Integrated *in silico* and biological validation of the blocking effect of C_ot-1 DNA on Microarray-CGH

SEUNG-HUI KANG[1,5], CHAN HEE PARK[1,2,5], HEI CHEUL JEUNG[1,5], KI-YEOL KIM[1,4], SUN YOUNG RHA[1,2,3,5] and HYUN CHEOL CHUNG[1,2,3,5]

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

Received February 16, 2007; Accepted March 22, 2007

Abstract. In array-CGH, various factors may act as variables influencing the result of experiments. Among them, C_ot-1 DNA, which has been used as a repetitive sequence-blocking agent, may become an artifact-inducing factor in BAC array-CGH. To identify the effect of C_ot-1 DNA on Microarray-CGH experiments, C_ot-1 DNA was labeled directly and Microarray-CGH experiments were performed. The results confirmed that probes which hybridized more completely with C_ot-1 DNA had a higher sequence similarity to the Alu element. Further, in the sex-mismatched Microarray-CGH experiments, the variation and intensity in the fluorescent signal were reduced in the high intensity probe group in which probes were better hybridized with C_ot-1 DNA. Otherwise, those of the low intensity probe group showed no alterations regardless of C_ot-1 DNA. These results confirmed by *in silico* methods that C_ot-1 DNA could block repetitive sequences in gDNA and probes. In addition, it was confirmed biologically that the blocking effect of C_ot-1 DNA could be presented via its repetitive sequences, especially Alu elements. Thus, in contrast to BAC-array CGH, the use of C_ot-1 DNA is advantageous in controlling experimental variation in Microarray-CGH.

Introduction

Conventional comparative genomic hybridization (CGH) was developed for genome-wide screening of chromosomal alterations. In CGH, a test sample is compared with a reference sample by labeling its respective genomic DNAs with different color markers. However, the application of this method is limited due to the resolution range of 3-10 Mb (1). In contrast to CGH, microarray technology allows for simultaneous assessment of the expression of a large number of genes at a single gene level. For this reason, application of microarray technology has been utilized in diverse fields such as disease classification, mutation and genotyping analysis and biomarker development (2).

Array-based CGH (array-CGH) combines microarray techniques with a conventional CGH allowing for the simultaneous investigation of a large quantity of gene alterations. Depending on the types of the probe used, array-CGH can be classified to one of the three following types: bacterial artificial chromosome (BAC)/phage artificial chromosome (PAC), cDNA, or oligonucleotide array CGH. Array-CGH with a BAC clone has been popular since its initial development. At first, whole genome array-CGH consisted of 2,400 BAC clones and had a resolution of approximately 1 Mb (3). At the present time, it has expanded to include 30,000 BAC clones for wider coverage of the genome (4). However, BAC clone-based array-CGH has problems such as experiment time, cost and limitation in resolution. Microarray-CGH using cDNA or oligonucleotides is one method that improves upon the resolution problem (5,6).

In Microarray-CGH experiments, a number of factors may act as inappropriate variables on experimental results such as reagents, technique and chip quality (7,8). C_ot-1 DNA is one of the commonly used reagents in Microarray-CGH experiments and is approximately 300 bp in length. It is a gDNA extracted from human placental DNA and treated by a process of extraction, shearing, denaturing, and reannealing. This DNA contains many repetitive DNA sequences such as Alu I and Kpn I (9,10) and is used to block non-specific cross-hybridization between repetitive sequences of gDNA and probes during hybridization. However, although C_ot-1 DNA should decrease experimental variations from its blocking effect, a report has indicated that single copy impurities in C_ot-1 DNA cause unpredictable cross-hybridizations between single copy sequence and repetitive sequence (11). That is, C_ot-1 DNA could contribute to an increment of intensity variations in Microarray-CGH experiment.

In the present study, C_ot-1 DNA was directly labeled with Cy5 and Cy3, and then homotypic experiments were performed to investigate the effect of C_ot-1 DNA on Microarray-CGH

Correspondence to: Professor Hyun Cheol Chung, Cancer Metastasis Research Center, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul 120-752, Korea
E-mail: unchung8@yumc.yonsei.ac.kr

Key words: C_ot-1 DNA, Microarray-CGH, sequence similarity, Alu repetitive elements

