Entropy-based analysis of the non-linear relationship between gene expression profiles of amplified and non-amplified RNA

JI HYE SHIN[1,2], CHAN HO PARK[3], YEON JU YANG[1,2], SANG CHUL KIM[1,2], MIN YOUNG SEO[1], SANG HWA YANG[1], SUNG BAE CHO[3], HYUN CHEOL CHUNG[1,2,4,5] and SUN YOUNG RHA[1,2,4,5]

[1]Cancer Metastasis Research Center, [2]Brain Korea 21 Project for Medical Science, [3]Department of Computer Science, Yonsei University; [4]Yonsei Cancer Center, Yonsei Cancer Research Institute, [5]Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Korea

Received July 10, 2007; Accepted September 10, 2007

Abstract. Two critical issues in microarray-based gene expression profiling with amplified RNA are its reliability and reproducibility compared to the non-amplified RNA. In this study, the non-linear relationship between the two methods was evaluated with the entropy in addition to the linear relationship using correlation coefficients. The correlation coefficients within the amplification method and between the two methods were significantly high, 0.98 and 0.88, respectively. Comparing the entropy as increasing fold-change difference (k), the average entropy value was reduced to 0.02 in the cell line and 0.09 in the tissue samples, indicating that the number of different genes between the two methods was decreased. In addition, the threshold of k according to the percentage of p estimated from entropy values could be used to provide the cut-off line on gene selection. The quantity discordance rate of 0.3-5.4% and the common outlier proportion of 84.2-94.3% between the two methods were detected, according to the expression levels. In summary, we showed a high similarity between the two methods using non-linear as well as linear comparison. Furthermore, we proved that the entropy as the measure of non-linear relationship is useful for analyzing the similarity of replicated microarray data sets.

Introduction

Recent advances in genomics have shown that in order to understand the molecular signature of cancer biology, it is essential to examine changes in the gene expression level on a genome-wide scale. A cDNA microarray is a rapid and comprehensive approach for monitoring the expression levels of thousands of genes among diverse samples simultaneously (1-3).

One of the major obstacles of a microarray is that the standard procedure requires a large amount of total RNA, typically >40 μg. However, most clinical samples such as tissues from a core needle biopsy are small in quantity. RNA amplification has been devised to overcome this limitation in various *in vitro* experiments (4-6). Several modifications have been introduced to the original protocol and have been applied to microarray research (7-9).

The most critical issues in using RNA amplification for microarray experiments are the reliability and reproducibility obtained by using amplified RNA compared with using the non-amplified total RNA. However, a large range of variation can occur because RNA amplification involves multi-step *in vitro* reactions. Zhao *et al* systematically evaluated every step of the amplification protocol (10). In order to estimate the consistency of the results, Nygaard *et al* reported a detailed statistical analysis of cDNA microarray data obtained from amplified samples compared with non-amplified samples (11). In most reports, the similarity between the two methods was estimated based on the Pearson correlation (10,12-16). Pearson correlation analysis can only quantify the linear dependencies between the measured data sets. It does not explain if the data sets are independent relations. However, information theory such as entropy can provide a general measure of the dependencies, and not just the linear dependencies (17).

This study applied the measure of entropy on gene expression data in order to further compare the similarity between the amplification and non-amplification methods.

Materials and methods

Cell line and tissue samples. The normal and tumor tissues were obtained from one colorectal cancer patient during a surgical resection (Severance Hospital, Yonsei University College of Medicine, Seoul, Korea). The tissue samples were immediately frozen in liquid nitrogen and stored at -150˚C until needed.

Twelve human cancer cell lines, A-549 (lung adeno-carcinoma), AGS (gastric adenocarcinoma), Caki-2 (kidney

Correspondence to: Dr Sun Young Rha, Department of Internal Medicine, Yonsei University College of Medicine, 134 Seodaemun-Ku, Shinchon-Dong 134, Seoul 120-752, Korea
E-mail: rha7655@yuhs.ac

Key words: entropy, non-linear relationship, cDNA microarray, RNA amplification

Table I. Experimental design and representative names.

