In Vitro Transcription Amplification and Labeling Methods Contribute to the Variability of Gene Expression Profiling with DNA Microarrays

Changqing Ma,^{*†} Maureen Lyons-Weiler,^{*}
Wenjing Liang,^{*} William LaFramboise,^{*}
John R. Gilbertson,^{*†} Michael J. Becich,^{*†} and
Federico A. Monzon^{*†}

From the Department of Pathology^{*} and Center for Pathology Informatics,[†] University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania

The effect of different amplification and labeling methods on DNA microarray expression results has not been previously delineated. To analyze the variation associated with widely accepted T7-based RNA amplification and labeling methods, aliquots of the Stratagene Human Universal Reference RNA were labeled using three eukaryotic target preparation methods followed by uniform replicate array hybridization (Affymetrix U95Av2). Method-dependent variability was observed in the yield and size distribution of labeled products, as well as in the gene expression results. A significant increase in short transcripts, when compared to unamplified mRNA, was observed in methods with long *in vitro* transcription reactions. Intramethod reproducibility showed correlation coefficients >0.99, whereas intermethod comparisons showed coefficients ranging from 0.94 to 0.98 and a nearly twofold increase in coefficient of variation. Fold amplification for each method positively correlated with the number of genes present. Our experiments uncovered two factors that introduced significant bias in gene expression data: the number of labeled nucleotides, which introduces sequence-dependent bias, and the length of the *in vitro* transcription reaction, which introduces transcript size-dependent bias. This study provides evidence that variability in expression data may be caused, in part, by differences in amplification and labeling protocols. (J Mol Diagn 2006; 8:183–192; DOI: 10.2353/jmoldx.2006.050077)

Analysis of gene expression with DNA microarrays has allowed reclassification of tumors based on unique molecular profiles with potentially important prognostic and therapeutic implications.^{1,2} However, there are still significant hurdles for gene expression profiling to achieve routine acceptance within the clinical laboratory. A frequent criticism for the clinical use of this technology is the

lack of concordance among results obtained using different array platforms.³

It is believed that the major causes for platform-dependent differences in gene expression are attributable to variations in array design, probe deposition, probe sequence, and gene annotation.⁴ Although these are major causes of variability in gene expression data, there are other methodological differences that can introduce minor but systematic biases. In this context, little attention has been paid to methodological differences such as the amplification and labeling reactions of different manufacturers. Linear, high-fidelity amplification is critical because it ensures accurate replication of the size, distribution, and complexity of the initial mRNA population. Several studies have suggested that systematic biases are introduced by variations in amplification technique that could impact expression results regardless of the choice of array platform.^{5,6} These results challenge the common underlying assumption that representation of transcripts in a sample remains unchanged by the amplification and labeling protocols used before hybridization.

The most widely used RNA amplification and labeling technique presently in use is the T7-based method developed by Gelder and colleagues (Eberwine method).⁷ A growing number of T7-based amplification systems are now commercially available, and most incorporate modifications from the original technique. The goal of the present study is to specifically test the effect of variations in amplification and labeling protocols on gene expression results. To achieve this goal, we compare three widely used, commercially available target amplification methods.^{8,9} We delineate the variation introduced by each one and determine its potential impact on gene expression data.

Supported by the Pennsylvania Department of Health (Pennsylvania Cancer Alliance Bioinformatics Consortium grant ME-01740 to M.J.B.) and the College of American Pathologists Foundation (scholar's award to F.A.M.).

Accepted for publication November 18, 2005.

This work was performed at the Clinical Genomics Facility of the University of Pittsburgh Cancer Institute.

Supplemental material for this article can be found on <http://jmoldx.amjpathol.org/>.

Address reprint requests to Federico A. Monzon, UPMC Shadyside, Dept. of Pathology, 5230 Centre Ave., Rm. WG22, Pittsburgh, PA 15237. E-mail: monzonfa@upmc.edu.

Table 1. Comparison of Target Amplification and Labeling Methods

	CodeLink	Affy	Affy4h	Enzo
Starting total RNA	1 μ g	1 μ g	1 μ g	5 μ g
Reverse transcription reagents	CodeLink		Invitrogen	
Reverse transcription incubation	2 hours		1 hour	
Second-strand cDNA synthesis reagents	CodeLink		Invitrogen	
Second-strand incubation	2 hours		2 hours	
Biotinylated ribonucleotides	Biotin-11-UTP	Biotin-conjugated Uridine analog		Biotin-CTP
<i>In vitro</i> transcription incubation	14 hours	16 hours	4 hours	Biotin-UTP
Purification and fragmentation reagents	CodeLink	Affymetrix		4 hours

Materials and Methods

RNA Sample

In our experimental design, a single total RNA sample is used to focus on the variation introduced by the differences in amplification methods without the interference from biological variation. The Universal Human Reference (UHR) RNA (Stratagene Corp., La Jolla, CA) was used for all amplification reactions. Aliquots of the total RNA sample were prepared according to the manufacturer's protocol. Quality of the RNA was assessed by OD260/280 in a ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE) and by capillary electrophoresis with the Agilent 2100 bioanalyzer (Agilent Technologies, Inc., Palo Alto, CA). Purification of mRNA was performed with the Oligotex Direct mRNA mini kit (Qiagen Inc., Valencia, CA) as suggested by the manufacturer.

Target Preparation Methods

Methods compared in this study will be described briefly in this section. For details readers are referred to the manufacturers' manuals and selected references.^{8–11} Table 1 summarizes the major differences and similarities among the three target labeling kits utilized.

