Journal of Plant Biology, December 2006, 49(6) : 498-506

Comparison between SAGE and cDNA Microarray for Quantitative Accuracy in Transcript Profiling Analyses

Eun Mi Eom[1§], Ji Yeon Lee[1§], Hye Sang Park[1], Youn Jung Byun[1], Young Mie Ha-Lee[2], and Dong Hee Lee[1*]

[1]Department of Life Science, Ewha Womans University, Seoul 120-750, Korea
[2]Research Institute for Basic Sciences, Yonsei University, Wonju 220-710, Korea

Array-based hybridization and the serial analysis of gene expression (SAGE) are the most common approaches for high-throughput transcript analysis. Each has advantages and disadvantages. The cDNA array allows rapid screening of a large number of samples but cannot detect unknown genes. In contrast, SAGE can detect those unknown genes or transcripts but is restricted to fewer samples. Combining these two methods could provide better high-throughput analysis that allows rapid screening of both previously known and unknown genes. For this, we have generated two cDNA microarrays (from human and plant systems) based on SAGE data. The results from both of these were analyzed for their correlation and accuracy. One specialized cDNA microarray, putatively named Gastricchip, was constructed with 1744 probes, including 858 cDNA fragments based on SAGE data from gastric-cancer tissues. The other microarray, putatively named Coldstresschip, was constructed with 1482 probes, including 1209 cDNA fragments based on SAGE data from cold-stressed *Arabidopsis*. The hybridizations for these microarrays with relatively small sized and mostly low-level expressed gene probes were evaluated by four different labeling methods. Using primarily for these customized microarrays, the Genisphere 3DNA SubmicroEX protocol, an indirect labeling technique, produced the lowest background but the highest signal recovery, with a 1.4 S/B cut-off and high reproducibility (R=0.89-0.95). These cDNA microarray data were closely correlated with the SAGE data (R=0.47-0.56), especially for genes with higher expression levels (R=0.66-0.70), demonstrating that results from SAGE and a cDNA microarray are comparable and that combinatorial approach provides more efficient and accurate gene-expression patterns. In particular, identity of the genes on both sets of data is assured and hybridization for cDNA microarray is efficient.

Keywords: gene-expression profile, microarray, SAGE, specialized cDNA microarray

Although many techniques can be used for high-throughput transcript analysis, the most common are array-based hybridization and SAGE. DNA microarray technologies (cDNA- or oligonucleotide-based) permit the systematic evaluation of quantitative transcription profiles. The microarray is an excellent method for rapidly screening large numbers of samples and genes. However, although such technology allows for extensive analysis of expression patterns for many genes, it can examine only the sequences that have already been identified. In contrast, SAGE does not require prior knowledge, and represents an unbiased, comprehensive representation of those transcripts (Velculescu et al., 1995; Zhang et al., 1997). Furthermore, SAGE can quantitatively identify low-abundance transcripts and detect relatively small differences in their expression. SAGE, however, is capable of analyzing just a limited number of expression profiles at one time due to the technical difficulties and the requirement for sufficient RNA samples.

One approach to circumventing their disadvantages is to combine these two techniques. First, SAGE is used to identify unbiased sequences implicated in a certain process, and then microarrays are implemented to rapidly verify these expression patterns in a large number of samples. This combinatorial method has been applied to obtaining the expression profiles for human cancers. For example, Nacht et al. (1999) have reported that a subset of the differentially expressed genes identified by SAGE can then be spotted on an array and screened with breast tumors and normal breast epithelial cells. Other research has focused on using the 'Colonchip' for colorectal carcinoma (Takemasa et al., 2001) and the "Ovachip" for epithelial ovarian cancer (Sawiris et al., 2002). These specialized cDNA microarrays were constructed and utilized by selecting genes that are not redundant but are preferentially expressed in specific situations. In fact, one advantage of preparing such specialized arrays is that they do not include irrelevant genes that could contribute as noise during the data analysis.

Because of its high-throughput nature, cDNA microarray technology is vulnerable to systematic variations introduced during experimental procedures (Kerr et al., 2000; Finkelstein et al., 2002). Although a number of statistical algorithms have been developed to normalize microarray data and to control experimental variations (Tseng et al., 2001), high-quality input images are still the prerequisite for obtaining significant new output. This requires reproducible processes, e.g., labeling of cDNA targets, hybridization, and washing of slides, to retain consistently high intensity over low-background images. Among them, the labeling method is a significant factor. Several protocols, including direct and indirect labeling of cDNA targets, have been utilized (Hegde et al., 2000). The target cDNA labeling method preferentially used is a direct process that incorporates fluorescence-modified nucleotides during target cDNA synthesis. This method, however, requires large amounts of starting RNA, up to 100 μg, which makes it suitable only when the quantity of starting material is not limited.

§Equally contributed.
*Corresponding author; fax +82-2-3277-2385
e-mail lee@ewha.ac.kr

In this study, we prepared two specialized microarrays based on SAGE information, one for human gastric cancer and the other for cold-stressed *Arabidopsis*. Our objective was to quantitatively compare the accuracy of transcript profiling obtained from cDNA microarray with that gained via SAGE. In addition, we wanted to solve potential problems associated with falsely matching gene IDs in SAGE to the ones in microarrays when constructs are built independently. To remedy this, we created a cDNA microarray with probes prepared by the GLGI method (i.e., Generation of Longer cDNA fragments from SAGE tags for Gene Identification).