Experimental method	Total RNA input (μg)	Experiment name		
		Cell line	Normal tissue	Tumor tissue
Non-amplification	50.0	Non-amp C	Non-amp N	Non-amp T
	4.0	1R-amp C4	1R-amp N4	1R-amp T4
One-round amplification	2.0	1R-amp C2	1R-amp N2	1R-amp T2
	1.0	1R-amp C1	1R-amp N1	1R-amp T1
Two-round amplification	0.5	2R-amp C	2R-amp N	2R-amp T

C, cell line RNA; N, normal tissue RNA; T, tumor tissue RNA; Non-amp, non-amplification; 1R-amp, one-round amplification; 2R-amp, two-round amplification; 4, 2 and 1, input amount of total RNA used in the amplification reaction.

carcinoma), HeLa (cervix adenocarcinoma), HCT 116 (colon carcinoma), HL-60 (acute promyelocyte leukemia), HT29 (colorectal adenocarcinoma), HT1080 (fibrosarcoma), MDA-MB-231 (breast adenocarcinoma), MOLT-4 (T-lymphoblast), SK-HEP-1 (liver adenocarcinoma) and U-87 MG (glioblastoma) were obtained from the American Type Culture Collection (Rockville, MD, USA). The cell lines were maintained in MEM media (Gibco, Grand Island, NY, USA) supplemented with 10% fetal bovine serum, 100 U/ml penicillin and 0.1 mg/ml streptomycin in a 5% CO_2 atmosphere at 37˚C.

Total RNA preparation. The total RNAs from the tissue samples and cultured cells were extracted using Trizol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. The extracted total RNAs from the tissue samples were purified using an RNeasy Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. The quality and quantity of the total RNA were determined using a spectrophotometer, GeneSpec III (Hitachi, Tokyo, Japan) and a Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, USA).

The Yonsei reference RNA (Cancer Metastasis Research Center, Yonsei University College of Medicine, Seoul, Korea) for the reference sample was prepared by pooling the equivalent amounts of the total RNA from the above 11 cancer cell lines except for HT-29. For the test sample of the cell line, 4 cell lines, AGS, HL-60, HT29 and HT1080, were randomly selected, and equivalent amounts of the total RNA were pooled.

RNA amplification. The amplification of the total RNA was performed based on a slight modification of a previously described protocol (12).

One-round amplification. Total RNA (4, 2, 1 and 0.5 μg) was used in the one-round amplification. The RNA template was mixed with 2 μg of the oligo-dT/T7 primer (5'-GGC CAGTGAATTGTAATACGACTCACTATAGGGAGG CGG-3', Genotech, Daejun, Korea) and denatured at 65˚C for 10 min. The first-strand cDNA was synthesized by reverse transcription with 4 μl of a 5X first-strand buffer (Invitrogen, Carlsbad, CA, USA), 2 μl of 100 mM DTT, 2 μl of 10 mM dNTP mix (Invitrogen), 2 μl of RNasin (Promega, Madison, WI, USA) and 2 μl of SuperScript II (Invitrogen) at 42˚C for 1 h. For the second-strand cDNA synthesis, 30 μl of a 5X second-strand buffer, 3 μl of 10 mM dNTP mix, 10 U of DNA ligase, 4 U of DNA polymerase I and 2 U of RNase H were added to the first-strand cDNA product and then incubated at 16˚C for 2 h. The double-stranded cDNA was extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), and precipitated with ethanol in the presence of 1 μl linear acrylamide (0.1 μg/μl, Ambion, Austin, TX, USA). The dried pellet was re-suspended in 9 μl of RNase-free water. The mRNA was transcribed from the double-stranded cDNA using a T7 Megascript Kit (Ambion). Briefly, 2 μl each of 75 mM NTP, 2 μl of an enzyme mix and 2 μl of a 10X reaction buffer were added to 8 μl of the double-stranded cDNA. The reaction mixture was then incubated at 37˚C for 5 h. The amplified mRNA was cleaned using an RNeasy Mini Kit according to the manufacturer's instructions. The quantity of the amplified RNA was measured by a spectrophotometer, and agarose gel electrophoresis and a Bioanalyzer 2100 were used to assess the integrity.

Two-round amplification. For two-round amplification, 0.5 μg of the one-round amplified RNA was diluted to a final volume of 11 μl with RNase-free water and primed with 1.5 μg of the random primers (Invitrogen). Prior to the second-strand synthesis, the cDNA was incubated with 1 μl of RNase H at 37˚C for 20 min and then heated to 95˚C for 2 min. The second-strand synthesis was then primed with 0.5 μg of an oligo-dT24/T7 primer by incubation at 65˚C for 10 min and 42˚C for 10 min. The remaining steps were identical to the one-round amplification.

cDNA microarray. The cDNA microarray was performed using a human cDNA chip (CMRC-GT, Seoul, Korea) containing 7,507 clones in a reference design. The test samples, the colorectal normal and tumor tissue RNAs and 4 mixed cell line RNAs, were labeled with Cy5 and individually co-hybridized with the Cy3-labeled Yonsei reference RNA (CMRC, Seoul, Korea). The experimental design and the representative names are listed in Table I.

