

# Reports

## Identification of artifactual microarray probe signals constantly present in multiple sample types

Shihong Mao<sup>1</sup>, Aletheia Lima Souza<sup>1,2</sup>, Robert J. Goodrich<sup>1</sup>, and Stephen A. Krawetz<sup>1</sup>

<sup>1</sup>Center for Molecular Medicine and Genetics, Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA and <sup>2</sup>Department of Animal Science, Federal University of Ceara, Fortaleza, Ceara, Brazil

BioTechniques 53:91-98 (August 2012) doi 10.2144/0000113903

Keywords: microarray; constantly present probe

Supplementary material for this article is available at [www.BioTechniques.com/article/113903](http://www.BioTechniques.com/article/113903)

The detection, identification, and quantitation of transcripts have evolved from simple Northern analysis, cDNA cloning, and sequencing to RT-PCR, microarrays, and now digital gene expression using ultra-high-throughput RNA sequencing (RNA-Seq). During the course of our studies we observed that some microarray probes show very high signal intensity values yet are discordant when compared with RNA-Seq. A total of 99 probes from approximately 30,000 were identified as consistently discordant in four human tissues or cell lines. Interestingly, this set of discordant probes appears array-dependent. Among the 99 probes identified, 70 constantly exhibited a high signal in all 713 available samples surveyed using the Illumina HumanHT-12v4 platform. Some were discordant with additional probes that annotated the same genes. Absence of a number of these transcripts was confirmed by quantitative RT-PCR (qRT-PCR). Our findings suggest that one must be cautious, as some array probes do not capture the level of the target.

Microarray technology has matured into a routine high-throughput means of simultaneously querying the levels of thousands of transcripts. Although this technique remains a valuable transcriptomic tool, it is also well known that the accuracy of microarray results reflects several variables. These include experimental noise, the level of background, and platform sequence probe design. To date, many laboratory methods and analytical procedures have been developed to normalize the data so as to reduce the impact of these factors (1–4). With the exception of a few cases (5–7) where probe signals have been independently confirmed, the veracity of the signal intensities (SIs) of the remaining probes is assumed.

Ultra-high-throughput RNA sequencing (RNA-Seq) is rapidly emerging as an attractive alternative platform to microarrays for quantitative transcriptome profiling. The accuracy and the consistency between these two platforms have been widely reported (8–13). Many reports suggest that RNA-Seq is more accurate than microarrays (9,12,13) when validated using a third methodology. For example, using proteomic validation, Fu and colleagues (9) concluded that RNA-Seq

is more accurate. While in general the correlation coefficient between RNA-Seq and microarrays are consistent (10,12), the correlation coefficient for very low or very highly expressed genes is variable (12). Although this discordance in microarray data has been noticed, the underlying cause has yet to be determined. During the course of our studies, a series of discordant observations were noted when microarray and RNA-Seq data were compared. To reconcile this observation, we defined a set of criteria to determine the extent of discordance among currently available data retrieved from Gene Expression Omnibus (GEO). The analysis presented below reveals an apparent set of platform-specific discordant probes.

### Materials and methods

#### Sample collection and RNA isolation

Semen samples were received from two donors denoted S<sub>1</sub> and S<sub>2</sub>. Each semen sample was equally divided in half to yield a total of four samples (S<sub>1</sub><sup>SCLB</sup>, S<sub>1</sub><sup>PS</sup>, S<sub>2</sub><sup>SCLB</sup>, and S<sub>2</sub><sup>PS</sup>). Two methods were independently utilized for removing somatic cell contaminants. Samples S<sub>1</sub><sup>PS</sup> and S<sub>2</sub><sup>PS</sup> were enriched

by PureSperm gradient centrifugation, and samples S<sub>1</sub><sup>SCLB</sup> and S<sub>2</sub><sup>SCLB</sup> were subject to somatic cell lysis (14). RNA was isolated using the sperm RNA isolation protocol as previously described (15). RT-PCR analysis using intron spanning primers confirmed the absence of DNA in all samples. Testis RNA was obtained from commercial libraries (lot no. 054P010702031A; Applied Biosystems/Ambion, Austin, TX, USA).

#### Generation of microarray data and determination of expressed probes

RNA (50 ng) from each sample was subject to two rounds of amplification using the Message Amp II system (Ambion). Biotin-UTP was incorporated during the second round of amplification. Biotinylated aRNA (750 ng) from each sample was hybridized to HumanHT-12v4 bead arrays (Illumina, San Diego, CA, USA). Arrays were washed, stained, and scanned according to the manufacturer's instructions. Initial data analysis and background subtraction was performed using GenomeStudio version 1.7.0 (Illumina). SI and a coupled *P* value were determined for each probe. Only those probes with a *P* value of less than 0.01 were considered further.

