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

Introduction

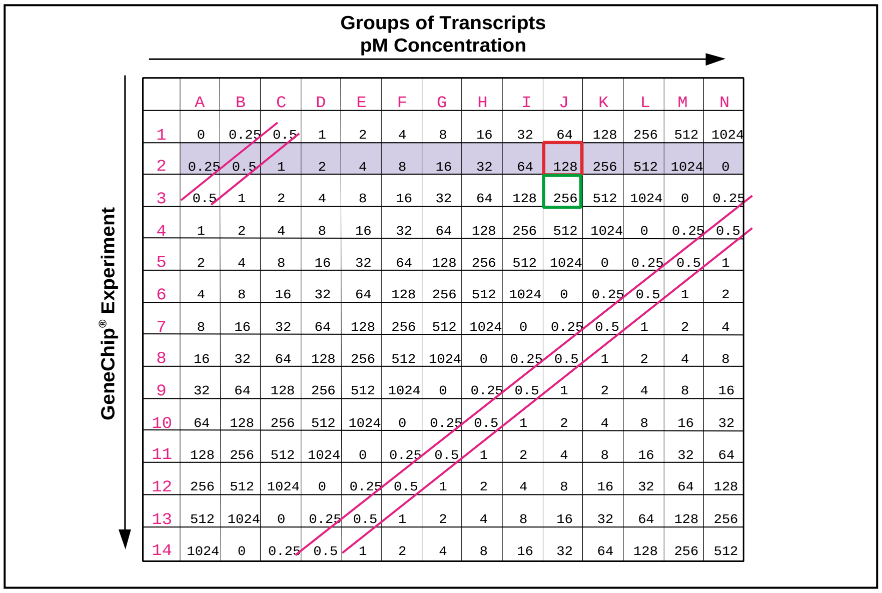
Affymetrix has designed new algorithms for monitoring GeneChip® expression data. These *statistical algorithms*, created in response to customer input, were designed to accommodate the typical distribution of data found in microarray experiments. The new *statistical algorithms* employ standard statistical 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 optimization, 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)


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

Figure 1. ***Latin Square used in algorithm development.***

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.

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 eight different transcripts, labeled and spiked into a mixture of labeled human RNA. After hybridization to Human Genome U95Av2 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.

