PRODUCT APPLICATION FOCUS

. . . a forum for manufacturers to describe the current and potential applications of new research instruments or products.

Cross-Hybridization of Closely Related Genes on High-Density Macroarrays

N.A. Miller, Q. Gong, R. Bryan, M. Ruvolo, L.A. Turner, and S.T. LaBrie
Incyte Genomics, St. Louis, MO, USA

BioTechniques 32:620-625 (March 2002)

ABSTRACT

DNA macroarrays are used in many areas of molecular biology research for applications ranging from gene discovery to gene expression profiling. As an increasing number of specialized macroarrays containing genes related by function or pathway are becoming available, a question that needs to be addressed is the level of hybridization signal specificity between highly similar genes that can be achieved. We have examined the ability of our LifeGrid™ macroarrays to distinguish hybridization signals between closely related genes.

We determined the level of cross-hybridization among genes ranging from 52% to 94% sequence identity. Fragments of genes from five protein families were arrayed onto nylon filters. The filters were subsequently hybridized with a ^{33}P-labeled probe prepared from a pool of synthetic mRNA transcripts containing a representative of each protein family. We found that fragments containing sequences with up to 94% sequence identity displayed relatively little cross-hybridization. We conclude that this macroarray system is very specific and that hybridization signals from closely related genes can be reliably measured.

INTRODUCTION

DNA macroarrays have become very useful in the elucidation of mRNA expression patterns and the identification of novel genes. Expression macroarrays generally consist of nylon membranes spotted with PCR products produced from cDNA clones. They are hybridized with radioactively or fluorescently labeled probes transcribed from mRNA. The information gathered from these arrays can be used to characterize complex biological interactions and processes at the level of transcription.

Specialized expression macroarrays containing genes related by function, tissue, or pathway are becoming widely adopted by researchers studying mRNA expression patterns (1,8,11). The results from hybridization of these filters provide researchers with a tremendous amount of data that can be used to identify genes up- and down-regulated in almost any imaginable sample combination, including different developmental stages, disease states, and environmental exposures.

The correct selection of differentially regulated genes for further functional study is dependent on the quantitative accuracy that can be achieved by the array (2). The validity of the results from a hybridization experiment depends on several factors. These include variations in the amount of DNA deposited in each spot, the quality of the mRNA used to synthesize the probes, the efficiency of the labeling reaction, and nonspecific hybridization of probe to the array (2,5,12,13). Because of the high number of closely related genes on specialized arrays, we decided to examine the signal resulting from cross-hybridization between genes with a high level of sequence identity.

The level of cross-hybridization between genes with high sequence identity is also of interest because arrays are not always available for the species of interest, so cross-species hybridizations are often done. Orthologous genes between human and mouse and between human and rat both have a mean of approximately 85% sequence identity (9,10). It is important to know whether the hybridization system has the ability to identify altered gene expression across species.

Similar studies have been done that analyze cross-hybridization on microarrays. In one study in which glass slides were arrayed with PCR products generated from cDNA clones, cross-hybridization was observed in genes with 70%–80% sequence identity (4). In another study, cross-hybridization occurred when an arrayed 50-mer oligonucleotide had 75% or greater sequence identity to the cDNA probe and when a 50-mer oligonucleotide with 50%–75% overall sequence identity to the probe had a greater than 15-bp stretch of complementary sequence (7).

In this study, we compare hybridization levels among related members in five gene families to determine the level of specificity of the LifeGrid™ Expression Macroarray system (Incyte Genomics, St. Louis, MO, USA).

MATERIALS AND METHODS

Sequencing

For the purpose of this study, full-length sequences of the clones were obtained by combining 3′ and 5′ sequencing reads done on a megaBACE™ sequencer (Amersham Biosciences, Piscataway, NJ, USA). The clones that did not have complete full-length sequences after the first round of sequencing were finished by primer walking. Briefly, another round of sequencing was done using primers that were designed toward the ends of the initial 5′ and 3′ reads. The sequence reads were assembled with Phrap (3).

