A Global View of miRNA Expression Using a Microarray Platform Featuring Probes for More than 15,000 Putative and Validated miRNAs

Edward Sekinger, Sachin Sah, Nan Leng, Timothy S. Davison, Laura Elizondo, Sylvie Beaudenon, Charles Johnson, Mike Wilson. | Asuragen, Inc., 2150 Woodward Ave., Austin, Texas 78744

SUMMARY POINTS

- Asuragen, Inc. has designed a custom Affymetrix microarray containing comprehensive representation of more than 15,000 verified and predicted miRNAs;
- The microarray has good concordance (R > 0.9) with the Ambion miRNA Bioarray and TaqMan® qRT-PCR results;
- Expression patterns for 1670 unique probes were observed across 10 tissues with each tissue displaying a distinct miRNA expression pattern;
- Tissue-specific putative miRNAs were identified from the expression dataset;
- Putative miRNA gene clusters are significantly enriched on chromosome 1, relative to known miRNAs.

INTRODUCTION

Asuragen, Inc., a molecular biology service provider and molecular oncology diagnostic company, has designed a custom, Affymetrix microRNA (miRNA) microarray containing probes derived from multiple sources, including a proprietary database of putative miRNAs. To our knowledge, this is the most comprehensive miRNA discovery platform available for commercial and collaborative investigations.

The Sanger miRBase v9.0 stores data for 452 unique human miRNAs, and current predictions are that vertebrate genomes encode as many as 1000 unique miRNAs based on genomic sequence and computational miRNA prediction algorithms (Berezikov, et al., Cell 120; 21-24, 2005). Since each miRNA is thought to modulate multiple genes, it is estimated that miRNAs could affect the expression of more than 30% of human and mouse proteins. Importantly, evidence is mounting that miRNAs are involved in the regulation of embryonic development, cell proliferation, apoptosis, fat metabolism, cell differentiation, neurodevelopment, viral disease, and cancer. As one step toward identifying the entire complement of human miRNAs, we sought to measure the expression patterns of novel, small RNAs that were originally identified by sequence-based prediction algorithms and computational analysis of the genome. As a result, this research tool could enable the comprehensive assessment of miRNA expression patterns that prove valuable for achieving a better understanding of miRNA-regulated processes and the discovery of diagnostic biomarkers involved in cancer.

MATERIALS AND METHODS

Human First Choice™ total RNAs were purchased from Applied Biosystems (AB). The small, miRNA-containing fraction (<40 nts) was isolated by polyacrylamide electrophoresis using the flashPAGE™ fractionation system (AB). Purified small RNAs were enzymatically biotinylated at the 3'-termini and hybridized to the array. Hybridization, washing, staining, imaging, and signal extraction were performed as recommended by the manufacturer, except that the 20X GeneChip® Eukaryotic Hybridization Control cocktail was omitted. At least two different probes were designed for each unique miRNA sequence; additionally, the pre-miRNA sequences of 276 verified miRNAs were represented by probes targeting the predicted loop region of the pre-miRNA. A Latin square spike-in study was designed using ten synthetic RNA oligonucleotides titrated in serial 2-fold dilutions to cover a range from 8.9 – 0.035 fmol; ten concentration unique spike pools were generated. Subsequently, each pool is mixed with human bladder total RNA, that was previously verified to be negative for each of the ten synthetic spikes, and used in a cyclic Latin square design consisting of 10 experiments. Methods for background correction and normalization of the data were assessed in order to achieve the best combination of performance for accuracy and precision. To do this, a systematic comparison was performed on 78 combinations of background adjustment, variance stabilization and array scaling or global normalization approaches. Performance specifications were derived from the Latin square experiment to determine the sensitivity, specificity, lower limit of detection, and response linearity, following methods described by Irizarry et al. Bioinformatics (22(7); 789-794 (2006)). Concordance measurements were determined by comparing the results to Ambion Bioarrays (manufactured with Codelink™ technology by GE Healthcare) that have probes to 385 miRNAs; or qRT-PCR assays (ABI TaqMan MicroRNA Assay).

