

REDI (REDuction of Invariant Probes) Analysis

This Technical Note provides an overview of Expression Analysis' success in improving detection of differentially expressed transcripts by removing invariant and unresponsive probes from the most frequently used Affymetrix microarrays.

Probe Performance on Affymetrix GeneChips®

Affymetrix arrays contain thousands of sets of short oligonucleotide (25mer) probes. In general, each probe set represents a different gene transcript and typically consists of eleven perfect match (PM) probes as well as corresponding mismatch probes. Ideally, each PM probe in the set would perform equally well and respond linearly to increasing amounts of the gene target. Although the majority of PM probes perform as desired, some probes are noticeably less responsive to the target concentration, leading to variable performance within the set. This variable probe performance can impact data analysis as expression values are based on a summary of the individual probe hybridization intensities.

Properties of hybridization affinities to short oligonucleotide probes are well-known and frequently modeled in microarray data analysis. However, the reasons for poor probe performance are varied, and sometimes are related to sequence errors. We have combined *in silico* sequence analysis of the PM probes with empirical observation obtained from Expression Analysis' own repository of Affymetrix microarray data to identify probes that hinder the discovery of differential expression and to develop a method that excludes these poorly performing probes from analysis.

The Data Set and Method of Analysis

In order to distinguish probe performance from biologically-related changes in transcript abundance, we examined a large set of unrelated hybridizations to the same GeneChip. Over a two-year period, targets were prepared from more than 3,000 human RNA samples and hybridized to Human Genome (HG) U133A GeneChips. The RNA samples were derived from a variety of human tissues and cells lines that had been exposed to multiple experimental conditions. After normalization at the probe level, the median and range of hybridization intensities were calculated for each of the ~240,000 PM probes on the HG-U133A GeneChip. Similar large sets of hybridizations to the Mouse430A and Rat230A GeneChips were also analyzed. From this data, we were able to distinguish probes that had variable intensity over a larger dynamic range from those probes whose response range was no better than "random" 25mer sequences. After validating the initial set using *in silico* sequence analysis (with the RefSeq library) and within-probe set correlation analysis, we created a list of poorly performing probes.

Perfect Match Probe Performance Category	Number on U133A GeneChip	Percent of U133A GeneChip
Responsive	169,154	68.2 %
Unresponsive	26,731	10.8 %
Invariant	50,914	20.5 %
AFFX Quality Control	1,166	0.5 %

Table 1: Characterization of PM Probes on HG-U133A.

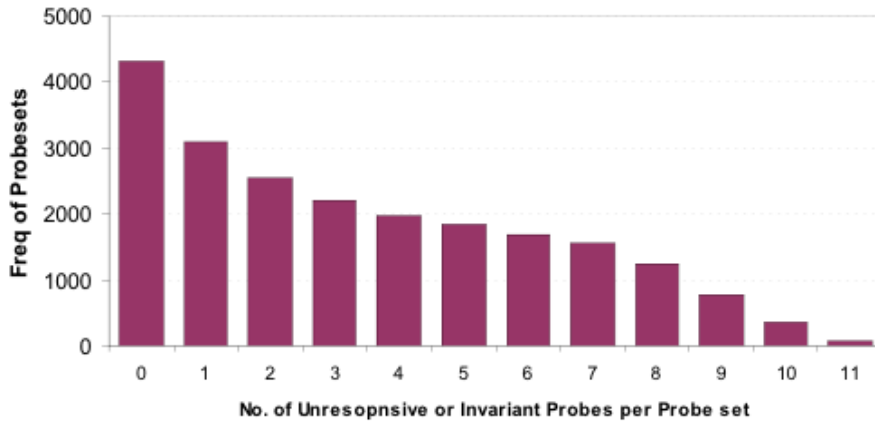
The Data Set and Method of Analysis

As expected, most of the PM probes (~70%) had a large intensity range relative to their median intensity and so appear responsive to changes in target concentration. However, some probes were either unresponsive (no hybridization signal) or invariant (same hybridization signal) across the hybridizations analyzed. This subset of poorly performing probes varied from 25 to 30% of the GeneChip depending on the chip type.



Poorly Performing Probes are Present in a Large Number of Probe Sets

Figure 1: Number of Poorly Performing Probes Per Probe Set on HG-U133A.