## RNA library preparation and sequencing

Standard RNA-Seq libraries were prepared using the TruSeq Sample Prep kit (part no. RS-930-2001; Illumina) according to the manufacturer's protocol. Briefly, 500 ng purified RNA were fragmented with divalent cations under elevated temperature. First-strand cDNA synthesis was then performed with random hexamers and reverse transcriptase. Second-strand cDNA synthesis was performed using RNase H and DNA Pol I. Following cDNA synthesis, double-stranded products were end-repaired, and PE adaptors (Illumina) were ligated onto the cDNA products followed by 15 cycles of PCR enrichment. All samples were subject to paired-end sequencing using the GAIIx Genome Analyzer (Illumina) for 34 cycles. Image analysis, base calling, FASTQ generation, and demultiplexing were performed using the genome analyzer pipeline software CASAVA version 1.8.1.

## Short read mapping, assembly, and estimating transcript abundance

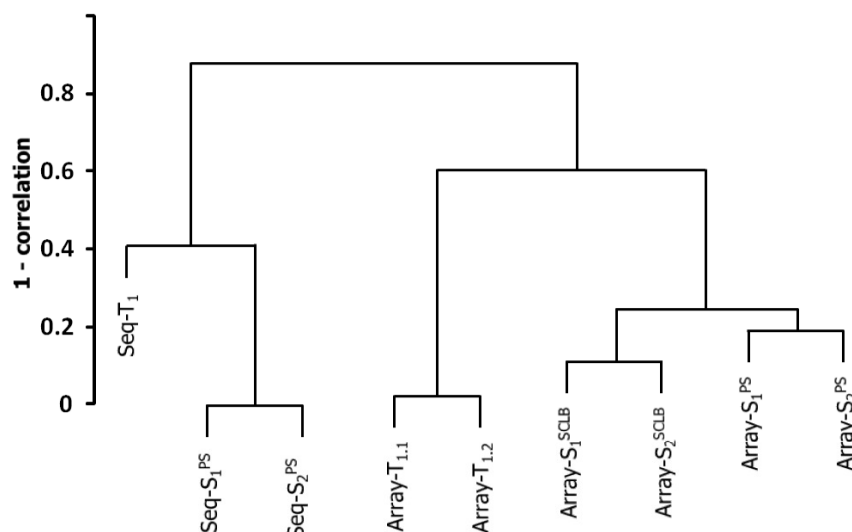
Short reads were mapped to National Center for Biotechnology (NCBI) build 37.2 of the human reference genome using version 1.3.2 TopHat paired-end base default parameters (16). Alignment results were confirmed independently using Novoalign (v.2.07; Novocraft Technologies, Selangor, Malaysia). The relative abundance of each transcript was calculated using Cufflinks version 1.1.0 (17) and presented as fragments per kilobase exon per million fragments mapped (FPKM). The relative abundance of each transcript was also confirmed using Genomatrix RegionMiner (18).

## Primer design for quantitative RT-PCR

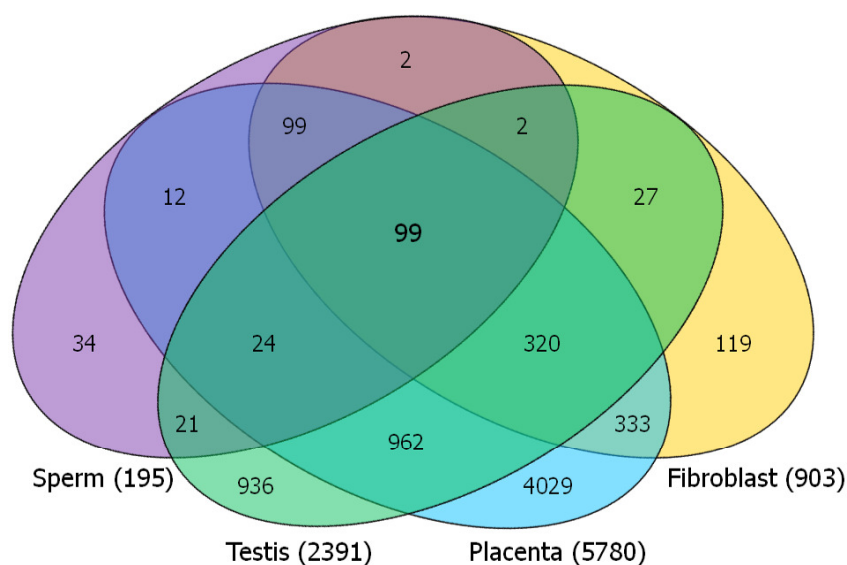
In this study, the primer pairs for quantitative RT-PCR (qRT-PCR) were designed as close to the microarray probes as possible. The ideal primer pairs encompassed or contained the microarray probes. Primers were designed as previously described (19). In brief, the sequences of the gene of interest were retrieved from ENSEMBL genome browser. Candidate primer sequences were generated using Oligo7 (Molecular Biology Insight, Cascade, CO, USA) or Primer3Plus (<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>). The specificity of the primer alignment along the genome was evaluated using the BLAST and BLAT prior to qRT-PCR as described (19).