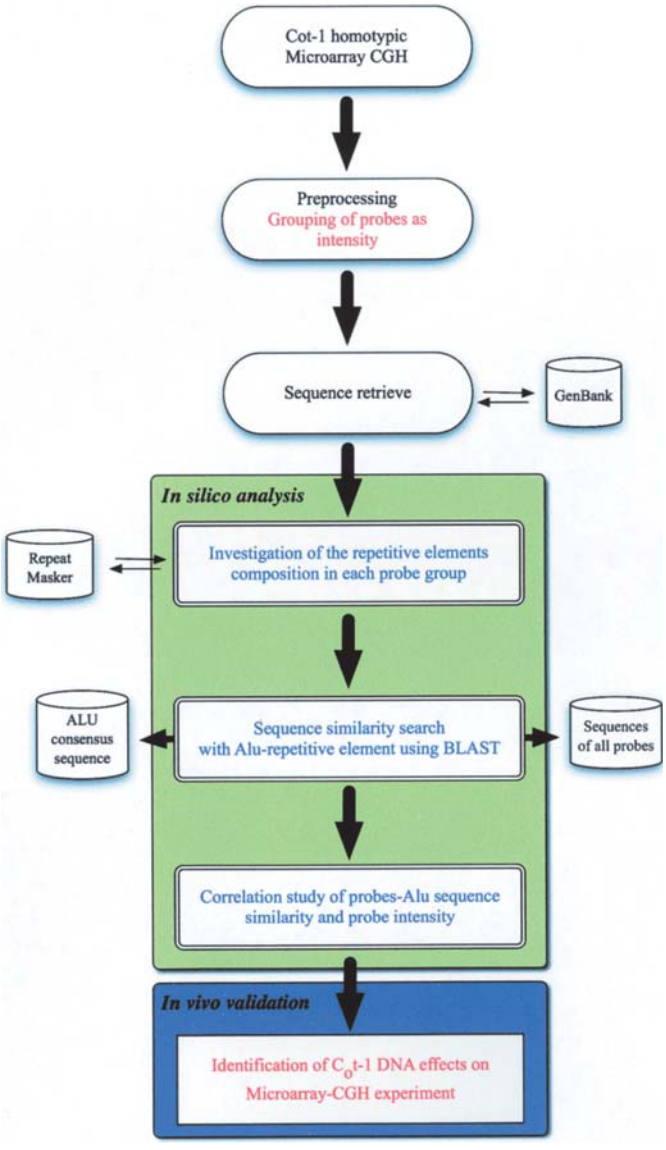

Figure 1. Analysis scheme. Microarray-CGH was performed using a homotypic C_ot-1 DNA labeling method in which Cy5 and Cy3 channels were labeled with identical C_ot-1 DNA samples. In the sequence retrieval procedure, NCBI GenBank was used to download selected sequences for data analysis. Alu elements were selected for correlation study using RepeatMasker. A correlation study was conducted using a sequence similarity search by stand-alone BLAST. Lastly, the effect of C_ot-1 DNA on Microarray-CGH was identified via a correlation study in four different experimental sets of sex-mismatched dye-swap arrays.

experiments. Through serial *in silico* and experimental analysis (Fig. 1), it was predicted that Alu elements in C_ot-1 DNA could play a key role in blocking the cross hybridization of repetitive sequences.

Materials and methods

DNA extraction. Fresh frozen tissue (100 mg) was minced and incubated with 400 μl of DNA lysis buffer (10 mM Tris pH 7.6, 10 mM EDTA, 50 mM NaCl, 0.2% SDS, 200 μg/ml Proteinase K) at 42˚C for 12-24 h. The incubated sample was boiled for 10 min at 100˚C to inactivate enzymatic activity and then treated with an equal volume of phenol/chloroform/isoamylalcohol to isolate nucleic acids. DNA was precipitated with 100% ethyl alcohol containing 1/3 volume of 10 M ammonium acetate and 2 μl of glycogen. After rinsing with 70% ethyl alcohol, DNA was dried at room temperature and subsequently dissolved in ultra-pure water. DNA concentrations were determined using UV absorption spectroscopy and then stored at -20˚C until further experiments.

Microarray-CGH experiment and data processing. Microarray-CGH was processed as described previously (12). Briefly, 8 ug of C_ot-1 DNA (Applied Genetics Laboratories, Inc.) was purified with a QIAquick PCR purification kit (Qiagen) according to the manufacturer's instructions. Labeling with either Cy3-dUTP or Cy5-dUTP was performed with the Bioprime labeling kit (Invitrogen) and unincorporated nucleotides were removed using a PCR purification kit (Qiagen). Eluted probes were mixed and supplemented with 20 ug of poly-A RNA (Sigma), 100 ug of yeast t-RNA (Gibco-BRL) and 288 ul of 1 M TE buffer (pH 8.0). The probe mixture was concentrated using a Microcon-30 (Millipore).

A number of differently designed sex-mismatched experiments were carried out using placental genomic DNA. Briefly, 6 ug of extracted placental genomic DNA was digested at 37˚C for 2 h by *Dpn*II (NEB) and cleaned with a QIAquick PCR purification kit (Qiagen) according to the manufacturer's instructions. Concentrated DNA mixtures were mixed with 15.3 μl 20X SSC (pH 8.0) and 2.7 μl 10% SDS to obtain the total volume of 90 μl. The mixture was then denatured for 2 min and hybridization was performed using a 17K human cDNA microarray (CMRC-GenomicTree Co.). The microarray was hybridized at 65˚C for 16 h in a hybridization chamber (GenomicTree Co.) where humidity was maintained with 1.5X SSC. Afterwards, the slides were washed for 2-5 min with 2X SSC containing 0.1% SDS, followed by 1X SSC containing 0.1% SDS, 0.2X SSC, and finally rinsed twice with 0.05X SSC. After washing, slides were centrifuged at 600 rpm for 5 min.

Slides were scanned using the GenePix 4000B scanner (Axon Inc.). Images were stored as TIFF type files and analyzed by GenePix Pro 4.1 software (Axon Inc.). Foreground intensity (Cy5) and background intensity (Cy3) were calculated against each spot and gridding was conducted using the Genepix Array List (GAL) file to produce a final GenePix Result (GPR) file. In the GPR file, the data of flagged spots and spots without a GenBank accession number were also removed.