Affymetrix Eukaryotic Target Preparation

Two *in vitro* transcription (IVT) labeling kits compared in this study are used to prepare biotin-labeled cRNA targets for Affymetrix GeneChip arrays: the Enzo BioArray high-yield RNA transcript labeling kit (Enzo) and the GeneChip expression 3'-amplification reagents for IVT labeling (Affy). For first- and second-strand syntheses, these two methods use reagents from Invitrogen Corp. (Carlsbad, CA) and follow the same experimental steps. Hence, major distinctions between the two methods exist in the IVT step. Twelve UHR RNA aliquots were labeled by each method and five of each were hybridized to arrays. We also performed additional experiments using a modified version of the Affy method in which IVT reactions incubated for only 4 hours at 37°C (Affy4h).

First-Strand and Second-Strand cDNA Synthesis: All reagents are from Invitrogen Corp. unless otherwise specified. Recommended amounts of total RNA (Table 1) in 8 μ l of nuclease-free water were spiked with 2 μ l of diluted poly(A) RNA control (Affymetrix, Santa Clara,

CA) and then incubated with 2 μ l of 50 μ mol/L T7-Oligo (dT)₂₄ primer (Affymetrix) at 70°C for 10 minutes and cooled on ice. Poly(A) RNA controls were diluted to appropriate concentrations immediately before performing the experiment to maintain the same proportionate final concentration of the spike-in controls to the total RNA. First-strand cDNA was synthesized by adding 4 μ l of 5 $\times$ first-strand buffer, 2 μ l of 0.1 mol/L dithiothreitol, 1 μ l of 10 mmol/L dNTP, 1 μ l of Superscript II reverse transcriptase, and incubating at 42°C for 1 hour. Second-strand cDNA was synthesized by adding 91 μ l of nuclease-free water, 30 μ l of 5 $\times$ second-strand buffer, 3 μ l of 10 mmol/L dNTP, 1 μ l of *Escherichia coli* DNA ligase, 4 μ l of *E. coli* DNA polymerase I, 1 μ l of RNase H, and incubating at 16°C for 2 hours. Two μ l of T4 DNA polymerase were added, and the reaction was incubated at 16°C for 5 minutes. Reactions were stopped by adding 10 μ l of 0.5 mol/L ethylenediaminetetraacetic acid. Double-stranded cDNA was purified using the Sample Cleanup Module (Affymetrix).

Synthesis of Biotin-Labeled cRNA with the Enzo Kit: Purified double-stranded cDNA was used in the IVT reaction using the Enzo BioArray high-yield RNA transcript labeling kit (Affymetrix) at 37°C for 4 hours in a 40- μ l reaction volume, containing 4 μ l of 10 $\times$ HY reaction buffer, 4 μ l of 10 $\times$ biotin-labeled ribonucleotides, and 4 μ l of 10 $\times$ dithiothreitol, 4 μ l of 10 $\times$ RNase inhibitor mix, 2 μ l of 20 $\times$ T7 RNA polymerase and variable amounts of RNase-free water.

Synthesis of Biotin-Labeled cRNA with the Affy Kit: Purified double-stranded cDNA was used in the IVT reaction using the GeneChip expression 3'-amplification reagents for IVT labeling kit (Affymetrix) at 37°C for 16 hours in a 40- μ l reaction volume, containing purified double-stranded cDNA, 4 μ l of 10 $\times$ IVT labeling buffer, 12 μ l of IVT labeling NTP mix, 4 μ l of IVT labeling enzyme mix, and variable amounts of RNase-free water. Ten additional labeling reactions incubated for only 4 hours were also performed (Affy4h method).

Fragmentation and Hybridization for Enzo and Affy Protocols: One μ l of purified biotin-labeled cRNA was then analyzed for purity and concentration by ND-1000 spectrophotometer and Agilent 2100 bioanalyzer. For the cRNA prepared by the Affy4h method, purified cRNA from two reactions were pooled to achieve the required amount of cRNA for hybridization. Fifteen μ g of purified cRNA were incubated with the adequate amount of frag-

mentation buffer (Affymetrix) at 94°C for 35 minutes. A 1- μ l aliquot was used to assess complete fragmentation by capillary electrophoresis.

GE Health Care CodeLink Expression System Target Preparation

Twelve biotin-cRNA samples were prepared by the CodeLink method using the CodeLink expression assay reagent kit (GE Health Care, Piscataway, NJ). All reagents used are from this kit unless otherwise specified. One μ g of total RNA in 8 μ l of nuclease-free water were spiked with 1 μ l of working solution of bacterial control mRNAs and 2 μ l of diluted poly(A) RNA control (Affymetrix), then incubated with 1 μ l of T7-oligo (dT) primer at 70°C for 10 minutes and cooled on ice. First-strand cDNA was synthesized by adding 2 μ l of 10 $\times$ first-strand buffer, 4 μ l of 5 mmol/L dNTP mix, 1 μ l of RNase inhibitor, 1 μ l of reverse transcriptase, and then incubating at 42°C for 2 hours. Second-strand cDNA was synthesized in a 100- μ l reaction volume by adding 63 μ l of nuclease-free water, 10 μ l of 10 $\times$ second-strand buffer, 4 μ l of 5 mmol/L dNTP mix, 2 μ l of DNA polymerase mix, 1 μ l of RNase H, and then incubating at 16°C for 2 hours. Double-stranded DNA was purified using the QIAquick PCR purification kit (Qiagen).

IVT reaction was performed by mixing purified double-stranded DNA with 4 μ l of 10 $\times$ T7 reaction buffer, 4 μ l of T7 ATP solution, 4 μ l of T7 GTP solution, 4 μ l of T7 CTP solution, 3 μ l of T7 UTP solution, 7.5 μ l of 10 mmol/L biotin-11-UTP (Perkin-Elmer Corp., Wellesley, MA), and 4 μ l of 10 $\times$ T7 enzyme mix and then incubating for 14 hours at 37°C; final reaction volume was 40 μ l. Biotin-labeled cRNA products were purified with the RNeasy mini kit (Qiagen). Fifteen μ g of cRNA from each sample were fragmented following the recommended procedures in the CodeLink target preparation manual.

