

Experimental Selection and Performance of 60-mer Oligonucleotide Probes for Profiling Global Gene Expression

Patrick J. Collins, Tri B. Doan, Peter G. Webb, Sandra L. Tang, Keith Butler, Anya Tsalenko, Stephanie Fulmer-Smentek & Karen W. Shannon.

Agilent Technologies
M/S 25U
3500 Deer Creek Road
Palo Alto, CA 94304

Ordering Information

www.agilent.com/chem/dna
u.s. and canada 1 800 227 9770
japan +0120 477 111
europe: marcom_center@agilent.com
global: dna_microarray@agilent.com

© Agilent Technologies, Inc. 2003

Research Use Only

LifeSeq is a trademark or registered trademark of Incyte Genomics, Inc. in the U.S. and other countries. Rosetta Resolver is a U.S. registered trademark of Rosetta Inpharmatics. Information, descriptions and specifications in this publication are subject to change without notice. Printed in the U.S.A.

Printed in the U.S.A.
May 1, 2003
5988-9189EN

Synopsis: Agilent has developed a novel experimental method for the selection and validation of highly specific oligonucleotide microarray probes. This probe optimization strategy is based on the assumption that, if multiple probes for a particular gene are specific for that gene, then they will show similar differential expression behavior. Optimized 60-mer probes to represent genes are selected from clusters of computationally determined candidate probes that perform similarly across a number of differential gene expression experiments. This method also enables verification of computationally predicted gene assemblies. Agilent's Human 1A Oligo microarray consists of 88% experimentally validated optimized probes that hybridize to all known splice-variants of the genes to which they were designed. Data are presented which demonstrate the power of this experimental method for selection of microarray probes which provide extremely accurate and reliable gene expression data.

ABSTRACT

In recent years, DNA microarrays have become key tools in providing an understanding of cellular gene expression patterns. Gene transcripts have typically been represented on microarrays by whole cDNA molecules or by multiple short oligonucleotides, but in general, selection of these sequences is not based on their performance in expression profiling experiments. We report here on the development of the Human 1 Oligonucleotide microarray using a novel experimental method for selection of 60-mer oligonucleotide probes and demonstrate the performance of these probes in profiling global gene expression. One probe is present on this microarray for each of more than 17,000 full-length genes in Incyte Genomics' LifeSeq Foundation database. These probes were designed to consensus regions across all gene transcripts, so are splice variant non-specific. Each is annotated with information on protein function, as well as data describing its performance in validation experiments. Probes are printed in a 22,575 feature microarray using Agilent's SurePrint technology. Data are presented which demonstrate the performance of these microarrays with regard to their consistency, accuracy and sensitivity. The results presented demonstrate the power of this experimental method for selection of microarray probes to provide extremely accurate and reliable gene expression data.

OBJECTIVE

Development of a probe selection method to:

1. Identify multiple specific 60 mer probes per gene and utilize empirical performance data to select one Optimum Probe to represent each gene
2. Enable verification of computationally predicted gene assemblies
3. Demonstrate the consistency and accuracy of the Optimum Probes in profiling gene expression in a high density microarray format (Human 1A Oligo Microarray, Agilent P/N G4110A).

Experimental Procedures

Ten candidate 60 mer oligonucleotide probes were computationally selected for each gene using Agilent's probe design algorithms. These probes were synthesized in situ using SurePrint inkjet technology on a 22,575 feature microarray format. For all experiments, candidate probe microarrays were hybridized with cRNA targets representing pairs of RNAs from Table 1. These targets were generated using Agilent's Linear Amplification Kit such that one RNA in each pair was labeled with cyanine 5 and the other was labeled with cyanine 3. Four replicate hybridizations were performed for each experiment using the reagents provided in Agilent's In situ Hybridization Kit Plus according to the recommended procedure. Hybridized microarrays were scanned using Agilent's dual-laser DNA Microarray Scanner and data extracted using Agilent's Feature Extraction software. Replicate log ratio values derived for each candidate probe from each experiment were then combined into one value using an error-weighted method.

CAST clustering algorithms (Ben-Dor, A. et al., (1999) J. Comput. Biol 6:281-297) were used to identify probes within each gene that exhibit the same performance across experiments in the log ratio values derived from them. A score was determined for each cluster of probes for each gene and this was used to rank clusters. The Optimum Probe to represent each gene was identified based on sensitivity and reproducibility data from the highest ranked cluster.