Multiple sequence alignment of probe sequences. Eight Alu-warning sequences reported in the NCBI GenBank were used (U14574.1, U14573.1, U14572.1, U14571.1, U14570.1, U14569.1, U14568.1, U14567.1) for the alignment of Alu-repetitive elements. JalView software was used for the analysis of multiple sequence alignment (MSA) and the analysis of alignment results (13).

Grouping of probes and conduction of RepeatMasker. To divide probes into groups with different intensities, we selected probes with Cy5 and Cy3 intensities larger than zero. A single intensity value was then calculated as follows: log intensity = $\log_2$RG (equation 1) (R, Cy5 intensity; G, Cy3 intensity).

Table I. Experimental design.

Experiment	Repetitive sequence blocking agent	Labeled sample		Condition
		Cy5	Cy3	
Homotypic direct labeling[a] of C_ot-1 DNA	None	C_ot-1 DNA	C_ot-1 DNA	C_ot-1 experiment
Sex-mismatched dye-swap labeling[b] (Placental gDNA)	With C_ot-1 DNA	Male	Female	Condition 1
		Female	Male	Condition 2
	Without C_ot-1 DNA	Male	Female	Condition 3
		Female	Male	Condition 4

[a]Homotypic C_ot-1 DNA labeling conditions for sequence-based analysis. C_ot-1 DNA was used as a sample in both Cy5 and Cy3 channels. No blocking agent was added. [b]Sex-mismatched dye-swap conditions for the identification of a C_ot-1 DNA effect on Microarray-CGH. In conditions 1 and 2 (with C_ot-1 DNA), male placental gDNA was used as the Cy5 channel, female placental gDNA as the Cy3 channel, and both were inversely hybridized for the dye-swap. In conditions 3 and 4 (without C_ot-1 DNA), male placental gDNA was used as the Cy5 channel, female placental gDNA as the Cy3 channel, and both were inversely hybridized for the dye-swap.

Probes were divided into 6 groups according to their log intensity values (equation 1). Probes with a log intensity >0 but ≤5 were designated as group A, and those with 5-10 were group B, 10-15 were group C, 15-20 were group D, 20-25 were group E and >25 were group F. To search the repetitive elements in each probe group, all sequences were submitted into RepeatMasker, which screens DNA sequences for interspersed repeats and low complexity DNA sequences (Smit AFA, Hubley R and Green P: RepeatMasker Open-3.0. 1996-2004. http://www.repeatmasker.org).

Correlation analysis of Alu similarity and probe intensity. To examine the alteration of probe intensity due to sequence similarity for Alu-elements, the sequence similarity search of probe sequences was conducted with stand-alone BLAST provided by the NCBI (14). The Alu sequence used for BLAST searches was the consensus sequence obtained from the 8 Alu-warning sequences by MSA and is as follows: 5'-GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAG CACTTTGGGAGGCCGAGGCGGGCGGATCACNNGA GGTCAGGAGATCGAGACCATCCTGGCTAACACGG TGAAACCCCGTCTCTACTAAAAANTACAAAAANNT TAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAG CTACTCGGNGAGGCTGAGGCAGGAGAATGGCGTGA ACCCNGGGAGGCGGAGCTTGCAGTGAGCCGAGATC GCGCCACTGCACTCCANNNNNNGCCTGGGCGACAN GAGCGAGACTCCGTCTCAAAAAAAA-3'. In the BLAST analysis, all options were set to the default except for the e-value (1000 vs 0.001). Finally, an adjusted score was calculated using the alignment score and the e-value obtained from the BLAST results (equation 2): adjusted score ($Score_{adj}$) = $\log_2$ (score/e-value).

Statistical analysis of sequence similarity. For the 6 probe groups, the rates of the number of sequences having the adjusted score and the number of total sequences were calculated and validated statistically by the chi-square test. The difference of the distribution of intensity among each intensity group was analyzed by analysis of variance (ANOVA). By using multiple comparisons, 6 groups were regrouped into two larger groups and the difference of the intensity between the two groups was analyzed by Student's t-test. In four Microarray-CGH experiments, the intensity alteration of each Cy5 and Cy3 intensity was validated via the Wilcoxon rank-sum test.

Results

Selection of the subject probes for homotypic C_ot-1 Microarray-CGH analysis. Among a total of 17,664 Microarray-CGH probes, 1,869 un-annotated probes that did not have a GenBank accession number were excluded from subsequent analysis. After initial screening of the data, probes having Cy5 and Cy3 intensities >0 were selected and thus, the final analysis was performed using 15,024 probes.

Multiple sequence alignment to find the consensus region of probe sequences. In order to find the consensus sequences in the probes hybridized with C_ot-1 DNA, multiple sequence alignment (MSA) was performed using the sequence of 82 probes with Cy5 and Cy3 intensities >15,000. Following examination of the GenBank annotation, we found that among 82 clones, 47 contained Alu-repetitive elements. In addition to Alu, the 82 clones contained 15 other repetitive elements, namely, OFR, LTR5, MER1, MER5, MER9, MER6, MER15, MER22, MSR1, L1, LTR7, KER, HGR, TAR1 and XTR.