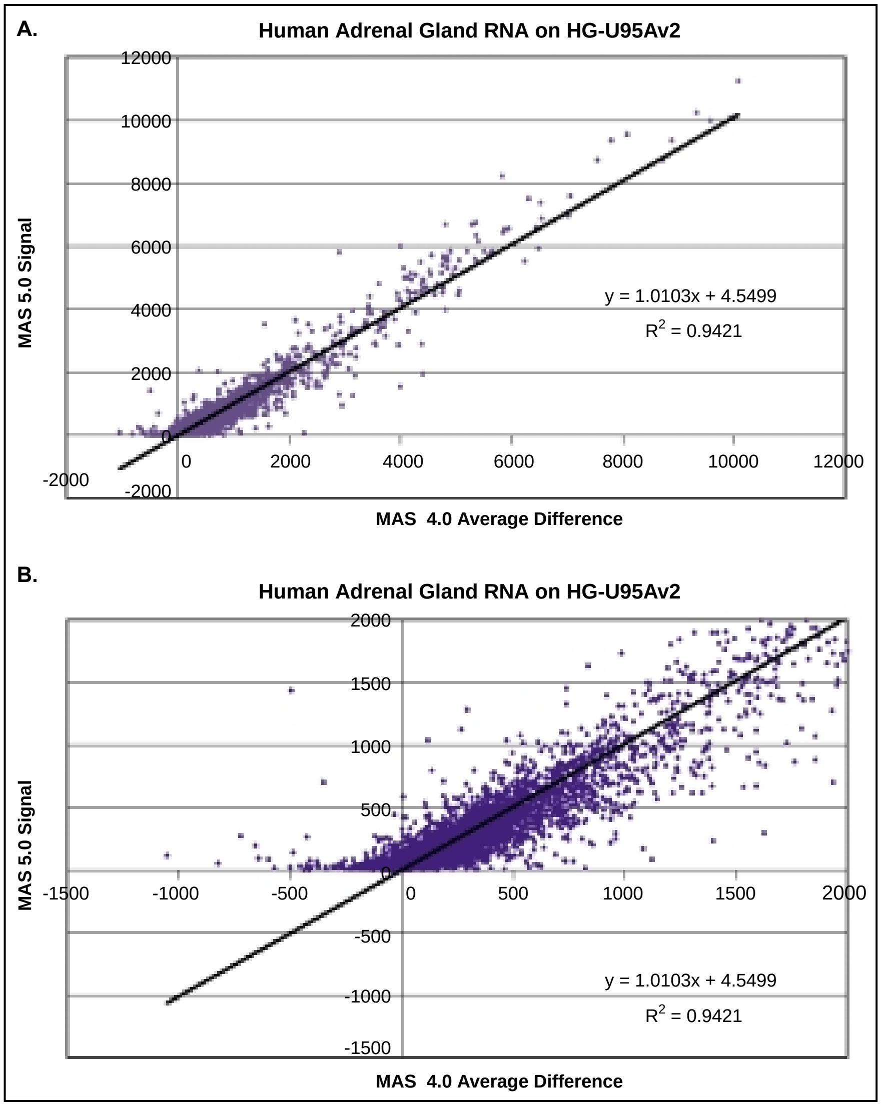


Figure 2. ***Expression values in experiments with human adrenal gland RNA.***

A. Human adrenal gland RNA was labeled and hybridized to a Human Genome U95Av2 GeneChip® 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.

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 U95A 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 regression ratio 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 regression ratios 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.*

The occurrence of negative values in MAS 4.0, seen in the two left quadrants 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 the lower two quadrants of Figure 2B.

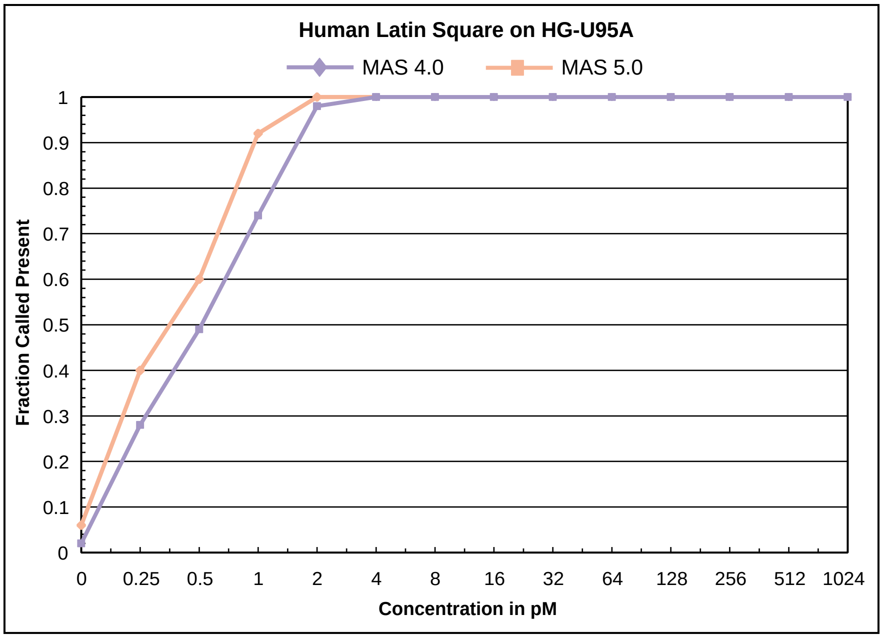


Figure 3. ***Expression calls in human Latin Square experiment.***

Analyses performed on the human Latin Square experiments in Figure 3 are shown here. The *x*-axis represents the different concentrations of spiked transcripts, and the *y*-axis represents the fraction of Present (P) calls. The MAS 5.0 analysis was performed at the default setting where transcripts with p values ≤ 0.04 are assigned Present calls.

