BIOMEDICAL AND ENVIRONMENTAL SCIENCES **23**, 384-390 (2010)

Assessment of Comparative Genomic Hybridization Experiment by an *in situ* Synthesized CombiMatrix Microarray with *Yersinia pestis* Vaccine Strain EV76 DNA[1]

YUAN-HAI YOU[#], PENG WANG[#], YAN-HUA WANG, HAI-BIN WANG, DONG-ZHENG YU, RONG HAI, AND JIAN-ZHONG ZHANG[2]

National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China

Objective The quality of microarray data influences the accuracy of comparative genomic analyses to a large extent. To ensure that the results obtained by using an *in situ* synthesized microarray are accurate, data quality is to be assessed by evaluating the melting temperature (*T*m) of probes, probability of false synthesis rates, and fragmentation of labeled targets. **Methods** DNA from the *Yersinia pestis* vaccine strain EV76 was used for microarray analyses. Microarray results were confirmed by PCR. Statistical and bioinformatics methods were employed to perform microarray data analyses and evaluation. **Results** Correlation coefficients of the three datasets were above 0.95 after two-time stripping and hybridization with a labeled DNA with the size of fragmentation being 200 bp - 2 kb, which showed that the hybridization results were highly reproducible. Correlation coefficients were lower with the values ranging from 0.87 to 0.92 between the datasets generated from hybridization with different sizes of the labeled DNA fragment. For the relationship between *Tm* and signal intensity, there was a different distribution of *Tm* in the lowest 300 or 3 000 probes with a range of 70 ℃-72 ℃ and the highest 300 or 3 000 probes with a range of 72 ℃-74 ℃. **Conclusion** The results of this study suggest that the initial microarray design may affect the accuracy of final analyses and that the probe *T*m and the size of the labeled fragment may be the two factors of the greatest importance.

Key words: Array CGH; Data quality; Assessment

INTRODUCTION

Microarray-based comparative genomic hybridization is a powerful tool for genomic analyses. DNA gains and losses in the entire genomes can be obtained with a single microarray experiment. Currently, at least four types of microarray fabricating technologies have been developed. Of these methods, the *in situ* synthesized microarray is preferable because of its high density, high accuracy, and high throughput. The fabrication system is flexible and easily controlled and has been widely used. Although some data in microarray databases may be questionable[1-2], few papers have addressed the reliability of *in situ* microarray data or investigated the factors that contribute to the data quality. Some models and algorithms have been developed for data quality assessment; they mainly assess the expressional level of accuracy[3-7]. A systematic assessment from the initial wet experiment to final data analyses is lacking. In the case of CombiMatrix microarrays, data reliability may be affected by the probe synthesis quality, probe design, target DNA fragmentation, and algorithm used by the analytic software. In this study, we used DNA from *Yersinia pestis* for microarray analyses and the assessments hence made should improve the reliability of the CombiMatrix CustomArray™ analyses.

MATERIALS AND METHODS

DNA Isolation and Optimization of Fragmentation

The *Yersinia pestis* vaccine strain EV76 was

[1]This research was supported by a grant from the National High Technology Research and Development Program of China(863 Program, No. 2006AA2Z4A7).

[2]Correspondence should be addressed to Jian-Zhong ZHANG. Tel: 86-10-58900707. Fax: 86-10-58900700. E-mail: zhangjianzhong@icdc.cn

Biographical note of the first author: Yuan-Hai YOU, male, born in 1978, master of public health; and Peng WANG, male, PhD. the National Institute for Communicable Disease Control and Prevention, Chinese Centre for Disease Control and Prevention.

[#] The authors contributed equally to this paper.

0895-3988/2010
CN 11-2816/Q

used in this study. The bacteria were cultivated in the nutrient agar at 28 ℃ for 48 h, and then the genome DNAs were extracted by using the DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer's instructions. The concentration of genomic DNA was adjusted to 117ng/μL with nuclease-free water, and the final volume was 30 μL. DNA was sonicated (SONICS, VIBRA CELL, USA) on ice at optimized sonication conditions to sizes of 200 bp to 2 kb for the first three hybridizations and 200-800 bp for the fourth hybridization. The results were visualized by 1% agarose gel electrophoresis.