For sequence comparison of the 47 clones with the Alu sequence, 8 Alu consensus sequences derived from each Alu family were selected from the NCBI GenBank. The result of the MSA indicated that, for the most part, the 8 sequences were homologous (Fig. 2a). MSA of the 8 Alu sequences and the 47 clones confirmed that all of the aligned regions of each sequence were indeed an Alu sequence (Fig. 2b). On the other hand, in probes with intensities <0 that were used

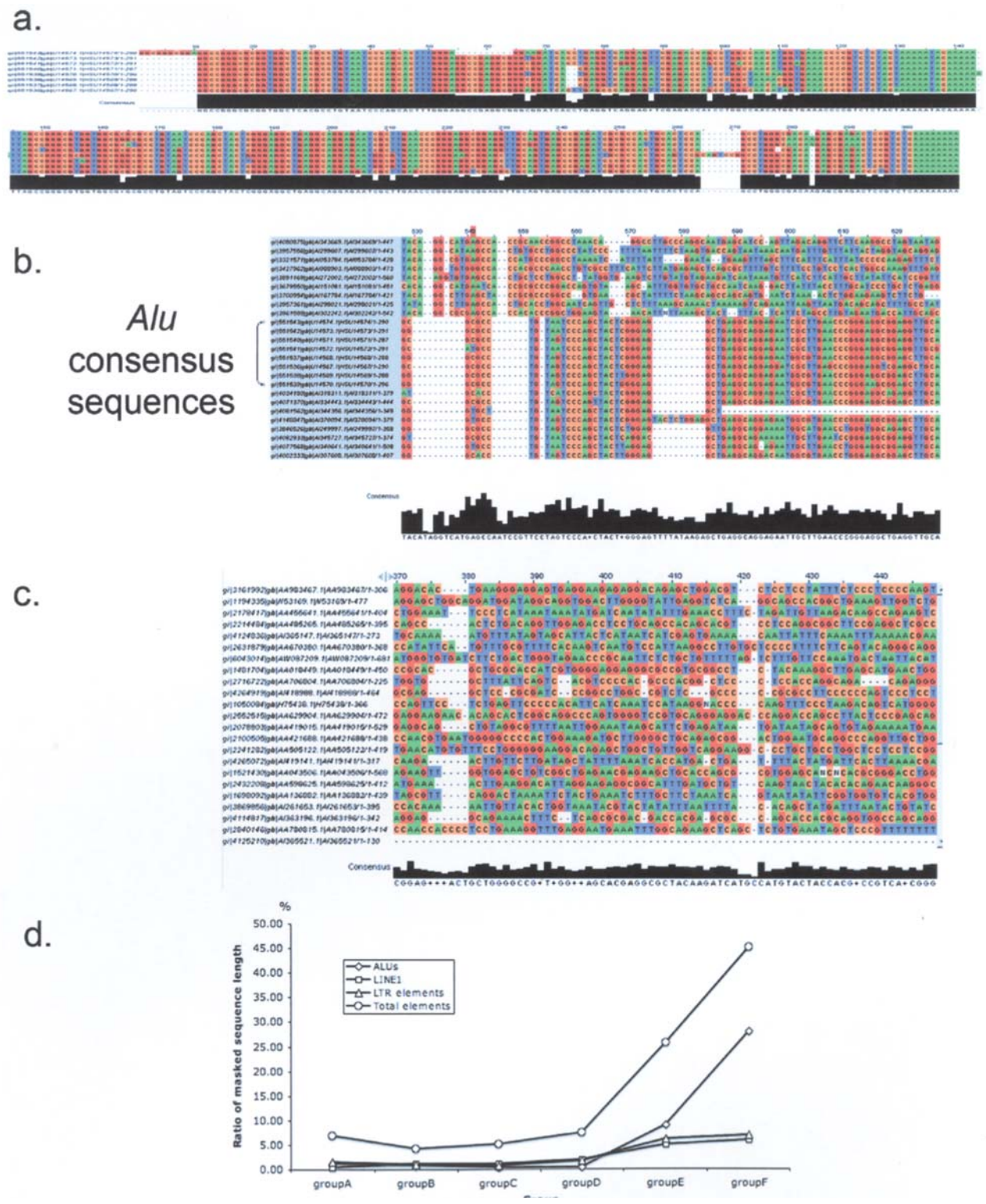

Figure 2. Multiple sequence alignment of probe and Alu sequences. (a) Multiple sequence alignment of Alu-warning sequences. (b) Multiple sequence alignment of probe sequences that have a high intensity (>15000) with Alu sequences. (c) Multiple sequence alignment of probe sequences having low intensities (<0) with Alu sequences. (d) Ratio of masked sequence length of the major repetitive elements. Ratio was calculated from the results of RepeatMasker (ratio = masked sequence length of repetitive element/total sequence length).

as negative controls, none of the consensus regions were detected by MSA (Fig. 2c). Therefore, it was confirmed that in the hybridization of the C_ot-1 DNA, cases exhibiting high intensities were due to the sequence similarity between the Alu-repetitive sequence present in C_ot-1 and the sequence contained in the probe.

Investigation of the repetitive element composition in each probe group. Although Alu elements were the major repetitive elements in probes having high intensities, we checked all the other repetitive elements in entire probes. First, probes were divided into 6 groups according to their log intensity (equation 1). To investigate the repetitive element composition

Table II. Comparison of repetitive elements in probe groups by RepeatMasker.