Our results suggest that the poorly performing probes can affect more than 80% of the probe sets on the HG-U133A, Rat230A, and Mouse430A GeneChips. In fact, from Figure 1, we see that approximately half of the HG-U133A probe sets have four or more unresponsive or invariant PM probes.

Sequence Errors Account for Some of the Poor Performance

The reasons for the observed poor probe performance are related to dynamics of probe behavior (apart from hybridization affinity to their intended target) and to sequence errors during probe design. In some cases, poorly performing probes cross-hybridize too readily. Other probes do not align at all to updated genomic sequence data. For example, four of the eleven probes in the 200654_at probe set (P4HB gene) demonstrate poor performance on HG-U133A and HG-U133_Plus_2.0 GeneChips. Sequence alignment revealed that the four poorly performing probes were located in the 3' end of the corresponding transcript, where the sequence used during probe design is no longer homologous to updated genomic sequence. Some amount of probe sequence errors are expected as genomic sequence quality improves over time relative to static probe selection, and has been noted by other investigators⁽¹⁻⁴⁾.

Figure 2: Comparison of Updated Genomic Sequence and Sequence Used During Probe Design.



Genomic sequence (top) for the P4HB gene is aligned with sequence used during design of the P4HP probe set, 200654_at (bottom). The locations of PM probes are underlined and highlighted in red or blue to indicate responsive or poor performance, respectively, on the HG-U133A GeneChips.

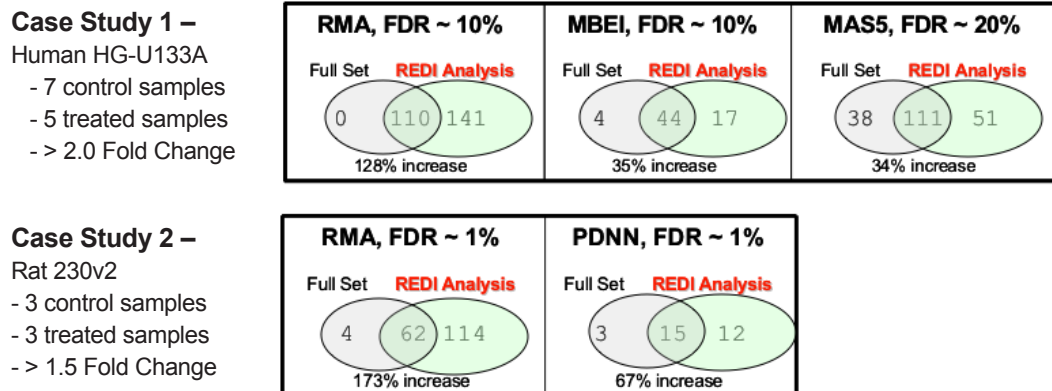
The REDI Method

As the reasons for poorly performing probes are often foundational and are *not* related to probe signal modeling assumptions, almost all current summary expression indices are negatively impacted by their presence. By eliminating the contribution of these probes during signal computation, we remove an artifact within the probe set that tends to make disparate biological groups appear more homogeneous. At Expression Analysis, we used our knowledge of individual probe behavior to develop the REDUction of Invariant probes (REDI) data analysis method that ignores the poorly performing probes during summary signal calculations. In this two-step approach, the hybridization intensities from all probes on the GeneChip are used during probe normalization (because data from a large number of invariant or unresponsive probes are actually desirable for quantile and other types of normalization). After normalization, data from the poorly performing PM probes are removed from the probe set signal computation step. Our REDI method can be incorporated into multiple expression metric algorithms, including Microarray Suite version 5 (MAS5), PLIER, RMA, GC-RMA, PDNN and dChip/MBEI.

Impact on Gene Lists

We show the results of using the REDI method in two case studies to illustrate the impact of poorly performing probes on microarray experiments. Case study 1 contained seven control and five treated samples hybridized to HG-U133A GeneChips. Case study 2 consisted of three control and three treated samples hybridized to Rat230A GeneChips. Multiple data analysis methods were used to identify differentially expressed genes between the control and treated samples. These analyses were initially performed with the full set of probes on the Affymetrix GeneChips and then repeated using the REDI method that removes the poorly performing probes. As shown in Figure 3, the size of the human differential gene lists were increased by more than 30% when REDI analysis was used and the rat differential gene lists grew by more than 50% across a variety of signal measures. For example, 110 differentially expressed genes were detected in case study 1 using RMA with the full set of probes. An additional 141 differentially expressed genes were detected when the poorly performing probes were removed using the REDI method with RMA.