RESULTS

A custom Affymetrix microarray was designed to represent a comprehensive set of verified and predicted miRNAs. The performance of the array was assessed by a Latin Square spike-in experiment, and the following results were observed: lower limit of detection (<33 amol), dynamic range ($10^{2.7}$), and reproducibility (CV<10%); (Table 2). Relative accuracy was assessed by comparing the results of differential miRNA expression to an alternative array platform and to TaqMan qRT-PCR. Results from both alternative technologies correlated well (R>0.9) with the custom array (Figure 1). Taken together, these results support the claim that the performance characteristics and the extended content of the array are suitable for enabling rapid discovery of novel miRNA expression patterns. Toward this goal, a survey of 10 normal tissues demonstrate that the array is effective at detecting the expression of an extended set of putative miRNAs (Table 3 and Figure 4) over the content that is available through the Sanger Institute's miRBase. Figure 5 illustrates the genetic locations of both Sanger and putative miRNAs on chromosome 1 wherein a significant enrichment of putative miRNAs is dispersed throughout the entire chromosome.

Data Source	Mature miRNAs	Unique Mature miRNAs
Sanger V8.2	3834	1314
Cummins, *et al.*[1]	133	2
Xie, *et al.*[2]	144	129
Berezikov, *et al.*[3]	975	692
Predicted[4]	13843	12888
Total		**15025**

Table 1. Microarray Content.
The custom Affymetrix microarray contains 31110 DNA probes representing 15025 verified and predicted miRNAs from the sources listed. Unique mature miRNA sequences are the non-redundant subset of all the mature miRNAs obtained from each of the sources. A miRNA is considered unique if there is at least one nucleotide base difference in pair-wise comparison to all mature miRNAs.
1 Cummins, et al., Proc Natl Acad Sci USA 103(10):3687-92, 2006);
2 Xie, et al., Nature 434(7031):338-45, 2005);
3 Berezikov, et al., Cell 120(1):21-4, 2005);
4 Rosetta Genomics' proprietary predicted sequences.

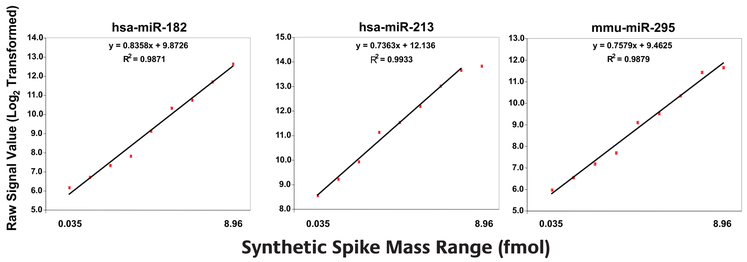
A.

Specification Metric Summary	Result
Sensitivity	< 33 amol
Linearity	$R^2 > 0.98$
Dynamic Range	$10^{2.71}$
Reproducibility	CV < 10%
Specificity (FPR)	< 1%

B.

RNA spike	Slope	R-squared	amol
hsa-miR-182	0.84	0.99	46
mmu-miR-201	0.76	0.99	37
hsa-miR-213	0.74	0.99	34
mmu-miR-291a	0.82	0.98	53
mmu-miR-293	0.86	1.00	33
mmu-miR-294	1.00	0.99	68
mmu-miR-295	0.76	0.99	41
mmu-miR-322	0.81	0.99	34
hsa-miR-323	0.74	0.98	35
hsa-miR-325	0.79	0.99	46

Table 2. Microarray Performance.
(A). Results for the specification metrics were derived from Latin Square experiments; specificity (false positive detection rate; FPR) was determined by measuring the signal level detected above a defined threshold for negative control probes in a complex small RNA background;
(B). Lower limits of detection and linearity are displayed for all 10 synthetic RNA spikes used in the Latin Square experiments; Representive plots of 3 synthetic RNA spikes used in the Latin Square experiments are presented below.