## Data sets

Sperm RNA data sets that describe both the microarray data GSE39526 and



**Figure 1. Unsupervised clustering of the transcript profile generated from HumanHT-12 microarray and RNA-Seq using two sperm RNA samples and one testis RNA sample.** Each sample in the array was assessed as a set of biological replicates. Seq, indicates RNA-Seq; Array, indicates microarray; S, indicates sperm sample; and T, indicates testis sample. The RNA profiles from the same platform (Array or Seq) are highly correlated, whereas the cross-platform correlation coefficient is low.



**Figure 2. Four-way Venn diagram of the discordant probes from four tissues/cell lines.** The number of discordant probes in each of the four tissues/cell lines is: sperm, 195; testis, 2391; human placenta, 5780; and fibroblast, 903. All four share a total of 99 discordant probes. The remaining discordant probes are tissue- or cell line-specific.

RNA-Seq data GSE39527 are available from GEO GSE39528. Additional HumanHT-12v4 microarray data sets were acquired from GEO (GPL10558). Human placenta and foreskin fibroblast cell line RNA-Seq data sets were obtained from GSE30554 (20).

## Results and discussion

The correlation of the RNA profiles obtained by microarray analysis was initially compared with that obtained from

RNA-Seq using a group of sperm data sets. The repertoire of RNAs in the mature male gamete is limited relative to other cell-types, and this reduced complexity should limit the possibility of cross-hybridization of the transcripts to several probes. However, a series of discordant probes of high SI by microarray, but of low abundance by RNA-Seq or qRT-PCR, were resolved. To determine whether this was a tissue/cell-specific phenomenon, the extent of discordance among other data sets was then determined.

**Table 1. Primer sequences design summary and qRT-PCR amplification**

Gene	Primer sequence	Primer locations	qRT-PCR
AHR	F- 5'-ATTTCAGCGTCAGTCACTGG-3' R- 5'-ACGACATATGAAGCACCTCT-3'	Spans intron, does not contain the probe region	No
ANKRD30B	F- 5'-AGAAAGTGCCAGCTTAATGTCC-3' R- 5'-GTTTTGCACCTCGATGACTG-3'	Does not contain the probe region	No
MCM8	F- 5'-TCAAAGTCTCAAATGCGGAAG-3' R- 5'-CATTCCATGCTTACACCCAT-3'	Spans intron, does not contain the probe region	No
PPP2R3A	F- 5'-CTCCAGAATTGACTTAGCCCTA-3' R- 5'-GCCCAGTGGTTACAATACCTGA-3'	Flanks the probe region	No
PTPLAD2	F- 5'-ACATGAAATAACTGCACATACCC-3' R- 5'-TGAATCTATTGCAGCTAATCTCC-3'	Flanks the probe region	No
RNF213 (KIAA1618)	F- 5'-CTGTGTTCTGCCATGACCCAGCTA-3' R- 5'-AGAACTAGATTTCAGCGCCAT-3'	Flanks the probe region	No
ZNF223	F- 5'-ATAAGAGACTCCATTGCCGAA-3' R- 5'-TATCAAGATTCAAGCGCCTC-3'	Flanks the probe region	No
PRM2	F- 5'-TCTCACTATAGGCGCAGACACT-3' R- 5'-CTTCTCGGCGGCAACTCA-3'	Flanks the probe region	Yes

### GAIIX RNA-Seq and HumanHT-12 correlation in sperm and testes

We aligned the paired-end RNA-Seq reads to the human reference genome. The RNA-Seq metrics are presented in Supplementary Table S1. The majority of reads that could not be aligned are due to their low QC scores. The reads that aligned to multiple positions on the genome were not considered for further analysis. Among the 40% unique aligned reads, ~42% aligned to exon regions, which


is about 12-fold enrichment compared with the genome. The size of each RNA fragment was calculated from the distance spanned from each of the paired reads (mean = ~160 bp, SD = ~50; see Supplementary Figure S1).