	Group A	Group B	Group C	Group D	Group E	Group F	Total
Interspersed elements							
SINEs	510	4397	11773	36442	75256	79532	207910
ALUs	357	2210	4952	11765	68539	77046	164869
MIRs	153	2187	6821	24677	6717	2409	42964
LINEs	328	4811	19345	79788	42162	18038	164472
LINE1	137	3481	14841	62344	37023	15944	133770
LINE2	119	1160	4060	16590	5033	2018	28980
LTR elements	450	2764	11310	59975	46176	18652	139327
DNA elements	223	1006	9849	28025	15889	2742	57734
Unclassified	0	318	323	268	74	461	1444
Non-interspersed elements							
Small RNA	0	0	179	794	866	68	1907
Satellites	0	0	0	356	113	509	978
Simple repeats	231	0	6214	25957	9823	1945	44170
Low complexity	258	0	10112	21024	4779	1744	37917
Masked ratio (%)[a]	6.92	4.21	5.22	7.50	25.85	45.33	10.81
Masked sequence length	2000	13296	69105	252769	195138	123682	655990
Total sequence length	28903	315875	1324302	3369943	754768	272832	6066623

[a]Masked ratio was calculated as follows: Masked sequence length/Total sequence length. To search the repetitive elements in each probe group, all sequences were submitted into RepeatMasker. These results confirmed that the major components of C_ot-1 DNA were ALU, LINE, and LTR elements. Among these elements, Alu elements contributed most significantly to the hybridization of C_ot-1 DNA with the probe sequences.

in each probe group, all of the probe sequences in each group were then submitted to the RepeatMasker. The masked sequence-length ratio of Alu and other major repetitive elements for the total input sequence was calculated (Table II). The ratio of the masked sequence length of total probes was highest in group F and decreased in order of intensity. The major repetitive elements were Alu, Line1, and LTR elements. In these elements, Alu had a longer masked sequence length (164869 bp) than any of the other elements and the ratio of masked sequence length of Alu showed the same pattern as the pattern of total elements (Fig. 2d). Via *in silico* analysis, Alu elements were found to be the major representative repetitive elements of C_ot-1 DNA for the correlation between spot-signal intensity and sequence similarity.

Correlation analysis of spot-signal intensities and Alu-element similarities. We investigated the correlation of signal intensity and Alu-element similarity on the sequence of the entire probes. First, using BLAST, the Alu similarity of the entire sequence of probe groups was examined. In addition, in each group, the number of probes with similarity to the Alu consensus sequence was examined (Table III). Among a total of 15,024 probe sequences, 1,657 probes (11%) exhibited similarity with the Alu-sequence; the ratio was highest in group F (57.1%), followed by group E (26.5%). In the remaining groups (A, B, C, and D), the incidence of aligned probes with similarity was not differential, ~5-10%. On the other hand, when we decreased the e-value to 0.001, the sequences of 806 probes (5.4%) were successfully aligned. Excluded sequences with this high e-value (e-value >0.001) were found mainly in the intensity groups C, D, and E.

Table III. Probe grouping based on sequence similarity.

Intensity group	Log intensity	N_t[a]	N_a[b]	Rate[c] (%)
A	0-5	73	7	9.6
B	5-10	759	73	9.6
C	10-15	3297	193	5.9
D	15-20	8332	500	6.0
E	20-25	1890	500	26.5
F	>25	673	384	57.1
Total		15024	1657	11.0

[a]N_t, number of total sequences in each group. [b]N_a, number of aligned sequences thet were included in BLAST results. [c]Rate of aligned sequences in each intensity group (calculated from N_a/N_t).

Next, the correlation between intensity and the adjusted score was examined. It was found that in groups A-D, which had a log intensity <20, there were low $Score_{adj}$ values and no