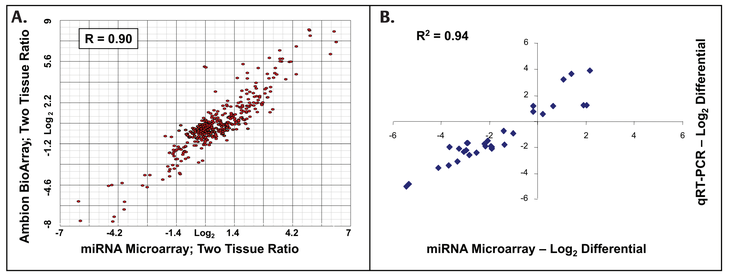


Figure 1. Microarray Concordance.
(A). Placenta and Testis miRNA $\log_2$ ratios for the Asuragen custom miRNA microarray (x-axis) and Ambion Bioarray (y-axis) are plotted. The $\log_2$ miRNA differential between tissues for each platform was first obtained and the resulting $\log_2$ values are used to produce the ratios plotted in the graph.
(B). A scatter plot of the $\log_2$ differential expression between the Asuragen custom miRNA microarray and qRT-PCR data was generated; the first step of this plot required generation of the $\log_2$ difference between signals obtained from placenta and bladder miRNA profiles for 30 different miRNAs by either qRT-PCR or microarray methods. Subsequently, the $\log_2$ differences for the 30 miRNAs investigated in this study are plotted against each detection method to generate the above graph.

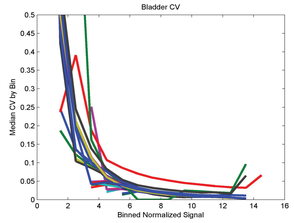


Figure 2. Plots of the Coefficients of Variation for Process Replicates.
Coefficients of Variation (CV) (y-axis) are plotted vs. mean binned $\log_2$ signal (x-axis) for 78 combinations of background subtraction, stabilization, scaling and normalization algorithms. Data collected from 10 full-process replicate arrays of fractionated human bladder total RNA was mean centered to a signal of $\log_2$ (100) to calibrate the reference scale before CV calculations were performed. For most data-processing approaches, the CV falls below 5% for $\log_2$ signals >6

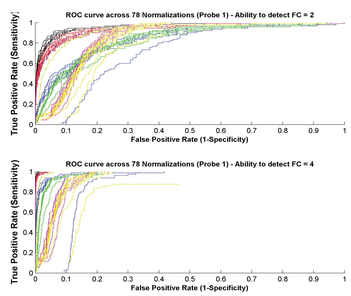


Figure 3. Receiver Operator Characteristic (ROC) Curves for all Spike-ins with Full Null-FC Distribution.
ROC curves color-coded by background correction (BC) and variance stabilization (VS) methods (black – no BC with VS; red - no BC no VS; blue – percentile BC with VS; magenta – percentile BC no VS; green – GC-content BC with VS; yellow – GC-content BC no VS) are displayed for fold-changes of two (top) or four (bottom) true positive and false positive differences identified from spike-in probes and a bladder RNA null-background, respectively. Differences in area under the curve (AUC) are representative of the ability to discriminate true from false differences in expression in a balance of sensitivity and specificity. The color-banding indicates the significant impact background correction and choice of variance stabilization have on the performance of the array platform in the context of sensitivity and specificity, with the inclusion of background correction having a detrimental impact. The impact is reduced but still apparent when detecting larger fold-change differences.

Tissue	Sanger 8.0 Total detected	Sanger 8.0 Unique by tissue	Additional Content[1] Total detected	Additional Content[1] Unique by tissue
Breast	190	1	998	46
Cervix	204	2	728	8
Heart	179	7	817	22
Kidney	207	2	784	24
Liver	152	2	615	9
Adipose	176	0	935	34
Ovary	203	3	715	26
Prostate	183	0	851	7
Rat Embryo	202	14	708	48
Lymph Node	171	0	969	50
Total Unique Probes:	**263**	**31**	**1407**	**274**
Putative miRNA prediction rate performance for the ten tissues was 10.3%.				

Table 3. Putative miRNA Content Performance.
Array signals from 10 different tissues were VSN transformed and tabulated according to the detection criteria described in the Figure 4 legend. A miRNA was determined to be unique with respect to the set of the ten tissues assayed if the filtering criteria were met only for the tissue indicated.
[1] Additional content refers to probes designed from Cummins, Xie, Berezikov and Proprietary prediction databases.