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 the balance between sensitivity and specificity. For example, higher confidence in expression calls (greater specificity) can be achieved by accepting fewer Present calls (lower sensitivity). Conversely, greater sensitivity may be achieved at the expense of lower specificity. The inverse relationship between sensitivity and specificity in MAS 5.0 was examined by monitoring detection calls in the set of experiments described below.

As shown in Figure 3, a greater number of accurate Present calls is made by MAS 5.0 at concentrations below 4 pM using the default p-value setting. However, the small number of false positives is also slightly greater for MAS 5.0 in this experiment, as seen at the 0 pM concentration. The sensitivity and specificity may be varied in a predictable fashion by varying the default setting for *p*-value cutoff.

Figure 4 illustrates the linearity of dose response from the 14 spiked transcripts and the upper and lower confidence limits. It also demonstrates that signal calculation in the *statistical algorithms* (MAS 5.0) generates signal values that accurately reflect the true concentration.

As with Figure 2, no signal value in Figure 4 is ever negative or zero. This allows signal values to be easily transformed to a logarithmic scale.

The effects of altering tunable confidence parameters on Present calls were studied in a yeast Latin Square experiment. The tunable parameter controlling the number of Present calls is termed a1 (alpha1). Altering a1 results in a shift in the p-value cutoff used to make a Present call.

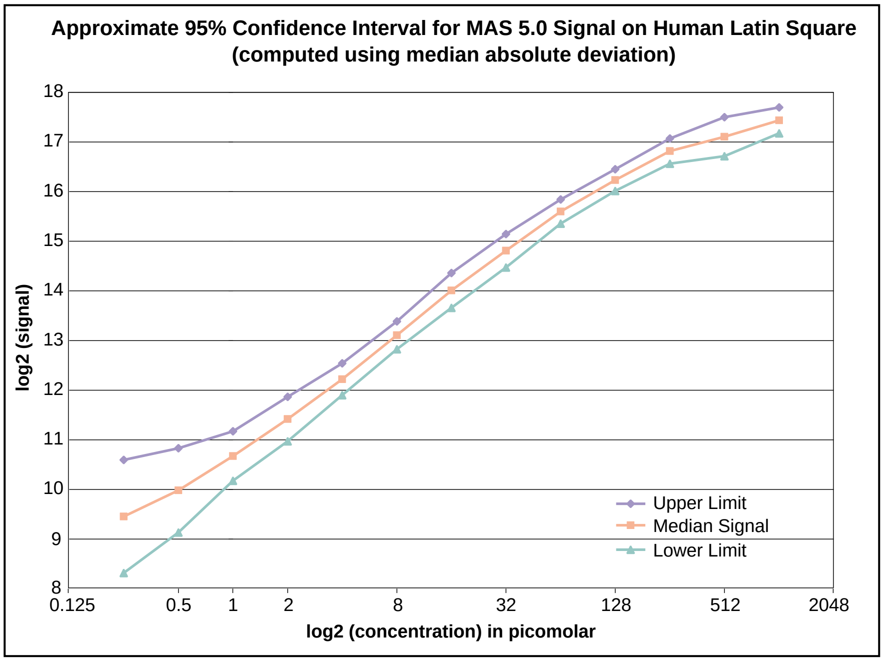


Figure 4. ***Signal value compared to true concentration value.***

Three-fold replicates of the Latin Square design were performed for 14 human transcripts using 42 human U95A arrays. One outlier array was discarded and the remaining arrays were used to examine the relationship between signal and concentration. The observed variation between replicates of the same transcript at each concentration was used to estimate a 95% confidence interval. The median signal of all transcript-concentration pairs is shown in the figure.

