AFFYMETRIX®

Technical Note

GeneChip® Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis

The rapid development in microarray technology has launched a revolution in biology, enabling researchers to simultaneously monitor nearly each and every mRNA transcript in an entire cellular sample. Unprecedented, highly parallel discoveries have been made possible, including the ability: to follow changes in the expression levels of many genes simultaneously[1], or to identify distinctive expression patterns characteristic of physiological and pathological states[2], and to screen for very subtle changes in response to a particular therapeutic treatment[3], to name a few.

There are varieties of microarray technologies available that use different attachment chemistries, different strategies in choosing immobilized molecules, different assay conditions and different data analysis tools.

The most effective microarray platform should offer the following qualities:

- **Sensitivity** to identify the rarely expressed transcripts in a complex background.
- **Specificity** to discern between different family members.
- **Absolute quantitation** to enable robust databases to be built with a minimum number of samples.

The hybridization efficiency of two nucleic acid strands results from the combination of many factors; their cumulative effect determines the sensitivity and specificity of a microarray experiment. These effects have been investigated in extensive detail at Affymetrix[4] and include:

- **Sequence-dependent factors** such as length, extent of complementarity and the overall base composition.
- **Sequence-independent factors** such as concentration of the probe and target, time, temperature, cation concentration, valency and character, pH, dielectric and chaotropic media, surface characteristics of the solid support, and density spacing of the probe molecules synthesized on the surface.
- **Sample-dependent complex background signal.** Any probe sequence, even under conditions optimized for the above factors, has the potential to interact with other complementary sequences present in the sample, even if the complementarity is short and partial. This may result in increased nonspecific background signal and reduced specificity and sensitivity. Furthermore, samples in which the composition and concentrations of the transcripts differ may result in different background signals that are not easy to extrapolate.

The Affymetrix Approach

Affymetrix has designed the following strategies to specifically address the issues, and to ensure that GeneChip® arrays feature the optimal solution for mRNA expression analysis.

Using the combination of a pair of 25-mer probes that are designed to be perfect match (PM) and mismatch (MM) to the targets offers the balance of highest sensitivity and specificity in the presence of a complex background—especially for low abundance transcripts.

Given the complexities associated with any microarray system, Affymetrix uses strategies that specifically address critical issues to discern between specific and non-specific binding with the PM-MM probe strategy. Based on extensive experimentation, we have found that 25-mers are ideal for utilization of the PM-MM strategy.

- **PM-MM probe pairs offer significantly higher sensitivity at low target concentrations than using PM probes alone.**

MM probes have the ability to quantify background signals that are present in the complex sample. Background discrimination becomes even more important when the specific signal intensity is relatively low and close to the background signal intensity.

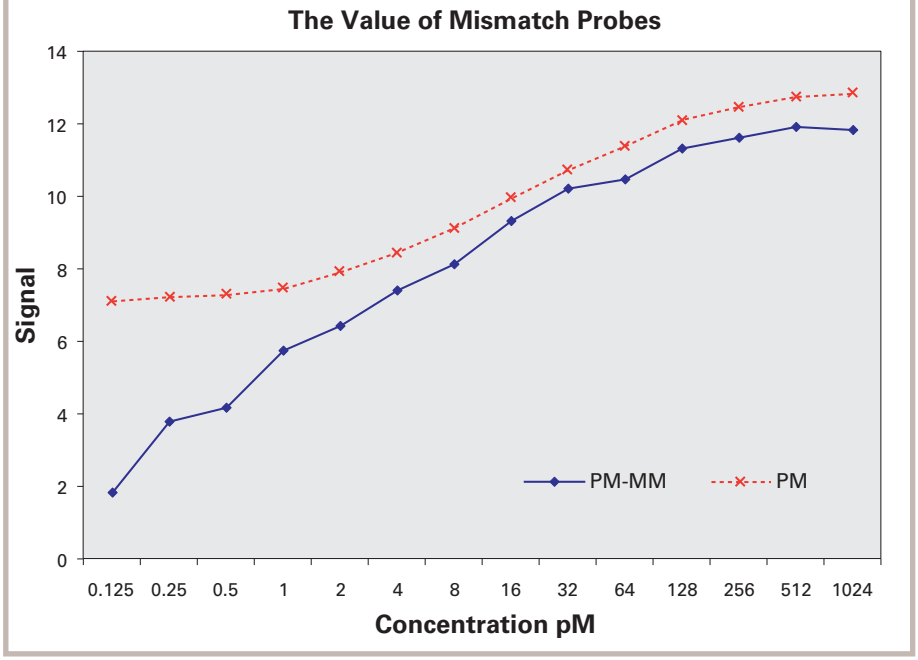
In Figure 1, mixtures of cloned transcripts were spiked at various concentrations into tissue samples where the transcripts were originally absent. Labeled samples were then hybridized on the GeneChip® Human Genome U95 Array. The value of MM probes becomes apparent below 8 pM where the PM-only probe sets lose sensitivity and do not respond to changes in the transcript concentration. The improved assay sensitivity with PM-MM probe pairs allows detection of transcripts representing 1:100,000 – 1:300,000 of total transcripts in a sample.

