Published online July 1, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 11 ***e93***
DOI: 10.1093/nar/gnh084

A method for cross-species gene expression analysis with high-density oligonucleotide arrays

Wan Ji*, Wenli Zhou, Keqin Gregg, Nan Yu, Scott Davis and Sara Davis

ViaGen Inc., 12357-A Riata Trace Parkway, Suite 100, Austin, TX 78727, USA

Received January 22, 2004; Revised April 7, 2004; Accepted May 26, 2004

ABSTRACT

DNA microarrays have been widely used in gene expression analysis of biological processes. Due to a lack of sequence information, the applications have been largely restricted to humans and a few model organisms. Presented within this study are results of the cross-species hybridization with Affymetrix human high-density oligonucleotide arrays or GeneChip® using distantly related mammalian species; cattle, pig and dog. Based on the unique feature of the Affymetrix GeneChip® where every gene is represented by multiple probes, we hypothesized that sequence conservation within mammals is high enough to generate sufficient signals from some of the probes for expression analysis. We demonstrated that while overall hybridization signals are low for cross-species hybridization, a few probes of most genes still generated signals equivalent to the same-species hybridization. By masking the poorly hybridized probes electronically, the remaining probes provided reliable data for gene expression analysis. We developed an algorithm to select the reliable probes for analysis utilizing the match/mismatch feature of GeneChip®. When comparing gene expression between two tissues using the selected probes, we found a linear correlation between the cross-species and same-species hybridization. In addition, we validated cross-species hybridization results by quantitative PCR using randomly selected genes. The method shown herein could be applied to both plant and animal research.

INTRODUCTION

Modulation of gene expression represents one of the fundamental mechanisms in biology underlining both physiological developments and pathological derailments. Microarrays have become instrumental in comprehensive investigation of gene expression, providing important new insights into the molecular mechanisms of biological processes (1,2). Despite its potential to revolutionize the genetic analysis of biological systems, the powerful method has been largely restricted to human and a few model species, mainly due to the lack of sequence information. Manufacturing high-density DNA microarrays is time-consuming and expensive, involving sequencing gene transcripts on a large scale from various tissues, clustering the expressed sequence tags (ESTs) to generate unique transcripts, annotating the transcripts, designing probes, and synthesizing or spotting the probes on glass plates (3,4). Ideally such a huge undertaking could be avoided and some existing microarrays could be transplanted for cross-species analysis.

Recent studies have suggested that all mammalian species diverged from a common ancestor about 100–65 million years ago (5). The short geological time span plus the preservation of protein functions result in a high degree of nucleotide sequence conservation among mammalian species. Makalowski and Boguski (6) have compared 2820 orthologous rodent and human sequences and revealed a high resemblance of gene transcripts between the two distantly related mammals. They showed that the average identity between human and mouse in protein-coding nucleotide sequences (CDS) is 85.9%, in 5′-untranslated region (5′-UTR) is 69.7%, and in 3′-untranslated region (3′-UTR) is 71.0%. Many efforts have been spent previously on cross-species gene expression analysis by using conserved heterologous probes and adjusting hybridization conditions, such as temperature and buffer components (7). Such empirical approaches, however, are inadequate and impractical for a microarray due to its huge variety of sequences and its exorbitant cost. We decided to explore a unique feature of Affymetrix GeneChip®—each gene was represented by multiple and, most times, non-overlapping oligonucleotide probes, 11 to 20 of them in general. We hypothesized that sequence conservation within mammals is high enough for one or several of the probes to be the same or nearly the same as its orthologs in human. Then, the selected probe(s) would provide valuable information for gene expression analysis in cross-species hybridization. To investigate the feasibility, we chose three mammals, cattle, pig and dog, for testing.

MATERIALS AND METHODS

Microarrays and RNA

Human Genome U133A GeneChip® and Mouse Genome 430A GeneChip® microarrays were purchased from Affymetrix (Affymetrix, Santa Clara, CA). Human and mouse heart and liver total RNAs were purchased from Clontech (Clontech, Palo Alto, CA). Heart and liver of cattle, dog and pig were obtained from freshly slaughtered carcasses. Total RNAs were isolated from heart and liver tissues by a

*To whom correspondence should be addressed. Tel: +1 512 401 5904; Fax: +1 512 401 5919; Email: wan.ji@viagen.com

method described Gauthier *et al*. (8). One hundred micrograms of total RNA were treated with 1.0 U of DNAse I (Amplification grade, Gibco-BRL, Bethesda, MD) at 37°C for 15 min. The RNAs were further purified using RNeasy Mini columns (Qiagen, Chatsworth, CA).

Preparation of cDNA

Complementary DNA (cDNA) was prepared with cDNA Synthesis System Kit purchased from Roche Diagnostics (Roche Diagnostics GmbH, Mannheim, Germany). In short, a mixture, containing 20 μg of total RNA, 200 pmol of oligo$[(dT)_{24}$T7promotor$]_{65}$ primer, and ddH_2O in a volume of 21 μl, was incubated at 70°C for 10 min, then placed on ice. The first-strand cDNA was synthesized by adding the following reagents to the mixture, 8 μl of 5× RT buffer, 4 μl of 0.1 M DTT, 4 μl of 10 mM dNTP, 1 μl of RNase inhibitor (25 U/μl), and 2 μl of AMV reverse transcriptase (25 U/μl). The reaction was incubated at 42°C for 60 min and terminated by cooling on ice.

The second-strand cDNA was synthesized with a reaction mixture containing 40 μl of the first-strand cDNA reaction, 72 μl of ddH_2O, 30 μl of 5× second-strand buffer, 1.5 μl of 10 mM dNTP, 6.5 μl of second-strand enzyme blend consisting of DNA polymerase I (80 U), *Escherichia coli* ligase (20 U), and RNase H(4 U). The reaction was incubated at 16°C for 2 h. The reaction was stopped by adding 17 μl of 0.2 M EDTA (pH 8.0).

The dscDNA preparation was digested with 1.5 μl of RNase I (10 U/μl) at 37°C for 30 min to remove residual RNA, and subsequently treated with 5 μl of Proteinase K (0.6 U/μl) at 37°C for 30 min.

The dscDNA preparation was extracted sequentially with 200 μl of phenol, 200 μl of phenol/chloroform/isoamyl alcohol (25/24/1), and twice with 200 μl of chloroform/isoamyl alcohol (24 : 1). The supernatant was saved and mixed with 0.6 vol of 5 M NH_4OAc, and then with 2.5 vol of chilled alcohol. It was kept at −60°C for 1 h to precipitate dscDNA. The mixture was centrifuged at 10 000 g for 10 min. The pellet was washed with 300 μl of 80% alcohol, and then air-dried. The dscDNA was dissolved in 1.5 μl of ddH_2O.

Preparation of biotin-labeled cRNA

Biotin-labeled nucleotide Bio-11-CTP and Bio-16-UTP were purchased from Enzo Biochem (Enzo Biochem, New York). T7 RNA polymerase MEGAscript T7 Kit was purchased from Ambion (Ambion, Austin, TX). Reaction mixture contained 2.0 μl of 10× T7 RNA polymerase buffer, 2.0 μl of 75 mM ATP, 2.0 μl of 75 mM GTP, 1.5 μl of 75 mM CTP, 1.5 μl of 75 mM UTP, 3.75 μl of 10 mM Bio-11-CTP, 3.75 μl of 10 mM Bio-16-UTP, 2.0 μl of 10× T7 RNA polymerase enzyme mix, and 1.5 μl of cDNA (as prepared above). The reaction mixture was incubated at 37°C for 5 h. The labeled cRNA was purified with RNeasy Mini kit and eluted in 50 μl of ddH_2O. It was fragmented at 95°C for 35 min in a solution containing 40 mM Tris-acetate (pH 8.1), 100 mM KOAc, and 30 mM MgOAc. The fragmented cRNA was used either immediately for chip hybridization or stored in a −80°C freezer.

Hybridization with Affymetrix GeneChips®

A hybridization mix consisting of 50 μg of fragmented cRNA, 125 μl of 2× MES buffer (0.2 M MES pH 6.7, 2M NaCl, 0.02% Triton), 6.25 μl of acetylated BSA (20 μg/μl), 2.5 μl of herring sperm DNA (10 μg/μl), 2.5 μl of biotinylated Control Oligo (5 nM, Affymetrix, Santa Clara, CA), and ddH_2O in a total volume of 250 μl, was heated at 95°C for 5 min and then allowed to equilibrate at 45°C

Microarray chips were first treated at 45°C for 5 min with 250 μl of prehybridization solution consisting of 125 μl of 2× MES buffer and 125 μl of ddH_2O. They were then incubated with the hybridization solution at 45°C for 16 h in a rotary agitation hybridization oven.

GeneChip washing, staining and scanning

After the removal of hybridization solution, the chips were washed ten times with 6× SSPE-T (0.9M NaCl, 0.06M NaH_2PO_4 pH 6.7, 6 mM EDTA and 0.01% Triton) using the Affymetrix Fluidics Station. Chips were rinsed once with 0.1× MES and then incubated with 0.1× MES at 45°C for 15 min in a rotary oven. The chips were rinsed with 1× MES before being stained with 220 μl of staining solution containing 205 μl of 1× MES, 23 μl of acetylated BSA (20 μg/μl), and 2.3 μl of phycoerythrin-strepavidin conjugate (1 mg/ml) (Molecular Probes, Eugene, OR). Chips were incubated in staining solution at 35°C for 15 min in a rotary oven. The chips were subsequently washed with 6× SSPE-T ten times on an Affymetrix Fluidic Station. The chips were immediately scanned on a HP GeneArray Scanner.

GeneChip image quantification and data processing

GeneChip images were quantified and gene expression values were calculated by Affymetrix Microarray Suite Version 5.0 (MAS 5.0) (9). Individual electronic mask was generated by Affymetrix MAS 5.0. Tools for generating a combined mask are present but not functional in the current version of MAS 5.0. We wrote a supplement software program in Perl to carry out the operation. Gene expression ratios were calculated and their scatter plots were drawn by Microsoft Excel 2002 (Microsoft, Redmond, WA).