This parameter was varied in this analysis to determine the effect on the number of Present calls at different concentrations.

Calls at higher concentrations are typically unaffected by altering a1. At concentrations below 8 pM in this experiment, a decrease in stringency (i.e., increase in a1 from 0.04 to 0.1) results in an increasing number of Present calls. As Figure 5 shows, a1 may be varied to obtain sensitivities using MAS 5.0 that are greater or lower than those obtained with MAS 4.0. However, an increase in a1 produces a corresponding increase in the number of false positive calls, as seen at the 0 pM concentration. It should be noted that false positives can easily be flagged by their high *p*-values.

Comparison Calls Generated by MAS 4.0 and MAS 5.0 Show Concordance

Analyses were performed to study the concordance between comparison calls generated by MAS 4.0 and MAS 5.0. To assess "No Change" calls, comparison analysis was performed on replicate Human Genome U95A arrays from the human Latin Square experiment, where transcript group concentrations were identical between experiments. Therefore, a No Change call is expected for each transcript group between replicates. Results are shown in Figure 6.

The results show that for both MAS 4.0 and MAS 5.0, over 95% (at some concentrations 100%) correct No Change calls are made.

The concordance between Increase calls made by MAS 4.0 and MAS 5.0 are shown in Figure 7. The same transcript group was compared in pairs of experiments to evaluate the calls for a 2-fold and 4-fold change in concentration. For example, in the Latin Square represented in Figure 1, transcript group B was compared between Experiments 1 and 2, then 2 and 3, then 3 and 4, and so on, to determine the fraction

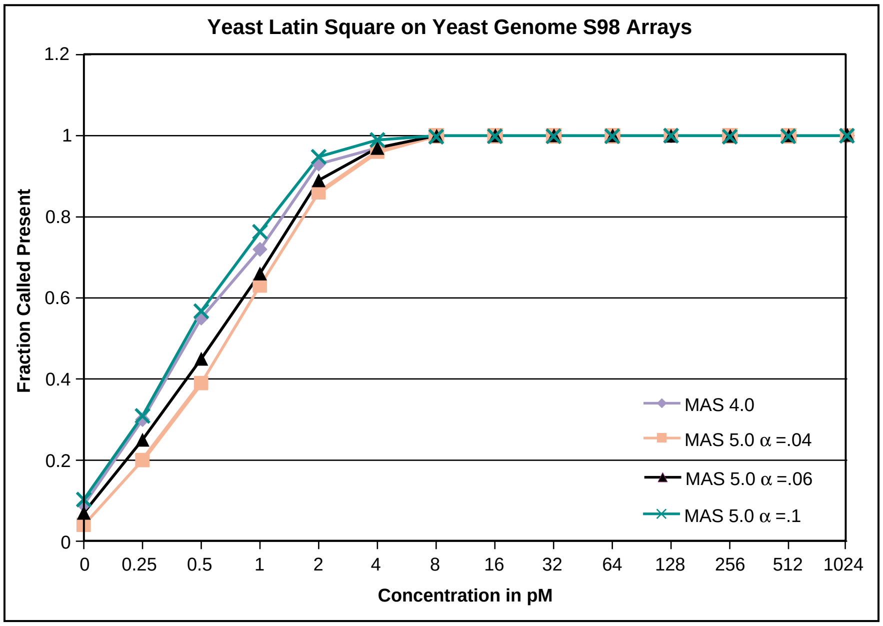


Figure 5. ***Expression calls in yeast Latin Square experiment: Altering p-value cutoff.***

Fourteen different sets of yeast transcripts were spiked into a mixture of labeled human RNA at concentrations ranging from 0 to 1024 pM and hybridized to 14 Yeast Genome S98 arrays. The *x*-axis represents the concentration of spiked transcripts. The *y*-axis represents the fraction of Present calls. The dark blue curve represents analyses performed with MAS 4.0, whereas the other three curves represent analyses performed at three different a1 settings of MAS 5.0: 0.04, 0.05 and 0.1.