Approximately 30,000 HumanHT-12 platform probes were of identical sequence to the current genome build and thus could also be detected by RNA-Seq. A summary of the unsupervised hierarchical clustering of RNA profiles as determined by microarray (SI) and GAIIX deep sequencing (FPKM) is presented in Figure 1. Biological replicates of samples assayed on the same platform, Array or Seq, were highly correlated. Interestingly, many probes that display a very high microarray SI value exhibited a very low FPKM when assessed by RNA-Seq. A series of criteria were developed to assess the generality of this observed discordance.

Two criteria were defined to identify discordant probes, such that if any hybridizing probe ( $P < 0.01$ ) satisfied either criterion, the probe was deemed discordant. First, for each probe and annotated gene above background, an FPKM of  $<1$  was considered discordant. This corresponds to an average of less than one fragment among one million aligned fragments mapped onto a 1-kb exon of the transcript and is considered as background arising from sequencing error(s) or a statistical mapping error. Second, the standard ratio (SR) was considered. That is, the average ratio of SI:FPKM from the group of genes with the highest SI array values and RNA-Seq FPKM when multiple samples are considered. If the ratio of the SI value and FPKM from any probes was 100-fold higher than the SR, the observation was considered discordant. Based on these criteria, a total of 195 and 2391 discordant probes were identified from sperm and testis, respectively (see Supplementary Table S2).

### qRT-PCR validation in sperm samples

Seven genes that exhibited discordant levels between microarray and RNA-Seq results in these samples were selected for verification by qRT-PCR (Table 1). The positions of the qRT-PCR primer pairs that were designed in relation to the microarray probe locations on the genome and qRT-PCR amplification results are summarized in Table 1. All were not detected by qRT-PCR (Table 1). The FPKM values and coverage from each transcript isoform of these seven genes is listed in Supplementary Table S3. The SI values of each corresponding probe, FPKM for each RNA-Seq sample, and the number of fragments mapped to the probe regions of each gene, with the PRM2 transcript providing a positive control, are shown in Table 2. The SIs of these seven discordant genes detected by microarray are very strong. In comparison, these transcripts were underrepresented



**OTCbiotech**  
an operational technologies company

Do You Use Antibodies in your Research?

**Try Aptamers for improved research results!**

OTC Biotech is offering for sale its Best DNA Aptamers against:

- Pathogenic Bacteria • Viruses • Parasites
- Toxins • Clinical Analytes • Small Molecules

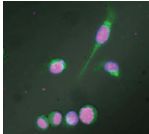
**Aptamers (Synthetic DNA Oligonucleotides):**

- ✓ Are less expensive than antibodies
- ✓ Have higher affinity than antibody counterparts
- ✓ Provide more specific binding vs. antibodies
- ✓ Provide more reproducible results vs. antibodies from lot-to-lot
- ✓ Work well in ELISA, blotting, flow cytometry, IHC, and other common immunologic assay formats
- ✓ Common 5' biotin, primary amine, and fluorescent dye modifications are available with your order

**Visit us at: [www.OTCbiotech.com](http://www.OTCbiotech.com)**

The OTC BioTech Catalogue has 92 Aptamer selections against 42 targets currently available for your review. To place an order:  
Call **800-677-8072** or email [sales@otcbiotech.com](mailto:sales@otcbiotech.com)

**Experience the Aptamer Difference in Your Research!**





**Table 2. SI, FPKM values, and the number of fragments that mapped to the probes regions from two sperm samples.**

Probe ID	Gene symbol	Microarray SI values				RNA-Seq (FPKM)		RNA-Seq (no. fragments)	
		$S_1^{PS}$	$S_1^{SCLB}$	$S_2^{PS}$	$S_2^{SCLB}$	$S_1^{PS}$	$S_2^{PS}$	$S_1^{PS}$	$S_2^{PS}$
ILMN_1812640	AHR	1228.2	3282.1	1687.4	7814.3	0	0.4	0	0
ILMN_1730678	ANKRD30B	1036.6	3516.4	1304.3	9083.0	3.6	2.3	0	0
ILMN_1798581	MCM8	759.5	3721.8	2079.2	10925	1.1	0	0	0
ILMN_1656393	PPP2R3A	99.1	261.7	101.7	674.9	1.8	0	0	0
ILMN_1690114	PTPLAD2	277.5	1432.0	228.5	3886.4	0	0.9	0	0
ILMN_1731203	RNF213	25.3	85.0	39.9	240.9	1.3	0.2	0	0
ILMN_1815578	ZNF223	115.8	571.1	267.3	1860.6	0	0.1	0	0
ILMN_1681772	PRM2	6925.0	2130.2	18178	5318.6	6558	9337	11461	37335