Evaluation of Amplification Products

cRNA yield for all methods was assessed in a ND-1000 spectrophotometer (Nanodrop Technologies). Fold amplification was calculated by dividing the total cRNA yield by the estimated mRNA content (2% of total RNA) in the initial starting total RNA of each reaction. mRNA or cRNA size distribution was obtained by capillary electrophoresis with the Agilent 2100 bioanalyzer (Agilent Technologies, Inc.) using the Smear Analysis function of the 2100 Expert software version B.01.02.S1136 (Agilent Technologies, Inc.). Six transcript size regions (0 ~ 0.2 kb, 0.2 ~ 0.5 kb, 0.5 ~ 1.0 kb, 1.0 ~ 2.0 kb, 2.0 ~ 4.0 kb and 4.0 kb ~ max) were defined in the electropherograms and then used to determine the percentage of area under the curve for each size interval. Six individual mRNA samples were evaluated to determine the size distribution of unamplified transcripts. All size distribution data were corrected for rRNA contamination. It is important to note that size distribution in the Agilent bioanalyzer is relative to the

fluorescence intensity and does not reflect the actual number of transcripts of a given size.

Hybridization, Washing, Staining, and Data Processing

Five cRNA samples from each method were hybridized to Affymetrix GeneChip HG-U95Av2 arrays, which contain 12,625 probe sets representing ~10,000 full-length genes. Briefly, 15 μ g of fragmented cRNA were mixed in a hybridization cocktail with control oligonucleotide B2 (Affymetrix), eukaryotic hybridization controls (Affymetrix), herring sperm DNA (Promega Corp., Madison, WI), acetylated bovine serum albumin (BSA) solution (Invitrogen Corp.), 2 $\times$ hybridization buffer (made from MES-free acid monohydrate) (Sigma-Aldrich Corp., St. Louis, MO), MES sodium salt (Sigma-Aldrich Corp.), 5 mol/L NaCl (Ambion, Inc., Austin, TX), 0.5 mol/L ethylenediaminetetraacetic acid (Sigma-Aldrich Corp.), molecular biology grade water, 10% Tween 20 (Calbiochem, San Diego, CA), and 10% dimethyl sulfoxide (for Affy and Affy4h methods only), and variable amounts of water to a final volume of 300 μ l. Two hundred μ l of hybridization cocktail was hybridized on each array at 37°C for 16 hours. Each array was then washed, and stained with streptavidin-phycoerythrin in a GeneChip Fluidics Station 400 (Affymetrix) and scanned by a GeneChip Scanner 3000 (Affymetrix) as recommended by the manufacturer. Quality control (QC) parameters were derived from the MAS 5.0 algorithm of the GCOS software (version 1.1; Affymetrix). Numerical gene expression data were derived from the raw intensity files using two distinct algorithms: the MAS 5.0 and the MBEI algorithm from the dChip software (<http://www.dchip.org>).^{12,13} Gene expression data has been submitted to the National Center for Biotechnology Information's Gene Expression Omnibus with accession number GSE3254.

Analysis of Gene Expression Data

Present (P) and absent (A) calls are based on the detection calls made by the GCOS software. For the purposes of this study, we defined a transcript (probe set) as truly present in the UHR RNA if it was identified as P at least three times in five replicates of any amplification labeling method.

Data from MBEI PM-only model¹² of the dChip software was used for all of the transcript lists analyses. The Avadis Pride software package v3.3 (Strand Genomics, Redwood City, CA) was used for annotation, filtering, and integration of gene expression data. Michael Eisen's Cluster and TreeView software tools (<http://rana.lbl.gov/EisenSoftware.htm>)¹⁴ were used to perform hierarchical clustering and view clustering results. Coefficient of variance (CV) for each transcript across samples was calculated by dividing the SD of its intensity values over the mean and expressed as a percentage (%CV).

Table 2. cRNA Yield, Fold Amplification for Each Method, and Quality Control Parameters from the Hybridizations to HG-U95Av2 Chips (Mean $\pm$ SD)

	CodeLink	Affy	Affy4h	Enzo	P value*
cRNA yield (μ g) [†]	83.80 $\pm$ 41.11 [‡]	37.35 $\pm$ 5.41	10.79 $\pm$ 1.70	31.16 $\pm$ 5.94	<0.0001
Fold amplification [†]	4189.75 $\pm$ 2055.73 [‡]	1867.67 $\pm$ 270.38	554.81 $\pm$ 88.37	322.84 $\pm$ 60.44	<0.0001
Median array Intensity (raw)	128.60 $\pm$ 46.31	109.00 $\pm$ 10.65	150.40 $\pm$ 11.94	207.08 $\pm$ 47.18	0.0019
Background	44.45 $\pm$ 8.89	50.03 $\pm$ 2.40	66.84 $\pm$ 4.41	77.02 $\pm$ 18.18	0.0005
RawQ (noise)	1.57 $\pm$ 0.28	1.76 $\pm$ 0.13	2.18 $\pm$ 0.09	2.53 $\pm$ 0.49	0.0004
% of P calls	54.50 $\pm$ 0.70	51.20 $\pm$ 2.80	54.60 $\pm$ 2.30	48.40 $\pm$ 3.30	0.0026
No. of present genes	7476	7207	7455	6869	NA

*One-way analysis of variance with Bonferroni correction.

[†]*n* = 12. For all other rows, *n* = 5.

[‡]Standard deviation for CodeLink method reflects lot-to-lot differences not seen in other amplification kits.