Quantitative PCR (TaqMan® reaction)

A mixture of 1.0 μg of total RNA, 2.0 μM oligo$(dT)_{24}$ primer, and ddH_2O in a volume of 11.0 μl was heated to 70°C for 10 min and cooled to 0°C. Added to the mixture was 4.0 μl of 5× first-strand buffer, 2.0 μl of 0.1 M DTT, 1.0 μl of 10 mM dNTP, 200 U of Superscript II reverse transcriptase (Life Technology, Rockville, MD), and ddH_2O to a final volume of 20 μl. The mixture was incubated at 42°C for 1 h to synthesize first-strand cDNA. Reactions were terminated by adding 80 μl of TE buffer followed by purification with QIAquick PCR Purification columns (Qiagen, Chatsworth, CA). The first-strand cDNA was eluted in 500 μl of ddH_2O. The relative concentrations of target genes were quantified by the TaqMan® reaction using an ABI Prism® 7000 thermal cycler. TaqMan® probes and forward and reverse primers were designed using the Primer Express 2.0 software program (Applied Biosystems, Foster City, CA). The sequences of PCR primers and TaqMan probes are listed in Table 1. The PCR primers and FAM/TAMRA-labeled TaqMan® probes were synthesized by Integrated DNA Technology Inc (Coralville, IA). TaqMan® reactions were performed in triplicate for each gene in a mixture containing 0.5–5 μl of the first-strand cDNA (depended upon target

Table 1. Oligonucleotide sequences for TaqMan reactions

TaqMan target	Forward primer	Reverse primer sequence	Tagman probe
CD63 protein	ATTTCGTTGTGAAGGACATCCA	CAGCCAGGCCGCAATCT	ACTGAGGGCTGTGTGGA
Membrane-type matrix metalloproteinase 1	TGGAAATTCAACAACCAGAAGCT	GTCCCGCAGGGCTGACTT	AAGGTTGAGCCGGGC
Very long-chain acyl-CoA dehydrogenase	CTGTCCAGGGCCTCAAGATC	GATACACCAGCTGTCACAGAGCAT	CTGAGTGAAGGCCACC
Glutathione peroxidase	TCGAAAAGTGCGAGGTGAATG	CCCGAAGGAAGGCGAAGA	AGAAGGCGCATCCG
Beta-actin	CTGGAACGGTGAAGGTGACA	CCACACTGTAGAACTTTGGGAATG	CAGTCGGTTGGATCGA
Ubiquinol-cytochrome *c* reductase core protein I	TGCTGGGCGCACACTTC	CGTTTCGGAGGAGGTTTTTG	CGACATGATGTTCGTCC
Apolipoprotein E	CTGCTCAACACCCAGGTCATT	GTAGGCCTTCACCTCCTTCATG	AGGAACTGACGGCGCT
Skeletal muscle actin, alpha 1 (ACTA1)	GTTCATCGGCATGGAATCG	CGCACTTCATGATGCTGTTGT	CGGGAATCCATGAGAC
Fibrinogen beta-chain	GGAGCGTACACCTGGGACAT	CCAGGAGCCCTGCCAGTT	CGGCACAGACGATGG
Beta-2-glycoprotein I	CCACGCGATGTTTGGAAAT	TGCGTCCAGTTCCCATGTT	ACACCGTTACCTGCACG
Tensin mRNA	TGACCTGGGAGTTGTCTTTGG	TGCCGTATTCTGGAAATCGAT	CTTGACGATGCCTTCA
Inhibin A subunit	GATGGTGCCCAACCTTCTCA	CACAGCGGGATTCCCTTAGA	CCAGCACTGTGCTTG
Complement cytolysis inhibitor	GCCAACCTCACGCAGAATG	GTCGGAGCCGTGGGAATT	CGACCGCTACTATCT
Formiminotransferase cyclodeaminase	CCTGGCCTGCCGATCTG	AATAGGCACCAAACACACCTGTCT	CTGCAGGTGGCAGC
Delta-tubulin (LOC51174)	GAACTCACTGACATACAGCACGTTT	AATGAGCATCTGTCTCAGATGTTTG	CTTGGGCTGGCCTC
T-box 4 (TBX4)	CCCCACTGAGCTGTAACATGTG	CATCGTCTGAACGCTGTAGCTT	ACGTCTGTCTCACCGTAC

transcript concentration), 1× TaqMan® buffer, 5 mM $MgCl_2$, 200 μM dNTPs, 300 nM of each forward and reverse primers, 150 nM TaqMan® probe, and 0.625 U of AmpliTag Gold DNA polymerase (Applied Biosystems) in a volume of 50 μl. TaqMan PCR was performed on an ABI 7000 fluorescence detection thermal-cycler with the following protocol: 95°C 10 min, then 40 cycles of 95°C 15 s, 60°C 1 min. Average threshold cycle (Ct) was calculated from triplicates of each gene and used for subsequent analysis.

RESULTS

We first compared the signals of cross-species hybridization with that of same species through examination of a few individual genes. Figure 1 shows the chip images of ribosomal protein L37 of human, cattle, dog and pig. In the figure, 'PM' stands for perfect-match probes and 'MM' for mismatch probes. The MM probe has a mutation in the middle of a 25 bp oligonucleotide. It serves as a control for hybridization specificity. Hybridization signal of a probe is calculated by the difference PM − MM. When hybridization signal was positive and relatively strong, we labeled it with an asterisk. Figure 1 demonstrates that ribosomal protein L37 is about equally expressed in human heart and liver with all 11 probes generating hybridization signals, while in cross-species hybridizations, cattle, dog and pig have much fewer probes generating hybridization signals, 4, 4 and 2 respectively. However, with the few informative probes, we could still make the same conclusion that ribosomal protein L37 is about equally expressed in heart and liver in the three mammals. It can be seen that the asterisk distribution is species specific. For example, both heart and liver of cattle have asterisks positioned at 3, 4, 5, 11, dog at 1, 3, 4, 5, and pig at 4 and 5, indicating the signals were derived from the genetically distinct specimens rather than random noises.

Figure 2 shows the hybridization images of troponin C. It is highly expressed in human heart but not in liver. All 11 probes hybridized with human heart, while only about half of that hybridized with the hearts of cattle, dog and pig. Again, the images demonstrate species-specific hybridization pattern. The few informative probes of cross-species hybridization could lead us to the same conclusion—troponin C is highly expressed in heart but not in liver. (Note: the non-specific hybridizations of the livers are magnified by Affymetrix MAS 5.0 for image examination, which is not adjustable for individual gene inspection.)

Figure 3 is the hybridization images of apolipoprotein CIII. It is highly expressed in human liver but not in heart. Ten out of eleven probes hybridized with human liver, while only two hybridized with the liver of other mammals. With the informative probes, we could still reach to the same conclusion that apolipoprotein CIII is expressed in liver but not in heart in all three mammals. Together, Figures 1–3 demonstrate that in cross-species hybridization, a few informative probes could generate valuable information for comparative gene expression analysis.

To understand the mechanism of the cross-species hybridization we compared the sequences of Affymetrix GenChip probes with their mammalian orthologs presently available in NCBI nucleotide databases, shown in Figure 4. Figure 4 demonstrates that signals from cross-species hybridization are true correlated with target mutations. In general, the more the mutations, the lesser the hybridization. It is, however, not just the number of mutations that ultimately determines cross-species hybridization signals, the relative positions of the mutations plays an equally important role. For example, both targets 10 and 11 of cattle ribosomal protein L37a have two mutations: while target 11 generated good hybridization signals, target 10 did not; both target 2 and 4 of cattle troponin C have two mutations: while target 4 generated good hybridization signals, target 2 did not. It appears that a relatively long stretch of perfect-matched oligonucleotides, $\geqslant$16 bp long, is essential for stable hybridization and signal generation.

To select informative probes for signal calculation, we arbitrarily installed a mask, PM − MM > 200 and PM/MM > 2, to screen off poorly hybridized probes. Figure 5 compares overall

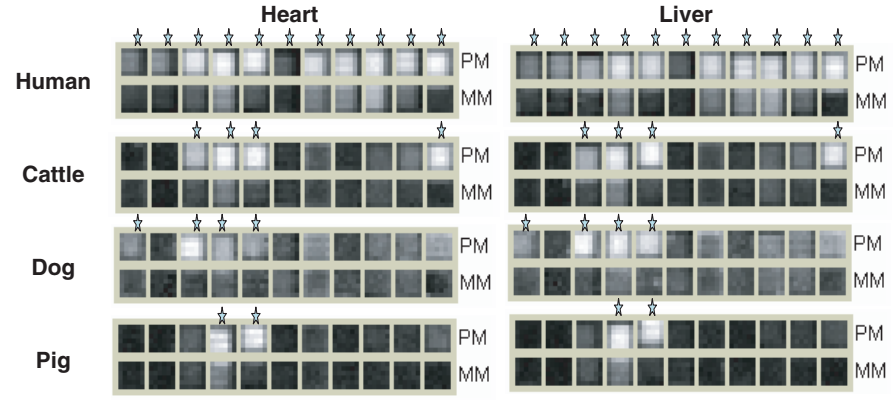

Figure 1. Hybridization images of heart and liver ribosomal protein L37a of human, cattle, dog and pig. PM and MM denote perfect match probes and mismatch probes, respectively. The probes that produce a relatively high hybridization signal are labeled with an asterisk.