Identity Scoring

One clone from each protein family was designated as the parent clone and is the clone to which the other clones in the family were compared by pairwise alignment. The sequences of each family member were compared to the parent by (*i*) their percent sequence identity to the parent clone over the region of alignment and (*ii*) their LALIGN score. The LALIGN score is a number generated by a homology-scoring algorithm that factors in both the number of overlapping bases and the sequence identity between two sequences (6). Table 1 lists the percent identity, number of bases of each clone that align with the appropriate parent probe, LALIGN scores, and GenBank® accession numbers for all genes used in this study. (These clones generally contain truncated cDNA inserts and so will not exactly correspond to the sequence reported in GenBank.)

Macroarray

PCR products were made from all target cDNA sequences using vector-specific primers (14). The products were quantified using the PicoGreen® quantitation system (Molecular Probes, Eugene, OR, USA), and their concentrations were normalized to approximately 200 µM. The dsDNA PCR products were denatured with 1.8 N NaOH and spotted in quadruplicate onto a nylon membrane (Millipore, Bedford, MA, USA) by a MegaGridder gridding robot (Incyte Genomics) using custom pins and protocols. The DNA was cross-linked to the nylon support by UV irradiation at 80 mJ (Bio-Rad Laboratories, Hercules, CA, USA).

PCR, mRNA Synthesis, and Probe Generation

Synthetic mRNA probe templates for each parent clone were produced. First, we PCR-amplified each cDNA insert with gene-specific primers (excluding all vector sequence). In a second round of PCR, a polyA tail was added to the 3′ end, and the T7 promoter sequence (5′-NNNNNNTAATAC-