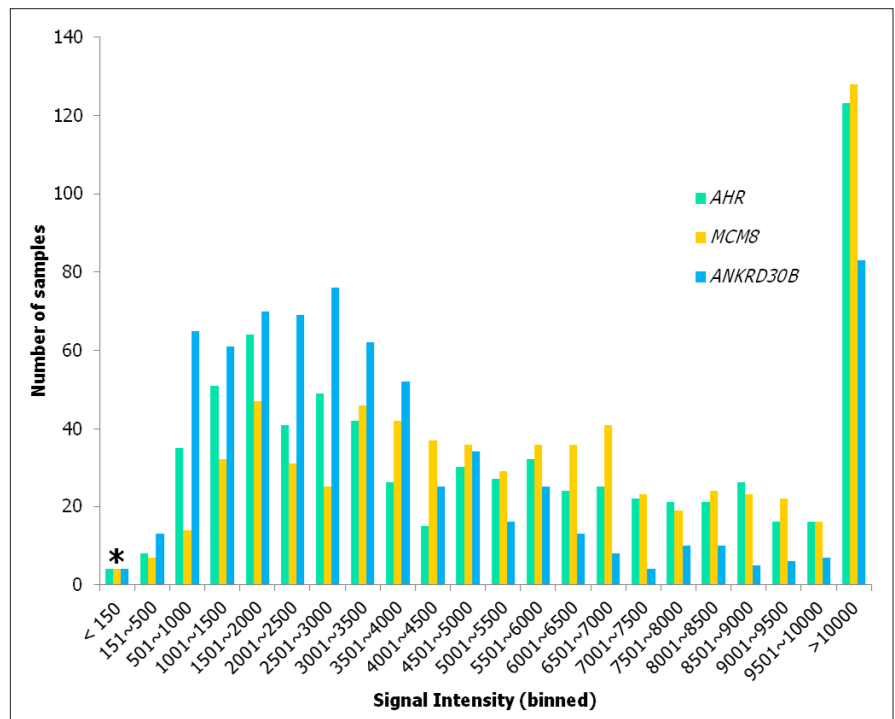
For each probe, the corresponding microarray SI, RNA-Seq FPKM, and the number of fragments mapped from RNA-Seq to the probe sequence region are shown. The probe sequence region was considered as 300 bp upstream and downstream of the position of the microarray probe. The number fragments for each probe were summed based on the extended 650 bp region.

in the RNA-Seq data sets, and no sequence reads could be mapped to the microarray probe regions of these seven genes.

### Discordant genes in other HT-12v4 data sets

In order to determine whether discordant probes were correlated with specific tissues or procedures that could be attributed to individual laboratories, the SIs of HumanHT-12 probes generated from different laboratories were examined using the RNA-Seq FPKM statistic. Based on the above criteria, 5780 discordant probes were identified in the human placenta samples, and 903 discordant probes were similarly identified in the human skin fibroblast cell lines. As illustrated in Figure 2, a four-way comparison of sperm, testis, placenta, and fibroblast cells showed that 99 probes were consistently identified in all of the four tissues or cell lines at  $P < 3.4 \times 10^{-6}$  (21). It appears that the other discordant probes are tissue-specific and/or reflect specific experimental conditions and were not considered further. The list of the 99 probe sequences is detailed in Supplementary Table S4. The set of discordant probes from all four data sets included probes that correspond to AHR, ANKRD30B, and MCM8 (Table 3), which were shown to be absent from sperm by qRT-PCR.

The HumanHT-12v4 bead array can be retrieved from GEO as platform GPL10558. A total of 38 series of publications and experiments consisting of 718 samples were available in GEO. The accession number of each data series in GEO and the number of samples in each series is provided in Supplementary Table S5. The SI values of AHR, ANKRD30B, and MCM8 were first determined since they had been confirmed as absent by qRT-PCR. The stringent criterion of SI of 150 (equal to  $P < 0.01$ ) was set as the threshold was then utilized to assess whether a given probe hybridized. As shown in Figure 3, among the 718 available samples, only five samples exhibited a SI below threshold and were not considered further as



**Figure 3. Distribution of the number of samples as a function of SI values of three genes.** The average SI and sd from 718 samples is  $5891.1 \pm 4632.7$  for AHR;  $6626.5 \pm 5001.4$  for MCM8; and  $4607.0 \pm 4737.0$  for ANKRD30B. A stringent threshold of 150 for SI was set to determine if the probe is hybridized; five samples were under this threshold, as indicated by the asterisk. In over 95% of the samples, their SIs were  $>500$ .