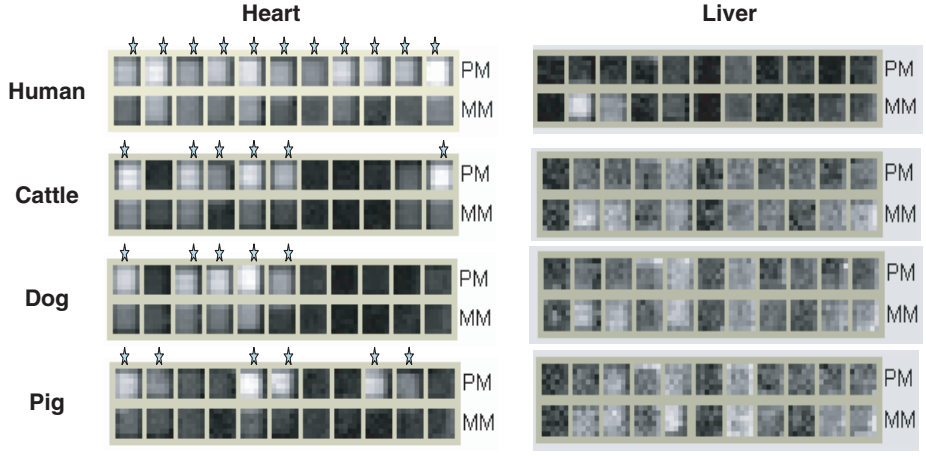

Figure 2. Hybridization images of heart and liver troponin C of human, cattle, dog and pig. PM and MM denote perfect match probes and mismatch probes, respectively. The probes that produce a relatively high hybridization signal are labeled with an asterisk.

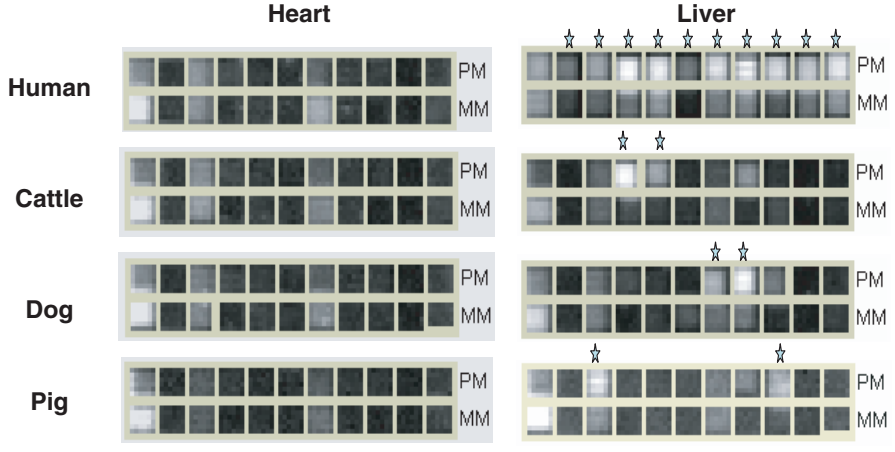

Figure 3. Hybridization images of heart and liver apolipoprotein C-III of human, cattle, dog and pig. PM and MM denote perfect match probes and mismatch probes, respectively. The probe that produce a relatively high hybridization signal are labeled with an asterisk.

chip signal before and after mask installation. Of three species and two tissues, there is on average more than a 7-fold signal increase after the mask installation, meaning the sensitivity of cross-species hybridization is greatly increased. This large enhancement is primarily due to the fact that in cross-species hybridization, most probes do not produce hybridization signals because of their mutations. By installing the mask, we simply removed the interfering noise of the mutated probes

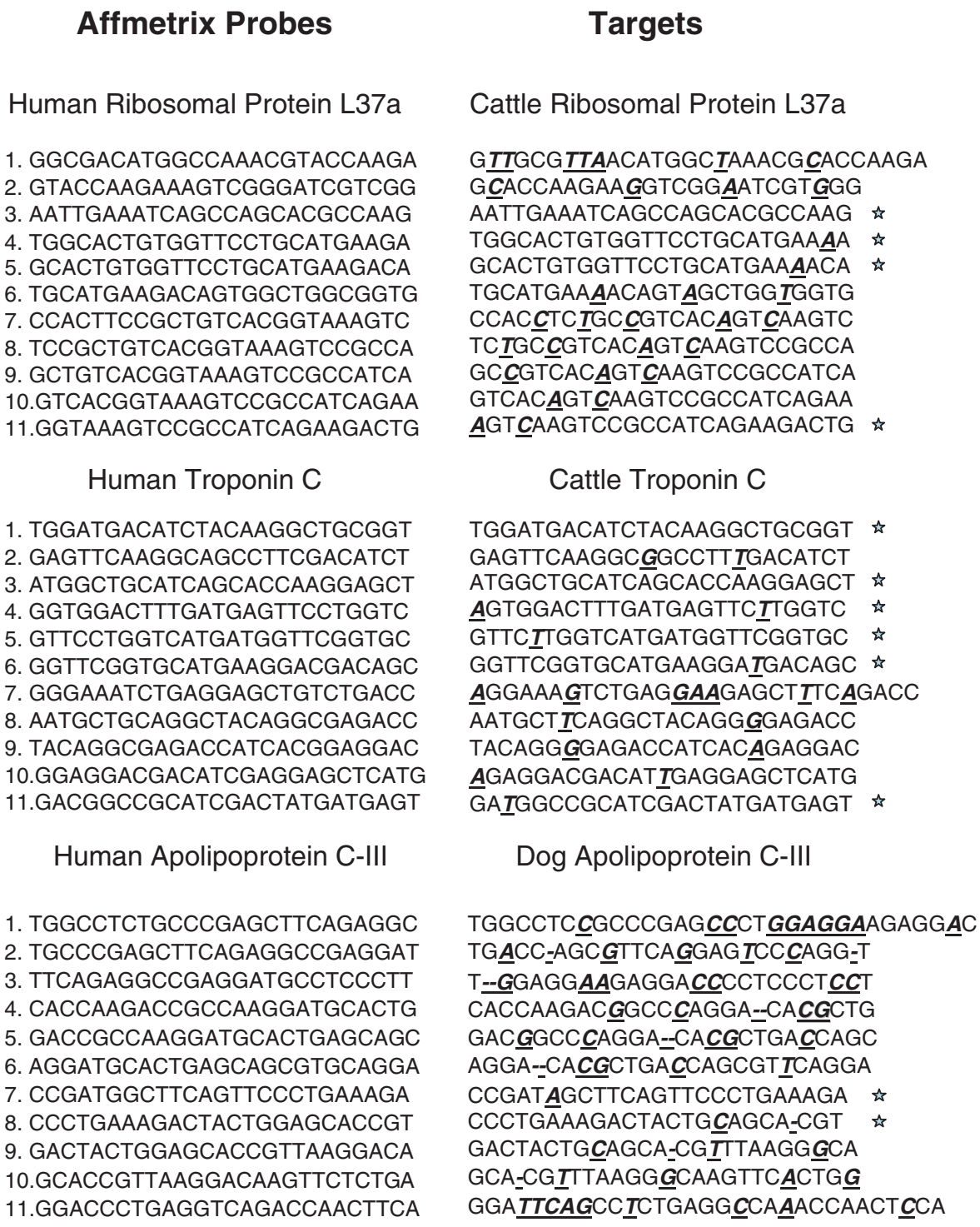

Figure 4. Sequence comparison of GeneChip probes and their heterologous targets. The mutated nucleotides are labeled with bold, italic and underlined fonts. The probes that produce a relatively high hybridization signal are labeled with an asterisk.

and let probes with good hybridization signals represent the gene hybridization.

While we greatly increased sensitivity of cross-species hybridization with the electronic mask, we might have inadvertently reduced the specificity of cross-species hybridization, allowing unspecific hybridization signals to represent true gene expression, which, in the same-species hybridization, could be avoided or alleviated by averaging with other probes. To verify whether the cross-species hybridization truly reflects the concentrations of gene transcripts, we need to test the method on reference solutions wherein gene transcript quantities are known or identifiable by alternative methods. High-density cDNA microarrays and high-density oligonucleotide microarrays have been scrutinized and have

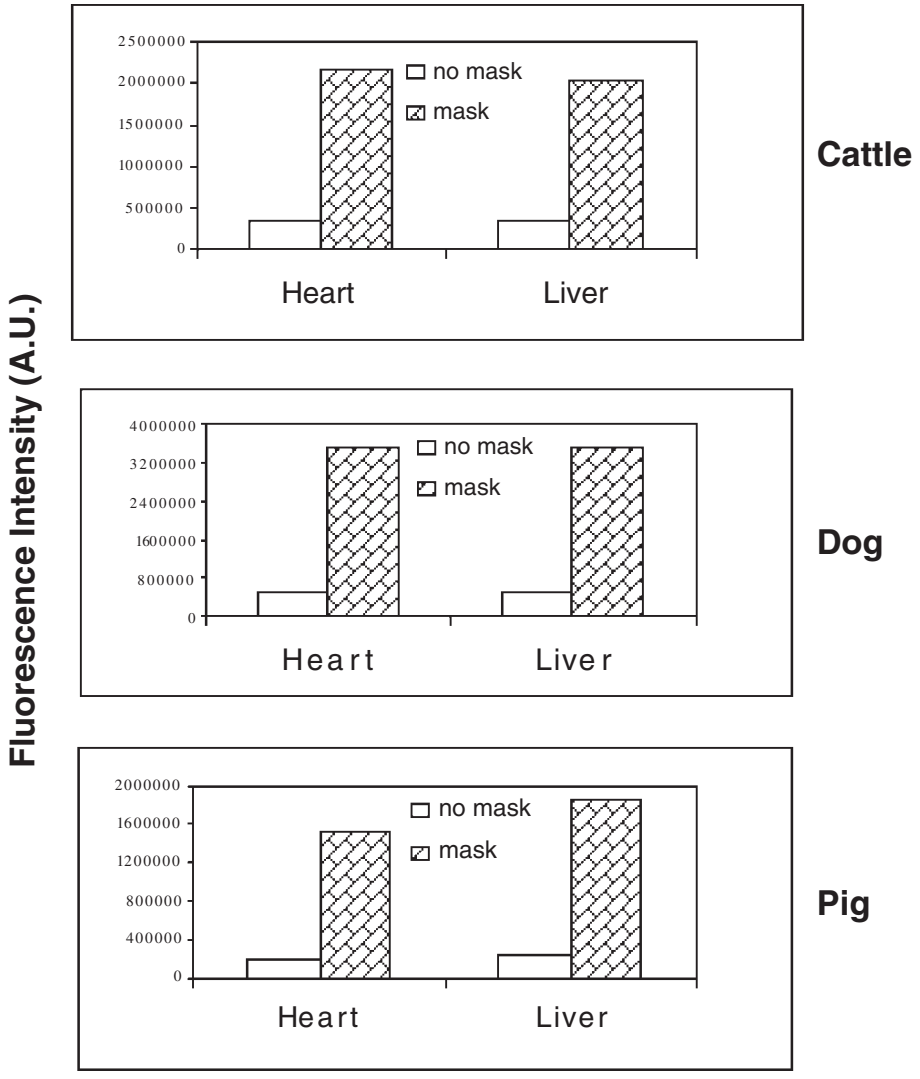

Figure 5. The increases of chip hybridization signals after applying of electronic mask.