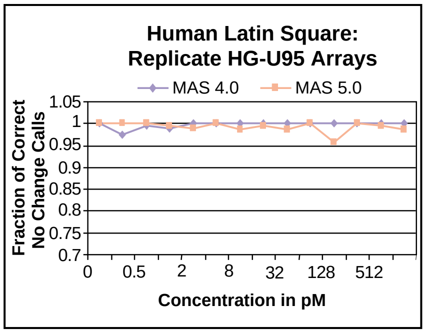


Figure 6. ***Human Latin Square: "No Change" Calls.***

Replicate experiments from the Human Latin Square set were analyzed to assess the "No Change" calls. Of the 14 groups of experiments, 11 groups had three replicates, one group had two replicates and two groups had 12 replicates each. The *x*-axis represents the concentration of the spiked transcripts. The *y*-axis represents the fraction of "No Change" calls.

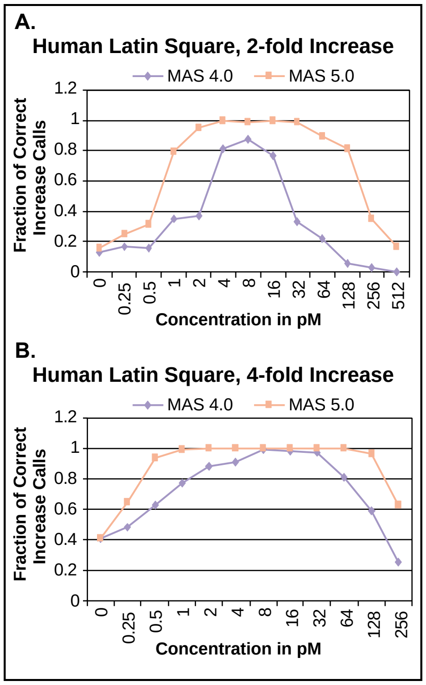


Figure 7. ***Increase calls in a human Latin Square.***

A. The fraction of Increase calls at the 2-fold change was plotted on the *y*-axis. The spiked transcript concentration is shown on the *x*-axis.

B. The fraction of Increase calls at the 4-fold change was plotted on the *y*-axis. The spiked transcript concentration is shown on the *x*-axis.

correctly assigned an Increase call at the expected 2-fold change. For the 4-fold change, transcript group B was compared between Experiments 1 and 3, 2 and 4, 3 and 5 etc., such that all transcript groups were evaluated.

In this data set, MAS 5.0 shows greater accuracy than does MAS 4.0 at both 2-fold and 4-fold Increase levels. For example, at the 2 pM spike in Figure 7B, MAS 5.0 gives a nearly 10% greater fraction of correct Increase calls than does MAS 4.0. The drop in accuracy seen at very high concentration is due to the saturation of probe sets by the vast excess of spiked transcripts.

To understand the difference in the performance of calls generated by MAS 4.0 and MAS 5.0 on a biological sample, we assessed the results using an independent method. We used the power of replicates and the well-known student's t-test as a reference.

We analyzed a number of replicates in two tissues and then asked, "How well do the results from any single comparison match results from the group as a whole?"

The group was built from two sets of experiments—mouse brain and mouse heart—containing six replicates for each tissue. Each replicate was compared to every other replicate with both MAS 4.0 and MAS 5.0 (36 comparisons of 12488 probe sets). The t-test was then used to evaluate whether the mean of the replicates in one tissue was significantly different from the mean of the replicates in the other tissue. The results are shown in Figure 8.

Overall, the Comparison calls generated by MAS 4.0 and MAS 5.0 are highly similar. The No Change calls of both MAS 4.0 and MAS 5.0 peak at approximately 0.9, as we would expect. The Decrease calls (D) peak at 1 and the Increase calls (I) peak towards zero for both algorithms, while the lines follow each other very closely. The additional benefit of MAS 5.0 is that we now have p-values to assess statistical significance of the comparisons of every gene.