Two-class unpaired comparisons of gene expression data from two methods were performed with the Significance Analysis of Microarrays (SAM)¹⁵ software tool v1.21 (<http://www-stat.stanford.edu/~tibs/SAM/>). All gene expression profile comparisons with SAM were performed at a false discovery rate (FDR) of less than 0.03% (Delta level of 3.0), except the comparison between Affy and Affy4h data, which was performed at a FDR of 0.32% (Delta = 2.0). STATA software v8.01 (STATA Corp., College Station, TX) was used for all other statistical analysis including correlation studies, Mann-Whitney tests, analysis of variance (all QC data), and regression analysis. SigmaPlot v.8.0 (SPSS Inc., Chicago, IL) and Microsoft Excel (Microsoft, Redmond, WA) were used for all plots.

For each method A to method B comparison of intensity values with SAM, transcripts that showed significantly increased values in method A over B were labeled as "affected by A." Conversely, transcripts significantly increased in method B, therefore decreased in method A, were labeled "affected by B." For the Enzo versus Affy4h comparison, we calculated differences in cytosine content in the target sequence of transcripts affected by these methods. The target sequence of a transcript is defined as the region interrogated by all probes in a probe set in the Affymetrix HG-U95Av2 array. Differences in cytosine content were calculated as the ratio of cytosine (c) to uracil (u) and expressed as guanine/adenine (G/A), thus reflecting the actual mRNA sequence. For the Affy versus Affy4h comparison, transcript sizes reported correspond to the target mRNA sizes reported by the array manufacturer. Both transcript lengths and probe sequence information were obtained from the NetAffx website (www.affymetrix.com).

Results

cRNA Yields

More than 30 μ g of cRNA were obtained with the Affy, Enzo, and CodeLink methods in almost all reactions (Table 2). The Affy4h method yielded $\sim$ 10 μ g on average. The CodeLink method had the highest cRNA fold amplification and showed more variability in cRNA yields, which was mostly based on lot-to-lot differences of the amplification kit (Table 2). Lot-to-lot variability in amplification yield was not observed in the Enzo or Affy methods.

Hybridization Performance

All hybridizations met quality control (QC) criteria as defined by the array manufacturer; however, some significant differences were noted (Table 2). Compared to hybridization results from Affy and CodeLink methods, the Enzo method had statistically significant higher background (one-way analysis of variance; Affy versus Enzo, *P* = 0.005; CodeLink versus Enzo, *P* = 0.001), rawQ value (noise) (Affy versus Enzo, *P* = 0.004; CodeLink versus Enzo, *P* = 0.001), and average median array intensities (raw) (Affy versus Enzo, *P* = 0.002; CodeLink versus Enzo, *P* = 0.012).

There were no significant differences across samples in the 3'/5' ratios of GAPDH, Lys, and Phe (Table 3). However, the 3'/5' ratios for β -actin, Dap, and Thr were significantly higher in the samples labeled with the CodeLink method compared to Affy, Enzo, and Affy4h methods (β -actin and Thr, *P* < 0.001 for all methods;

Table 3. 3'/5' Ratios for Housekeeping Genes and Bacterial Poly(A) RNA Spike Controls (Mean $\pm$ SD)

	CodeLink	Affy	Affy4h	Enzo	Transcript Length (kb)
Housekeeping genes					
GAPDH	1.16 $\pm$ 0.12	1.11 $\pm$ 0.06	1.07 $\pm$ 0.10	0.91 $\pm$ 0.05	1.27
β -Actin	5.08 $\pm$ 1.73	1.44 $\pm$ 0.13	1.20 $\pm$ 0.08	1.10 $\pm$ 0.13	1.76
Poly(A) RNA spike controls					
Lys	3.04 $\pm$ 0.24	3.16 $\pm$ 1.46	2.97 $\pm$ 1.21	2.87 $\pm$ 0.58	1.00
Phe	1.75 $\pm$ 0.22	1.65 $\pm$ 0.11	2.45 $\pm$ 0.56	2.45 $\pm$ 0.14	1.32
Dap	5.85 $\pm$ 1.12	2.34 $\pm$ 0.40	3.11 $\pm$ 0.92	2.47 $\pm$ 0.95	1.82
Thr	4.43 $\pm$ 0.36	2.02 $\pm$ 0.12	2.35 $\pm$ 0.22	2.37 $\pm$ 0.67	1.98

Figure 1. Concordance on present/absent transcript calls among the four methods studied. Total number of transcripts in the U95v2 array is 12,592. Complete agreement among methods is represented by a **white background** and further divided into present and absent calls. Disagreement among methods is **shaded**.

Dap, $P = 0.004$, 0.006 , and 0.011 for each method, respectively). Interestingly, control transcripts that showed increased 3'/5' ratios in the CodeLink method are all nearly 2 kb long, while the controls not affected by this bias (GAPDH, Lys, and Phe) are all less than 1.5 kb long. Additionally, rRNA sequences were identified as present by the MAS5 algorithm in all but the Enzo labeling method. Interestingly, intensity values for the rRNA probe sets in the Enzo method were not significantly decreased as compared to the others, which indicates that the present/absent calls are not directly related to a higher abundance of labeled rRNA transcripts (Supplemental Figure 1 at <http://jmd.amjpathol.org/>) and most likely reflect the higher noise and background observed with this method (Table 2).

