for GeneChip® array data. This new, statistically based approach provides:

- Calculation of statistical significance for detection and change calls (*p*-values) and confidence limits for log ratio values (fold change).
- Easily tunable parameters that enable the user to vary the stringency of the analyses.
- Elimination of negative expression values observed with the *empirical algorithms*.
- Easily referenced standard statistical techniques.

This technical note reviews the design and testing for Affymetrix® new *statistical algorithms* and explores performance characteristics of the *statistical algorithms* versus the previous *empirical algorithms*.

EXPERIMENTAL DESIGN: HOW THE NEW ALGORITHMS WERE SELECTED AND OPTIMIZED

To select components for the new algorithms and test for significance and sensitivity, a "training" data set was required. To conduct this comprehensive testing, each transcript group was spiked into a labeled mixture of RNA from a tissue source in an experimental design known as a Latin Square. A Latin Square is used to accurately monitor the detectability of transcripts over a range of concentrations. It also allows the statistical analysis of patterns and variability in repeated measurements in a systematic fashion, thus revealing patterns in the data and allowing rigorous comparisons. The Latin Square experimental design used in the development of the *statistical algorithms* enabled thorough testing of a large set of transcripts over a broad range of concentrations (Figure 1).

Figure 1. Each row (numbered 1 through 14) represents a GeneChip® experiment. Each column (labeled A through N) represents a distinct set of transcripts. Each set contains a pool of transcripts distinct from the transcripts in any other set. In other words, no transcript is present in more than one set. Additionally, no set of transcripts is present at the same concentration in more than one experiment. For example, in Experiment 2 (shaded in blue), every transcript in Set J (boxed in red) is spiked along with the complex background into the target hybridization mixture at 128 pM. Then in Experiment 3, the same set of transcripts, (i.e., Set J boxed in green) is spiked in at 256 pM, twice the previous concentration, and is present at a different concentration in each subsequent experiment.

Latin Square used in algorithm development.

Groups of Transcripts
pM Concentration

GeneChip® Experiment

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
2	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024	0
3	0.5	1	2	4	8	16	32	64	128	256	512	1024	0	0.25
4	1	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5
5	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1
6	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2
7	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4
8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8
9	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16
10	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32
11	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
12	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128
13	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
14	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

The Latin Square experimental design was used extensively in the algorithm development process to test a wide range of data sets, including transcript groups of *E.coli, S. cerevisiae* and *H. sapiens.* In the human Latin Square, each transcript group was designed to contain one distinct human transcript from 0 to 1024 pM in concentration. These were spiked into a labeled mixture of human RNA where these 14 transcripts showed no expression. In total, 12 transcripts were used to compute the results; two of the transcripts were removed from the final calculations due to low quality. In the yeast Latin Square, each transcript group contained about eight different transcripts, labeled and spiked into a mixture of labeled human RNA. After hybridization to Human Genome U95A or Yeast Genome S98 arrays respectively, Affymetrix® Microarray Suite (MAS) 4.0 containing the *empirical algorithms* and MAS 5.0 containing the new *statistical algorithms* were used to analyze the data.

In addition to Latin Square experiments, more conventional data sets were generated and analyzed where RNA from different sources was labeled and hybridized to GeneChip® probe arrays, followed by analysis with MAS 4.0 and MAS 5.0.