been shown repeatedly to be able to quantify relative concentrations of gene transcripts since their inception eight years ago, especially in the case of same-species hybridization (10,11). It has also been well documented that equivalent organs of different mammals perform similar physiology function, e.g. pig and human islets, which forms the basis of xenotransplantation—organs are transplanted between different mammal species (12,13). Therefore, it is reasonable to assume that equivalent organs of different mammals have similar gene expression profile. As a result, gene transcript concentrations of our investigative mammals could be approximately determined by measuring the equivalent organs of humans using Affymetrix human GeneChips. Then, the accuracy of a cross-species hybridization could be determined by comparing with a same-species hybridization with equivalent organs using Pearson correlation coefficient or Euclidean distance as quantitative measures (14). Since most gene expression experiments are designed to find differentially expressed genes, the correlation of gene expression ratio between two sets of equivalent tissues, two from human and two from other mammals, would be more meaningful.

Figure 6 is a compilation of scatter plots of Ln[Cattle(Heart/Liver)] versus Ln[Human(Heart/Liver)]. In the plot, we used logarithms of ratios, so that a/b and b/a would carry the same weight when calculating Pearson correlation coefficients. We used a serial combination of PM − MM and PM/MM to select informative probes from cattle hybridization and to calculate gene expression with them. Figure 6 provides both sensitivity and specificity of a cross-species hybridization, because not only does it inform us how many genes are detected by cross-hybridization (sensitivity) but also how accurate relative gene expression values are (specificity, correlation coefficient *r*). It demonstrates that both sensitivity and specificity of cattle hybridizations were determined by electronic masks we chose. In general, masks of lower PM − MM or PM/MM resulted in a higher number of genes being detected, i.e. increased sensitivity, while higher PM − MM or PM/MM resulted in a higher level of correlation coefficient, i.e. increased specificity. At PM − MM > 300 and PM/MM > 2.0, we detected 2972 gene transcripts in cattle heart and liver, which is comparable with the same-species hybridization, wherein we detected 2151 transcripts in human heart

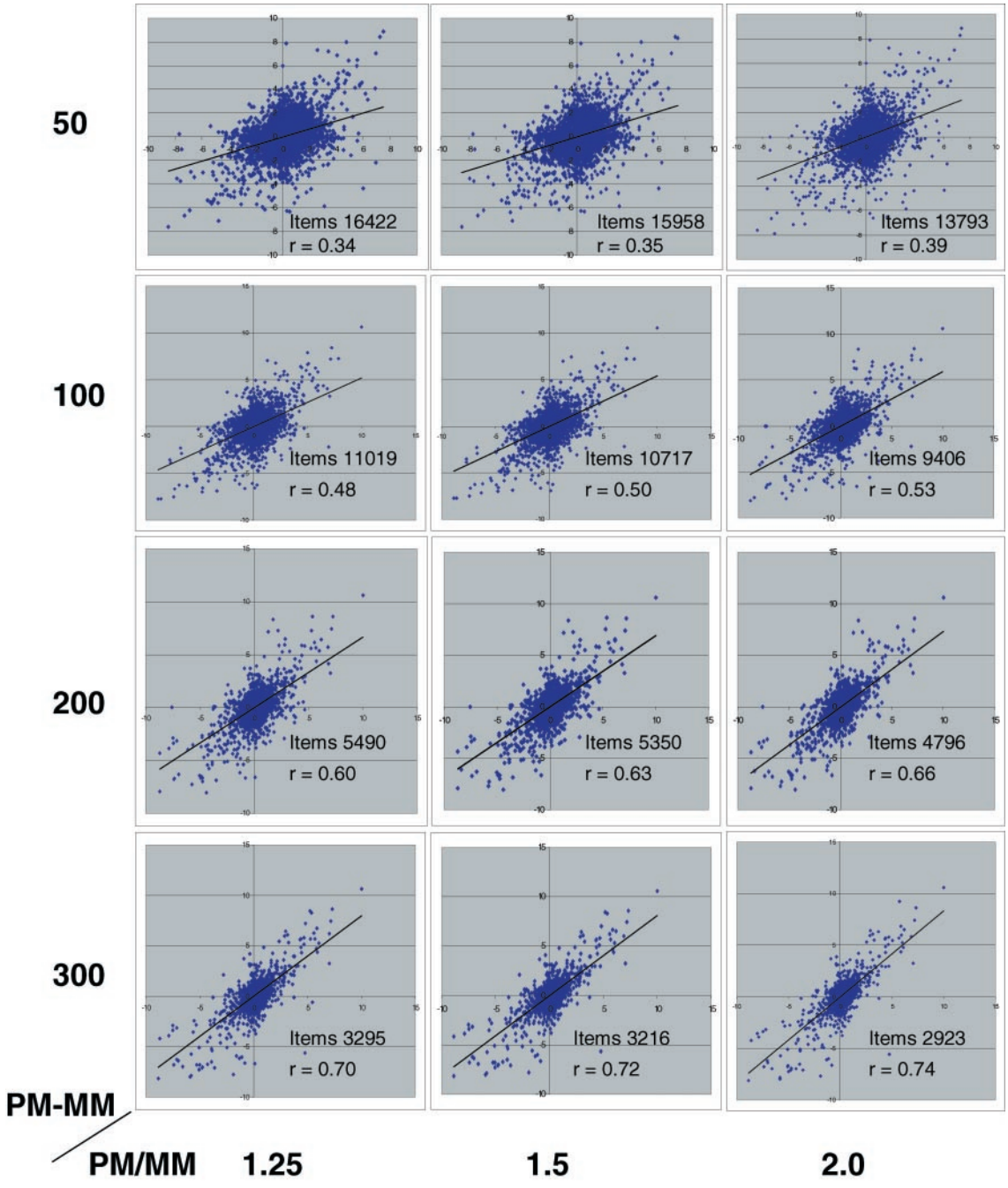

Figure 6. Scatter plots of Ln(Cattle Heart/Liver) versus Ln(Human Heart/Liver). Electronic masks with various combinations of PM − MM and PM/MM were applied. A scatter plot was drawn for each mask as shown.

and liver at PM − MM >100 using MAS 5.0 default setting. Across the 2972 genes, the correlation coefficient between cattle and human is 0.792, which is quite high considering the number of genes involved in the calculation, the intrinsic variation between separate chip experiments (15), and the natural physiological differences between human and cattle. Similar conclusions could also be reached from dog and pig experiments (data not shown).

To confirm the cross-species hybridization we randomly selected from the Genechips' 16 targets that have been sequenced and annotated for cow, dog and pig in NCBI nucleotide database (www.ncbi.nlm.nih.gov). We used the TaqMan reaction to confirm the cross-species hybridization data (16,17). The results are listed in Table 2. We drew $(Ct_{heart} - Ct_{liver})$ or ΔCt against $Log(heart/liver)_{GeneChip}$ in Figure 7. The results demonstrate a linear relationship. As well articulated by Ji *et al*. (18), the ΔCt measures log concentration ratio between two samples. Therefore, in general, the two independent methods, cross-species GeneChip hybridization and TaqMan reaction, concur on gene transcript measurement.

To further validate the cross-species hybridization we prepared cRNAs from mouse heart and liver and hybridized them

Table 2. Comparison of GeneChip signals and TaqMan Ct values

Affymetrix Probe	Hybridization target, NCBI accession no.	Annotation	GeneChip detected probe no.	GeneChip signal heart	GeneChip signal liver	TaqMan Ct heart	TaqMan Ct liver
200663_at	AJ012589	Bovine CD63 protein	2	609	528	24.7 ± 0.3	24.3 ± 0.0
160020_at	AF290429	Bovine membrane-type matrix metalloproteinase 1	2	834	780	34.7 ± 0.5	33.6 ± 0.4
200710_at	U30817	Bovine very-long-chain acyl-CoA dehydrogenase	4	374	584	26.7 ± 0.5	27.5 ± 0.1
200736_s	X13684	Bovine glutathione peroxidase 1	2	390	133	27.5 ± 0.3	28.8 ± 0.5
200801_x	AY141970	Bovine beta-actin (ACTB)	2	485	819	25.5 ± 0.0	23.7 ± 0.1
201903_at	NM_174629	Bovine ubiquinol-cytochrome *c* reductase core protein I	4	446	75	32 ± 0.1	35.8 ± 0.1
203381_s	X61171	Bovine apolipoprotein E	2	12	646	36.1 ± 0.0	27.7 ± 0.2
203872_at	NM_174225	Bovine skeletal muscle actin, alpha 1 (ACTA1)	4	551	4	25.5 ± 0.4	Undet.[a]
204988_at	V00110	Bovine fibrinogen beta-chain	4	3	1018	35.2 ± 0.3	19.2 ± 0.1
205216_s	X60065	Bovine beta-2-glycoprotein I	1	1	1764	Undet.	20.7 ± 0.2
221246_x	AF225897	Bovine tensin mRNA	2	337	19	29.3 ± 0.2	33.4 ± 1.0
210141_s	M13273	Bovine inhibin A subunit	1	475	1	27.5 ± 0.1	39.3 ± 0.7
208791_at	M84639	Swine complement cytolysis inhibitor	1	8	1058	29.1 ± 0.4	26.0 ± 0.5
220604_x	L16507	Swine formiminotransferase cyclodeaminase	1	1	385	Undet.	29.3 ± 0.1
221326_s	AF416724	Canine delta-tubulin (LOC51174)	1	578	269	35.7 ± 0.1	38 ± 0.4
220634_at	AY185179	Canine T-box 4 (Tbx4)	1	1597	4	31.4 ± 0.4	32.5 ± 0.7

with both Human Genome U133A and Mouse Genome MOE430A GeneChip. Thus, the same transcripts were analyzed by both the same-species and cross-species hybridizations. Due to the differences of individual transcript annotations and arrangements, the human genes on chip U133A and their mouse orthologs on MOE430A cannot be aligned on whole genome scale. We manually and randomly pick 53 transcripts that have the same descriptions on both the human and the mouse GeneChip. We listed their gene expression values in Table 3. In Table 3, the gene expression of the same-species hybridization on MOE430A were calculated by the algorithm employed in Affymetrix MAS 5.0 using all 11 probes, while the expression by the cross-species hybridization on U133A were calculated by algorithm described in this manuscript using PM − MM > 200 and PM/MM > 2.0 mask. After masking off poorly hybridized probes, there were, on average, only two to three informative probes left for each gene. We compared gene expression ratios of heart to liver between the same-species and the cross-species in Figure 8. Figure 8 demonstrates a strong linear relationship between the two measurements with a correlation coefficient of 0.93 across 53 randomly selected genes. Since the data were derived from 28 genes, a Student *t*-test was performed by calculating Equation 1 (19):