The set of present genes, as defined in the Materials and Methods section, consisted of 8281 transcripts, equivalent to 65.76% of all probe sets on a HG-U95Av2 array. The Enzo method had the lowest number of present probe sets (Table 2). There was a positive correlation ($R^2 = 0.9553$) between fold amplification and the number of present transcripts in samples from the Affy, Enzo, and CodeLink methods. Furthermore, this correla-

tion is maintained as the stringency of the present transcript definition goes from at least three of five replicates to four of five and five of five (data not shown). Interestingly, despite having relatively low-fold amplification, the number of present probe sets in data from the Affy4h method is almost identical to the CodeLink method. The four methods showed 83.3% agreement in present/absent calls for all transcripts interrogated by the HG-U95Av2 array (Figure 1). Of these, 6183 (74.66%) were identified as present by all four methods. Only 2098 were discordant between methods, and from the discordant set less than 10% (of all transcripts on the array) were identified as present by only one method.

Size Distribution of cRNA Products

Table 4 shows the distribution of cRNA products for each method. These data are derived from the electropherogram profiles of the IVT products. All methods yielded cRNA with different size distributions when compared to the nonamplified mRNA in the Universal Human Reference RNA sample, with the Enzo method being most similar to it. One-way analysis of variance shows that the most significant difference among all methods was seen in the abundance of transcripts sizes between 0 to 200 bp ($P < 0.0001$) and 200 to 500 bp ($P < 0.0001$, except Enzo, $P = 0.014$).

Long incubation methods (Affy-16 hours and CodeLink-14 hours) produced a significantly higher abundance of short cRNA transcripts (<1000 nucleotides) in comparison to short incubation methods (Enzo and Affy4h) (t -test: 0 to 200 bp: $P < 0.0001$; 200 to 500 bp: $P < 0.0001$; 500 to 1000 bp: $P = 0.0002$). In contrast, short incubation methods (Enzo and Affy4h) produced a higher percentage of longer cRNA transcripts (>2000 nucleotides) (2000 to 4000 bp: $P < 0.0001$; 4000-max: $P = 0.0002$), indicating a shift toward smaller transcripts in long incubation methods. When comparing Affy and Affy4h methods (same reagent kit, but different IVT reaction length), these significant differences were also present (t -test: 0 to 200 bp: $P = 0.0002$; 200 to 500 bp: $P = 0.0043$; 1000 to 2000 bp: $P = 0.0047$; 2000 to 4000 bp: $P = 0.0004$). When we plot the transcript abundance for each size range against fold amplification for each method, methods with high-fold amplification showed a trend to enriching small transcripts (<1000 bp), whereas methods with low-fold amplification generated a higher

Table 4. Size Distributions of mRNA and cRNA Samples (Percent in Size Region)

Base pair	Non-amplified	Long IVT Reactions		Short IVT Reactions		<i>P</i> value*
	UHR mRNA (<i>n</i> = 6)	CodeLink (<i>n</i> = 6)	Affy (<i>n</i> = 4)	Affy4h (<i>n</i> = 10)	Enzo (<i>n</i> = 6)	
0 ~ 200	4.2 ± 1.35	11.5 ± 1.03	10.4 ± 0.52	8.2 ± 0.97	6.9 ± 0.46	<0.0001
200 ~ 500	4.5 ± 1.02	15.4 ± 1.30	14.0 ± 1.39	11.2 ± 1.06	7.1 ± 0.62	<0.0001
500 ~ 1000	17.6 ± 9.11	26.3 ± 1.92	21.8 ± 1.53	21.4 ± 0.89	16.4 ± 0.85	0.0002
1000 ~ 2000	28.3 ± 10.40	25.6 ± 1.16	23.4 ± 1.22	25.1 ± 0.57	24.7 ± 0.96	0.6162
2000 ~ 4000	29.2 ± 13.38	15.1 ± 1.45	19.4 ± 1.04	22.2 ± 0.96	28.4 ± 0.60	<0.0001
4000 ~ max	16.0 ± 7.54	6.12 ± 1.88	11.08 ± 3.01	11.9 ± 1.77	16.5 ± 1.13	0.0002

Note: Intensity of shading in each cell is directly proportional to its value.
*Long IVT versus short IVT (Student's *t*-test).

Table 5. Intra- and Inter-Method Pair-Wise Correlation Coefficients

	Affy	CodeLink	Enzo	Affy4h
Affy	0.9958			
CodeLink	0.9795	0.9916		
Enzo	0.9650	0.9494	0.9953	
Affy4h	0.9890	0.9703	0.9662	0.9962

abundance of long transcripts (Supplemental Figure 2 at <http://jmd.amjpathol.org/>).

Reproducibility of Gene Expression Measurements

Pair-wise Pearson correlation coefficients of normalized gene expression measurements, within and between methods, were calculated using the set of present transcripts. Gene expression data showed excellent intra-method reproducibility and sensitivity, with correlation coefficients >0.990 for all methods (Table 5). The Affy and Affy4h methods had the highest intermethod correlation coefficient ($r = 0.989$), whereas the Enzo and CodeLink data correlated with each other the least ($r =$

0.949). With unsupervised hierarchical clustering, the arrays formed distinct clusters based on target preparation methods confirming that intermethod variability is greater than intramethod variability (data not shown).

Variability of Gene Expression Measurements

CVs for each present transcript were calculated across all replicates within a method (intra-assay) or across all four methods (interassay). As seen in Figure 2a, all methods had average CVs of less than 12%, with Affy having the highest ($10.45 \pm 6.64\%$) and Affy4h the lowest ($7.41 \pm 4.81\%$). Intermethod variability was almost double that of the intramethod (mean, $19.93 \pm 9.87\%$). Figure 2, b–d, shows examples of the variability seen between methods for selected transcripts. CV plots for all transcripts in each method are presented in Supplementary Figure 3 (<http://jmd.amjpathol.org/>). As has been shown in other studies, variability was higher for transcripts in the low-intensity region.^{4,13}

Paired comparisons between all methods with the SAM algorithm revealed significant changes in transcript measurements, showing that cRNA targets prepared by the four studied methods have significant, reproducible, and consistent differences (Table 6). Because all experiments

Figure 2. Variability in gene expression data. **a:** Intra- and interassay CV for all present transcripts. The **solid line** on each box represents the median CV and the **dashed line** represents the mean CV. **b:** Example of two transcripts with high-intensity values in hybridization results showing no change across methods. **c:** Example of one transcript with low-intensity values showing difference between Affy versus Affy4h. **d:** Example of low expressor affected by the Enzo method in the Enzo versus Affy4h.