The next technical note will contain more results on Comparison calls and their performance in MAS 5.0, together with discussions of tunable parameters and confidence limits.

Conclusions

In conclusion, our extensive validation studies, some results from which are shown above, allow us to state with confidence the following:

- The new *statistical algorithms* in MAS 5.0 perform as robustly and precisely as the previous algorithms, while providing the additional value of statistical significance (p values) and confidence limits, thereby offering GeneChip® users the ability to evaluate the significance of their results.
- Negative expression values (signals) have been eliminated; only positive values are generated in MAS 5.0.

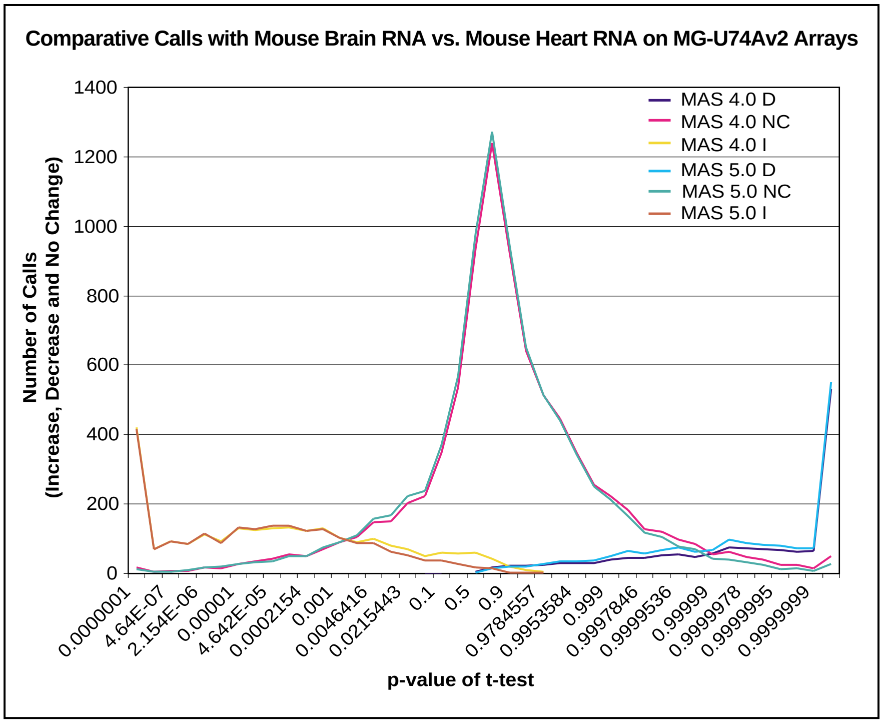


Figure 8. ***Comparison Calls with Mouse Brain RNA vs. Mouse Heart RNA on U74Av2 arrays.***

Two sets of experiments were analyzed, each consisting of six replicate mouse U74Av2 arrays hybridized to samples derived from mouse brain RNA or from mouse heart RNA. These two sets of replicate experiments were analyzed with MAS 4.0 and MAS 5.0. Thirty-six pair-wise comparisons were performed between the two sets of experiments. The *y*-axis represents the number of Increase, Decrease or No Change calls made by MAS 4.0 or MAS 5.0, for all 36 pair-wise comparisons, (i.e., for 36 x 12488 probe sets). A student's t-test was performed on the two sets of experiments and a p-value was generated for each comparison of individual probe sets. This p-value was plotted on the *x*-axis. The graph assesses the distribution of Comparison Calls made by MAS 4.0 and MAS 5.0 in comparison to a t-test.

AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd.,

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
Tel: +44 (0)1628 552550
Fax: +44 (0)1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

www.affymetrix.com

For research use only.
Not for use in diagnostic procedures.

Part No. 701097 Rev 3

AFFYMETRIX®