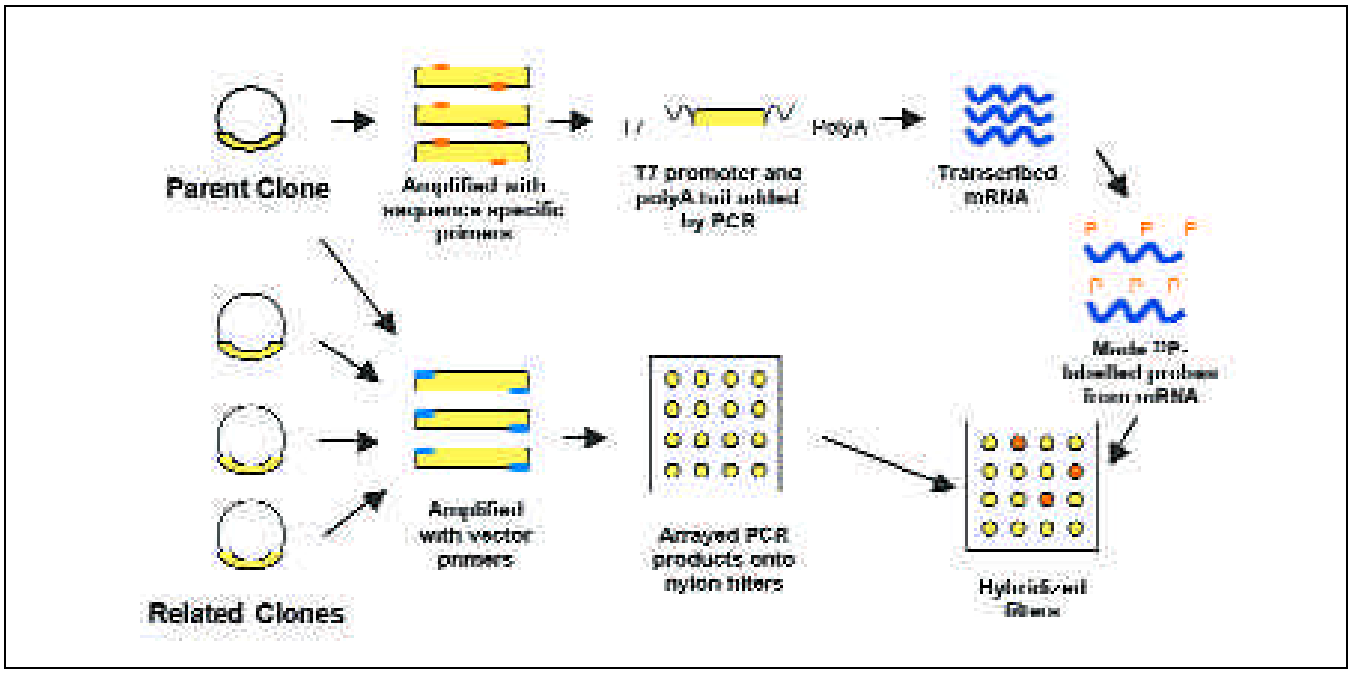

Figure 1. Schematic diagram of target DNA and probe generation and hybridization. Parent clones were amplified with sequence-specific primers to incorporate a polyA tail and a T7 promoter sequence. The resulting DNA was used to produce mRNA transcripts that were used to generate ^{33}P-labeled probes. Target DNA was synthesized from all clones by PCR with vector-specific primers. The products were normalized to 200 µM and gridded in a quadruple-spotted pattern onto nylon filters for hybridization with the ^{33}P-labeled probe.

Table 1. Protein Family Members, Identity Relationships, and Cross-Hybridization Levels

Family Name	Clone	GenBank Accession Number	Percent Identity	Overlap (bp)	LALIGN Score	Percent Signal Intensity	
						Simple	Complex
p450s	**1**	**g177181**	**100**	**484**	**2388**	**100**	**100**
	2	g181301	94	479	2145	39.9	0.6
	3	g181301	93	484	2079	10.2	0
	4	g177181	78	173	494	7.6	0
	5	g263688	79	476	1488	3.9	1.0
	6	g177181	87	479	1870	9.2	4.5
	7	g181359	70	432	981	0.2	0.9
	8	g181359	66	330	612	3.7	1.4
	9	g180986	59	106	134	0.2	0
	10	g181359	72	68	159	0.3	3.18
	11	g1008046	56	427	373	0.5	0.2
Chemokines	**1**	**g2916795**	**100**	**497**	**2485**	**100**	**100**
	2	g288396	68	453	860	6.5	0.1
	3	g187447	68	457	873	20.9	0
	4	g2326515	67	112	227	6.6	0.2
	5	g4097420	65	406	643	4.4	0.2
	6	g178017	63	180	266	0.8	0.2
G Proteins	**1**	**g6164866**	**100**	**170**	**821**	**100**	**100**
	2	g3360459	65	37	58	9.4	0.1
	3	g3360459	65	37	58	5.2	0.2
	4	g7022043	67	170	336	21.5	0.2
	5	g7022042	58	60	55	6.1	0.3
	6	g3329379	68	50	76	7.4	2.5
Metalloproteases	**1**	**g349925**	**100**	**1430**	**7131**	**100**	**100**
	2	g992403	85	1430	5100	8.9	3.1
	3	g687242	60	89	81	4.3	9.6
	4	g238586	59	768	793	5.4	1.1
Other Proteases	**1**	**g36060**	**100**	**951**	**4748**	**100**	**100**
	2	g565646	52	679	678	3.4	1.3
	3	g7302349	54	252	77	3.5	2.3

For each clone the GenBank accession number, the percent sequence identity of each clone with the parent probe, and number of overlapping bases with the parent probe is listed (parent clones are in bold). In addition, the LALIGN score, which takes into account both the percent identity and base pairs of overlap of each clone with the appropriate parent probe sequence, is included. The hybridization intensities for each clone relative to the parent when probed with a simple probe (100 ng each parent mRNA) and with a complex probe (250 ng placental mRNA plus 0.25 ng each parent mRNA) are listed as a percentages in the last two columns.