Table 6. SAM Analysis Results from Paired Comparison of All Methods

Method A	Method B	FDR (%)	Number of affected transcripts (%*)	Affected in A (%*)	Affected in B (%*)
Affy	CodeLink	0.0202	1633 (19.7)	864 (10.4)	769 (9.3)
Affy	Affy4h	0.3187 [†]	2029 (24.5)	1335 (16.1)	694 (8.4)
Affy	Enzo	0.0109	5070 (61.2)	2577 (31.1)	2493 (30.1)
CodeLink	Affy4h	0.0151	3090 (37.3)	1445 (17.4)	1645 (19.9)
CodeLink	Enzo	0.0140	4976 (60.1)	2407 (29.1)	2569 (31.0)
Affy4h	Enzo	0.0082	5085 (61.4)	2585 (31.2)	2500 (30.2)

*Percentage of all present transcripts (8281).

[†]Delta for this comparison was set at 2.0, all others at 3.0.

started with the same total RNA and were hybridized to the same array type, these differences are introduced by the target preparation (amplification) methods. For each method A to method B comparison of intensity values with SAM, transcripts that showed significantly increased values in method A over B were labeled as “affected by A.” Conversely transcripts significantly increased in method B were labeled “affected by B.” The comparison between Enzo and Affy4h methods had the highest number of affected transcripts; whereas the comparison between Affy and Affy4h had the lowest even at a less stringent level. For all comparisons, each method accounted for approximately half of the affected transcripts.

Because there are multiple factors that could contribute to the observed intermethod differences, we performed two focused comparisons that allowed us to isolate the sources of variation: the Enzo versus Affy4h comparison was used to analyze the effect of double-nucleotide labeling, and the Affy versus Affy4h comparison was used to analyze the effect of long IVT reaction time. From all of the methods studied, Affy and CodeLink are the most similar in terms of workflow; however, comparison between these two methods still showed affected transcripts that could not be explained by the variation sources discussed above. Lists of genes affected for Enzo/Affy4h, Affy/Affy4h, and Affy/CodeLink comparisons are presented in Supplemental Tables 1 to 3 (<http://jmd.amjpathol.org/>).

Sources of Variation

Dual Labeling

The Enzo method uses double-nucleotide labeling (biotin-CTP and biotin-UTP) whereas others use single labeling (Table 1). Samples labeled by the Enzo method had higher average unnormalized fluorescence intensity values than all other methods (Table 2). As seen in Table 6 for the Enzo/Affy4h comparison, 61.4% of all transcripts have significantly different gene expression values and are therefore affected by the method-dependent variation.

We hypothesized that if this method-dependent variation is a direct result of the double-nucleotide labeling, then the transcripts that show higher gene expression values with the Enzo method will have a higher cytosine content in the transcript sequence interrogated by the probe set, because this nucleotide is only labeled by this method. This was expressed as the G/A ratio of the target

transcript sequence as defined in the Materials and Methods section. The average G/A ratio of transcripts showing elevated expression in Enzo data were 1.166 ± 0.485 , which is significantly higher than those of transcripts increased by the Affy4h method (0.773 ± 0.305 ; Mann-Whitney test: $z = -32.477$; $P < 0.00001$). When transcripts that are affected significantly by the two methods are categorized according to their G/A ratio, we found that 93.7% of transcripts with ratios >2.0 show significantly higher values with the Enzo method and 84.70% of genes with ratios <0.5 show higher values with the Affy4h method (Figure 3).

Incubation Time

Given that the Affy and Affy4h methods only differed in the length of IVT incubation time (Table 1), comparison of these two methods provides an insight on how this factor affects gene expression data. In this comparison, 24.5% of all present transcripts are significantly different between Affy and Affy4h methods with a Delta of 2.0 (FDR = 0.3187%).

Based on the transcript size shift observed with long IVT reactions, we hypothesized that transcripts with significantly higher expression values in samples labeled

Figure 3. Role of double-nucleotide labeling in the variability of gene expression data. Enzo- and Affy4h-affected transcripts (totally 5085) were divided into groups based on the guanine to adenine (G/A) ratio in the target sequence, reflecting cytosine to uracil incorporation into the cRNA transcript. In each group, the number of affected transcripts by each method and the corresponding percentages of total number of affected transcripts in this group were obtained. The percentage of affected transcripts of each method in each group is plotted.

Figure 4. Role of transcript length in the variability of gene expression data. Affected transcripts from SAM analysis of the Affy versus Affy4h comparison were grouped based on their transcript lengths. The proportion of affected transcripts for each method in each transcript length interval is plotted. Please note only the transcripts shorter than 1.5 kb were of interest to this analysis.

with a long (overnight) IVT are more likely to be short transcripts. Therefore, we investigated if genes <1.5 kb would be preferentially amplified by a long IVT labeling method. Figure 4 shows the percentage of transcripts <1.0 kb that are selectively increased in the Affy method in comparison to the Affy4h. These data show an inverse relationship between transcript length and the percentage of transcripts whose expression values were increased by the long IVT. Linear regression analysis shows an R^2 of 0.9291, indicating a strong association between the increase of transcript length and the decrease of the proportion of long IVT affected transcripts. This association could not be found when a comparison of both long IVT methods (Affy/CodeLink) was done (Supplemental Figure 4; <http://jmd.amjpathol.org/>).