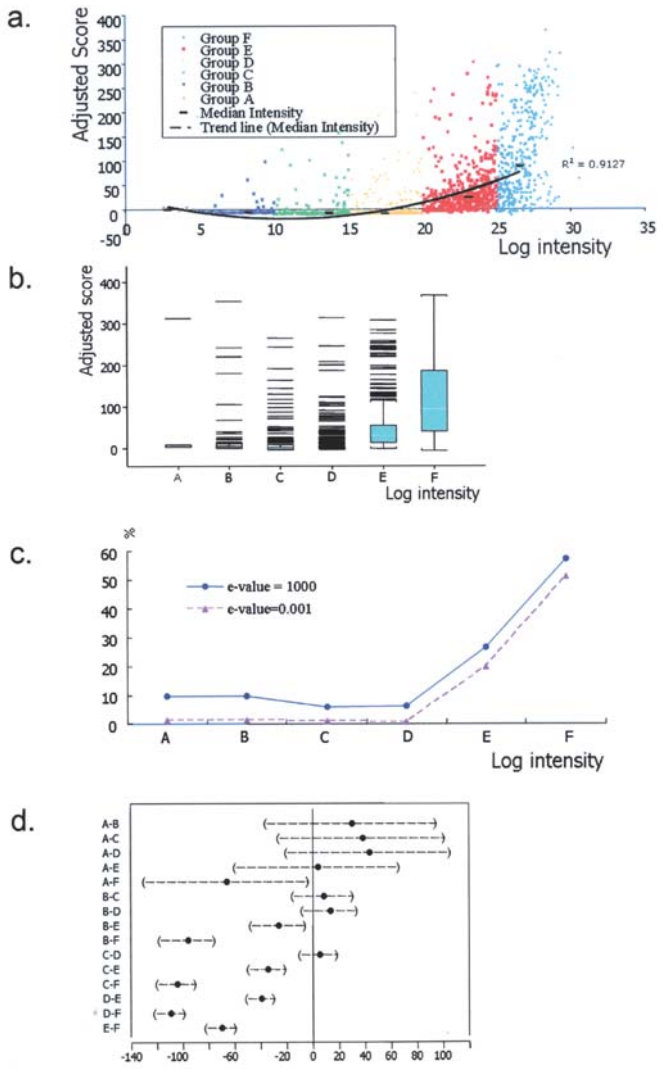

Figure 3. Correlation analysis between spot-signal intensities and Alu-element similarities. (a) Scatter plot of log intensity-adjusted score. Log intensity was calculated using equation 1 and adjusted score was calculated using equation 2. (b) Box plot of $Score_{adj}$ per intensity group (c) Rates of aligned sequences in each intensity group. Solid line and dotted line are rate curves of e-value 1000 and 0.001 in BLAST option, respectively. (d) Multiple comparison analysis result. Groups A-F were compared to each other for Alu-element similarity scores. As a result, the six groups were divided into two groups; A-D and E-F.

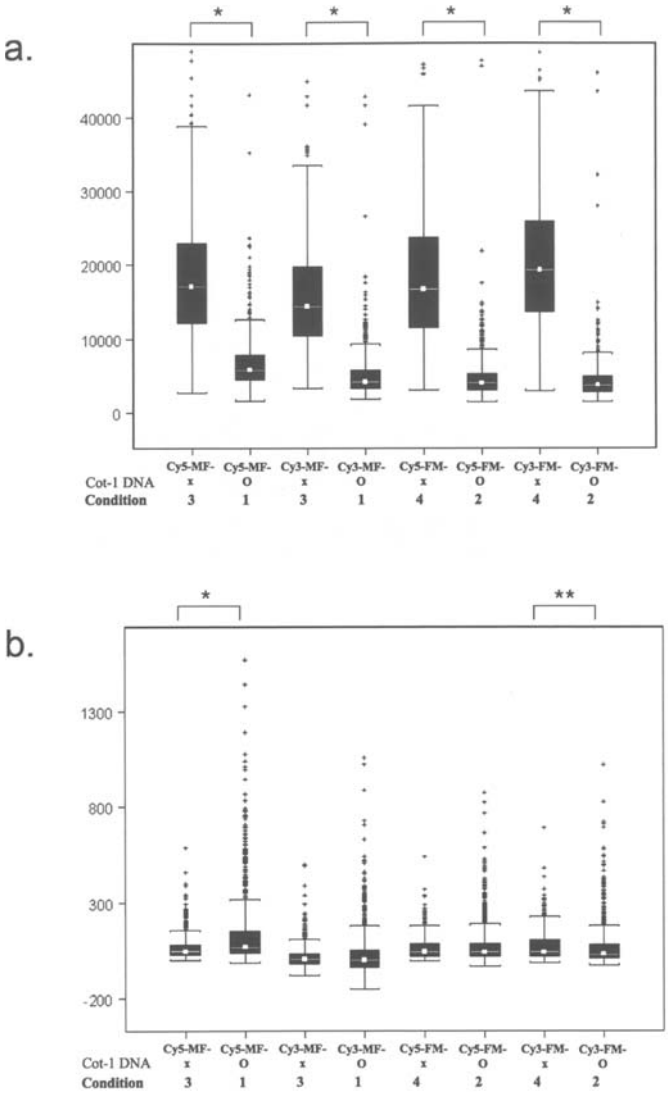

Figure 4. Identification of C_ot-1 DNA effect on Microarray-CGH. The effect of C_ot-1 DNA was examined in actual Microarray-CGH experiments using a sex-mismatched dye swap method. Intensity alterations in each Cy5 and Cy3 channel of four conditions were represented via box-whisker plot. (a) Intensity alterations of group F. (b) Intensity alterations of group B.

difference among the groups was found (Fig. 3a and b). Conversely, in groups E and F which had log intensities >20, numerous sequences with a high $Score_{adj}$ were identified. Sequences with a high Alu similarity were gradually increased, therefore, the median value of $Score_{adj}$ in those groups showed a good correlation by the quadratic equation (Fig. 3a).

In each intensity group, the rates of aligned sequences among groups A-D did not exhibit a great difference. However, in groups E and F, which had high intensities, the ratio was increased to 27% and 57%, respectively (Fig. 3c, Table III). Although the rates of aligned sequences were smaller in e-value 0.001 than e-value 1000, the rates among probe groups showed similar pattern. In regards to the number of the total sequences belonging to the intensity group, the ratio of aligned sequences in a group was increased for high intensity groups (chi-square test, $p<0.001$). Such score differences among intensity groups were validated (ANOVA test, $p<0.001$). In addition, when multiple comparison analysis was also performed (Fig. 3d), groups A, B, C, and D hardly showed any differences. Groups E and F, however, were different from the previous 4 groups. When regrouping all 6 groups into two groups (E-F and A-D), the scores of these two groups were significantly different from each other (Welch's t-test, $p<0.001$).