MATERIALS AND METHODS

Preparation of cDNA Probes for Human Gastric Cancer Research

Four separate sources were used to prepare the cDNA fragments for the Gastricchip (for gastric cancer in humans). The first one was 152 GLGI clones from SAGE tags previously obtained in this laboratory (Lee et al., 2003), for which PCR was performed with a pair of primers: 5′-GCCAGGGTTTTCCCAGTCACGA-3′ and 5′-ACAGGAAACAGCTATGACCATG-3′. The reaction contained 5 μL 10X PCR buffer, 0.2 mM of each dNTP, 5 pmole of each primer, and 1 unit of *Tag* DNA polymerase (Bio-Line) in a 50 μL reaction volume. Conditions included an initial 94°C for 2 min; followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min; then a final extension at 72°C for 5 min. The second experimental source was 708 GLGI products prepared directly, without the cloning step. When multiple GLGI PCR products were generated from a SAGE tag, each DNA band separated on the agarose gel was picked and re-amplified by second-round PCR, which entailed the same conditions as in the first round. When the PCR products were too small (<100 bp), longer PCR fragments, i.e., with additional ~200-bp-long stuffer DNA fragments were generated by ligation PCR with three primers. The third source was 15 Korean Unigene clones. For the fourth source, the following DNAs were used as controls: a pZErO-2 (Invitrogen, USA) plasmid to assess non-specific hybridization to the cloning vector; 9 *Arabidopsis* cDNAs to monitor non-specific background hybridizations; 8 housekeeping genes, and 7 known gastric cancer-related genes.

Preparation of cDNA Probes for *Arabidopsis* Cold-Stress Research

The cDNA probes used in preparing the Coldstresschip for *Arabidopsis* were genes selected from two SAGE data sets developed in this laboratory (Jung et al., 2003; Lee and Lee, 2003). The first set included 1089 genes that were up-regulated or down-regulated by at least 6-fold, as well as 120 genes that were differentially expressed at slightly <6-fold but with $p < 0.01$. Secondly, 88 induced genes from cold-treated leaves identified by SSH were also included. Thirdly, 20 known cold-induced genes, 18 cold-related transcription factor, and 107 pathogen-related genes were added. Finally, the following DNAs served as controls: the pZero-2 plasmid to assess non-specific hybridization to the cloning vector; 25S rDNA, 18S rDNA, and 7 housekeeping genes as internal quantification standards; 10 human cDNAs to monitor non-specific background hybridizations; and an 18-microarray control set assembled to provide access to a uniform set of clones for use in our plant-DNA microarray experiments. To prepare the cDNA fragments, four different sources were used. First, the inserts in 170 GLGI clones and 88 SSH clones were amplified by PCR as described above. Second, 1011 GLGI products were used directly, without the cloning step. When multiple GLGI products were generated from a SAGE tag, a fraction of the DNA fragments on the specific bands that had separated on the agarose gel were picked and used as template in second-round PCR. Conditions and primers were the same for each round. When the PCR products were too small (<100 bp), new upstream, gene-specific primers were designed by searching the sequence information that matched the corresponding SAGE tag, and were then used for amplification of the larger products. Third, we PCR-amplified 28 clones (their selection based on our SAGE tag information) plus 107 known pathogen-related genes purchased from the *Arabidopsis* EST collection (ABRC at Ohio State University), and a 20-microarray control-gene set bought from The European Arabidopsis Stock Centre (NASC) at the University of Nottingham. For the fourth source, we obtained 20 well-known cold-induced genes and 18 cold-related transcription factor genes by PCR, with cDNA as template.

Fabrication of cDNA Microarrays

All the amplified DNA fragments were purified using Sephadex, according to the NSF protocol (http://soybeangenomics.cropsci.uiuc.edu/protocols/). These were analyzed before arraying for quality control on an agarose gel. DNA microarrays were printed by GenomicTree (Korea). Briefly, the arrays were produced using an OmniGrid™ Microarrayer (GeneMachines, USA) with stealth pins (TeleChem, USA) that withdraw a volume of about 250 nL and deposit a spot volume of about 1 nL, with a diameter of approximately 130 μm. Printing was done with a silanized glass slide (CMT-GAPS™, USA). Each slide was crosslinked via 300 mJ of shortwave UV irradiation (Stratalinker; Stratagene, USA) and stored in a desiccator.

RNA Isolation

Total RNA from human gastric tissue was prepared using Tri-reagent according to the manufacturer's instructions. For *Arabidopsis*, frozen leaf samples were homogenized in the presence of liquid nitrogen. Total RNA was isolated using the RNeasy Plant Mini Kit (Qiagen, Germany), following the manufacturer's protocol.

Direct Probe Labeling and Microarray Hybridization

Total RNA was labeled by direct incorporation of Cy3- or Cy5-conjugated deoxy UTP (Perkin Elmer Life Sciences) during cDNA synthesis. The overall hybridization was performed according to the lab protocol of Brown (http://cmgm.stanford.edu/pbrown). Briefly, 100 μg each of total RNA from normal and cancerous gastric tissues was mixed

with 4 μg oligo-dT primers (5′-TTTTTTTTTTTTTTTTTTT(A/C/G)(A/C/G/T)-3′) in 15.4 μL of water, then denatured at 65°C for 10 min. To this, the remaining components were added to obtain the following reaction mixture, in a total volume of 30 μL: 1X Superscript II reverse transcriptase buffer (Life Technologies, UK); 0.01 M DTT; 0.5 mM each of dATP, dCTP, and dGTP; 0.2 mM dTTT; 3 nmol of either Cy3-dUTP or Cy5-dUTP; and 2 μL of Superscript II reverse transcriptase. After incubation at 42°C for 2 h, the unincorporated nucleotides were removed using QIAquick columns (Qiagen), and the reaction products from two samples (one with Cy3-labeling, the other with Cy5-labeling) were combined. Afterward, 20 μL of 1 μg μL^{-1} human Cot-1 DNA (Life Technologies), 2 μL of 10 μg μL^{-1} poly A^+ RNA (Sigma), and 2 μL of 10 μg μL^{-1} yeast tRNA (Life Technologies) were added, and the samples were placed in a Microcon-30 filter (Millipore, USA). Following centrifugation at 14000g for 12 min, the labeled cDNAs were collected by placing the sample reservoir upside down in a new collection tube and spinning it for 30 s. Afterward, 10 μL of 20X SSC, 20 μL of formamide, and 2 μL of 2% SSC were added to the 8 μL of labeled cDNAs. The samples were denatured by placing them in a 100°C water bath for 3 min, then centrifuged at 14000g for 2 min, and hybridized in the GT-Hyb-chamber II™ (GenomicTree, Korea). After the hybridization chamber was submerged in a 42°C water bath for 16 h, the slides were submersed in a pre-warmed (42°C) washing solution (1X SSC, 0.2% SDS) that was stirred gently for 4 min. The racks were then carefully washed in a second solution (0.1X SSC, 0.2% SDS) for 4 min, with gentle stirring. They were transferred to a third solution (0.1X SSC) and washed twice, for 2.5 min each. Following this last washing, the slides were immediately dried by centrifugation (5 min at 600 rpm) and signal intensities were measured by scanner.