Discussion

This study demonstrates specific biases in gene expression data introduced by commercially available T7 RNA polymerase-based amplification reagent kits and protocols. Although T7 amplification is generally regarded as linear, several studies have shown differences in gene expression between amplified cRNA (single or double round) and nonamplified mRNA.^{5,6,16-18} Our results extend those obtained in other studies and show that these differences in gene expression results can be dependent on two important factors: the number of labeled nucleotides in the amplification kit and the length of IVT reaction, which translates to a transcript size-dependent bias. Our experimental design, focusing on a single RNA sample, allowed us to identify and characterize these specific biases. A similar approach was recently used by Daly and colleagues¹⁹ in their characterization of the variability in the Affymetrix array platform in a clinical context. This approach focuses on characterizing the variability inherent to the method or platform that can potentially be used in a clinical setting.

Our results suggest that, when choosing or designing labeling kits for clinical applications, attention should be paid to the number of biotinylated ribonucleotides used for labeling at the IVT step. When comparing single versus double-nucleotide labeling with normalized data, we found that ~30% of the present genes had substantially higher gene expression values in Enzo (double nucleotide) compared to Affy4h (single nucleotide), suggesting the data sets generated from methods using two labeling nucleotides are not directly comparable to data sets derived by using a single labeling nucleotide. It has previously been shown that incorporation of biotin-CTP is not as efficient as biotin-UTP.⁸ Our results are in agreement with these findings, because we found differences when the guanine/adene (G/A) ratio of the targeted sequence was higher than 2, indicating that at least two incorporated biotin-CTPs per biotin-UTP are necessary to significantly increase the amount of fluorescent signal per transcript. It is essential to note that Enzo and Affy4h methods only differ at the IVT reaction, with all other steps and reagents being identical (Table 1). Therefore, the IVT reaction is the only source of the observed variation between these two methods. Although other components in this reaction might also vary, the most significant difference between these two methods is the number of biotinylated nucleotides. The possibility that differences in enzyme concentration or buffer composition between the two IVT kits may contribute to the observed variation cannot be formally excluded in our experiments. However, the correlation between the number of guanines in the target sequence and higher expression values in the Enzo method is a strong argument for the role of double-nucleotide labeling as a source of this variation.

We also demonstrate that the distribution of transcripts shifts toward shorter cRNA products in protocols with long IVT incubations, suggesting enhanced amplification of short transcripts. This is further corroborated by the fact that short transcripts were more likely to be increased in cRNA samples from long IVT labeling methods. Interestingly, Spiess and collaborators²⁰ reported a similar cRNA size shift with long IVT incubation, but suggested that degradation of cRNA molecules by T7 RNA polymerase accounted for this observation. However, in our results, long incubations consistently gave higher yields, contrasting with the decrease in cRNA yield after 5 hours observed in their study. Furthermore, in their description of exonuclease activity of T7 RNA polymerases, Sastry and Ross²¹ indicated that this activity is only unmasked in paused/arrested transcription complexes and that the kinetic balance during normal transcription is balanced toward polymerization. We speculate that the degradation and/or decrease in IVT yields seen by Zhao and colleagues¹⁸ and Spiess and colleagues²⁰ with IVT reactions exceeding 4 hours could be a result of paused transcription complexes due to depletion of reaction components. New IVT kits that are designed for longer incubation times seem to overcome this problem. Although the degree of amplification correlated with the increase in short cRNA transcripts, we were unable to assess the role of enzyme concentration between protocols with identical incubation times because

the kit manufacturers would not provide this proprietary information.

In this study, the number of transcripts identified as P in a sample, was directly related to the degree of amplification achieved in all methods but one (Affy4h). This suggests that transcripts actually present in a sample are not always amplified successfully, which contributes to the variability within and between assays. In fact, as seen in other studies,⁴ variability in gene expression measurements was most pronounced in the low-fluorescence intensity range (ie, in the low-expressor transcript range) as would be expected if low-abundance transcripts are not efficiently amplified each time. It is interesting to note that the Affy4h method, in which we used pooled reactions due to low-fold amplification, yielded similar P calls as the CodeLink platform, which showed the highest fold amplification. These results suggest that multiple labeling reactions may be more effective at amplifying low-expressor transcripts, because more transcription initiation events may occur with multiple short-term incubations. Further testing of this hypothesis is currently underway in our laboratory.

In the present study, all methods provided low intra-method CVs, but intermethod variability was considerably higher. Intramethod variability reflects random errors created during the performance of a specific method, whereas intermethod variability comprises both random experimental errors and systematic biases. Average CVs across any two methods ranged from 15.65 to 20.44% approximating the average CV across all methods of 19.93%. Other studies have reported correlation coefficients for the CodeLink and Affymetrix platforms between 0.59 to 0.79.^{4,22,23} In our study we obtained higher correlation coefficients between these two platforms, which could reflect the fact that all samples were hybridized to the same array type, therefore isolating only the variability contributed by the labeling method.

Another significant difference observed between labeling methods was underrepresentation of 5' probes from genes larger than 1.5 kb with the CodeLink method. This phenomenon was observed by Baugh and colleagues,⁶ and was demonstrated to be related to inefficient reverse transcription. Indeed, when comparing the CodeLink method against all others, which share a common reverse transcription step, the former requires a longer incubation period (2 hours versus 1 hour) that may lead to depletion of dNTPs and early termination of reverse transcription reactions yielding 5' truncated cDNA products. It is also possible that IVT further contributes to 5' underrepresentation when the T7 RNA polymerase fails to transcribe full-length transcripts. It is likely that the majority of gene expression results are not affected by this phenomenon, because most probes in current array designs are 3' biased, but this factor should be taken into account for probes that interrogate the 5' region of selected transcripts.