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \mathbf{1}$$

where r is coefficient of correlation and n is sample size. The result of $t_{51} = 18.1$ yields a very high level of statistical significance ($p < 8.28 \times 10^{-24}$), rejecting the null hypothesis H_0 (that there is no linear relation between the ratios obtained by the two methods).

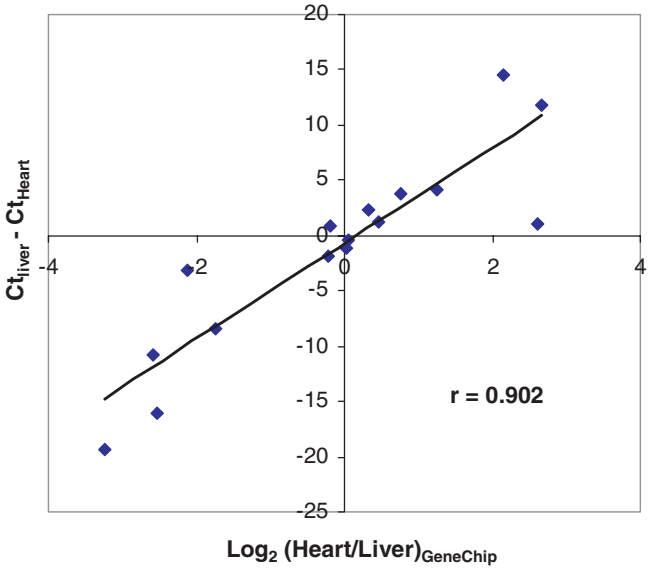

Figure 7. ($Ct_{liver} - Ct_{heart}$) versus $Log_2(heart/liver)_{GeneChip.}$ Gene targets were randomly selected from cattle, dog and pig GeneChip hybridization. Cts were determined by TaqMan reaction using ABI 7000. Note: We use Ct = 40 to represent undetectable genes. The line in the graph represents the line of linear regression between the two sets of data.

DISCUSSION

In this study we tested cross-species hybridization of three distantly related mammals on Affymetrix human GeneChip®. We designed a method for its practical applications, demonfstrating that both sensitivity and specificity of cross-species hybridization could be substantially increased by selecting informative probes with electronic masks. This method is easy to implement, requiring no additional equipment or modification of experimental procedures. It largely expands the scope of microarray applications, from human and closely related Primates (20,21) to Carnivora and Artiodactyla. The results are highly significant for agricultural and animal researches, where many investigative species are not equipped with sequence information. In the near future, it is possible that one GeneChip® of a model species could be employed for all of its relatives, e.g. human for all mammals, rice for all grasses, with one generic procedure demonstrated in this study.

We have demonstrated the validity of cross-species hybridization by three independent methods, by TaqMan quantitative

Table 3. Fifty-three human and mouse orthologs from HG_U133A and MOE430A GeneChips

HG-U133A GeneChip ID#	Mouse Heart	Mouse Liver	GeneChip Probe	MOE430A GeneChip ID#	Mouse Heart	Mouse Liver	
205738_s_at	761.4	18.7	Homo sapiens fatty acid binding protein 3	1416023_at	1541.0	13.4	Mus musculus fatty acid binding protein 3
201406_at	315.9	342.0	Homo sapiens ribosomal protein L44	1416807_at	96.8	124.8	Mus musculus ribosomal protein L44
217073_x_at	27.4	1697.8	Homo sapiens apolipoprotein AI	1438840_x_at	5.7	2770.9	Mus musculus apolipoprotein A-I
219466_s_at	5.3	4483.6	Homo sapiens apolipoprotein A-II	1417950_a_at	16.4	2635.8	Mus musculus apolipoprotein A-II
205892_s_at	0.3	489.9	Homo sapiens fatty acid binding protein 1	1417556_at	8.5	2159.8	Mus musculus fatty acid binding protein 1
204988_at	4.4	1100.2	Homo sapiens fibrinogen, B beta polypeptide	1428079_at	0.8	1912.0	Mus musculus fibrinogen, B beta polypeptide
212581_x_at	1405.6	949.3	Hs.glyceraldehyde-3-phosphate dehydrogenase	1418625_s_at	2882.9	1874.2	Mm.glyceraldehyde-3-phosphate dehydrogen
203400_s_at	53.6	575.9	Homo sapiens transferrin	1425546_a_at	28.7	1638.5	Mus musculus transferrin
221700_s_at	341.2	422.1	Homo sapiens UBA52	1416624_a_at	1972.5	1618.5	Mus musculus Uba52
204892_x_at	834.7	1772.2	Hs.eukaryotic translation elongation factor 1 alpha 1	1424635_at	1217.2	1583.8	Mm.eukaryotic translation elongation factor 1
207076_s_at	1.6	245.1	Homo sapiens argininosuccinate synthetase	1416239_at	4.1	1474.1	Mus musculus argininosuccinate synthetase
210328_at	3.3	532.5	Homo sapiens glycine N-methyltransferase	1417422_at	1.7	1415.9	Mus musculus glycine N-methyltransferase
200633_at	1054.2	1027.6	Homo sapiens ubiquitin B	1449436_s_at	1231.8	1359.1	Mus musculus ubiquitin B
201254_x_at	310.0	348.6	Homo sapiens ribosomal protein S6	1416141_a_at	1194.1	1358.5	Mus musculus ribosomal protein S6
201665_x_at	905.1	1330.1	Homo sapiens ribosomal protein S17	1459986_a_at	1553.7	1301.4	Mus musculus ribosomal protein S17
202025_x_at	113.3	1480.5	Homo sapiens acetyl-Coenzyme A acyltransferase	1416947_s_at	46.6	1146.8	Mus musculus acetyl-Coenzyme A acyltransfe
201035_s_at	1903.6	1893.2	Hs.L-3-hydroxyacyl-Coenzyme A dehydrogenase short chain	1455972_x_at	1262.1	1107.3	Mm.L-3-hydroxyacyl-Coenzyme A dehydroge
201094_at	185.6	523.5	Homo sapiens ribosomal protein S29	1415833_x_at	1200.6	1100.6	Mus musculus ribosomal protein S29
203335_at	190.2	227.6	Homo sapiens phytanoyl-CoA hydroxylase	1460194_at	307.6	1066.1	Mus musculus phytanoyl-CoA hydroxylase
206177_s_at	29.4	1250.5	Homo sapiens arginase 1	1419549_at	0.8	1059.4	Mus musculus arginase 1
205754_at	2.5	690.6	Homo sapiens coagulation factor II	1418897_at	12.3	1043.0	Mus musculus coagulation factor II
217232_x_at	3778.5	423.3	Homo sapiens hemoglobin, beta	1417184_s_at	3022.4	1005.1	Mus musculus, hemoglobin, beta
206345_s_at	1.1	1222.0	Homo sapiens paraoxonase 1	1418190_at	16.4	974.2	Mus musculus paraoxonase 1
213738_s_at	947.7	240.8	Hs. ATP synthase, H+ transporting, alpha subunit, isoform 1	1423111_at	1751.7	891.9	Mm.ATP synthase, H+ transporting, alpha sut
205132_at	1111.8	0.9	Homo sapiens actin, alpha, cardiac muscle	1415927_at	2252.3	4.5	Mus musculus actin, alpha, cardiac
212361_s_at	1284.9	74.6	Hs. ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	1452363_a_at	2153.0	57.9	Mm. ATPase, Ca++ transporting, cardiac mus
214468_at	1627.8	0.5	Homo sapiens cardiac alpha-myosin heavy chain	1448826_at	1926.6	2.4	Mm. alpha cardiac myosin heavy chain
204938_s_at	1748.9	97.3	Hs. phospholamban	1450952_at	1821.0	10.5	Mm. phospholamban
215389_s_at	1057.5	29.1	Hs. troponin T2, cardiac	1418726_a_at	1684.1	0.1	Mm. troponin T2, cardiac
210046_s_at	1056.5	100.5	Hs. isocitrate dehydrogenase 2 (NADP+)	1454661_at	1652.6	774.2	Mm. isocitrate dehydrogenase 2 (NADP+)
206117_at	1883.2	1.7	Homo sapiens tropomyosin 1	1423049_a_at	1630.0	1.5	Mus musculus tropomyosin 1
204810_s_at	638.8	3.7	Homo sapiens creatine kinase	1417614_at	1502.5	12.7	Mus musculus creatine kinase
211025_x_at	926.5	558.4	Homo sapiens, cytochrome c oxidase subunit Vb	1456588_x_at	1436.8	843.1	Mus musculus cytochrome c oxidase, subunit
209904_at	892.0	0.9	Homo sapiens cardiac ventricular troponin C	1418370_at	1350.8	12.2	Mus musculus troponin C, cardiac
200966_x_at	559.6	20.6	Homo sapiens aldolase A	1416921_x_at	1331.5	140.5	Mus musculus aldolase 1, A isoform
200978_at	1726.5	775.2	Homo sapiens malate dehydrogenase	1448172_at	1317.9	574.2	Mus musculus malate dehydrogenase
200023_s_at	397.4	196.4	Hs. eukaryotic translation initiation factor 3, subunit 5	1427021_s_at	1299.4	1683.2	Mm. eukaryotic translation initiation factor 3, s
200737_at	286.2	227.9	Homo sapiens phosphoglycerate kinase 1	1438640_x_at	1038.3	811.5	Mus musculus phosphoglycerate kinase 1
203872_at	1109.9	8.0	Homo sapiens actin, alpha 1	1427735_a_at	1009.1	1.6	Mus musculus actin, alpha 1
200789_at	952.4	718.0	Homo sapiens enoyl Coenzyme A hydratase 1	1448491_at	1003.8	672.3	Mus musculus enoyl coenzyme A hydratase 1
209283_at	1042.3	18.5	Hs. crystallin, alpha B	1416455_a_at	972.0	1.2	Mm. crystallin, alpha B
208980_s_at	582.1	466.1	Hs. ubiquitin C	1432827_x_at	880.8	1069.1	Mm. ubiquitin C
201903_at	213.9	34.8	Hs. ubiquinol-cytochrome c reductase core protein I	1428782_a_at	854.6	314.4	Mm. ubiquinol-cytochrome c reductase core p
201066_at	267.3	52.6	Homo sapiens cytochrome c-1	1416604_at	846.7	435.4	Mus musculus cytochrome c-1
204540_at	1336.2	28.1	Hs. eukaryotic translation elongation factor 1 alpha 2	1418062_at	844.9	21.4	Mm. eukaryotic translation elongation factor 1
211528_x_at	437.0	574.6	Human sapiens b2 microglobulin	1452428_a_at	828.0	968.3	Mus musculus beta-2 microglobulin
200696_s_at	320.1	54.9	Homo sapiens gelsolin	1415812_at	815.7	20.6	Mus musculus gelsolin
202941_at	673.3	124.1	Hs. NADH dehydrogenase (ubiquinone) flavoprotein 2	1428179_at	728.9	278.6	Mm. NADH dehydrogenase (ubiquinone) flavc
222021_x_at	481.7	264.7	Hs. succinate dehydrogenase complex, subunit A, flavoprotein (Fp	1426688_at	656.2	431.8	Mm. succinate dehydrogenase complex, subu
200080_s_at	448.1	432.6	Hs. H3 histone, family 3A	1434127_a_at	627.0	510.5	Mm. H3 histone, family 3A
201779_s_at	166.9	142.7	Homo sapiens ring finger protein 13	1430713_s_at	612.4	298.4	Mus musculus ring finger protein 13
211072_x_at	529.8	148.5	Homo sapiens, tubulin alpha 1	1423846_x_at	609.1	87.1	Mus musculus, tubulin alpha 1
205736_at	1123.2	1.9	Homo sapiens phosphoglycerate mutase 2	1418373_at	582.5	1.7	Mus musculus phosphoglycerate mutase 2