Indirect Probe-Labeling and Microarray Hybridization

A limited quantity of total RNA from tissue was treated by an indirect high-density labeling method. Here, three kits -- 3DNA Array 50™ Expression Array Detection Kit, 3DNA Array 350RP™ Expression Array Detection Kit, and 3DNA Submicro™ Expression Array Detection Kit -- were utilized for target-labeling according to the protocol of the manufacturer (Genisphere, USA). Briefly, 25, 2, and 5 mg of total RNA for the Array 50™, Array 350RP™, and Submicro™ kits, respectively, were reverse-transcribed using reverse transcription (RT) primers tagged with either the Cy3- or Cy5-specific 3DNA capture sequence. The synthesized tagged cDNAs were then fluorescent-labeled by Cy3-3DNA or Cy5-3DNA, based on the complementary capture sequence obtained with the 3DNA capture reagents. The tagged cDNA was hybridized to a microarray in a 1X formamide-based hybridization buffer (25% formamide, 4X SSC buffer, 0.5% SSC, and 2X Denhardt's solution) at 45°C for 16 h. Afterward, the slides were washed with 2X SSC, 0.2% SDS at 50°C for 10 min; then 2X SSC at RT for 10 min; and 0.2X SSC for 10 min. After the slides were washed for 2 min at RT in 95% ethanol, a second hybridization was conducted with the 3DNA capture reagent at 45°C for 6 h in the GT-Hyb-chamber II™. This was followed by serial washing with the three previously described solutions and drying via centrifugation.

Data Acquisition and Analysis of cDNA Microarrays

After washing, the slides were immediately scanned using an ArrayWoRx (Applied Precision, USA). To maximize the camera's dynamic range without saturation, and to normalize the two channels with respect to signal intensity, the exposure setting was adjusted so that the intensity level of the brightest spot on a slide was 40 to 45%. Intensity values were quantified from the resultant pairs of TIFF files using ImaGene image analysis software, and were analyzed with the GeneSight software package (BioDiscovery, USA). The data were imported into Microsoft Excel spreadsheets for further analysis. To test the effect from low-intensity signals, we set the cutoff for significant S/B ratios at 1.4. After the direct and indirect methods were compared, spots with S/B ratios of <1.4 in both channels were excluded from further examination. Analyses were performed using mean signal-intensity values for each spot. For each slide, the local background was subtracted from the intensity, and the minimum intensity was raised to 20 by using a "floor" function. The mean intensity for each element was normalized by LOWESS methods (Yang et al., 2001).

RESULTS

Preparation of cDNAs and Construction of the Specialized cDNA Microarrays

We used two independent sets of SAGE data to construct our cDNA microarrays. One was from human gastric cancer research previously conducted in our laboratory. This SAGE analysis had revealed 858 candidate genes that were differentially expressed by >6-fold in cancerous tissue. cDNA fragments corresponding to these genes were obtained by the GLGI PCR method (Chen et al., 2000) using four different RNA samples as PCR template, depending on where

Table 1. Probes used for the specialized microarray from human gastric-cancer tissues.

Category		No. of cDNAs
I. SAGE clones		
	Induced genes[a]	517
	Repressed genes	341
II. SSH clones		
	Induced genes	860
III. Controls		
	House-keeping genes	8
	Well-known-induced genes	7
	Arabidopsis cDNA	9
	pZero-2 vector	1
	Stuffer	1
	Total	1744

[a]These selected genes were either induced or repressed in Patient 1 or 2 with advanced gastric cancer. If a reverse expression pattern was found in the cancerous tissues of Patients 1 and 2, this tag was included as 'induced gene'.

Table 2. Probes used for the specialized microarray from cold-stressed *Arabidopsis*.

Category		No. of cDNAs
I. SAGE clones		
72-h cold treatment		
Cold- induced genes		
	in leaf	354
	in pollen	153
Cold-repressed		
	in leaf	277
	in pollen	73
1-h cold treatment		
Cold- induced genes		
	in leaf	254
Cold-repressed		
	in leaf	98
II. SSH clones		
Cold-induced genes		88
III. Controls		
Well-known cold-induced genes		20
Well-known cold-related transcription factor		18
Well-known pathogen-related genes		107
Housekeeping genes		7
Human cDNAs		10
Microarray control set		20
pZero-2 vector		1
18S and 25S rDNA		2
Total		1482

they were primarily expressed. From these, 152 were cloned and sequenced for identification of corresponding tags. Others were directly used for preparing the microarrays without cloning or sequence confirmation. The quality of all PCR products was doubly verified by gel electrophoresis, before and after purification. The resulting microarray, with 1744 transcripts (Table 1), was spotted in duplicate on a single slide. It was arranged in 16 subgrids, each of which had 11 rows and columns, including blank spots. Initial hybridization (data not shown) revealed that the quality of individual spots and the consistency between duplicate spots were excellent.