GACTCACTATAGGGAG-3′) was added to the 5′ end of each parent DNA fragment. Synthetic mRNA was synthesized with a T7 transcription kit (Ambion, Austin, TX, USA). The mRNA was quantified with RiboGreen™ (Amersham Biosciences). For hybridization, the mRNA transcripts from each parent clone (100 ng/clone) were pooled, and labeled probes were generated with Moloney murine leukemia virus (MMLV) reverse transcriptase (Ambion) and [α-^{33}P]dCTP (3000 Ci/mmol, 10 µCi/µL) (NEN® Life Science Products, Boston, MA, USA) using modified oligo-dT primers, according to the Human LifeGrid 1.0 Product Manual (Incyte Genomics). To ensure that the entire length of the PCR product was transcribed, individual probes were generated from each parent clone and sized on an 8% polyacrylamide gel (data not shown). For hybridization experiments using a complex mRNA probe, 250 ng placental mRNA (BD Biosciences Clontech, Palo Alto, CA, USA) was combined with 0.25 ng each parent mRNA for probe generation.

Hybridization, Imaging, and Analysis

Filters were hybridized overnight at 42°C in NorthernMax™ Pre-hyb/hyb Buffer (Ambion), rinsed for 5 min with 2× SSC, washed twice for 30 min at 68°C in a 2× SSC, 1% SDS solution, and washed twice for 30 min at 68°C in a 0.6× SSC, 1% SDS solution. The filters were wrapped in plastic and exposed to phosphor screens overnight. The screens were then imaged on a STORM® PhosphorImager system (Amersham Biosciences). The images were analyzed using ArrayVision™ 4.0 software (Imaging Research, St. Catherine's, ON, Canada).

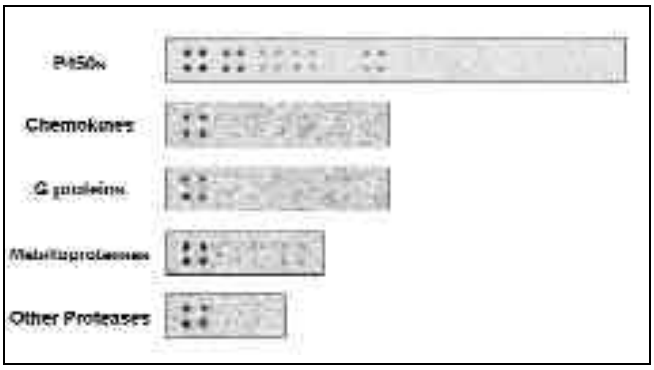

Figure 2. A representative hybridized filter of each gene family. The PCR products were spotted in quadruplicate and are in order from left to right.

RESULTS AND DISCUSSION

Five sets of clones (with 3–11 members) containing related sequences were chosen for the cross-hybridization study. The gene families in this study are chemokines, G proteins, cytochrome p450s, metalloproteases, and other proteases. The sequence identity of the selected genes ranges from 52% to 94%. Each cloned insert was amplified and gridded in quadruplicate on nylon macroarrays. To measure cross-hybridization, we chose a single member (the parent) from each of the five sets and generated synthetic mRNA for labeling. We pooled 100 ng synthetic mRNA from each parent (this amount is 10- to 100-fold higher than the highest expressed gene generally found in 500 ng mRNA) and prepared radiolabeled first-strand cDNA probes using a modified polyT primer and MMLV reverse transcriptase. The probes were applied to the macroarrays and processed using high-stringency conditions (Figure 1). Each experiment was repeated several times, and the hybridization intensities (average pixel density X area) of all spots representing each clone were averaged. Figure 2 contains a representative image of a hybridized filter for each gene family.

The percentages of cross-hybridization were calculated by dividing the average signal intensity of each family member into the average signal intensity of the appropriate parent clone. The average percent signal intensity and average percent sequence identity of each clone compared to its parent for each of the gene families from multiple experiments (sep-