- **PM-MM probe pairs offer specificity in the presence of complex background signals, allowing the optimal balance between sensitivity and specificity.** MM probes are effective internal controls, since they will hybridize to nonspecific sequences about as effectively as their counterpart PM probes. As a result, unpredictable background signal variations associated with samples from different sources as well as from cross-hybridization can now be efficiently quantified and subtracted with the utilization of the MM probes.

Because of the targeted design of short probes and multiple probes per sequence (Figure 1), it is possible to unambiguously detect closely related sequences. For example, the yeast histone genes HTA1 and HTA2 that share 93% identity at the DNA level can be detected with GeneChip® arrays (L. Wodicka and D. Lockhart, unpublished data).

Sensitivity/discrimination experiments showed that probes of approximately 25 nucleotides long provide a very effective balance between signal intensity and related sequence discrimination. This enables expression monitoring of thousands of targets in complex samples[5]. Such a powerful strategy, however, is most effective with shorter (e.g., 25-mer) probes, since a single base mismatch is sufficient to destabilize the hybridization. Conversely, a longer probe such as a 60-mer oligonucleotide, cannot efficiently take advantage of a similar strategy since it requires much longer stretches of mutations or deletions to achieve a similar discrimination effect[6]. In addition, 60-mer probes will often not be able to discriminate between closely related sequences, especially with relatively abundant transcripts.

Figure 1. Comparison of the assay sensitivity using PM probes only or PM-MM probe pairs.



Global performance evaluation on GeneChip® arrays vs. examination of a small subset of genes maximizes the number of genes detected on the arrays.

While it is often feasible to achieve very high discrimination for a particular target sequence by optimizing hybridization conditions, there can be no assurance that the same conditions are sufficiently stringent for the other thousands of probes represented on the same array. Instead of examining the hybridization behavior of only a small subset of sequences as described by Hughes *et al*[6], Affymetrix has taken an approach to optimize hybridization and detection conditions by surveying the hybridization results over essentially all probe pairs on an array.

We have used population distributions to monitor the global hybridization behavior of an array. It is assumed that a true stray signal has equal likelihood to hybridize more efficiently to either the PM or the MM probe in a pair. In contrast, under conditions optimized for global hybridization efficiency, more PM probes will have a higher intensity value than their corresponding MM probes when a complex target sample is hybridized to the array.

As shown in the histogram in Figure 2, the x-axis represents bins of the discrimination score metric used, log (PM/MM). The negative values at the far

left of the graph correspond to probe pairs where virtually all signal is in the MM probe. Conversely, the probe pairs for which all the signal is in the PM probe are reflected by positive values at the right end of the graph. The data in green correspond to complete stray signal distribution and as expected, the mode is centered on PM=MM (i.e. log (PM/MM) = 0). The population shown in red estimates the number of probe pairs bound to true target sequences in the mixture and is therefore skewed to the right. Assay conditions have been optimized in this experiment to maximize the size and displacement of this population (red) relative to the stray signal population (green).

Improving global discrimination scores through the optimization of assay conditions as described above has translated directly into assay performance improvements in the form of increased sensitivity for "present" calls and increase/decrease change calls, as well as reduction in false positives for both metrics. It ensures that the optimized assay conditions will accommodate the maximum number of probes represented on the array. This strategy cannot be accomplished by observing only a limited number of selected sequences.

Using multiple probes per sequence allows both absolute and comparative analysis that is essential for building a quality database with the minimum number of samples.

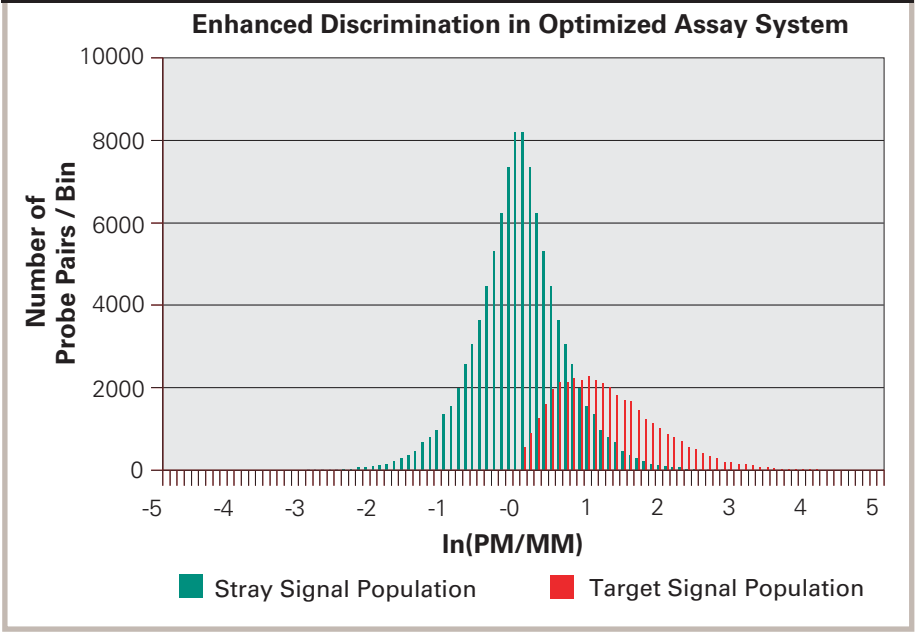
Cross-hybridization derived signal is an intrinsic and possibly significant component of the total signal for many probes interrogating a complex sample, resulting in inaccuracy in RNA quantitation and increased rate of false positives and miscalls. This was observed in the initial work using high density arrays to monitor expression[5], and is abundantly evident in the system presented by Hughes *et al*[6].