The cDNA probes for the specialized microarray for cold-stressed *Arabidopsis* were constructed with two datasets from our previous research. The resulting microarray, comprising 1482 transcripts spotted in duplicate, included 1209 GLGI products (Table 2) that were arranged in 16, 10 by 10 subgrids.

To assess the size distribution (Fig. 1) for the GLGI products from these two microarrays, we subtracted the average length of the poly A tail sequence from the size of the actual PCR products. Based on the sequences of the 152 cancer clones, we calculated the average size of that poly A tail to be 32.7 bp. Overall, our GLGI-amplified cDNAs ranged from 10 to 600 bp, much shorter than those of the typical cDNA microarray.

Optimization of Target-Labeling and Hybridization Methods

To obtain consistently low backgrounds and high-intensity signals from only small amounts of starting RNA, we tested

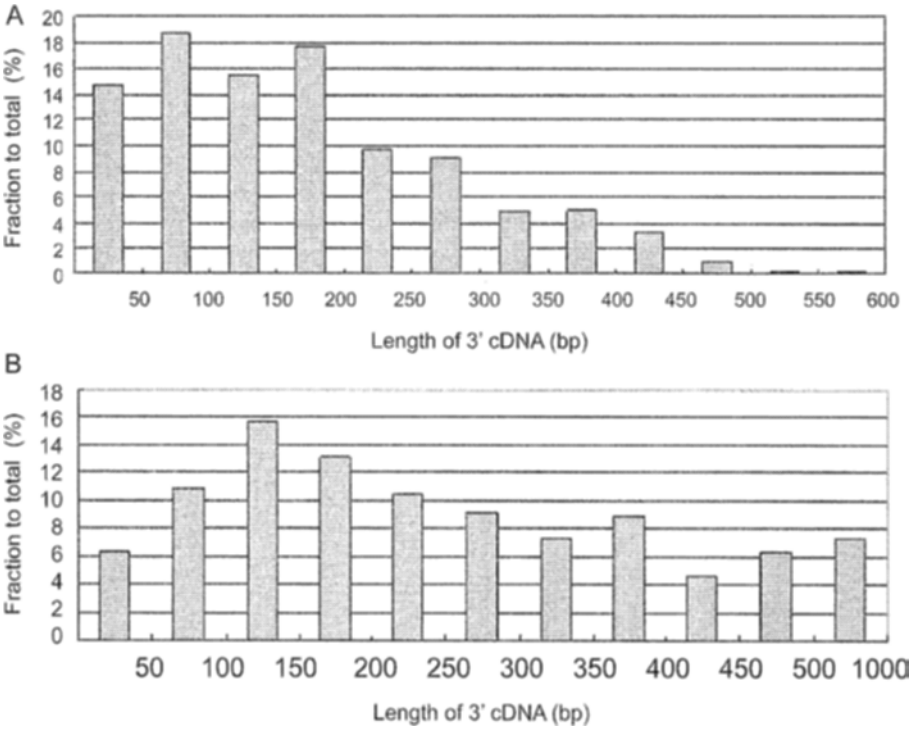

Figure 1. Size distribution of GLGI products obtained from SAGE tag. (**A**) Gastricchip. In all, 659 GLGI PCR products from human gastric tissue were analyzed by agarose gel electrophoresis and sizes were estimated. To determine transcript specific size, the average poly (A) size was subtracted from apparent PCR product sizes. Average size of GLGI-amplified cDNA was 180 bp. (**B**) Coldstresschip. Lengths of 556 GLGI PCR products were estimated by gel electrophoresis; average poly (A) size observed in 170 GLGI clones was subtracted from that size. GLGI-amplified cDNAs ranged from 6 to 900 bp.

Table 3. Comparison of hybridization signals obtained with four separate target cDNA labeling methods.
A. Signal intensity.

		350RP[a]		Array50[a]		SubmicroEX[a]		direct-1[b]	
		Cy3	Cy5	Cy3	Cy5	Cy3	Cy5	Cy3	Cy5
Total signal (S)	Average	515.4	892.8	1780.6	2474.9	1382.3	2037.5	2449.8	2349.0
	STD	1170.4	1611.3	5212.3	2576.3	2966.2	3084.1	3854.1	2745.8
	Max	14260.8	25977.4	54732.1	29753.2	28548.8	27830.6	53122.9	37617.2
	Min	88.9	128.2	244.4	1212.5	93.9	126.9	908.1	762.5
	Median (50%)	160.2	326.2	430.3	1514.7	351.6	664.4	1233.4	1225.0
BG (B)	Average	98.1	143.0	292.8	1253.2	122.2	151.3	1018.8	796.9
Net signal (S-B)[c]	Average	417.3	749.8	1487.8	1221.7	1260.2	1886.2	1431.1	1552.1
	STD	1165.0	1605.9	5040.8	2567.3	2936.6	3065.9	3815.3	2736.3
	Max	14121.2	25756.5	54269.9	28447.4	28303.5	27582.0	51953.9	36748.7
	Min	-6.1	-6.3	-19.3	-10.1	-8.0	-6.9	-52.9	-7.4
	Median (50%)	62.5	185.0	145.4	263.8	236.0	517.4	217.1	426.7
No. of probes with S/B>1.4		1194	1545	981	695	1508	1706	736	1008

B. Correlation coefficients for signal-intensity ratios from four labeling methods.

	350RP	Array50	SubmicroEX	direct-1	direct-2	direct-3
350RP	1.00					
Array50	0.43	1.00				
SubmicroEX	0.54	0.78	1.00			
direct-1[b]	0.42	0.72	0.80	1.00		
direct-2[b]	0.43	0.78	0.81	0.91	1.00	
direct-3[b]	0.38	0.68	0.72	0.87	0.85	1.00

[a]The labeling of total RNA was performed using an indirect high-density labeling method. 3DNA Array 50™ Expression Array Detection Kit (Array50), 3DNA Array 350RP™ Expression Array Detection Kit (350RP), and 3DNA Submicro™ Oligo Expression Array Detection Kit (SubmicroEX) was utilized for target labeling. [b]The labeling of total RNA was performed with fluorescently modified dNTP, using a direct-labeling method. Direct-1 and direct-3 was used for Cy3-labeled gastric normal RNA and Cy5-labeled gastric cancer tissues. Direct-2 was dye swap experiments.

four different labeling techniques, including one direct and three indirect 3DNA methods. The hybridization results were then compared. Optimization of target-labeling and hybridization was especially crucial because the microarrays constructed in this study contained relatively small-sized probes and mostly genes expressed at low levels.