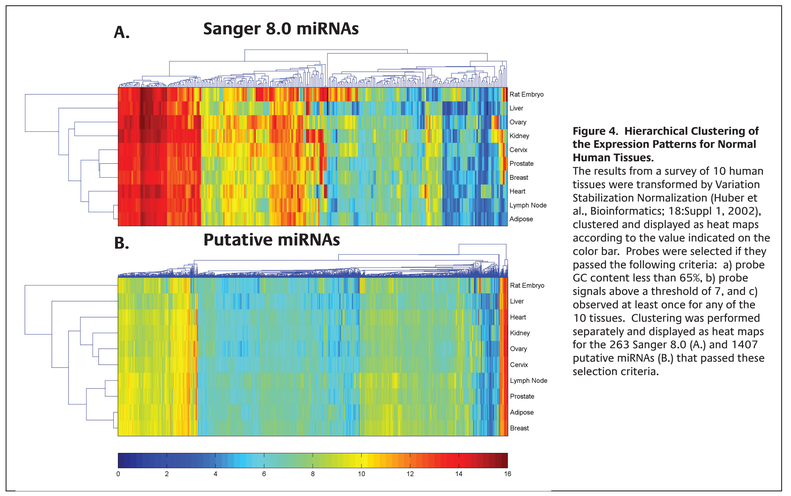


Figure 4. Hierarchical Clustering of the Expression Patterns for Normal Human Tissues.
The results from a survey of 10 human tissues were transformed by Variation Stabilization Normalization (Huber et al., Bioinformatics; 18:Suppl 1, 2002), clustered and displayed as heat maps according to the value indicated on the color bar. Probes were selected if they passed the following criteria: a) probe GC content less than 65%, b) probe signals above a threshold of 7, and c) observed at least once for any of the 10 tissues. Clustering was performed separately and displayed as heat maps for the 263 Sanger 8.0 (A.) and 1407 putative miRNAs (B.) that passed these selection criteria.

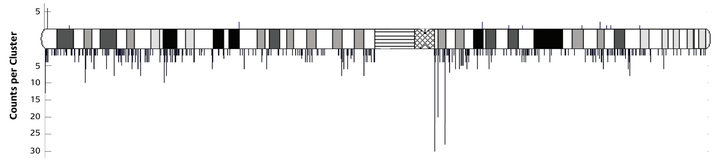


Figure 5. Chromosomal Location of Sanger 8.0 and Putative miRNA Content on Human Chromosome 1.
Chromosomal locations of the miRNA genes were mapped to the human genome (Human NCBI Build 36 (hg18)) if a perfect match was observed for each pre-miRNA sequence. Clusters are generated by collecting all miRNAs located within a 200,000 nucleotide window, and displayed separately for Sanger 8.0 (above the cytoband image of Chromosome 1) and putative content (below).

CONCLUSION

Asuragen, Inc. has designed a custom Affymetrix miRNA microarray that contains comprehensive content for 15025 unique miRNAs collected from public and private databases. Experimental specification metrics demonstrate the performance of the custom miRNA microarray to be very sensitive, (to <33 amol input RNA), dose-responsive in a linear fashion with 2.71 common logs of dynamic range, and reproducible (CVs <10%); (Table 2). The array demonstrates high concordance (R>0.9) when compared to results obtained from the Ambion Bioarray or TaqMan qRT-PCR (Figure 1). In an attempt to address genome-level estimations of miRNA content, 1407 potentially novel miRNAs were detected from a survey of 10 normal tissues. Each tissue resulted in detection of 3.5 to 5.6 more miRNAs than detected within the Sanger content alone. All ten tissues revealed more tissue-specific content than the Sanger content. This includes adipose, prostate, and lymph node for which no unique miRNAs are detected within the miRNAs represented in the Sanger database. In addition to the tissue-specific identification based on detection, the putative miRNA expression patterns suggest tissue-specific signatures (Figure 4), an observation that is also true for the Sanger miRNAs. Combined, the Sanger and putative miRNA probes could increase the accuracy of tissue classification. Lastly, known miRNAs have been suggested to cluster in genomic regions in a non-random fashion and, therefore, novel miRNAs might reside in the vicinity of these validated miRNA genes (Berezikov et al., Cell, 120:21-24, 2005). Preliminary analysis of putative miRNA locations reveal significant enrichment of miRNA gene clusters throughout chromosome 1 (Figure 5), with the exception of the centromeric/pericentromeric region. While this data is tantalizing, further, in-depth statistical analysis is required to determine if putative miRNAs have the propensity to cluster or flank known miRNAs, and if this clustering is organized or random. In conclusion, the comprehensive, miRNA Affymetrix microarray designed and provided by Asuragen, Inc. is a valuable research tool for assisting scientific groups in uncovering the miRNA expression patterns of normal and cancer tissues. Moreover, proper analysis of the array data can be used to develop diagnostic, prognostic, and therapeutic strategies for patient health care.