The *in silico* results presented above confirmed that in probe groups where the signal intensity was high, numerous probes with high Alu similarity were distributed. In addition, the correlation of intensity to the Alu similarity of probe sequence was also confirmed. Lastly, our results confirmed that hybridization of a probe with C_ot-1 was associated with a higher sequence similarity with the Alu repetitive element.

Identification of a C_ot-1 effect on Microarray-CGH. In C_ot-1 homotypic Microarray-CGH experiments, repetitive sequences

such as Alu may mediate an effect on the signal intensity of probes as we expected from *in silico* data. For this analysis, four sex-mismatched dye swap experiments were performed in Microarray-CGH experiments using C_ot-1 DNA (Table I). Thereafter, the intensity alteration between experiments with the use of C_ot-1 and without C_ot-1 was compared in intensity groups F and B.

Results indicated that when C_ot-1 DNA was used, the intensity variation of the probes in both Cy5 and Cy3 of group F decreased (Fig. 4a). In addition, when C_ot-1 DNA was not used, the distribution variation of intensity was greater than when using C_ot-1. The differences of intensity in all compared experiments showed significant changes (Wilcoxon rank-sum test, p-value <0.001). Otherwise, in probe group B, the intensity was not differential regardless of the use of C_ot-1 DNA (Fig. 4b).

These results confirmed that for probes with a high similarity to the Alu sequence, the use of C_ot-1 DNA reduced the variation of intensity due to the blocking effect on repetitive sequences. Thus, in Microarray-CGH using cDNA, technical variations can be controlled using C_ot-1 DNA.

Discussion

The applications of microarrays are very diverse and include tumor classification, prognosis prediction, drug target screening, SNP marker identification, mutation site screening and elucidation of signaling networks. Among these applications, array-CGH was developed by applying the advantages of microarrays on conventional CGH to examine changes in copy number at a genomic level. C_ot-1 DNA has been used in conventional CGH experiments in order to block genomic repetitive sequences present in samples (15). However, it has been reported that C_ot-1 DNA increases the intensity variation of array-CGH experiments (11,16), which is thought to be due to repetitive sequences present in C_ot-1 DNA that cause non-specific binding of samples and probes. This non-specific binding results in increased variation in quantitative measurement of array-CGH. Therefore, we examined the actual effects of C_ot-1 DNA on cDNA-based Microarray-CGH by applying the homotypic C_ot-1 DNA direct labeling method and *in silico* sequence-based analysis.

C_ot-1 DNA is produced from human gDNA and contains a large amount of repetitive sequences such as Alu and Kpn. For this reason, if C_ot-1 DNA was labeled directly and hybridized to cDNA probe-spotted Microarray-CGH, theoretically it should not hybridize. In this experiment, however, numerous probes were hybridized with C_ot-1 DNA. Such an experimental result implies the existence of homologous regions between repetitive sequences present in C_ot-1 DNA and probe sequences. In multiple sequence alignment results, the high intensity group contained LTR, MER, and various other repetitive sequences, in addition to the Alu repetitive sequence. To investigate repetitive elements in entire probes, therefore, probe sequences were submitted to RepeatMasker. The resulting data revealed that Alu elements had the longest masked sequences. In addition, the ratio of the masked sequence length of Alu elements to the total sequence length showed the same pattern as that of the total repeat elements. These results confirmed that the Alu element is the major representative repetitive element in Cot-1 DNA. For this reason, the Alu consensus sequence was used for the correlation study of intensity and similarity in our experiments.

The Alu repetitive sequence is a dimer consisting of two monomers approximately 280 bp in length (17); it is a type of short interspersed nuclear elements (SINEs) that comprise approximately 10-11% of the human genome (18). The Alu element is very abundant having some 1.2 million copies in the genome, as well as abundant subfamilies. Such Alu elements mediate an effect on normal gene expression (19,20) and thus, during cDNA synthesis, Alu sequences may be inserted into cDNA (21).

In the present study, C_ot-1 DNA showed a stronger hybridization to the probes with high sequence similarity to the Alu element. This result implies that a higher signal intensity of a probe is correlated with increased homology with the Alu sequence. By using stand-alone BLAST, the sequence similarity between the Alu sequence and probe sequences was analyzed. For the analysis, two types of variables were used. Log intensity (equation 1) was used as the probe intensity and the adjusted score (equation 2) was used as the sequence similarity score with the Alu element. The adjusted score was used due to the fact that even if two probes have the same score, it is not statistically significant if the e-value is high. The e-value threshold was assigned as 1,000 for BLAST analysis. Although the rates of aligned sequences were smaller in e-value 0.001 than e-value 1000 (Fig. 3c), the rates among probe groups showed similar patterns. This result means that more numerous sequences show homology to the Alu consensus sequence in e-value 1000 though the e-values of those sequences are high. For this reason, the e-value threshold for BLAST was assigned as 1,000 to obtain more probes for the correlation analysis. The correlation of these two variables was examined, and it was found that among groups A-D with low signal intensities, their adjusted score was not differential. However, in groups E and F, which consisted of high intensities, the pattern of the score increase was significant, exhibiting a strong correlation by the quadratic equation. This result is in agreement with the hypothesis that in probes with higher intensity, the association of the Alu element with the probe sequence should be increased. In the chi-square test, ANOVA, and Student's t-test, the distributions showed identical patterns.