The maximum signal intensities and the calculated standard deviations were the same for both the indirectly labeled probe (SubmicroEX) and the directly labeled probes (Table 3A), but relatively low backgrounds were produced. In addition, the indirect method resulted in a large proportion of spot intensities -- 86.3 and 97.6% for the Cy3- and Cy5-labeled probes, respectively -- that had S/B ratios of >1.4. In contrast, the direct method produced approximately 42.1% (Cy3-) and 57.7% (Cy5-) of the intensities with ratios larger than 1.4.

For each labeling method, the signal intensity calculated over the local background (S/B) was further examined by comparing correlation values (Table 3B). The data produced from the Array 50, SubmicroEX, and direct-1 methods were highly correlated with each other while that obtained from the 350RP™ kit was less so. This difference might be attributed to the fact that 350RP™ uses random RT primers and total RNA for cDNA synthesis, while the other three methods utilize an oligo dT primer for their RT reactions. Those random primers can generate labeled products when non-poly A RNA is abundant in the total RNA, thereby resulting in high-background signals for the selected probes.

Among the three labeling methods with high correlation coefficients, SubmicroEX had an additional advantage, in that only a small amount of RNA (2 to 5 μg) was required in order to generate sufficiently reproducible results. We demonstrated that our hybridization data from using 2 μg of total RNA with SubmicroEX were highly correlated with those obtained when we instead used 5 μg (data not shown), indicating that the smaller quantity was adequate to generate target cDNA. Moreover, the SubmicroEX method gave rise to low background and, thus, generated the highest number of spots with S/B ratios >1.4. Therefore, because this particular indirect-labeling method produced superior and consistent hybridization results, we utilized it for our subsequent experiments.

Evaluating Reproducibility

Each slide was hybridized with fluorescent cDNA targets prepared with total RNA from normal and cancerous tissues. Four independent hybridization experiments were conducted both to minimize the inherent variability of the microarray assay (Lee et al., 2000) and to ensure the reliabil-

Table 4. Correlation coefficient for four replicates of experiments to compare normal versus cancerous gastric tissues[a].

Replicate	N-Cy3 versus C-Cy5 (1-A)	N-Cy5 versus C-Cy3 (1-B)	N-Cy3 versus C-Cy5 (2-A)	N-Cy5 versus C-Cy3 (2-B)
N-Cy3 versus C-Cy5 (1-A)[b]	-			
N-Cy5 versus C-Cy3 (1-B)	0.902	-		
N-Cy3 versus C-Cy5 (2-A)	0.918	0.887	-	
N-Cy5 versus C-Cy3 (2-B)	0.951	0.925	0.927	-

[a]N, normal; C, gastric cancer; Cy3, Cy3-3DNA; Cy5, Cy5-3DNA.
[b]Numbers 1 and 2 in the parentheses represent replicate 1 and replicate 2; A and B indicate the dye-swap experiments.

ity of those microarray results. One pair of slides (Table 4: 1-A, 2-A) was probed with Cy3-labeled cDNAs from normal tissues and with Cy5-cDNA from cancerous tissues. The other pair was probed with reverse-labeled probes (Table 4: 1-B, 2-B) to overcome any artifacts caused by dye-related differences.

The correlation coefficients (0.887 to 9.51) from all four replicates were highly reproducible (Table 4). Hybridization in the dye-swap experiment was more variable than between replicates. When the effects of dye-swap and treatment were combined, the correlation coefficients were 0.90 and 0.89, respectively, again indicating highly reproducible results.

Reproducibility within single slides was assessed by comparing their *cv*, i.e., correlation variance. Each value was calculated by dividing the standard deviation by the mean of the repetitive spot intensities. This specialized cDNA microarray contained 16 replicates of 4 genes -- GAPD, β-actin, α-tubulin, and RPL29 -- and the *cv* for one slide ranged from 0.16 to 0.24. The average *cv* for all spots from the four replications was 0.34. Therefore, the lower values suggested that reproducibility was much better within than between slides. In fact, the majority of spots with signal intensities of 100 to 40,000 were generally reproducible, while genes with expression levels of <100 (comprising approximately 5% of all the genes) were somewhat variable. These less-reproducible spots were omitted from analyses that considered S/B ratios to avoid the possibility that genes with very low expression levels had biased and large ratio values.

As with the data produced in our assessment of a human system, similar results were obtained when we investigated the optimal procedures of hybridization for our *Arabidopsis* cDNA microarrays. This suggests that these defined hybridization protocols might be suitably applied to many other types of systems.

Comparison of Microarray and SAGE Gene Expression Data from a Human System

We compared the quantitative accuracy in transcript profiling, using expression data obtained from both cDNA microarrays and SAGE analysis. Here, we selected 569 genes that originated from the SAGE tag library and which produced signal intensities with S/B ratios >1.4. A logarithmic-scale scatter graph illustrated the intensity values from the microarray and the number of spots from SAGE (Fig. 2A). This graph showed that the results from our two analytical methods were very similar in terms of their absolute analyses, and that the correlation was closer when we examined genes with high expression levels. Microarray intensity scores were generally one or two orders of magnitude higher than found with the SAGE frequencies.