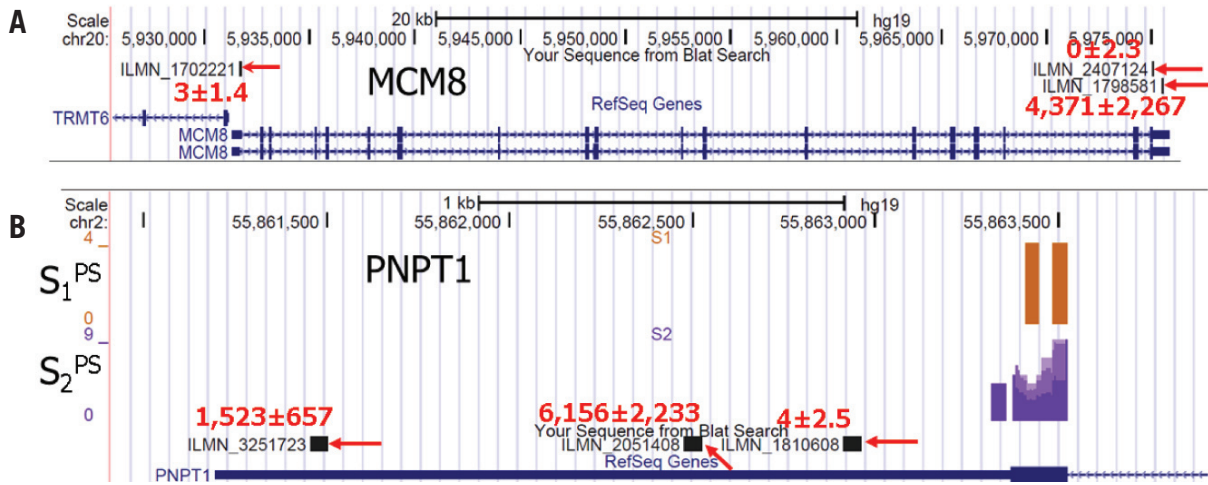
## STEM-CELLBANKER® Optimized stem cell cryopreservation

- Retains stem cell pluripotency, normal karyotype & viability
- Avoid controlled freezing & liquid nitrogen
- Serum free/ Pharmacopeia chemically defined

**amsbio**

**20** years  
of cell freezing experience

>> FIND OUT MORE [www.amsbio.com](http://www.amsbio.com) | [info@amsbio.com](mailto:info@amsbio.com)



**Figure 4. Positions of microarray probes with MCM8 and PNPT1.** (A) Three microarray probes (indicated by red arrows) annotate the MCM8 gene. One is located in the 5' untranslated region (UTR; ILMN\_1702221), while the other two are located in the 3' UTR. SI values (average ± sd of mean) are shown. Probe ILMN\_1798581 shows a very high SI, whereas both the other 3' probe (ILMN-2407124) and 5' probe (ILMN\_1702221) show no signal. (B) Three probes (indicated by red arrows) annotate the 3' UTR region of PNPT1. The aligned sequence reads in samples S<sub>1</sub><sup>PS</sup> and S<sub>2</sub><sup>PS</sup> are also shown. Probes ILMN\_3251723 and ILMN\_2051408 exhibit very high SIs, yet no sequence reads mapped to these positions. A few reads mapped to the exon area of PNPT1 (vertical brown and purple bars).

**Table 3. Statistics of SI and FPKM values of the seven genes from human placenta tissue and human skin fibroblast cell line**

ID_REF	Symbol	Human placenta				Skin fibroblast cell line			
		Microarray SI value			Seq	Microarray SI value			Seq
		Min	Max	Avg	FPKM	Min	Max	Avg	FPKM
ILMN_1812640	AHR	7559.7	15381.6	10766.6	11.3	685.8	3536.4	1799.7	3.0
ILMN_1730678	ANKRD30B	2134.8	3139.6	2650.5	0.0	342.6	3122.9	1177.6	0.0
ILMN_1798581	MCM8	7113.3	14190.4	9244.3	1.4	761.9	8828.8	2787.2	1.4
ILMN_1656393	PPP2R3A	214.1	545.1	417.1	1.2	35.6	250.8	121.6	3.7
ILMN_1690114	PTPLAD2	254.4	563.2	406.4	4.1	70.6	543.5	228.4	1.6
ILMN_1731203	RNF213	145.5	235.0	189.8	12.2	1.3	98.9	41.9	5.6
ILMN_1815578	ZNF223	321.6	646.4	431.9	7.1	46.1	538.6	203.0	1.0
ILMN_1681772	PRM2	-56.0	-23.6	-39.9	0.0	-37.8	3.1	-18.1	0.0