Probe labeling and hybridization. The cDNA microarray experiment was performed based on the protocol of CMRC,

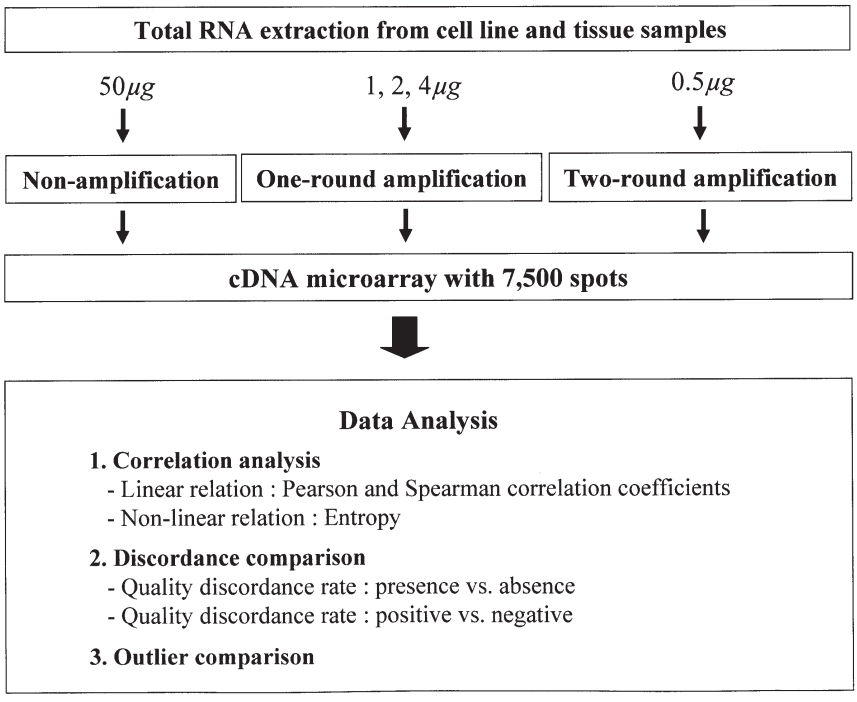

Figure 1. Scheme of the data analysis.

Yonsei University, Korea (18). Two micrograms of the amplified mRNA and 50 μg of the total RNA were labeled with Cy3- or Cy5-dUTP during reverse transcription. The RNA was mixed with 6 μg of the random primer (Invitrogen) or oligo-dT primer (Genotech), respectively, and incubated at 65˚C for 10 min. Eight microliters of the 5X first-strand buffer, 4 μl of 100 mM DTT, 2 μl of SuperScript II RT, 2 μl of 20X low-dT/dNTP mix, and 1 μl of RNasin were added to the RNA/random primer mixture and incubated at 42˚C for 2 h. The residual RNA was hydrolyzed by incubation at 65˚C for 30 min in 15 μl of a 0.1 M NaOH solution. The reaction was neutralized with 5 μl of 1.0 M HCl. The Cy3- and Cy5-labeled probes were purified using a QIAquick PCR Purification Kit (Qiagen). The purified probes were combined and mixed with 20 μg of Human Cot-1 DNA (Invitrogen), 20 μg of yeast tRNA (Invitrogen), and 20 μg of poly(A) RNA (Sigma, St. Louis, MO, USA). The final probe was concentrated to 80 μl using a Microcon YM-30 column (Millipore, Bedford, MA, USA) and then denatured at 100˚C for 2 min. The cDNA microarrays were pre-hybridized in a 3.5X sodium chloride/sodium citrate buffer (SSC), 0.1% sodium dodesyl sulfate (SDS), and 10 mg/ml bovine serum albumin (BSA) at 42˚C for 1 h prior to probe application. The probe from the amplified mRNA was hybridized in 25% formamide, 5X SSC and 0.1% SDS at 42˚C, and the probe from the total RNA was hybridized in 3.5X SSC and 0.3% SDS at 65˚C for 16 h. Following hybridization, the arrays were washed in 2X SSC with 0.1% SDS, 1X SSC with 0.1% SDS, 0.2X SSC, and 0.05X SSC, sequentially washed for 2 min each, and then spun dried at 500 x g.