The scatter graph also revealed differences between SAGE and intensity ratios from our microarray experiments (Fig. 2B), which were shown by a correlation coefficient of 0.558 for 569 probes. Generally, no great variations in results were found between these two methods, such that genes with manifold differences maintained higher correlations between them while those with lower-fold differences showed relatively low correlations.

Comparison of Microarray and SAGE Gene Expression Data from a Plant System

Expression data were evaluated from cDNA microarrays and SAGE analyses of *Arabidopsis* specimens. Intensity values from our microarrays were directly compared with the tag abundance from SAGE (Fig. 3A). Scores from the former were generally one or two orders of magnitude higher than from the SAGE frequencies. Intensity values for 1186 genes that were common to both experimental systems were calculated for cold-treated and untreated leaves. The abundance of corresponding tags in SAGE was presented as a logarithmic-scale scatter plot. Although some genes were classified as being up- or down-regulated via SAGE analysis but not via our microarrays, the majority showed patterns of induction or repression that were similar between these two methods.

To investigate the correlation between fold-differences in SAGE and intensity ratios in the microarray, we compared the fold ratios of 1186 probes that originated from the SAGE tag library and gave rise to signal intensities with S/B ratios >1.4. The correlation coefficient between our two methods was 0.47 (Fig. 3B). Most of the discord seemed to come from low-intensity signals close to background and/or from small fold-differences, as had also been found with our human system. Therefore, we examined 339 genes with >2-fold differential expressions in both the cDNA microarray and SAGE analysis, and determined that the correlation coefficient for those genes was 0.66 between the two methods (Fig. 3C). Subsequently, we selected the most highly induced and most highly repressed genes for comparison (Fig. 3D), and found, in general, no great fold-differences between methods. Although the correlation coefficient was 0.70, those genes with high fold-differences maintained the same inclination between them. However, the remaining genes, with lower fold-differences, showed less correlation.

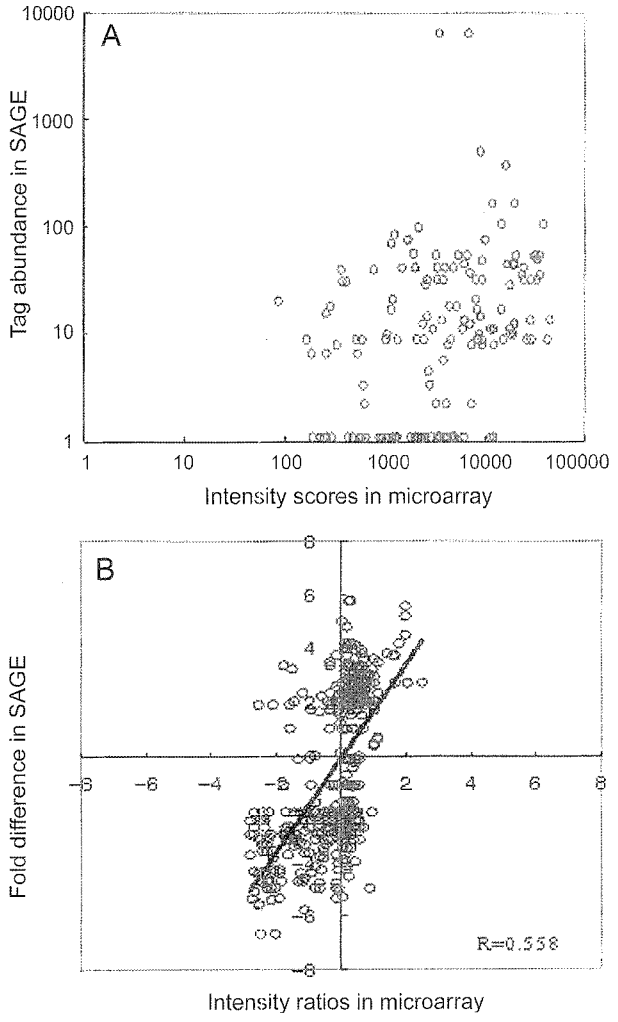

Figure 2. Gene expression data from microarray analysis and SAGE using human tissues. (**A**) Comparison of intensity scores in microarray and tag abundance in SAGE, as plotted in logarithmic scale on abscissa and ordinate, respectively. (**B**) Comparison of intensity ratios in microarray with fold-differences in SAGE. Each ratio and fold-difference was log-transformed and compared. Maximum correlation coefficient was 0.56 between microarray and SAGE methods.

The relatively poor correlation in the latter might have been attributed to variations in the baselines for each method.

DISCUSSION

In this study, we have constructed and evaluated the quality of two specialized cDNA microarrays based on SAGE data: one from research on human gastric cancer, the other concerning cold stress in *Arabidopsis*. With conditions optimized for the preparation of cDNA targets and data processing, this system gave high-quality, reproducible results in detecting low-abundance transcripts.

Several high-throughput techniques are used to monitor gene expression. Although a cDNA microarray is highly efficient for screening, it is limited to analyzing only previously identified genes. Therefore, by combining cDNA microarrays and SAGE, in series, we could prepare specialized microarrays that included sequences previously not implicated or identified. These arrays were capable of rapidly verifying expression patterns in a large number of samples. In this study, most of our probes were directly generated from target RNA, using GLGI-PCR that was based on SAGE tag sequences. With optimized target cDNA synthesis, we showed that these specialized microarrays were highly reproducible and that their results were strongly correlated with our SAGE data. These outcomes were unlike those previously reported, which relied on SAGE data for selecting the genes that could be adequately evaluated with those microarrays (Yang et al., 1999; Takemasa et al., 2001; Sawiris et al., 2002).