There are 12 samples in human placenta microarray data set and 44 samples in skin fibroblast cell line data set. There is one sample in placenta and one sample in foreskin fibroblast RNA-Seq data. The minimum, maximum, and average SI values of the seven genes from 12 and 44 samples were given in the table, as well as FPKM values from RNA-Seq. PRM2 were used for reference. Three probes (AHR, ANKRD30B, and MCM8) are discordant in both human placenta and skin fibroblast.

the samples tested were predominantly derived from mouse.

The SI values of the 99 discordant probes in all the remaining samples were assessed. Significant signal levels were detected for all 99 probes in greater than 95% of the samples, while 70 probes were detected in all 713 samples (Supplementary Table S4) independent of tissue or cell-type.

In some cases several differentially spaced probes were designed to interrogate a single transcript over an extended region. In the 99 probes identified, 74 are annotated by at least one additional probe. Ideally, in the absence of alternative splicing, the SIs of these probes should be similar, since they are representative of the same transcript. However, the SIs of the constantly present probes are different in this respect. For example, the HT-12 microarray platform uses three different probe sequences to query the level of MCM8. However in all available samples, only probe ILMN\_1798581 exhibits a signal level above background. This is consistent with the view that this probe does not accurately measure the transcript present.

(Figure 4A) as supported by the RNA-Seq (Table 2) and qRT-PCR (Table 1) data sets. Figure 4B shows the discordant probes in gene PNPT1. Probes ILMN\_3251723 and ILMN\_2051408 shows high SI values in all available samples, yet sequence reads could not be mapped to these probes (Figure 4B). They are annotated as constantly present probes in Supplementary Table S4. The discordant

probes within one gene support the view that some probes cannot correctly capture the mRNA levels. The above are consistent with the observations of Marioni and colleagues (8) who, using another platform, compared the consistency of Affymetrix array intensities and Illumina deep sequencing reads. A number of probes showed very high SIs values (Figure 3 in Reference 8) with very low

For Mac & PC

Since 1989

- Fully automatic PCR primers multiplexing,
- TaqMan & nested primers search and analysis
- Search for siRNA and molecular beacons
- Open Reading Frames analysis, including protein data, Batch processing and more!

[www.oligo.net](http://www.oligo.net)

number of sequence reads. Similarly, Malone and colleagues (11) (Figure 3 in Reference 11) showed that many probes have high SIs values and a very low number of reads.

A total of 99 constantly present microarray probes were identified based on the lack of comparable FPKM RNA-Seq values from four different tissues/cell lines. The nature of the constantly present probes has yet to be determined. The discordant probes do not correspond to ribosomal RNA (rRNA) or known microRNA (miRNA) or piwiRNA (piRNA) sequences. They do not form any common pathway nor do they have common biological functions or share a common motif. Mis-annotation is also unlikely, as analysis using the reannotation approach of Barbosa-Morais et al. (22) showed that almost all of the discordant probes mapped to the correct transcripts. Accordingly, the discordance may appear to reflect a yet unknown component of array platform technology. For example, Johnson and colleagues (23) created a custom designed series of probes to detect a suite of human genes inserted within the transgenic mouse genome. The probes were designed to be unique to the human genome. Surprisingly, comparative genomic hybridization (CGH) analysis showed that some probes were preferentially bound by mouse sequences.

The degree to which the constantly present probe may affect the downstream data analysis is significant. On one hand, for a single-class (e.g., up- or down-regulated data set), the effect will be particularly severe. The goal of one class data analysis is to mine the genes that are constantly expressed in each sample. Inclusion of data derived from constantly present probes will negatively influence analysis as the corresponding transcripts are always found to be present at significant levels. On the other hand, for a data set with two or more classes, fold change (4), *P* value, or permutation test (1,2) are typically used to identify the differentially expressed genes. In this case, the constantly present probes may not be noticed. However, as shown in Figure 3, the SI values of such constantly present probes can vary in different samples, and it is possible to mistake the result as significant. Irrespective, microarray technology remains a useful tool to initially survey annotated transcriptomes to infer the state of a cell, but care must be taken to avoid the constantly present probe.