Image scanning and data processing. The fluorescence signals on the microarrays were acquired using a GenePix 4000B scanner (Axon Instruments, Foster City, CA, USA). The scanned images were processed using GenePix Pro 4.0 software (Axon Instruments). Systemic errors were corrected by normalization of the $\log_2$-transformed data using an intensity dependent, within-print tip normalization based on the Lowess function (19,20). After normalization, 8096 spots in the cDNA microarray, we removed 750 genes with missing values at least one experiment and obtained 7346 genes for further analysis.

Data analysis. The scheme of the data analysis is shown in Fig. 1.

Correlation coefficients. The Pearson and Spearman correlation coefficients were measured to observe the characteristics of the relationship between the two data sets. The correlation coefficient, r, varies from –1 to +1 and the larger absolute value of r means a stronger correlation between two data sets. Supposing that X is the expression ratio of the total RNA and Y is that of the amplified RNA, the Pearson correlation coefficient is defined as follows:

$$r_{pearson} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{N})(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N})}}$$

The Spearman correlation coefficient uses the rank matrices (vectors) of x and y to measure the correlation level, which is represented as follows:

$$r_{pearson} = 1 - \frac{6\Sigma (Dx - Dy)^2}{N(N^2 - 1)}$$

where Dx and Dy are the rank matrices (vectors) of X and Y.

Entropy analysis. Entropy analysis measures the disorder of a system and is calculated by the following equation:

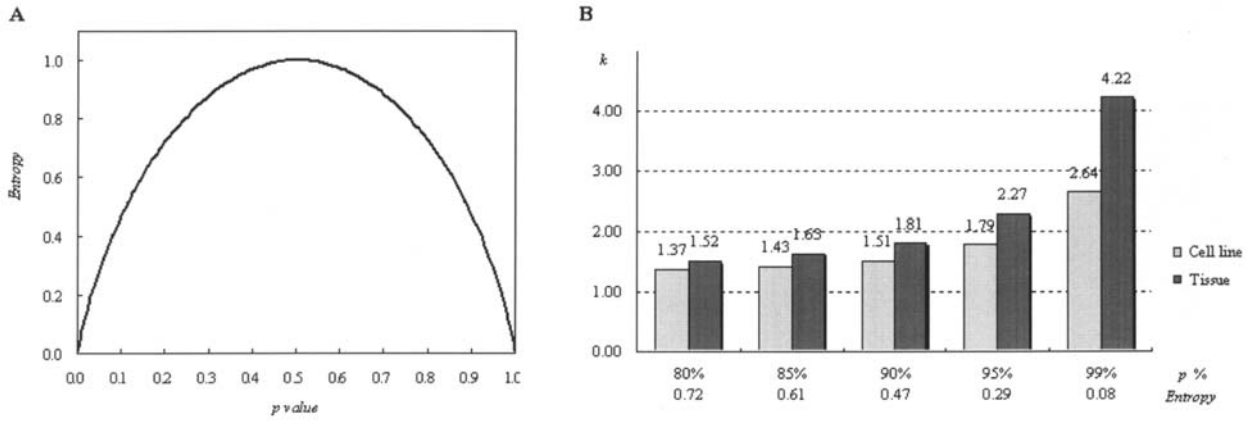

Figure 2. (A) Threshold levels of the fold-change difference according to the p value. The k value (fold-change difference) is the difference in the expression levels between the two methods. The p value is the proportion of the gene sets with a smaller change when divided by a certain k. (B) The k value according to the p percentage.

$$\text{Entropy} = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where p_i is the proportion of elements in the i-th class among c classes. In order to measure the entropy, all of the genes were divided into two groups by the fold-change difference (k), which is the difference in the expression levels between the two methods. If a gene showing a 2-fold up-regulation in the non-amplification method increased to 4-fold in the amplification method, the fold-change difference for the gene between the two methods is two, i.e. k=2. The p value is the proportion of the gene sets with a smaller change when divided by a certain fold-change difference (k). The distribution between the p value and the entropy is shown in Fig. 2A. As the p value approaches 0 or 1, the entropy approaches 0. A smaller entropy value indicates a smaller number of genes whose expression patterns are changed by less than a certain k value, meaning no difference.

Results

The efficiency and reproducibility of the RNA amplification method. In order to evaluate the efficiency and reproducibility of the RNA amplification process, cDNA microarray was performed using amplified RNA obtained from the independent amplification reactions using different amounts of the total RNA of the tissue and cell line samples. When 4, 2, 1 and 0.5 μg of the total RNA were used in the one-round amplification, there was a fold amplification (average ± SD) of $1.2\pm0.1\times10^3$, $1.9\pm0.1\times10^3$, $2.4\pm0.3\times10^3$ and $2.6\pm0.4\times10^3$, respectively, assuming that ~1% of the total RNA population was the mRNA. The two-round amplification beginning from 0.5 μg of the total RNA yielded an amplification of $1.68\pm0.02\times10^5$ fold. The efficiency of RNA amplification was proportional to the initial amount of total RNA added.