Results

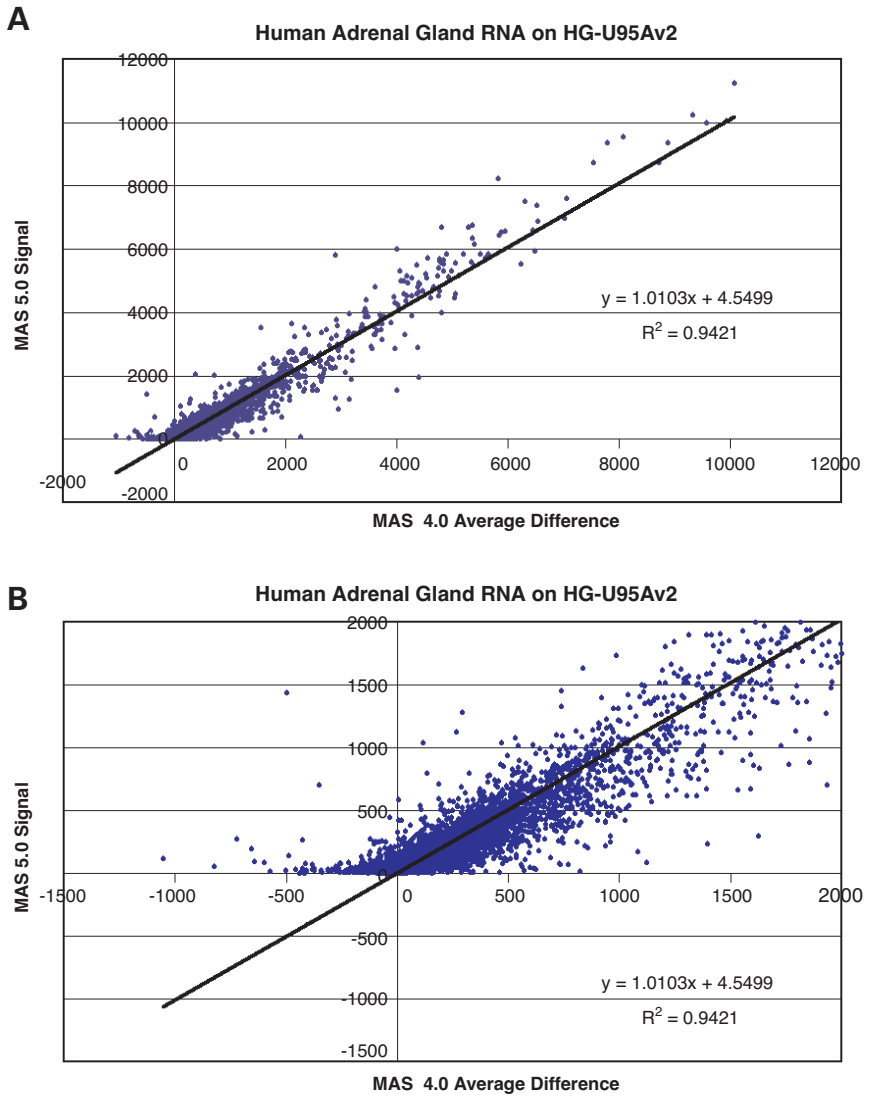
COMPARISON OF EXPRESSION VALUES GENERATED BY MAS 4.0 AND MAS 5.0

Analyses were performed to study the concordance between expression values generated by *empirical algorithms* (MAS 4.0) and *statistical algorithms* (MAS 5.0). Experiments performed with human adrenal gland RNA on human genome U95Av2 arrays are shown in Figure 2.

As shown in Figure 2A, when results generated with MAS 4.0 and MAS 5.0 are compared, the results are highly similar, with a squared correlation coefficient of 0.94. Similar results were obtained across multiple tissues and with replicates (data not shown). This similarity was also observed on arrays representing different organisms. The r^2 values for these comparisons were in the 0.92-0.93 range for *M. musculus, A. thaliana* and *S. cerevisiae*, and 0.96-0.97 for *D. melanogaster* and *E. coli*.

Figure 2. A. Human adrenal gland RNA was labeled and hybridized to a GeneChip® Human Genome U95Av2 probe array. The scanned image was analyzed with MAS 4.0, as well as with MAS 5.0. The Average Difference values for all 12,625 probe sets derived from MAS 4.0 were plotted on the x-axis against the Signal derived from MAS 5.0 on the y-axis. **B.** The lower left quadrant of A is enlarged to indicate the absence of negative signal values in MAS 5.0.

Expression values in experiments with human adrenal gland RNA.



The occurrence of negative values in MAS 4.0, seen in the second quadrant in Figure 2B, has been eliminated from MAS 5.0. The corresponding output in MAS 5.0, termed "Signal" has no negative and zero values as is evident in Figure 2B.

ADDED STATISTICAL QUALITY MEASURES AND TUNABLE PARAMETERS

An added benefit of the *statistical algorithms* in MAS 5.0 is the inclusion of probability values (statistical significance) associated with detection and comparison calls. This additional metric allows users to assess the significance of results, and if desired, adjust