Table 1. RNAs used in Hybridization Target Preparation

RNA	Type
Clontech Univ. Ref	Total
Stratagene Univ. Ref.	Total
Placenta	polyA ⁺
Spleen	polyA ⁺
Liver	polyA ⁺
Lung	polyA ⁺
Brain	polyA ⁺
MG63	polyA ⁺
K-562	polyA ⁺
HeLa	polyA ⁺

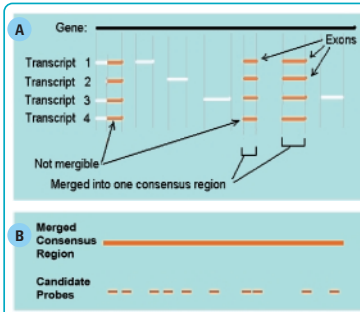


Figure 1: Candidate probe design. The source of gene sequences to which probes were designed was Incyte Genomics' LifeSeq Foundation database (September 2002 release). Exons which were common to all full-length transcripts were identified for each gene, and merged to form one or more splice-variant non-specific consensus regions (A). Ten candidate probes were computationally selected for each consensus region (B) using standard probe design algorithms and synthesized in 22,575 feature microarrays.

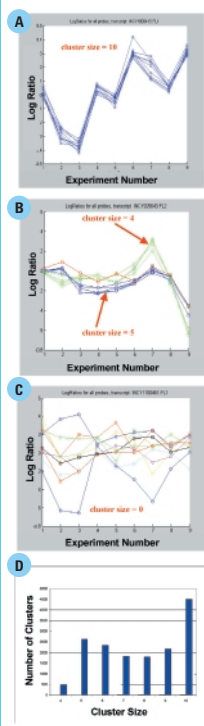


Figure 2: Clustering of candidate probe data. CAST clustering algorithms were used to identify probes within each gene that exhibit the same performance across experiments in the log ratio values derived from them. i.e., are co-regulated. Similarities between different probes was calculated as Pearson correlations between corresponding expression values. The resulting probe clusters were ranked according to the number of probes they contained. Figure 2A presents an example in which all probes behaved very similarly across all experiments and so fell into one cluster of ten probes. For the gene represented in figure 2B, five probes fell into the top-ranked cluster (blue), four into another cluster (green) while one outlier (red) did not consistently behave similarly as the probes in either of these two clusters. None of the candidate probes for the gene represented in figure 2C displayed co-regulation across the experiments so no clusters were identified. Figure 2D presents a distribution of cluster sizes for the top-ranked clusters for each gene for which a probe was empirically validated. Probes were computationally selected for any genes for which the top-ranked cluster consisted of less than four probes.

Figure 3: Selection of Optimum Probes from clustered probe data.

Candidate probe performance data for four individual genes are presented in figure A-D as log ratio vs. signal intensity plots. In each case, probes in the left panel are colored by experiment, and in the right panel are colored green if they fall into the top-ranked cluster and red if they do not. In A, all ten candidate probes behave very similarly across all ten experiments with regard to the log ratio values derived from them, and consequently form one cluster. All of these probes are in the top-ranked cluster so all are colored green in the right panel. For the genes in B and C, all the probes do not fall into one cluster as indicated by those that are colored red in the right panel. No clusters of co-regulated probes were identified for the gene in D, so no top-ranked probes are indicated in the panel on the right side. When a top-ranked cluster of at least four probes was identified for a gene, an Optimum Probe was selected from the cluster members based on empirical sensitivity and reproducibility data. Optimum Probes are indicated as asterisks for A-C. For genes such as the one presented in D, where no cluster of at least four probes was identified, Optimum Probes were computationally selected.