For the reproducibility within the independent microarray data sets, the Pearson and Spearman correlation coefficients were measured using the expression ratios of all the genes in the array (Table II). The average correlation coefficient (r) within the experiments using different starting amounts (1R-amp 4 vs. 1R-amp 2, 1R-amp 2 vs. 1R-amp 1, and 1R-amp 1 vs. 1R-amp 4) was 0.99 for the Pearson correlation analysis and 0.98 for the Spearman correlation analysis. The average r within the experiments according to the number of amplification reactions (1R-amp vs. 2R-amp) was 0.97 for the Pearson correlation analysis and 0.93 for the Spearman correlation analysis. These results indicate that the reproducibility of this amplification procedure was not affected by the input amount of total RNA within the range 0.5-4.0 μg.

Comparison of non-amplification and amplification methods with correlation coefficient. In order to evaluate the fidelity and reproducibility between the non-amplification and amplification methods, the Pearson and Spearman correlation coefficients of the microarray data sets obtained from the two methods were measured using the expression ratios of all the genes in the array. As shown in Table II, the average r between the two methods (Non-amp vs. Amp) in the tissues and cell line samples was 0.88 for the Pearson correlation analysis and 0.81 for the Spearman correlation analysis.

Comparison of non-amplification and amplification methods with the measure of entropy. The entropy was analyzed in order to estimate the similarity considering the non-linear relationship between the microarray data sets from the non-amplification and amplification methods. The entropy between the microarray data sets from the two methods using the cell line and tissue samples was measured at the fold-change difference (k) of 1.5, 2.0, 3.0 and 4. As k became larger, the average entropy value in all cases was reduced, indicating that the number of different genes between the two methods decreased (Table III). The average entropy in the tissues had a larger value than those in the cell line sample. Meanwhile the entropy values between the normal and tumor tissues were similar (data not shown).

A comparison of the change in the entropy according to the various amounts of RNA inputed showed that the entropy values at a certain k were similar (Table III). The entropy

Table II. Average correlation coefficients for non-amplification and amplification methods.

	Comparison	Pearson r	Spearman r
Within amplifications	1R-amp 1 vs. 1R-amp 2	0.99	0.98
	1R-amp 2 vs. 1R-amp 4	0.99	0.98
	1R-amp 4 vs. 1R-amp 1	0.98	0.97
	1R-Amp[a] vs. 2R-amp	0.97	0.93
Between non-amplification and amplification	Non-amp vs. 1R-amp 1	0.88	0.82
	Non-amp vs. 1R-amp 2	0.89	0.83
	Non-amp vs. 1R-amp 4	0.88	0.81
	Non-amp vs. 2R-amp	0.87	0.80

Expression levels of all genes in the array were used for the measurement of the correlation coefficients. The average correlation coefficients of the cell line and tissue samples in each condition are displayed. [a]Amp is the average value of 1R-amp 4, 1R-amp 2 and 1R-amp 1.

Table III. Entropy between non-amplification and amplification methods.

	Cell line				Tissue			
k	1.5	2.0	3.0	4.0	1.5	2.0	3.0	4.0
Non-amp vs. 1R-amp 1	0.52	0.21	0.05	0.03	0.73	0.37	0.16	0.09
Non-amp vs. 1R-amp 2	0.46	0.18	0.05	0.02	0.71	0.35	0.15	0.08
Non-amp vs. 1R-amp 4	0.64	0.22	0.05	0.02	0.75	0.40	0.18	0.11
Non-amp vs. 2R-amp	0.49	0.16	0.05	0.02	0.74	0.38	0.16	0.09
Non-amp vs. Amp[a]	0.53	0.19	0.05	0.02	0.73	0.38	0.16	0.09

Expression levels of all genes in the array were used for the measurement of entropy. The k value (fold-change difference) is the difference in the expression levels between the two methods. [a]Amp is the average value of 1R-amp 4, 1R-amp 2 and 1R-amp 1.

values became smaller with less variation as k increased. Considering the number of amplification reactions, the change in the entropy in the two-round amplification also was similar to the pattern of change in the entropy in the one-round amplification (Table III). These results suggest that the amplification method in the cell line or tissue samples had a high degree of similarity with the non-amplification method as the cut-off value of k increased, and was unaffected by the various amounts of RNA inputed as well as the number of amplification reactions.