## Acknowledgments

This work was supported in part by the Charlotte B. Failing Professorship to S.A.K. and in part by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and

Human Development Contract 25PM6 in collaboration with the LIFE Study Working Group, Division of Epidemiology, Statistics, and Prevention Research who provided semen samples for analysis. A.S. was supported by visiting scholar fellowship from Brazilian Research Council (CAPES). We are grateful to Graham Johnson and Edward Sandler for their review of the manuscript. S.M. designed the analysis work flow and analyzed the data and wrote the manuscript; A.L.S. designed primers, performed the qRT-PCR experiments, and reviewed the manuscript; R.J.G. prepared libraries for both RNA-Seq and HumanHT-12 bead arrays and reviewed the manuscript; and S.A. K. oversaw the project and edited the manuscript.

## Competing interests

The authors declare no competing interest.

## References

1. Tusher, V.G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116-5121.
2. Fisher, R. 1950. *Statistical Methods for Research Workers*, 11th ed. Oliver & Boyd, Edinburgh.
3. Shi, L., L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24:1151-1161.
4. Mariani, T.J., V. Budhraj, B.H. Mecham, C.C. Gu, M.A. Watson, and Y. Sadovsky. 2003. A variable fold change threshold determines significance for expression microarrays. *FASEB J.* 17:321-323.
5. Yu, H., K. Nguyen, T. Royce, J. Qian, K. Nelson, M. Snyder, and M. Gerstein. 2007. Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool. *Nucleic Acids Res.* 35:e8.
6. Brodsky, L., A. Leontovich, M. Shtutman, and E. Feinstein. 2004. Identification and handling of artifactual gene expression profiles emerging in microarray hybridization experiments. *Nucleic Acids Res.* 32:e46.
7. Nelson, D.C., D.J. Wohlbach, M.J. Rodesch, V. Stolc, M.R. Sussman, and M.P. Samanta. 2007. Identification of an in vitro transcription-based artifact affecting oligonucleotide microarrays. *FEBS Lett.* 581:3363-3370.
8. Marioni, J.C., C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. 2008. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509-1517.
9. Fu, X., N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, et al. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 10:161.
10. Brunskill, E.W., H.L. Lai, D.C. Jamison, S.S. Potter, and L.T. Patterson. 2011. Microarrays and RNA-Seq identify molecular mechanisms driving the end of nephron production. *BMC Dev. Biol.* 11:15.
11. Malone, J.H. and B. Oliver. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9:34.
12. Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57-63.
13. Xiong, Y., X. Chen, Z. Chen, X. Wang, S. Shi, J. Zhang, and X. He. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet.* 42:1043-1047.
14. Goodrich, R., G. Johnson, and S.A. Krawetz. 2007. The preparation of human spermatozoal RNA for clinical analysis. *Arch. Androl.* 53:161-167.
15. Goodrich, R., E. Anton, and S.A. Krawetz. Isolating mRNA and small noncoding RNAs from human sperm. In K. Aston and D. Carrell (Eds.), *Methods in Molecular Biology: Spermatogenesis and Spermiogenesis: Methods and Protocols*. Humana Press, Totawa. (In press).
16. Trapnell, C., L. Pachter, and S.L. Salzberg. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105-1111.
17. Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511-515.
18. Sultan, M., M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321:956-960.
19. Lima-Souza, A., E. Anton, S. Mao, W.J. Ho, and S.A. Krawetz. 2012. A platform for evaluating sperm RNA biomarkers: dysplasia of the fibrous sheath-testing the concept. *Fertil. Steril.* 97:1061-1066.
20. Cabili, M.N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J.L. Rinn. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25:1915-1927.
21. Mao, S., C. Wang, and G. Dong. 2009. Evaluation of inter-laboratory and cross-platform concordance of DNA microarrays through discriminating genes and classifier transferability. *J. Bioinform. Comput. Biol.* 7:157-173.
22. Barbosa-Morais, N.L., M.J. Dunning, S.A. Samarajiva, J.F. Darot, M.E. Ritchie, A.G. Lynch, and S. Tavare. 2010. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.* 38:e17.
23. Johnson, G.D., A.E. Platts, C. Lalancette, R. Goodrich, H.H. Heng, and S.A. Krawetz. 2011. Interrogating the transgenic genome: development of an interspecies tiling array. *Syst. Biol. Reprod. Med.* 57:54-62.

Received 2 April 2012; accepted 10 July 2012.

Address correspondence to Stephen A. Krawetz, 271 C.S. Mott Center, 275 E. Hancock Ave., Detroit, MI, USA. Email: steve@compbio.med.wayne.edu

To purchase reprints of this article, contact: biotechniques@fosterprinting.com