The results presented above showed that signal intensity was indeed associated with sequence similarity, although the effect of the variation on the entire signal intensity was not clear. To examine this, we used identical placental gDNA of both males and females to perform Microarray-CGH with a sex-mismatched dye swap method. The experiment was performed using an identical gDNA set under condition 1, 2, 3, and 4, depending on the presence or absence of C_ot-1 DNA. To examine the alteration of the signal intensity of probes in the presence or absence of C_ot-1 DNA, the Cy5 and Cy3 signal intensity of each pair was analyzed. We found that in the absence of C_ot-1 DNA, probes in group F showed higher intensity, as well as larger intensity variation, than those in the presence of C_ot-1 DNA. However, in group B, there were very low intensities and small variations regardless of the addition of C_ot-1 DNA. In other words, in probes with

high Alu similarity, due to the presence of C_ot-1 DNA, intensity stabilization and reduction of signal variation were detected. In regard to cDNA probes belonging to either group E or F, due to Alu consensus regions present in the probes, signal intensity in Microarray-CGH experiments may have appeared higher than the actual expression levels because of the use of gDNA as a sample. However, C_ot-1 DNA used in the same experiment would hybridize not only with samples but also with the above probes, thus reducing non-specific cross-hybridizations. In conclusion, in Microarray-CGH using cDNA, technical variation can be reduced using C_ot-1 DNA.

Probes in which signal intensity was decreased and signal variation was reduced after application of C_ot-1 DNA were well-hybridized with C_ot-1 DNA. In most of these probes, their sequence similarity with the Alu repetitive element was high. In addition, most of these probes were unknown genes that were not annotated or hypothetical proteins. Therefore, in cases where C_ot-1 DNA is not used, such probes may act as an artifact during data analysis influenced by repetitive elements.

Repetitive sequences present in the genome are an enormous hindrance to sequencing experiments which apply hybridization techniques. As various methods designed to remove the effect of such repetitive sequences have been developed, the use of blocking agents such as C_ot-1 DNA and the method of using probes to remove repetitive sequences are now available. In array-CGH, rather than using a BAC clone, the use of cDNA or oligonucleotide probes may be an alternative for preventing a repetitive sequence effect and for reducing experimental time and cost (5,6).

This study performed Microarray-CGH quality analysis using an *in silico* sequence-based assay with subsequent biological validation. Such *in silico* sequence-based assay may be applied effectively to the quality validation of a large quantity of Microarray-CGH probes.

Acknowledgements

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (0405-BC01-0604-0002).

References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818-821, 1992.
2. Il-Jin K, Hio-Chung K and Jae-Gahb P: Microarray applications in cancer research. Cancer Res Treat 36: 207-213, 2004.
3. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D and Albertson DG: Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29: 263-264, 2001.
4. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C and Lam WL: A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36: 299-303, 2004.
5. van den Ijssel P, Tijssen M, Chin SF, Eijk P, Carvalho B, Hopmans E, Holstege H, Bangarusamy DK, Jonkers J, Meijer GA, Caldas C and Ylstra B: Human and mouse oligonucleotide-based array CGH. Nucleic Acids Res 33: e192, 2005.
6. Carvalho B, Ouwerkerk E, Meijer GA and Ylstra B: High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. J Clin Pathol 57: 644-646, 2004.
7. Marshall E: Getting the noise out of gene arrays. Science 306: 630-631, 2004.
8. Li X, Gu W, Mohan S and Baylink DJ: DNA microarrays: their use and misuse. Microcirculation 9: 13-22, 2002.
9. Britten RJ, Graham DE and Neufeld BR: Analysis of repeating DNA sequences by reassociation. Methods Enzymol 29: 363-418, 1974.
10. Weiner AM, Deininger PL and Efstratiadis A: Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55: 631-661, 1986.
11. Newkirk HL, Knoll JH and Rogan PK: Distortion of quantitative genomic and expression hybridization by Cot-1 DNA: mitigation of this effect. Nucleic Acids Res 33: e191, 2005.
12. Seo MY, Rha SY, Yang SH, Kim SC, Lee GY, Park CH, Yang WI, Ahn JB, Park BW and Chung HC: The pattern of gene copy number changes in bilateral breast cancer surveyed by cDNA microarray-based comparative genomic hybridization. Int J Mol Med 13: 17-24, 2004.
13. Clamp M, Cuff J, Searle SM and Barton GJ: The Jalview Java alignment editor. Bioinformatics 20: 426-427, 2004.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997.
15. Traut W, Sahara K, Otto TD and Marec F: Molecular differentiation of sex chromosomes probed by comparative genomic hybridization. Chromosoma 108: 173-180, 1999.
16. Carter NP, Fiegler H and Piper J: Comparative analysis of comparative genomic hybridization microarray technologies: report of a workshop sponsored by the Wellcome Trust. Cytometry 49: 43-48, 2002.
17. Mighell AJ, Markham AF and Robinson PA: Alu sequences. FEBS Lett 417: 1-5, 1997.
18. Jurka J: Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev 14: 603-608, 2004.
19. Almenoff JS, Jurka J and Schoolnik GK: Induction of heat-stable enterotoxin receptor activity by a human Alu repeat. J Biol Chem 269: 16610-16617, 1994.
20. Britten RJ: DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 93: 9374-9377, 1996.
21. Claverie JM and Makalowski W: Alu alert. Nature 371: 752, 1994.