Threshold levels of the fold-change difference according to the p value in entropy analysis. Fig. 2 shows a plot of the threshold levels of k, according to the percentage of p estimated from the entropy values. At the p percentage of 90%, the threshold value of k was 1.51 in the cell line sample and 1.81 in the tissue sample. This means that 90% of the genes showed a change <1.51-fold in the cell line and <1.81-fold in the tissue sample. At the same p percentage, the threshold level of k in the cell line was lower than that in the tissue. Thus, the change in the expression level between the two methods was smaller in the cell line sample than in the tissue sample.

Comparison of the non-amplification and amplification methods with concordance. In order to evaluate whether or not the non-amplification and amplification methods provided similar microarray data, the sensitivity and the concordance of the spots detected from the two methods were analyzed. First, the spot quality on microarrays from the two methods in the cell line and tissue samples was investigated. Comparing the percentage of the present and absent signals in all spots, the average present signals from the non-amplification and amplification method were 96.2 and 98.9% in the cell line and 95.9 and 98.5% in the tissue samples, respectively. These results indicate that the amplification method has more present signals than the non-amplification method, suggesting that the amplification method has a better sensitivity than that of the non-amplification method. The signal intensity of the spots present in the microarrays from the two methods was compared in order to verify that this difference in sensitivity was associated with the signal intensity. A similar total of signal intensities between the two methods was observed (data not shown), suggesting that the improved sensitivity in the amplification method was not associated with the enhanced signal intensity, but with an increased coverage of the spots.

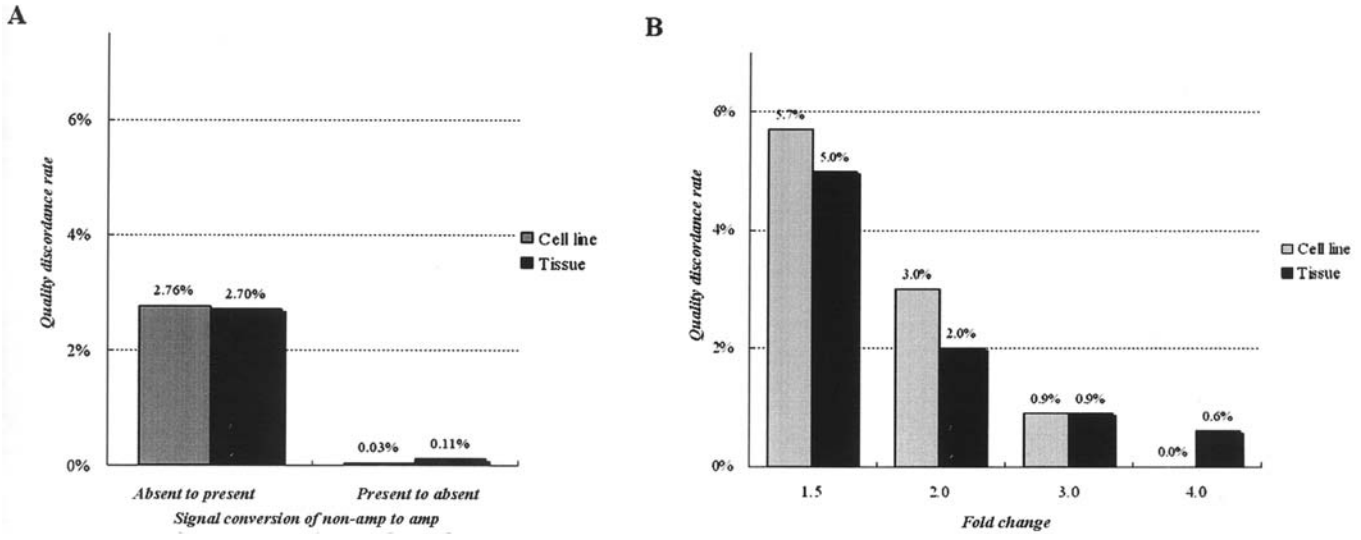

Figure 3. Discordance rates between the non-amplification and amplification methods. (A) Quality discordance rate. Data were obtained by comparing the present and absent signals of all the spots on the arrays. (B) Quantity discordance rate. Data were obtained by comparing the positive and negative signals of the spots with the common present signals in the two methods.