The gene expression of the same-species hybridization on MOE430A were calculated by the algorithm employed in Affymetrix MAS 5.0 using all 11 probes, while the expression by the cross-species hybridization on U133A were calculated by algorithm described in this manuscript using PM − MM > 200 and PM/MM > 2.0 mask.

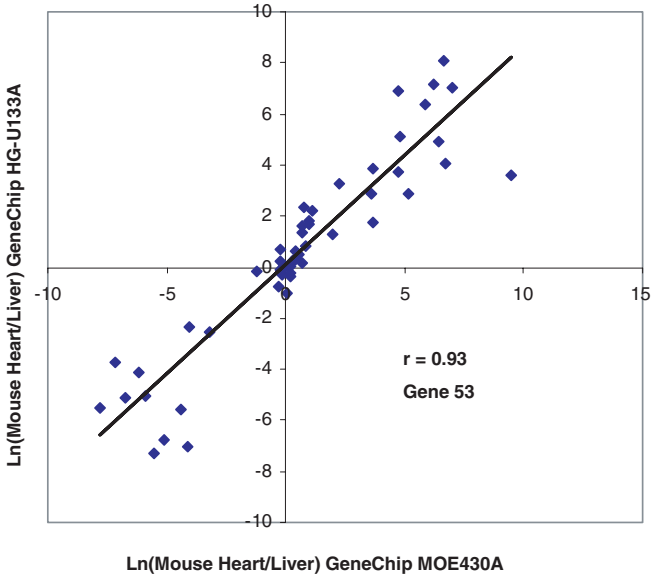

Figure 8. Correlation of the cross-species hybridization and the same-species hybridization on GeneChip. Complementary RNAs from mouse heart and liver were hybridized with both Human Genome U133A and Mouse Genome MOE430A GeneChip. The gene expression of the same-species hybridization on MOE430A were calculated by the algorithm employed in Affymetrix MAS 5.0 using all 11 probes, while the expression by the cross-species hybridization on U133A were calculated by algorithm described in this manuscript using PM − MM > 200 and PM/MM > 2.0 mask. The line in the graph represents the line of linear regression between the two sets of data.

PCR and by the same-species hybridization using randomly selected genes, and by hybridization with the same organs from different species. The notion that the same organs have similar gene expression patterns may not be true for all individual genes and for all organs. As demonstrated previously, primates have pronounced differences in gene expression in brains (22,23). Nevertheless, similarity of gene expression in heart and liver among different mammals has been indirectly demonstrated in this study.

Since gene expression values are continuous variables, we employed linear correlation coefficient to compare the cross-species hybridization with other established experimental procedures. We strenuously avoided using the artificial classifications, such as 'positive' and 'negative', to analyze our experimental data, because they are arbitrary and inevitably subjective. When comparing the correlation coefficient of the cross-species hybridization versus TaqMan with the published same-species hybridization versus TaqMan (24,25), we found they are quite similar or even better, meaning the cross-species hybridization is as reliable as the same-species hybridization. The result is somewhat unexpected, but is quite explicable by the fact that in the cross-species hybridization, we always select the probes with good hybridization signals for gene expression calculation, thus avoiding the interferences of the probes with low signal noises.

Why is it necessary to select informative probes for cross-species hybridization? The answer can be better understood by first examining the algorithm used by Affymetrix MAS 5.0 in calculating gene expression (9). In MAS 5.0, a gene expression value is calculated by the weighted average of all of its probe signals using One-step Tukey's Biweight Estimates. The weight carried by each probe is inversely related to its distance from the mean value of the probes. As shown in Figures 1–3, in the same-species experiment, when most probes generate hybridization signals, the algorithm provides a robust estimate of gene expression by reducing the influence of outliers. In cross-species hybridization, when most probes do not generate hybridization signals due to their target mutations, however, MAS 5.0 will treat the few informative probes as outliers. If not singled out by electronic masks, they will be overwhelmed by mutated ones. Therefore, the appropriate application of an electronic mask is essential for signal calculation in cross-species hybridizations.

Another related question is whether the few informative probes could truly represent gene expression. We have made a simple calculation on probe usage by each gene in cattle heart/liver experiment with a mask PM − MM > 200 and PM/MM > 2. The mask singled out 4745 genes with expression value >200 AU. Out of the 4745, 1192 genes have multiple probes and 3553 have single probes. On average each gene is represented by 1.5 probes. Statistically, randomly chosen 1.5 probes are not as reliable as 11 probes for gene expression calculation, which is mathematically evident from Central Limit Theorem (26). However, this is only true if the single or few oligos are randomly selected from the probe set. After close examinations of GeneChip hybridization, we found that there are usually large variations among individual probes of each transcript. Carefully selecting those with good hybridization specificity (PM/MM > threshold) and signals (PM-MM > threshold) can not only increase GeneChip sensitivity but also maintain its specificity. Moreover, in almost all other formats of high-density DNA microarrays, a gene is represented by a single probe, whether it is PCR-amplified cDNA or *in vitro*-synthesized oligonucleotide (27,28). Thus, the 'one gene one probe' microarray is a universally accepted standard, at least right now. There is a notion that while it is acceptable for longer cDNA array to work with one probe per gene, it is unacceptable for shorter oligonucleotides to work with less than 11 probes. The notion, however, is unsubstantiated. In fact, Agilent has been manufacturing oligonucleotide arrays for years with either three 25mer per gene or one 60mer per gene, resulting in hundreds of published papers. It is due to the advance of photolithographic technology that Affymetrix is able manufacture high-density oligonucleotide arrays with 11 probes per gene. The feature is almost impossible for a cDNA array to achieve with its probes synthesized by PCR and arrayed by dipping/spotting methodology. For accurate measurements, all cDNA arrays should have multiple probes per gene, because compared with the shorter oligonucleotide arrays, the longer cDNA probes have a much higher chance of cross-hybridizations or contaminations. Therefore, it is the technology not the requirement that ultimately determines the number of probes per gene in an array. We simply exploited the unique multiple-probe design of GeneChip, which is afforded by photolithographic technology, and selected part of it for cross-species hybridization analysis. Moreover, we have demonstrated that cross-hybridization results can be confirmed by both same-species hybridizations and TaqMan reactions. When using the cross-species hybridization, it is important to understand that the measured gene expression value will be different even if one-base change occurs between species. Therefore, it is invalid to use the

present method to compare the gene expression of different species, because they may have different genetic alterations. However, because the purpose of almost all gene expression analysis is to find relative gene expression changes among multiple tissues or multiple states in an investigative species, the absolute gene expression measurement is of little consequence. Just as both cDNA microarrays, which have probes several hundred base pairs long, and oligonucleotide microarrays, which have probes twenty-five base pairs long, can all be used for gene expression analysis, the cross-species hybridization, which utilizes even shorter probes on GeneChip, can also be used for gene expression analysis. It should also be stressed that good and clean chip hybridizations are essential for cross-species analysis and even more so than in the same-species hybridizations. The reason is that when fewer or even a single probe is used for gene expression analysis, unspecific hybridizations, such as 'snowflakes' or peeling of probes, which can usually be eliminated by Affymetrix default algorithm as outliers, will be calculated as real signals.