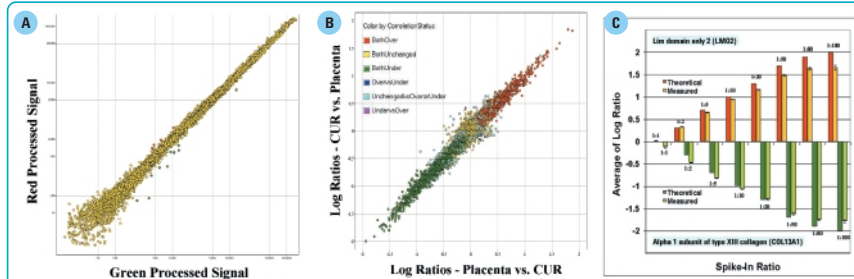
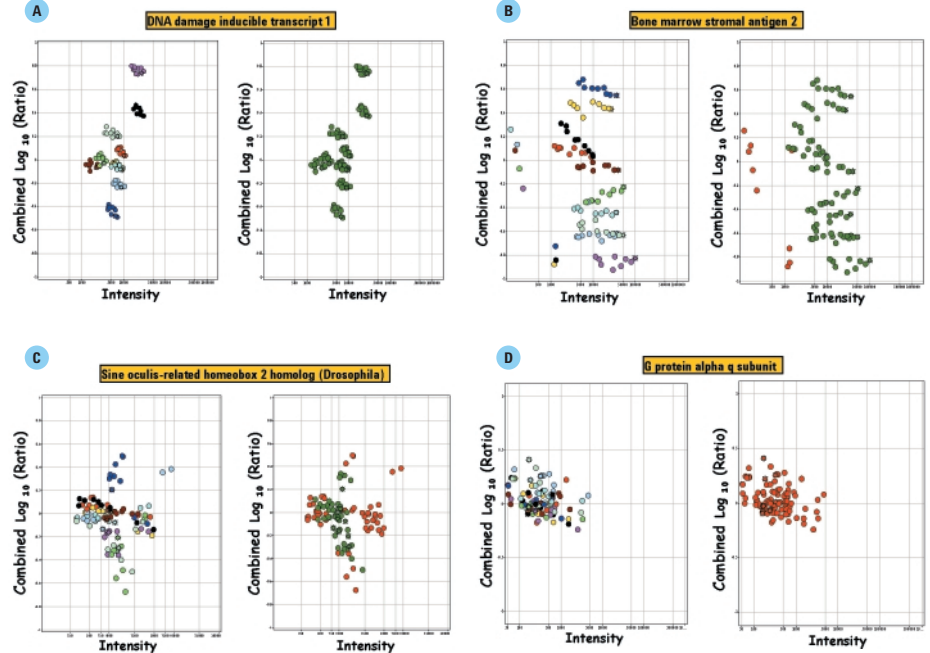


Figure 4: Performance of Optimum Probes in profiling global gene expression. The set of Optimum Probes was synthesized in a 22,575 feature microarray format, the Human 1A Oligo Microarray. The level of noise observed with this microarray is demonstrated using a self vs. self hybridization (A) in which the target consisted of cyanine 5- and cyanine 3-labeled cRNA, both prepared from HeLa polyA⁺ RNA. The data converge tightly to the diagonal corresponding to the expected differential of 1, with only 0.075% of log ratio values determined to be significantly different than 0 ($p < 0.001$) (data points colored red or green). The consistency of the system in measuring significant log ratios was determined by plotting data from "dye-swapped" hybridizations against one another (B). cRNA targets in one polarity consisted of Clontech Universal Reference -cyanine 3 and Placenta-cyanine 5 and the dyes were swapped for the reciprocal polarity. The significance of Log Ratio values was determined across four replicate hybridizations ($p < 0.001$). The tight convergence of the data to the diagonal illustrates the consistency in measuring log ratios, as well as the low level of dye-bias in the system. The accuracy of the probes on the Human 1A Oligo Microarray in measuring differential gene expression was demonstrated using a "spike-in" experiment (C). Synthetic cyanine 3- and cyanine 5-labeled targets were generated to two genes, LMO2 and COL13A1, which are not expressed at detectable levels in HeLa cells. Ten replicate probes are present on the microarray for both of these genes. The synthetic targets were spiked at eight different ratios into eight HeLa self vs. self hybridizations such as the one described above (A). The cyanine 5-labeled LMO2 target, was added to to each hybridization to a constant concentration of 2 pM, while the cyanine 3-labeled target was varied. For COL13A1, the cyanine 3-labeled target, was added to to each hybridization to a constant concentration of 2 pM and the cyanine 3-labeled target was varied. The values shown are averaged over the ten replicate probes and error bars represent one standard deviation.

Final Human 1A Oligo Microarray Specifications (P/N G4110A)

- 22,575 feature microarrays on 1" x 3" glass slides
- 17,986 60-mer probes representing 17,086 genes
- 15,032 probes (88%) selected based on empirical data and 2,054 (12%) computationally selected
- > 3,000 blank features for custom use

Conclusions

- The method described identifies 60 mer microarray probes which hybridize to all known splice-variants of the genes to which they were designed
- Highly specific Optimum Probes were experimentally selected for 88% of the genes tested
- This method enables verification of computationally predicted gene assemblies
- Optimum probes are highly consistent and accurate in profiling gene expression