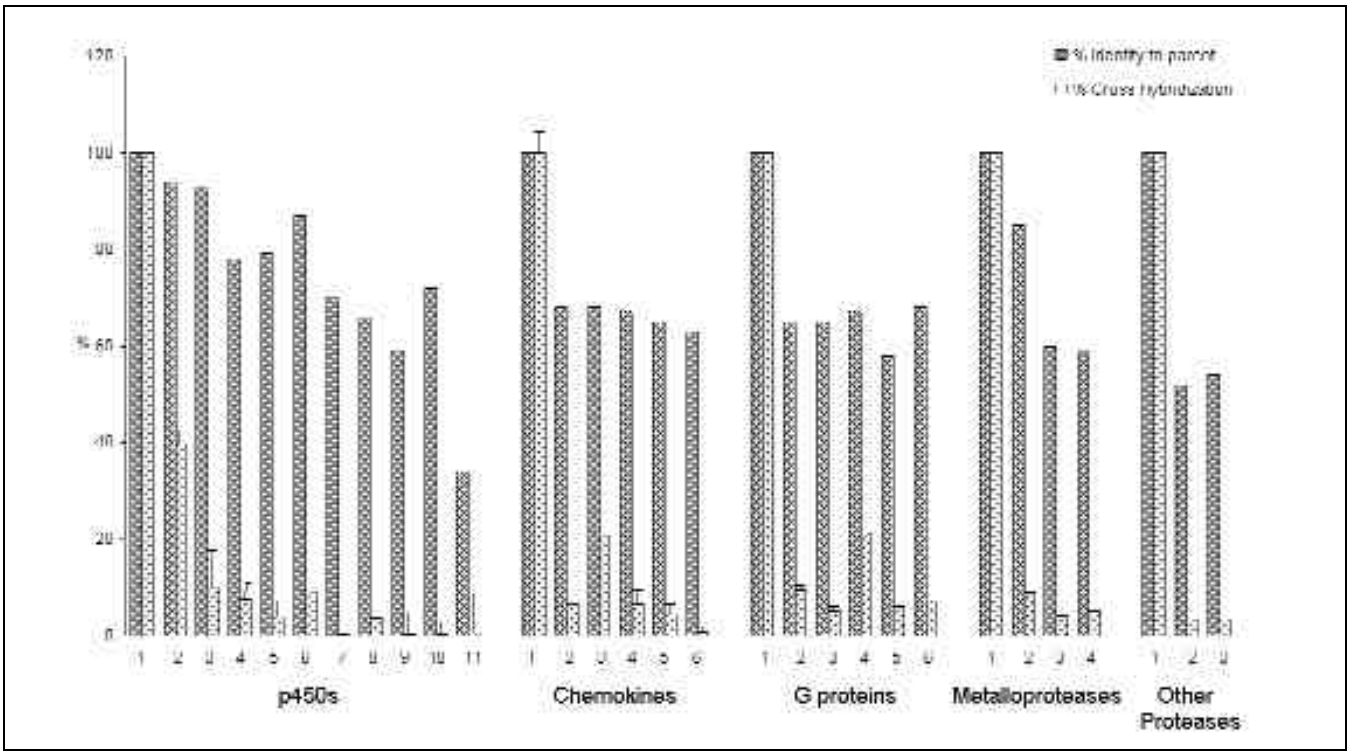

Figure 3. Percent identity and percent cross-hybridization between the parent clone and each member of the five protein families. The first clone in each family is the parent clone (identity and cross-hybridization = 100%). The percentage of cross-hybridization between the parent clones and the other clones in each family were calculated from the average hybridization intensity from each clone. The LALIGN scores of the clones can be found in Table 1. The error bars represent the standard error over 16 data points.