One problem in applying this combinatorial array technique is the preparation of probes. Two array systems are currently popular, i.e., the long cDNA microarray and *in situ* synthesis of oligonucleotide microarrays. The latter are reportedly more reliable for global screening (Li et al., 2002; Ramakrishnan et al., 2002), and their results are related to the specificity of the probe. For example, 30-mer probes can distinguish up to 90% sequence identity, whereas longer cDNAs have a lower sequence identification of no more than 80%. Despite this benefit, however, the wide use of a system such as Affymatrix would be limited because of its high cost.

As a compromise, more recent microarrays have been constructed with probes of 60- to 70-mer oligonucleotides (Xu et al., 2002). These slightly longer probes retain the ability to efficiently distinguish sequences, but with less expense. The GLGI PCR-generated cDNA fragments, based on SAGE tag sequences, are 50 to 300 nt long, which is more than those oligonucleotide probes but shorter than most cDNA probes, thereby providing better sequence identification. Furthermore, most of the SAGE GLGI-PCR products belong to untranslated regions (UTRs), in which unique sequences are more common than in the coding regions. Thus, specialized cDNAs prepared by GLGI-PCR that are based on SAGE tag sequences may possibly provide higher specificity than from a typical cDNA microarray.

We tested several labeling methods to obtain high-quality, reliable signals from microarrays with relatively small probes, most of which had originated from genes with primarily low expression. Although the direct-labeling method generated probes of good quality, the indirect method, using a dendrimer, gave better results. For example, using the SubmicroEX system, we obtained a high-intensity signal while maintaining a background that was as much as 10-fold less than that found with the direct method. Furthermore, because each dendrimer had a predetermined and quantified fluorescence intensity, and because each cDNA transcript was bound to a single dendrimer, the amount of signal generated was directly proportional to the number of cDNA molecules detected. Therefore, this eliminated the former problem of variable incorporation of the modified dNTP during the RT reaction, which prevents direct calculations of bound cDNA molecules.

The signal-to-background ratio quantifies how well one can resolve a true signal from the background of the system. Here, we included only data with S/B ratios >1.4 in order to remove the false signal from low-hybridization spots. Using the indirect SubmicroEX labeling method, over 90%

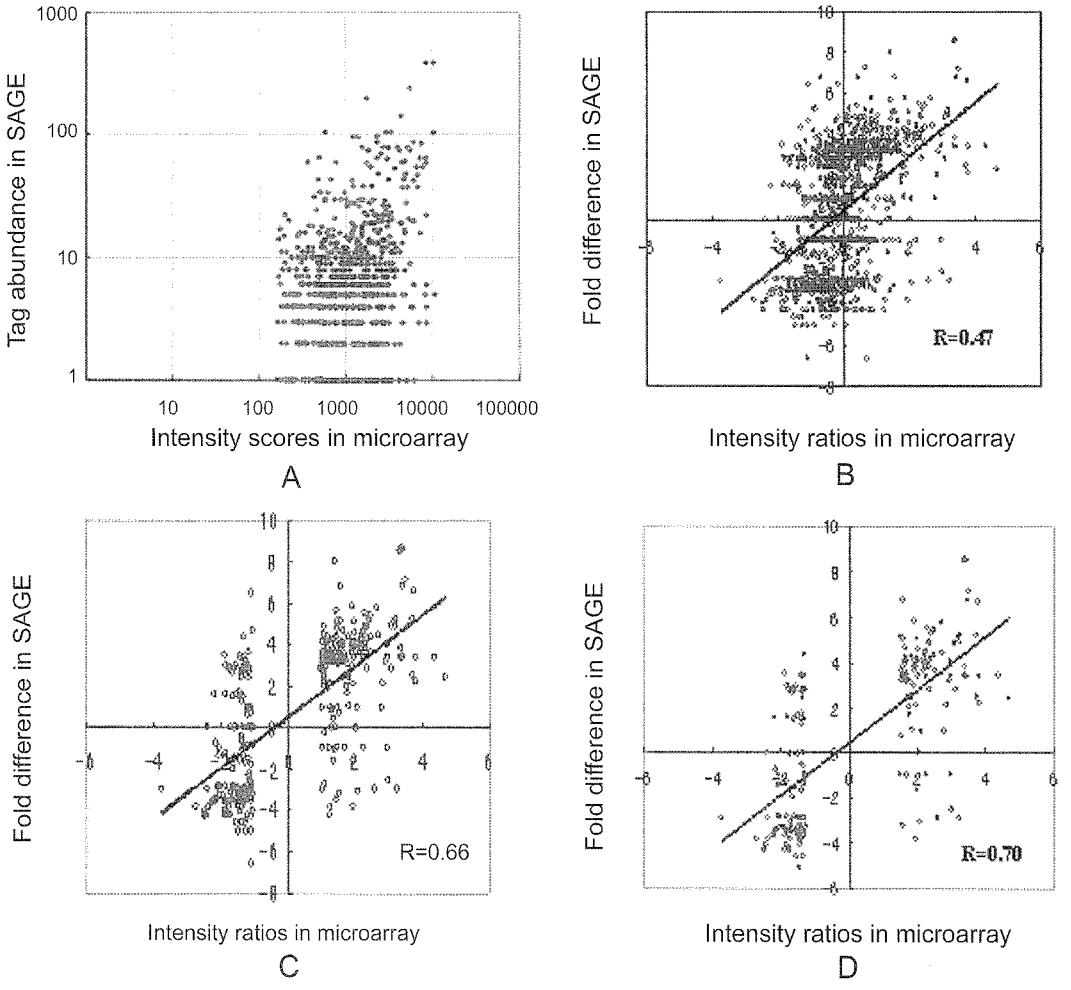

Figure 3. Comparison of gene expression data from microarray analysis and SAGE, using cold-stressed *Arabidopsis*. Two scatter plots are displayed in logarithmic scale. (**A**) Comparison of intensity scores in microarray and tag abundance in SAGE, as plotted in logarithmic scale on abscissa and ordinate, respectively. (**B**) Intensity values for 1186 genes, with a common set of genes included in both experiments from microarray. (**C**) Intensity values for 389 genes, with a common set of genes differentially expressed by >2-fold in both experiments. (**D**) Intensity values for top 100 increased transcripts and top 100 decreased transcripts.

of the spots were significant enough to analyze compared with <60% of spots that could be considered by the direct method. Likewise, this method gave highly reproducible and reliable data and required only a small amount, e.g., 2 μg, of target RNA. In contrast, many techniques necessitate having larger quantities of total RNA, up to 100 μg, for their starting materials in order to prepare the labeled-target cDNA. Fulfilling such a requirement is not easy, especially when trying to obtain samples from micro-dissected tissues.