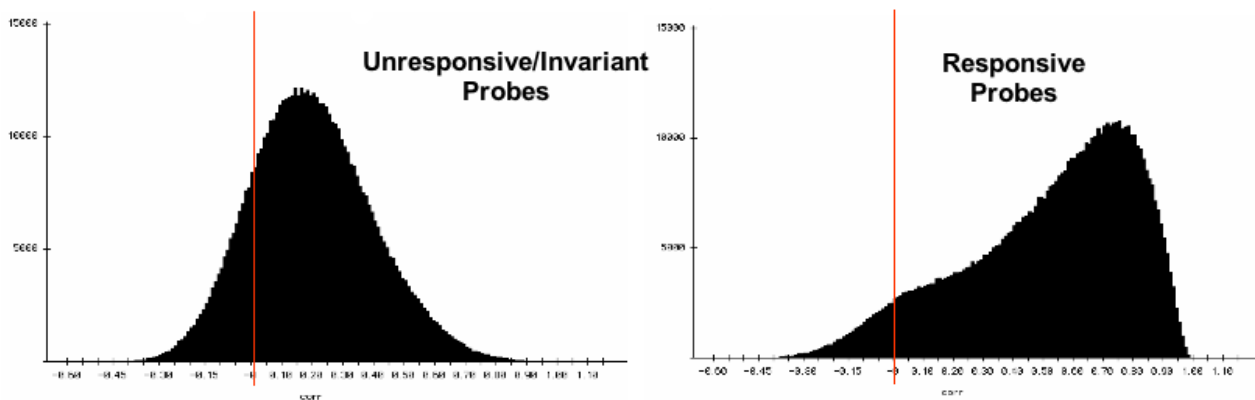
Figure 3: Number of Differentially Expressed Genes Identified With and Without REDI Analysis



Validation

To further validate our REDI method, we examined the correlation in probe intensity among responsive and poorly performing probes. Responsive probes should highly correlate with other responsive probes within the same probe set, while unresponsive/invariant probes should have little or no correlation with any other probe. For example, suppose as with the 200654_at probe set shown in Figure 3, that there are 4 poorly performing probes and 7 responsive probes in the probe set. Then there are 40 possible probe correlations that we expect to be near 0 (each unresponsive/invariant probe paired with one of the other 10 probes) and 21 correlations that we expect to be much larger (each responsive probe paired with another responsive probe), possibly in the range of 0.6 to 0.9. When we do this analysis for the entire HG-U133A chip, we have ~450,000 correlations using probes where at least one probe is unresponsive or invariant and ~820,000 correlations where both are responsive probes. As shown in Figure 4, the poorly performing probes tend to have much lower correlation values (median is 0.19), while the responsive probes having a median of 0.6 with a mode between 0.7 and 0.8. The prominent left tail in the histogram for responsive probes may be due to other interesting behaviors, including splice variants.

Figure 4: Histograms of Correlations among Probes in the same Probe Set



Summary

Expression Analysis has determined that (typically) greater than 25% of the PM probes on a given species chip are relatively unresponsive or invariant, irrespective of the labeled cRNA that is applied, even in the latest generation chips (HG-U133 family, Rat230 family, Mouse430 family). We contend it is much more difficult to accurately discern differential expression when many probes within a probe set have poor performance. We examined individual probe behavior in thousands of hybridizations and developed a new method of analysis that negates the effect of poorly performing probes. Using REDI analysis, we see a consistent increase in the number of differentially expressed genes based both on significance and magnitude of fold change thresholds.

- 1 J. Harbig, R. Sprinkle, and S. Enkemann. A sequence-based identification of the genes detected by the probe sets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Research* 33 (3):1-9, 2005.
- 2 B. Mecham, D. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T. Mariani. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics* 18 (2004):308-315, 2004.
- 3 B. Mecham, G. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Wetmore, T. Mariani, I. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research* 32 (9):1-8, 2005.
- 4 J. Zhang, R. Finney, R. Clifford, L. Derr, and K. Buetow. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* 85 (2005):297-308, 2005.