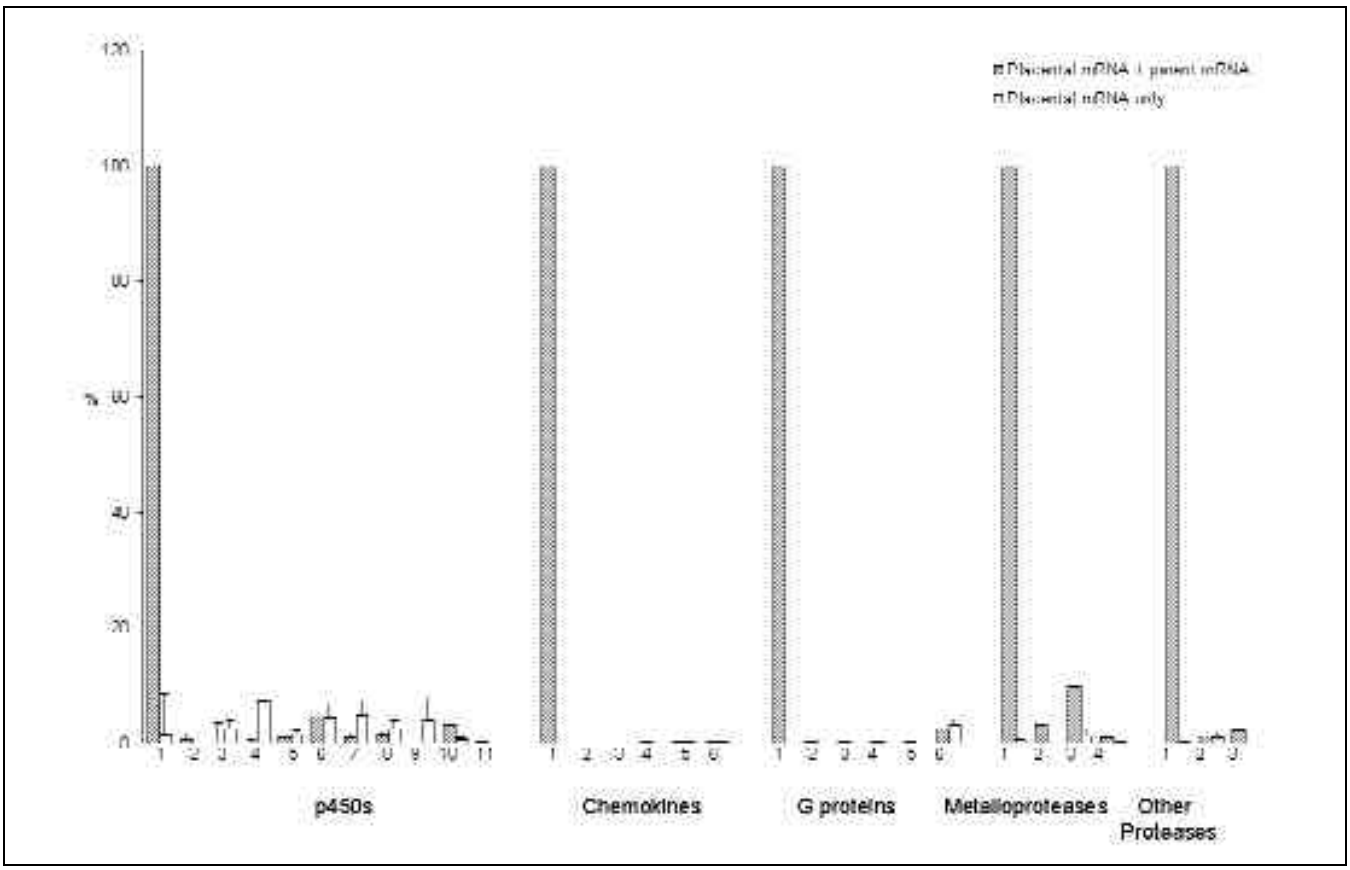

Figure 4. Percent cross-hybridization between the parent clones and each family member when hybridized as part of a complex probe. The first bar in each family represents the parent clone when hybridized with complex mRNA spiked with mRNA of the parent clone (1 transcript/1000). The rest of the bars in each family represent the percentage of hybridization relative to the first parent bar. The error bars represent standard error from 12 data points.

arate probe preparations and hybridizations) are listed in Table 1 and plotted in Figure 3. In all but two cases, clones with up to 93% sequence identity show about 10% or less cross-hybridization. Further examination of sequences of the two exceptions (chemokine family member 3 and G protein family member 4) did not result in an explanation for their unusually high cross-hybridization levels. Neither clone is a splice variant of the parent nor has a disproportionate guanine/cytosine content (data not shown).

In a typical hybridization experiment, a complex pool of mRNA is used to generate the probe. An experiment was done to ensure that similar results could be achieved with a complex probe in which the parent mRNAs are present at a physiological level. Filters were hybridized with probes that were generated from 250 ng placental mRNA, and from 250 ng placental mRNA spiked with 0.25 ng each parent mRNA (1 transcript/1000 is a moderate to high level in an mRNA population). The levels of hybridization of each clone are reported as percentages of the level of hybridization of the parent clone (hybridized with 0.25 ng specific mRNA) in Figure 4.