Our unique approach of combining two different technologies -- cDNA microarray and SAGE -- allowed us to directly compare their data when the same samples were used. Although some differences arose in the list of genes classified as up- or down-regulated, the majority of these genes showed similar patterns of induction or repression during assessment by both methods, and no great variations in fold-differences were generally found. The correlation coefficient between methods was 0.47 to 0.56. Moreover, those genes with high fold-differences maintained the same trends between them while genes with lower fold-differences showed relatively less correlation.

The quantitative differences in cancer/normal ratios that had been determined by SAGE and microarray analysis appear to be caused by variations in the detection system. For example, microarray software automatically calculates fold-change values according to its algorithm; when the noise level is greater than a baseline experimental score, the intensity score from the other experiment is divided by the noise value rather than by a baseline score. Consequently, fold-change scores calculated in this manner are just approximate values. The SAGE method has similar problems. When no tag is detected in one sample, a value of '1' is used to avoid having to perform any division by zero. Therefore, such SAGE folds also are presented as approximation values.

Most of the genes spotted in our microarray were based on SAGE tags with fold-differences between normal and cancer tissues. Overall expression patterns were similar and correlations were high between the two methods. In addition, the correlation in absolute analyses was much better for genes with higher expression levels, and the correlation in comparative analyses was greater for large-fold changes in

expression. For example, a high correlation-coefficient value, 0.804, was obtained when the top 100 increased transcripts and the top 100 decreased transcripts were compared among the methods (data not shown).

In conclusion, we have demonstrated here that we can generate highly reproducible, good-quality data when we serially combine an improved labeling method with a specialized microarray that is prepared with PCR products that are selected based on previous SAGE tag information. Such a system will be useful for high-throughput analysis and the detection of unidentified and low-expression genes.

ACKNOWLEDGEMENTS

This work was supported by grants from the Crop Functional Genomics Center (CG1211); KOSEF (R01-2004-000-10621-0), the Brain Korea 21 Project; and the National Core Research Center (NCRC) program (R15-2006-020) of Korea.

Received November 16, 2006; accepted December 8, 2006.

LITERATURE CITED

Chen JJ, Rowley JD, Wang SM (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc Natl Acad Sci USA 97: 349-353

Finkelstein D, Ewing R, Gollub J, Sterky F, Cherry JM, Somerville S (2002) Microarray data quality analysis: Lessons from the AFGC project. *Arabidopsis* Functional Genomics Consortium. Plant Mol Biol 48: 119-131

Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J (2000) A concise guide to cDNA microarray analysis. Biotechniques 29: 548-556

Jung SH, Lee JY, Lee DH (2003) Use of SAGE technology to reveal changes in gene expression in *Arabidopsis* leaves undergoing cold stress. Plant Mol Biol 52: 553-567

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7: 819-837

Lee JY, Eom EM, Kim DS, Ha-Lee YM, Lee DH (2003) Analysis of gene expression profiles of gastric normal and cancer tissues by SAGE. Genomics 82: 78-85

Lee JY, Lee DH (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in *Arabidopsis* pollen undergoing cold stress. Plant Physiol 132: 517-529

Lee ML, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 97: 9834-9839

Li J, Pankratz M, Johnson JA (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. Toxicol Sci 69: 383-390

Nacht M, Ferguson AT, Zhang W, Petroziello JM, Cook BP, Gao YH, Maguire S, Riley D, Coppola G, Landes GM, Madden SL, Sukumar S (1999) Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. Cancer Res 59: 5464-5470

Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M, Zhu X, Patterson M, Shippy R, Sendera TJ, Mazumder A (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucl Acids Res 30: e30

Sawiris GP, Sherman-Baust CA, Becker KG, Cheadle C, Teichberg D, Morin PJ (2002) Development of a highly specialized cDNA array for the study and diagnosis of epithelial ovarian cancer. Cancer Res 62: 2923-2928

Takemasa I, Higuchi H, Yamamoto H, Sekimoto M, Tomita N, Nakamori S, Matoba R, Monden M, Matsubara K (2001) Construction of preferential cDNA microarray specialized for human colorectal carcinoma: Molecular sketch of colorectal cancer. Biochem Biophys Res Comm 285: 1244-1249

Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH (2001) Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. Nucl Acids Res 29: 2549-2557

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487

Xu D, Li G, Wu L, Zhou J, Xu Y (2002) PRIMEGENS: Robust and efficient design of gene specific probes for microarray analysis. Bioinformatics 18: 1432-1437

Yang GP, Ross DT, Kuang WW, Brown PO, Weigel RJ (1999) Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. Nucl Acids Res 27: 1517-1523

Yang YH, Dudoit S, Luu P, Speed TP (2001) Normalization for cDNA microarray data, *In* M Bittner, Y Chen, A Dorsel, ER Dougherty, eds, Microarrays: Optical Technologies and Informatics. Vol 4266, SPIE BIOS 2001, pp 141-152

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW (1997) Gene expression profiles in normal and cancer cells. Science 276: 1268-1272