Because each sample can have its own mask, a question also arises as to which mask should be chosen when comparing multiple samples or tissues. Again, the solution could be reached by examining the purpose of installing an electronic mask. The electronic mask is intended to remove the mutated probes which interfere with signal calculation. As shown in Figures 1–4, some mutated probes generate low hybridization signals, but the low signal could also be produced by the samples that simply have low gene expression, such as troponin C in liver. The difference is that the mutated probes have low signals across all tissues, while the regulatory ones have low signals in only some of them. A simple mask will indiscriminately remove both kinds of probes. A combined mask, which was created from all individual masks using the Boolean logic function, AND, will specifically remove the mutated probes while keeping those of the regulatory ones. Therefore, when comparing multiple samples, we first created a combined mask, and then used it for gene expression calculation.

One particular concern about the method is its error rate. Because of the electronic mask, fewer than the 11 probes that were spotted on GeneChip are utilized for cross-species expression analysis. Questions arise whether the selected probes can still accurately measure gene expression; whether there are substantial differences between the selected few and the whole set of GeneChip probes with regard to gene expression measurement, and if there were, what the error rates are. These questions are very much related to a more fundamental one facing oligonucleotide arrays: can individual oligo probes 25 bp long provide necessary specificity for gene expression analysis? Do we have to average the signals of all 11 probes for that? Theoretically, these questions could be answered with human and mouse chips. Mouse samples would hybridize with all 11 probes on the mouse chip but only with highly conserved ones on the human chip. We could then compare the signals of the mouse chip with their human chip counterparts and find whether there is any correlation. Technically, however, such a scheme is very difficult to carry out on a large scale. Because very few human and mouse orthologs have been identified on human and mouse GeneChips, finding human and mouse orthologs has to be done manually, which is very slow and labor intensive, making whole genome comparison almost impossible. In addition, the mouse and human probes were designed separately from different regions of their representative sequences, sharing almost no common probes. Even though Affymetrix has identified some homologs among different chips and listed them in its website (http://www. affymetrix.com/support/technical/comparison_spreadsheets. affx), the homologs and orthologs are quite different things, which has been clearly explained by Makalowski and Boguski (6).

To answer the critical question of whether the selected probes with high hybridization signals could faithfully represent gene expression, we further analyzed the same-species hybridizations of human heart and liver. We imposed electronic masks on chip .cell files in the same way we did in cross-species hybridization. We reasoned that if the selected probes did not always hybridize with their intended targets and frequently produce unspecific noises, then we would not find good correlation between the signals generated from the selected probes and those from the whole sets.

Figure 9 compares signal ratios of heart/liver before and after masks. The comparisons are made for the transcripts that can be detected both before and after masks (>100 AU, arbitrary unit of fluorescence). It demonstrates very high correlations between the two. linear correlation coefficients, 0.98, 0.97 and 0.97 after mask (PM − MM > 100 PM/MM > 2.0, PM − MM > 200 PM/MM > 2.0 and PM − MM > 300 PM/MM > 2.0, respectively). On average, usable probes were reduced to 2.82, 2.57 and 2.54 probes per transcript in detectable ones (>100 AU) after mask, (PM − MM > 100 PM/MM > 2.0, PM − MM > 200 PM/MM > 2.0 and (PM − MM > 300 PM/MM > 2.0, respectively). Thus, with the much fewer probes, the gene expression ratios were kept almost constant, irrespective of the cutoff values of the masks.

Figure 10 shows Venn diagrams of differentially expressed genes (more than 2-fold) before and after mask installation. It demonstrates that the large majority of differentially expressed genes detected before mask installation can also be detected afterward. Out of 1237 differentially expressed genes before mask, 1068, 1075 and 1059 were detected after mask (PM − MM > 100 PM/MM > 2.0, (PM − MM > 200 PM/MM > 2.0 and PM − MM > 300 PM/MM > 2.0 respectively). Around 85% of the differentially expressed gene detected by full sets of probes can also detected by the selected probes with all three masks we tested. Figure 10 also demonstrates a large increase of detectable differentially expressed genes after mask installation. The main reason is that chip sensitivity is largely increased after the mask installation; many previously undetectable genes (<100 AU) became detectable (>100 AU). Even though these parts of the data (dotted green) cannot be directly verified by the data before mask, their validity can be indirectly inferred by statistical reasoning. If we consider the whole set of data after mask as a population and those co-detected before mask as a sample, then validity of the population can be inferred by the validity of the sample. Therefore, the mask can be used to increase sensitivity of GeneChip even in the same-species hybridization.

As shown in Figures 1–4, the number of informative probes a heterologous gene holds is influenced by the number of nucleotide mutations the gene has. The highly conserved cattle troponin C has more informative probes than the relatively variable dog apolipoprotein C-III. It is essential to understand the statistical nature of this relation, so that it could be

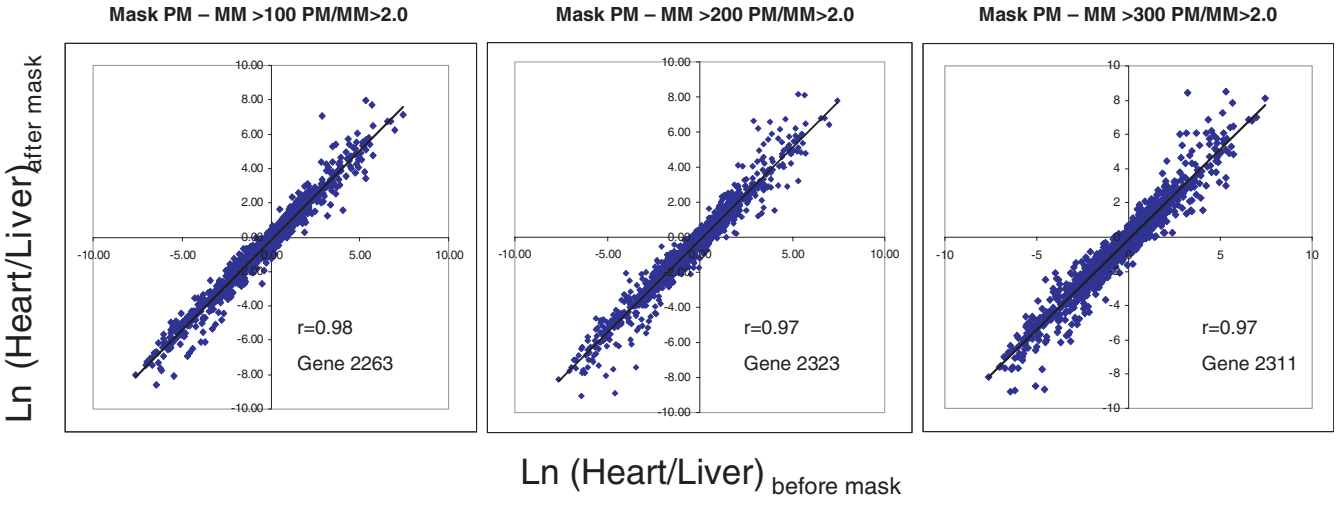

Figure 9. Correlation of gene expression ratios before and after masks. The gene expressions before masks were calculated by the algorithm employed in Affymetrix MAS 5.0 using all 11 probes, while the expressions after mask were calculated by algorithm described in this manuscript using the cutoff values indicated in the graphs.

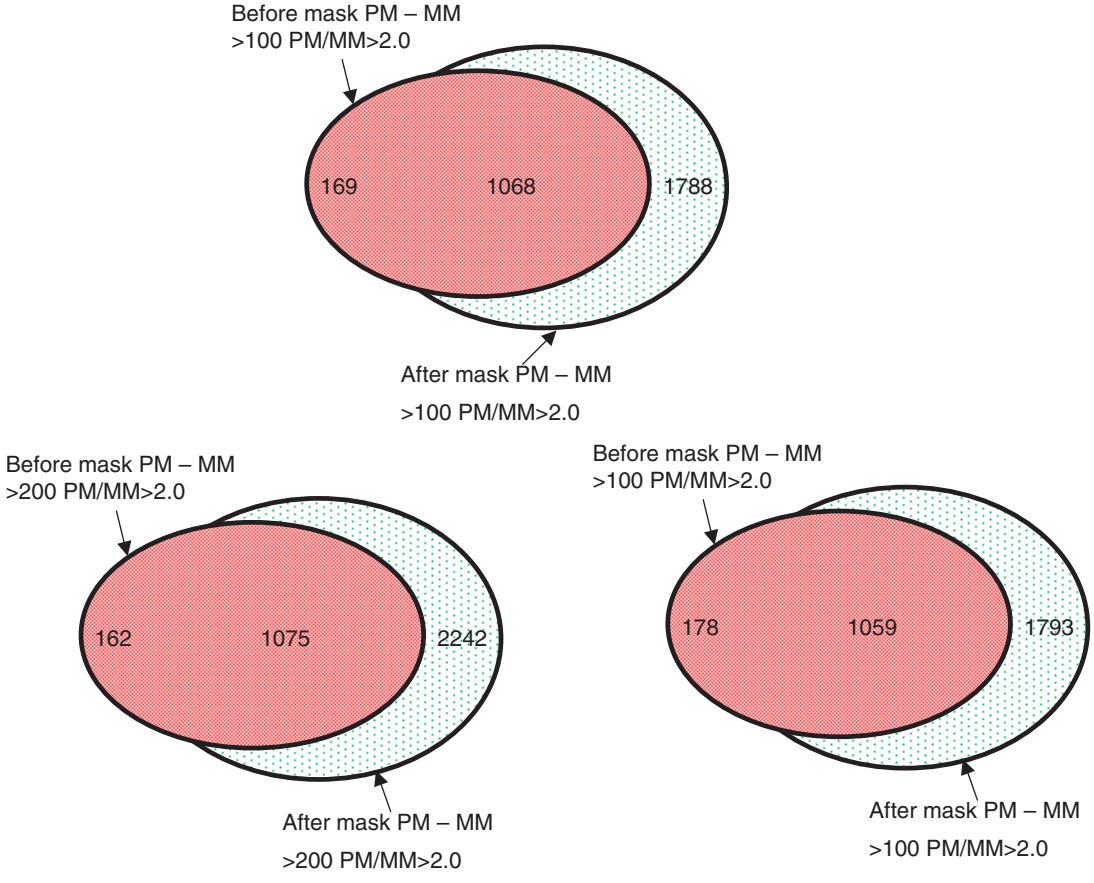

Figure 10. Venn diagrams of differentially expressed genes (heart/liver > 2-fold) before and after mask installation. The gene expressions before masks were calculated by the algorithm employed in Affymetrix MAS 5.0 using all 11 probes, while the expressions after mask were calculated by algorithm described in this manuscript using the cutoff values indicated in the graphs.