We consider these experiments to represent two extreme cases—an experiment with a very high amount of a single mRNA species and one with a lower, more typical level of mRNA. In our simple probe experiment, we found that if the transcript of one member of a gene family is present at a very high level, then it may cross-hybridize with the other family members when the sequence identity is greater than 90% over 100 bp. In the other scenario, a member of each gene family was represented at a normal level of about 1 in 1000 transcripts. We tested cross-hybridization at this level in the framework of a complex probe made from placental mRNA and found very little cross-hybridization. These results indicate that only under rare circumstances when a very similar transcript is very highly expressed will gene family cross-hybridization dramatically compromise results in this macroarray system.

These data suggest that this system is too stringent for cross-species hybridizations. By reducing the hybridization temperature or increasing the salt concentration of the hybridization buffer, it may be possible to obtain reproducible results from transcripts of orthologous genes. However, these results may be compromised by signals from transcripts of paralogous genes that happen to have similar levels of sequence identity to the arrayed gene fragments.

DNA macroarrays that are designed for the study of specific pathways and cellular functions are becoming increasingly available. The subsets of genes that make up these arrays often include gene families containing several closely related genes. This necessitates that studies be done to determine the ability of the array and hybridization procedures to distinguish between closely related genes. We have examined five gene families containing members with up to 94% sequence identity. Most clones with up to approximately 90% identity show relatively little cross-hybridization (<10%). The information from this study will be useful in designing clone sets for and interpreting data from specialized macroarrays. The data presented here demonstrate that the LifeGrid macroarray system is very specific and that hybridization signals from closely related genes can be reliably measured.

REFERENCES

1. **Bowtell, D.D.L.** 1999. Options available—from start to finish—for obtaining expression data by microarray. Nat. Genet. Suppl. *21*:25-32.
2. **Cole, K.A., D.B. Krizman, and M.R. Emmert-Buck.** 1999. The genetics of cancer-a 3D model. Nat. Genet. Suppl. *21*:38-41.
3. **Ewing, B. and P. Green.** 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. *8*:186-194.
4. **Girke, T., J. Todd, S. Ruuska, J. White, C. Benning, and J. Ohlrogge.** 2000. Microarray analysis of developing Arabidopsis seeds. Plant Physiol. *124*:1570-1581.
5. **Gonzalez, P., J. Zigler, Jr., D. Epstein, and T. Borras.** 1999. Identification and isolation of differentially expressed genes from very small tissue samples. BioTechniques *26*:884-892.
6. **Huang, X. and W. Miller.** 1991. LALIGN finds the best local alignments between two sequences. Adv. Appl. Math. *12*:373-381.
7. **Kane, M.D., T.A. Jatkoe, C.R. Stumpf. J. Lu, J.D. Thomas, and S.J. Madore.** 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. *28*:4552-4557.
8. **Lennon, G. and H. Lehrach.** 1991. Hybridization analyses of arrayed cDNA libraries. Trends Genet. *7*:314-317.
9. **Makalowski, W. and M. Boguski.** 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA *95*:9407-9412.
10. **Makalowski, W., J. Zhang, and M. Boguski.** 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res. *6*:846-857.
11. **Piétu, G., O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Samson, R. Houlgate et al.** 1996. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array library. Genome Res. *6*:492-503.
12. **Schuchardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel.** 2000. Normalization strategies for cDNA microarrays. Nucleic Acids Res. *28*:e47.
13. **Vernier, P., R. Mastrippolito, C. Helin, M. Bendali, J. Mallet, and H. Tricoire.** 1996. Radioimager quantification of oligonucleotide hybridization with DNA immobilized on transfer membrane. Anal. Biochem. *235*:11-19.
14. **Zhao, Z., H. Hashida, N. Takahashi, Y. Misumi, and Y. Sakaki.** 1995. High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. Gene *156*:207-213.

Address correspondence to Dr. Sam LaBrie, Incyte Genomics, 3160 Porter Dr., Palo Alto, CA 94304, USA. e-mail: labrie@incyte.com