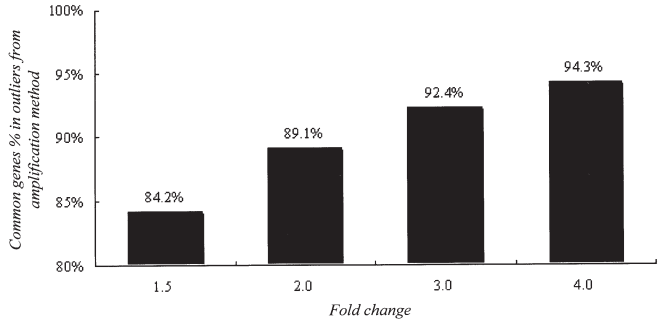

Figure 4. Comparison of the common outliers according to the change in the expression level.

The concordance of the two methods was estimated from the proportion of genes with two discordances, the 'quality discordance' as the presence and absence, and the 'quantity discordance' as the negative and positive. The quality discordance rate was measured by comparing the present and absent signals of all the spots on the arrays, providing the specific transcripts were amplified or lost during the amplification procedure. As shown in Fig. 3A, an average of 2.76% in the cell line and 2.70% in the tissue sample showed the conversion of the absent signals in the non-amplification to the present signals in the amplification. The average 0.03% in the cell line and 0.11% in the tissue sample showed the conversion of the present signals in the non-amplification to absent signals in the amplification.

In order to determine whether the gene expression patterns of the two methods were similar, the quantity discordance rate in the spots with the common present signals in the two methods was calculated at the fold-change in the expression ratio with 1.5, 2.0, 3.0 and 4.0. As shown in Fig. 3B, the quantity discordance rate at the 1.5-fold change was 5.7% in the cell line and 5.0% in the tissue samples. The discordance rate was decreased at the larger fold-change in the expression ratio. At the 4.0-fold change, the discordance rate was ~0%, suggesting that the expression pattern of the genes is preserved after the RNA amplification process.

Expression profiling using non-amplification and amplification methods. The outliers according to the change in the expression level were compared in order to investigate whether or not the two methods provide similar and reliable data in gene expression profiling. The outliers were selected using the common 7,107 spots with the present signals in the two methods according to change in the expression levels of 1.5-, 2-, 3- and 4-fold. Comparing the outliers between the two methods, an average 84.2-94.3% of the outliers in the amplification method was the same as the outliers in the non-amplification method, according to the expression levels (Fig. 4). This suggests that the gene expression patterns from the two methods are similar, even though the changes in the expression level were not exactly the same, and that the amplification methods could provide reliable data in selecting the differentially expressed genes.

Discussion

Even though RNA amplification is used more often for gene expression profiling with minute amounts of sample, there are few reports comparing the amplification with the non-amplification methods. In measuring the similarity of the two methods, all of these reports focused only on the linear relationships between the measured data sets. However, when non-linear relationships exist between the two objects of measurement, the correlation coefficients alone may not be a dependable source for comparing the two data sets. In this study, the non-linear relationship was further considered based on entropy in order to evaluate the similarity of the non-amplification and amplification methods.

First, the T7-based linear RNA amplification method was established, which generated efficient and reproducible results in agreement with previous reports. The similarity of

the amplification method for gene expression profiling was evaluated by comparing the correlation coefficients and entropy with the non-amplification method. The Pearson and Spearman correlation coefficients indicated that the correlation coefficient within the microarray data obtained from the amplification method was consistently high, confirming the high reproducibility of the amplification method. Meanwhile, the correlation between the non-amplification and amplification methods was slightly lower than that within the amplification method. It was also observed that, after considering the rank of expression ratio, the Spearman r value was lower than the Pearson r value in all samples. This suggests that slight variations occurred during the amplification method.

A comparison of the entropy showed that the average entropy in all cases decreased as k became larger, indicating that the genes differing between the two methods decreased. Some difference in entropy values between the sample types was observed at all k values. This might have been related to the difference in sample quality or the alteration in the degree of the expression ratios due to the RNA amplification process. The threshold of fold-difference change (k) according to the p percentage was measured and this data provided the cut-off line on gene selection. In conclusion, the result of the entropy analysis was similar to that of the correlation coefficient analysis, suggesting that there are slight variations based on the non-linear relationship as well as the linear relationship between the two methods.

In order to evaluate whether or not there were spots causing the poor correlation between the two methods, the quality and sensitivity of the microarray data obtained from the two methods were analyzed. For comparing the two methods, the improved coverage was observed in the amplification method. However, the total intensity of all the spots was similar. This suggests that an improved coverage is not associated with the enhanced signal intensity or the expression ratios of the spots on the array from the amplification method. The outliers identified using the amplification method were further compared with those identified using the non-amplification method. Comparing the outliers at the fold-change of 2, an average of 84.2% of the outliers identified by the amplification method were the same as the outliers identified by the non-amplification method. The percentage of the common outliers increased as the threshold of the fold-change was increased. In addition, the outliers identified only by the amplification method showed that the total intensity of the spots was either <500 or the expression ratios were close to 1.0. As most of the major microarray data analyses require outliers, which are the differentially expressed genes, the similarity in the outliers between the two methods may be sufficient for significant gene selection. Therefore, the amplification method in gene selection provides reliable results similar to the non-amplification method.