If only one probe is used to interrogate each sequence, it is absolutely essential that each selected probe meets all requirements to be a highly functioning probe. For example, the selected probe: a) does not exhibit cross-hybridization to other background sequences, b) does not fold on itself, and c) needs to hybridize efficiently to the target under the assay conditions. To compensate for many of these sources of noise, error and variance, Affymetrix uses the multiple probe pair sampling strategy in the GeneChip® array design.

- **Multiple probes offer the potential to calculate statistics that provide more confidence in microarray results.**
  Information from multiple probes per sequence aggregates in statistically robust ways, especially when MM probe results are incorporated into the analysis. Because of this, it is possible to use statistically based algorithms to provide confidence and probability information for absolute call results as well as comparison ratios. Interrogating sequences that correspond to multiple 25-mers on GeneChip® arrays vs. a single 60-mer[6] results in more reliable data that are robust and resistant to target sequence mutations.
- **GeneChip® arrays offer absolute expression analysis in addition to comparative analysis. Therefore, the results obtained are conducive to building large databases.**
  Redundancy in the number of probes used for each sequence in combination with statistical analysis software allows GeneChip® arrays to be used not only for comparative analysis, but also for absolute quantitative expression analysis. The use of multiple independent probe pairs to detect the same RNA molecule greatly improves the RNA quantitation accuracy by averaging and outlierrejection. This eliminates the necessity to include control samples in each experiment as used in the two-color assay.

  Ishii *et al*[7] compared the GeneChip® technology with SAGE and concluded, "These two methods correlated quite well in both absolute expression analysis and comparative analysis during differen-tiation." In addition, they commented, "This finding suggests that GeneChip® technology is reasonably reliable for quantitative analysis of expression profiling and would be appropriate as a common

Figure 2. Global evaluation of hybridization results on a GeneChip® array.



platform upon which to build a gene expression database."

In fact, Aach *et al*[8] has reported such an effort in systematically integ-rating and managing expression analysis results from several different studies into one database. They have found that, "Generation of ERAs (mRNA relative abundance estimates) is straightforward for data derived from Affymetrix oligonucleotide arrays and SAGE, but (two-color) microarray-derived data present a significant issue."

Interestingly, Hughes *et al*[6] indicated that the 'one probe per sequence' approach is only suitable for obtaining comparison results in two-color assays, but not for absolute expression quantitation. "This suggests that when the experimental design is focused on determining changes in transcript abundance between two samples, rather than absolute transcript abundance, multiple oligonucleotides are not needed in most cases."

Conclusion

In conclusion, Affymetrix has employed the unique PM-MM probe pair probe design approach to obtain the optimal balance of highest sensitivity and specificity in the presence of a complex sample background. The global assay optimization strategy involving investigating the hybridization results of all probes on an array as a single data set, allows the derivation of assay conditions that maximize the number of genes detected on an array. In addition, redundant sampling of each sequence with multiple probe pairs in a probe set provides robustness and reliability in the data obtained. In combination, the GeneChip® platform provides robust, reliable and reproducible quantitative expression analysis results with statistical significance for effectively building large expression databases.

REFERENCES

[1]Cho, R.J. *et al*., (1998) A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, 2:65-73

[2]Golub, T.R. *et al*., (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**:531-537

[3]Jin *et al*., (2001) Effects of Early Angiotensin-Converting Enzyme Inhibition on Cardiac Gene Expression After Acute Myocardial Infarction. *Circulation*, **103**:736-742

[4]Forman, J.E. *et al*., (1997) Thermodynamics of Duplex Formation and Mismatch Discrimination on Photolithographically Synthesized Oligonucleotide Arrays in Molecular Modeling of Nucleic Acids. (N.B. Leontis and J. Santa Lucia Jr., eds.) *American Chemical Society Symposium Series* **682**:208-228

[5]Lockhart, D.J. *et al*., (1996) Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. *Nature Biotechnology*, 14:1675-1680

[6]Hughes, T.R. *et al*., (2001) Expression Profiling Using Microarrays Fabricated by an Ink-Jet Oligonucleotide Synthesizer. *Nature Biotechnology*, **19**: 342-347

[7]Ishii, M. *et al*., (2000) Direct Comparison of GeneChip and SAGE on the Quantitative Accuracy in Transcript Profiling Analysis. *Genomics*, **68**:136-143

[8]Aach, J. *et al*., (2000) Systematic Management and Analysis of Yeast Gene Expression Data. *Genome Research*, **10**:431-445

AFFYMETRIX, INC.

3420 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
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
UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

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

Part Number 701009 Rev. 4