generalized for future cross-species hybridizations. We created a simple mathematic model for cross-species hybridization based upon the experimental results described above. From Figure 4, we concluded that a contiguous matched oligo of 16 bp long was sufficient to generate a specific hybridization signal. A similar conclusion has also been reached by Kane *et al*. (29) using an alternative method. If we assume general homology between human and an investigative mammal is *H*

and DNA mutation is random, then the probability of getting a 16 bp contiguous match is H^{16} and the probability of getting at least one mismatch is $1-H^{16}$. If we start from 5′ end of a probe and assume the first 16 bp oligo has mismatches, then the conditional probability of having mismatches for the second 16 bp oligo, which is only a one-base shift from the first one, is decided by the number of mismatches held by the first 16 bp oligo. If the first 16 bp oligo has only one mismatch, then the conditional probability of the second one having mismatches is $15/16 + (1-H)-15/16 \times (1-H) = 1-H/16$. If the first one has two or more mismatches, then the conditional probability of the second one having mismatches is 1. The probability distribution of the first 16 bp oligo having one or multiple mismatches is a binomial one described by Equation 2:

$$P(X = k) = C_{16}^{k} H^{16-k} (1 - H)^{k} \quad \mathbf{2}$$

where $P(X = k)$ represents the probability of the first 16 bp oligo having k mismatches.

Therefore, the conditional probability of the second 16 bp oligo having mismatches can be described by Equation 3:

$$P(2\text{nd} \mid 1\text{st}) = \frac{C_{16}^{1} H^{15} (1 - H)}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} (1 - H/16) + \frac{\sum_{k=2}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} \quad \mathbf{3}$$

where $P(2\text{nd} \mid 1\text{st})$ represents the conditional probability of the second 16 bp oligo having mismatches in the event that the first one has mismatches.

From the General Rule of Multiplication (30), we can estimate the probability of both the first and the second oligos having mismatches by Equation 4:

$$P(1\text{st} \cap 2\text{nd}) = P(1\text{st}) P(2\text{nd} \mid 1\text{st}) \quad \mathbf{4}$$

where $P(1\text{st} \cap 2\text{nd})$ represents the probability of both the first and the second oilgos having mismatches, $P(1\text{st})$ represents the probability of the first oligo having mismatches, and $P(2\text{nd} \mid 1\text{st})$ represents the conditional probability of the second oligo having mismatches in the event the first one has mismatches. For a 25 bp probe, there could be 10 overlapping but different 16 bp oligos. The probability of all the 10 oligos having mismatches can be described by Equation 5.

$$P(1\text{st} \cap \cdots \cap 10\text{th}) = (1 - H^{16}) \times \left[\frac{C_{16}^{1} H^{15} (1 - H)}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} (1 - H/16) + \frac{\sum_{k=2}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} \right]^{9} \quad \mathbf{5}$$

where $P(1\text{st} \cap \cdots \cap 10\text{th})$ is the probability of all the ten 16 bp oligos of the probe having mismatches. For an oligonucleotide microarray with N probes for each gene, the probability of all N probes having mismatches for all 16 bp oligos is given by Formula 6:

$$(1 - H^{16})^{N} \left[\frac{C_{16}^{1} H^{15} (1 - H)}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} (1 - H/16) + \frac{\sum_{k=2}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} \right]^{9N} \quad \mathbf{6}$$

where H is the nucleotide homology between the microarray and investigative species and N is the number of probes. The probability of finding at least one perfectly matched 16 bp oligo in all N probes is given by Formula 7:

$$1 - (1 - H^{16})^{N} \left[\frac{C_{16}^{1} H^{15} (1 - H)}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} (1 - H/16) + \frac{\sum_{k=2}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}}{\sum_{k=1}^{16} C_{16}^{k} H^{16-k} (1 - H)^{k}} \right]^{9N} \quad \mathbf{7}$$

From Formula 7 we can estimate the chance of getting informative probes in a cross-species hybridization. For example, Makalowski and Boguski (6) have shown that average identity between human and mouse in protein-coding nucleotide sequences (CDS) is 85.9%, in 3′-untranslated region (3′-UTR) is 71.0%. Then, the chance of finding at least one informative probe for a rodent gene on Human-U133 GeneChip in the 3′-UTR region is 16% and in CDS region is 69%. Therefore, we are able to find informative probes for many genes among the thousands on an Affymetrix GeneChip.

ACKNOWLEDGEMENTS

We thank Mr Michael Wanner and Mr Eric Sherrer for examination and correction of this manuscript.

REFERENCES

1. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet*., **21** (Suppl. 1), 33–37.
2. Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.
3. Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet*., **21** (Suppl. 1), 20–24.
4. Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.
5. Lewin,R. (1999) *Patterns in Evolution*. Scientific American Library Press, New York, pp. 72–77.
6. Makalowski,W. and Boguski,M.S. (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.
7. Sambrook,J., Fritsch,E.F. and Maniatis, T (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 11.7–11.19.
8. Gauthier,E.R., Madison,S.D. and Michel,R.N. (1997) Rapid RNA isolation without the use of commercial kits: application to small tissue samples. *Pflugers Arch*., **433**, 664–668.
9. Anonymous (2001) *Affymetrix Microarray Suite, Version 5.0*. Affymetrix, Santa Clara, CA.

10. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
11. Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P. (1996) Accessing genetic information with high-density DNA arrays. Science, **274**, 610–614.
12. MacKenzie,D.A., Hullett,D.A. and Sollinger,H.W. (2003) Xenogeneic transplantation of porcine islets: an overview. *Transplantation*, **76**, 887–891.
13. French,A.J., Greenstein,J.L., Loveland,B.E. and Mountford,P.S. (1998) Current and future prospects for xenotransplantation. *Reprod. Fertil. Dev.*, **10**, 683–696.
14. Kohane,I.S., Kho,A.T. and Butte,A.J. (2003) *Microarrays for an integrative genomics*. The MIT Press, Cambridge, MA, pp. 137–140.
15. Butte,A.J., Ye,J., Haring,H.U., Stumvoll,M., White,M.F. and Kohane,I.S. (2001) Determining significant fold differences in gene expression analysis. *Pac. Symp. Biocomput.*, **6**, 6–17.
16. Heid,C.A., Stevens,J., Livak,K.J. and Williams,P.M. (1996) Real time quantitative PCR. *Genome Res.*, **6**, 986–994.
17. Gibson,U.E., Heid,C.A. and Williams,P.M. (1996) A novel method for real time quantitative RT-PCR. *Genome Res.*, **6**, 995–1001.
18. Ji,W., Cai,L., Wright,M.B., Walker,G., Salgam,P., Vater,A. and Lindpaintner,K. (2000) Preservation of gene expression ratios among multiple complex cDNAs after PCR amplification: application to differential gene expression studies. *J. Struct. Funct. Genom.*, **1**, 1–7.
19. Fisher,L.D. and van Belle,G. (1993) *Biostatistics*. John Wiley & Sons, New York, pp. 345–417
20. Chismar,J.D., Mondala,T., Fox,H.S., Roberts,E., Langford,D., Masliah,E., Salomon,D.R. and Head,S.R. (2002) Analysis of result variability from high-density oligonucleotide arrays comparing same-species and cross-species hybridizations. *Biotechniques*, **33**, 516–522.
21. Nagpal,S., Karaman,M.W., Timmerman,M.M., Ho, VV., Pike,B.L. and Hacia,J.G. (2004) Improving the sensitivity and specificity of gene expression analysis in highly related organisms through the use of electronic masks. *Nucleic Acids Res.* **32**, e51.
22. Enard,W., Khaitovich,P., Klose,J., Zollner,S., Heissig,F., Giavalisco,P., Nieselt-Struwe,K., Muchmore,E., Varki,A., Ravid,R., Doxiadis,G.M., Bontrop,R.E. and Paabo,S. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.
23. Gu,J. and Gu,X. (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends Genet.*, **19**, 63–65.
24. Tan,F.L., Moravec,C.S., Li,J., Apperson-Hansen,C., McCarthy,P.M., Young,J.B. and Bond,M. (2002) The gene expression fingerprint of human heart failure. *Proc. Natl Acad. Sci. USA*, **99**, 11387–11392.
25. Smoot,L.M., Smoot,J.C., Graham,M.R., Somerville,G.A., Sturdevant,D.E., Migliaccio,C.A., Sylva,G.L. and Musser,J.M. (2001) Global differential gene expression in response to growth temperature alteration in group A Streptococcus. *Proc. Natl Acad. Sci. USA*, **98**, 10416–10421.
26. Kachigan (1991) *Multivariate Statistical Analysis*. Radius Press, New York, p. 89.
27. Tani,T.H., Khodursky,A., Blumenthal,R.M., Brown,P.O. and Matthews,R.G. (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. *Proc. Natl Acad. Sci. USA*, **99**, 13471–13476.
28. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.*, **19**, 342–347.
29. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
30. Freund,J. (1993) *Introduction to Probability*. Dover Publications, Mineola, NY, pp. 134–144.