In this study, the non-linear comparison of microarray data was further evaluated by measuring the entropy between the amplification and non-amplification methods. A slight variation in the amplification method was detected from the non-linear comparison as well as linear comparison. However, the spots with those variations were not considered as significant in general gene selection. The results showed that the two platforms were quite similar in nature but were not exactly the same, as expected. Considering the selection of the differentially expressed genes, the two systems might supply common results.

Acknowledgements

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare, Korea (no. 0405-BC01-0604-0002) for Dr H.C. Chung and by a grant of the IMT-2000 project, Ministry of Health and Welfare, Korea (no. 01-PJ11-PG9-01BT00A-0028) for Dr S.Y. Rha.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR and Caligiuri MA: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537, 1999.
2. Lockhart DJ and Winzeler EA: Genomics, gene expression and DNA arrays. Nature 405: 827-836, 2000.
3. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA and Frierson HF Jr: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 61: 7388-7393, 2001.
4. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M and Coleman P: Analysis of gene expression in single live neurons. Proc Natl Acad Sci USA 89: 3010-3014, 1992.
5. Phillips J and Eberwine JH: Antisense RNA amplification: A linear amplification method for analyzing the mRNA population from single living cells. Methods 10: 283-288, 1996.
6. Luo L, Salunga RC, Guo H, Bittner A, Joy KC, Galindo JE, Xiao H, Rogers KE, Wan JS and Jackson MR: Gene expression profiles of laser-captured adjacent neuronal subtypes. Nat Med 5: 117-122, 1999.
7. Mahadevappa M and Warrington JA: A high-density probe array sample preparation method using 10- to 100-fold fewer cells. Nat Biotechnol 17: 1134-1136, 1999.
8. Wang E, Miller LD, Ohnmacht GA, Liu ET and Marincola FM: High-fidelity mRNA amplification for gene profiling. Nat Biotechnol 18: 457-459, 2000.
9. Pabon C, Modrusan Z, Ruvolo MV, Coleman IM, Daniel S, Yue H and Arnold LJ Jr: Optimized T7 amplification system for microarray analysis. Biotechniques 31: 874-879, 2001.
10. Zhao H, Hastie T, Whitfield ML, Borresen-Dale AL and Jeffrey SS: Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. BMC Genomics 3: 31, 2002.
11. Nygaard V, Loland A, Holden M, Langaas M, Rue H, Liu F, Myklebost O, Fodstad O, Hovig E and Smith-Sorensen B: Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance. BMC Genomics 4: 11, 2003.
12. Feldman AL, Costouros NG, Wang E, Qian M, Marincola FM, Alexander HR and Libutti SK: Advantages of mRNA amplification for microarray analysis. Biotechniques 33: 906-914, 2002.
13. Hu L, Wang J, Baggerly K, Wang H, Fuller GN, Hamilton SR, Coombes KR and Zhang W: Obtaining reliable information from minute amounts of RNA using cDNA microarrays. BMC Genomics 3: 16, 2002.
14. Gomes LI, Silva RL, Stolf BS, Cristo EB, Hirata R, Soares FA, Reis LF, Neves EJ and Carvalho AF: Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray. Anal Biochem 321: 244-251, 2003.
15. Li Y, Ali S, Philip PA and Sarkar FH: Direct comparison of microarray gene expression profiles between non-amplification and a modified cDNA amplification procedure applicable for needle biopsy tissues. Cancer Detect Prev 27: 405-411, 2003.
16. Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y and Mao Y: Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. J Biotechnol 107: 19-28, 2004.

17. Steuer R, Kurths J, Daub CO, Weise J and Selbig J: The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18: S231-S240, 2002.
18. Kim TM, Jeong HJ, Seo MY, Kim SC, Cho G, Park CH, Kim TS, Park KH, Chung HC and Rha SY: Determination of genes related to gastrointestinal tract origin cancer cells using a cDNA microarray. Clin Cancer Res 11: 79-86, 2005.
19. Kim BS, Lee S, Rha SY and Chung HC: cDNA microarray experiment: Design issues in early stage and the need of normalization. Cancer Res Treat 35: 533-540, 2004.
20. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15, 2002.