Until now, reports on gene expression biases were limited to a description of this problem; however, we have performed a characterization of the factors that contribute to this variability. In summary, our results indicate that variability introduced by T7 RNA polymerase-based am-

plification methods can be explained, at least partially, by two factors: the number of biotinylated nucleotides used in the labeling reaction and the length of the IVT reaction. These biases are not corrected by intensity-based normalization techniques, such as the invariant set normalization method,¹² and therefore can generate discordant results even if the same sample is analyzed with different labeling methods. Although our results do not address the impact of these biases in the classification of samples by gene expression (gene expression profiling) or the question of which labeling method best reflects the actual transcriptome, they explain, in part, the discordant results seen between studies with similar experimental designs.³ Our results show that the bias introduced by the IVT method is insufficient to overcome biological variability. However, the fact that it introduces sequence- and transcript size-dependent variation in a systematic manner can lead to erroneous experimental results. This is relevant for researchers using gene expression profiling as a discovery tool and those performing meta-analysis of gene expression profiles from different studies.

It is expected that, newly developed sequence-based normalization methods could overcome these biases in gene expression data. As shown recently, concordance between different platforms has improved substantially thanks to advances in gene annotation and array design,²⁴ and high reproducibility among laboratories can be achieved when standardized protocols and array platforms are used.^{25,26} As shown by Dobbin and colleagues,²⁷ biological variability is maintained if a standardized operating protocol (SOP) is used. Therefore, studying the effect of different labeling protocols on the ability to detect biological variability would have been redundant to other studies.²⁶ However, it is expected that standard operating procedures to perform clinical tests based on gene expression profiles will be developed. To this end, data from our experiments could also be used to establish which microarray probes have acceptable performance across multiple labeling protocols, as suggested by Daly and colleagues.¹⁹ Our results emphasize the importance of standardization in target preparation methods to optimize gene expression analysis and achieve a consistency compatible with the clinical application of this technology.

Acknowledgments

We thank Uma Chandran, James Lyons-Weiler, and Jeffrey Kant for their helpful discussion and commentaries.

References

1. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* 2004, 101:8111-8116
2. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: A gene-expression

- signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, 347:1999–2009
3. Johnson K, Lin S: QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. *BioTechniques* 2003, Mar. Suppl:62–63
4. Tan PK, Downey TJ, Spitznagel Jr EL, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003, 31:5676–5684
5. Puskás LG, Zvara Á, Hackler Jr L, Van Hummelen P: RNA amplification results in reproducible microarray data with slight ratio bias. *BioTechniques* 2002, 32:1330–1340
6. Baugh LR, Hill AA, Brown EL, Hunter CP: Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* 2001, 29:e29
7. Gelder RNV, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH: Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci USA* 1990, 87:1663–1667
8. Dorris DR, Ramakrishnan R, Trakas D, Dudzik F, Belval R, Zhao C, Nguyen A, Domanus M, Mazumder A: A highly reproducible, linear, and automated sample preparation method for DNA microarrays. *Genome Res* 2002, 12:976–984
9. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Norton H, Brown EL: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech* 1996, 14:1675–1680
10. CodeLink Gene Expression System: Manual Labeled cRNA Target Preparation. Edited by GE Healthcare, Piscataway, 2004
11. Affymetrix: Affymetrix GeneChip Expression Analysis Technical Manual. Santa Clara, Affymetrix, 2001
12. Li C, Wong WH: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2001, 2:research 0032.1–0032.11
13. Li C, Wong WH: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 2001, 98:31–36
14. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998, 95:14863–14868
15. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001, 98:5116–5121
16. Dumur CI, Garrett CT, Archer KJ, Nasim S, Wilkinson DS, Ferreira-Gonzalez A: Evaluation of a linear amplification method for small samples used on high-density oligonucleotide microarray analysis. *Anal Biochem* 2004, 331:314–321
17. Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y, Mao Y: Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. *J Biotechnol* 2004, 107:19–28
18. Zhao H, Hastie T, Whitfield M, Borresen-Dale A-L, Jeffrey S: Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. *BMC Genomics* 2002, 3:31
19. Daly TM, Dumauld CM, Dotson CA, Farman MW, Kadam SK, Hockett RD: Precision profiling and components of variability analysis for Affymetrix microarray assays run in a clinical context. *J Mol Diagn* 2005, 7:404–412
20. Spiess A-N, Mueller N, Ivell R: Amplified RNA degradation in T7-amplification methods results in biased microarray hybridizations. *BMC Genomics* 2003, 4:44
21. Sastry SS, Ross BM: Nuclease activity of T7 RNA polymerase and the heterogeneity of transcription elongation complexes. *J Biol Chem* 1997, 272:8644–8652
22. Shippy R, Sendera T, Lockner R, Palaniappan C, Kayser-Kranich T, Watts G, Alsobrook J: Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* 2004, 5:61
23. Yauk CL, Berndt ML, Williams A, Douglas GR: Comprehensive comparison of six microarray technologies. *Nucleic Acids Res* 2004, 32:e124
24. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: Independence and reproducibility across microarray platforms. *Nat Methods* 2005, 2:337–344
25. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005, 2:345–350
26. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S, Zarbl H, Members of the Toxicogenomics Research Consortium: Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2005, 2:351–356
27. Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, Conley B, Buetow KH, Heiskanen M, Simon RM, Minna JD, Girard L, Misek DE, Taylor JM, Hanash S, Naoki K, Hayes DN, Ladd-Acosta C, Enkemann SA, Viale A, Giordano TJ: Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin Cancer Res* 2005, 11:565–572