Microarray Design, Labeling, Hybridization, and Stripping

A whole-genome CombiMatrix CustomArray™ 12K (Mukilteo, WA, USA) was used in this study. The array contained 12 000 *in situ* synthesized oligonucleotide probes. At least one probe was designed for each of the 4 080 ORFs identified in accordance with the sequences of all published *Yersinia pestis* genomes. Oligonucleotide probes were designed to have similar melting temperatures (*T*m) of 70 ℃-75 ℃ and a length of 35-40 bp. There were 500 factory-quality control spots on the array, which were excluded from further data analyses. One microgram of sonicated DNA from the EV76 strain was labeled with Cy5-ULS by using the Kreatech ULS array CGH Labeling kit (EA-005, Kreatech, Netherlands) according to the manufacturer's instructions and then hybridized to the microarray.

Experiments were repeated three times according to the CombiMatrix protocol PTL_004 and once on the basis of the PTL_006 protocol. Microarrays were pre-hybridized with 6×SSPE containing 0.05% Tween-20, 5× Denhardt's solution, and 100 ng salmon sperm DNA for 30 min at 50 ℃. The Cy5-ULS labeled DNA fragments were then hybridized in the hybridizing solution (6× SSPE and 0.05% SDS) by denaturing at 95 ℃ for 3 min and then incubating for 16 h at 50 ℃. Post-hybridization wash steps were 6× SSPET for 5 min at 50 ℃, 3× SSPET for 1 min, 0.5× SSPET for 1 min, and PBST for 1 min at room temperature. After hybridization and imaging, the microarray was stripped by using the CustomArray™ Stripping Solution according to the manufacturer's protocol.

Microarray Scanning and Data Analyses

Hybridized microarrays were covered with imaging solution and scanned with an Axon GenePix™ 4 000B. The stripped microarray was also scanned with the same PMT value to evaluate the background noise as well as the stripping efficiency. Data were extracted by using Microarray_Imager_5.9.3. The foreground median of each spot was taken into account in the following analysis. Datasets generated from hybridizations were referred to as EV_1, EV_2, EV_3, and EV_4. Correlation coefficients of the four datasets were calculated by using Microsoft Office Excel. Hierarchical clustering was completed with Mev_4_0 (Multiple Experiment Viewer, TIGR). *T*m frequencies of the 300 and 3 000 weak signals and the 300 and 3 000 top signals were calculated by Minitab15 (Minitab Inc, USA). To estimate the false synthesis rate at each position of the probe, we selected the 300 probes with the lowest signal intensity. Percentage histograms were also generated in Excel.

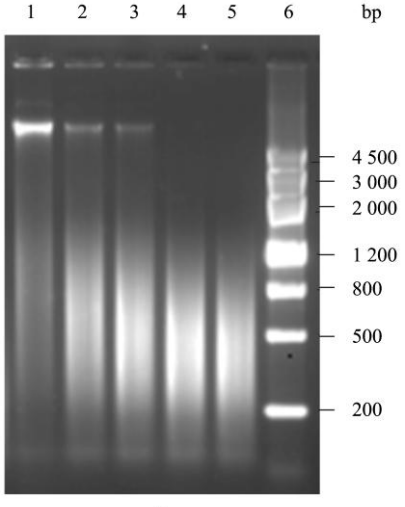
Conformance of Microarray Data

To confirm the results of microarray hybridizations, we randomly selected 46 ORFs from 300 ORFs to be amplified with PCR, as ORFs corresponded to probes with low hybridization signals on the microarray (Table 2).

RESULTS

Optimization of DNA Fragmentation

Different sonication conditions yielded different gel profiles (Fig. 1). With short sonication duration, the sizes of the DNA fragments ranged from 200 bp to 2 kb, and with longer sonication, the sizes ranged from 200 to 800 bp. We chose to use nine seconds of sonication for the first three hybridizations and 180 seconds for the fourth hybridization.



A

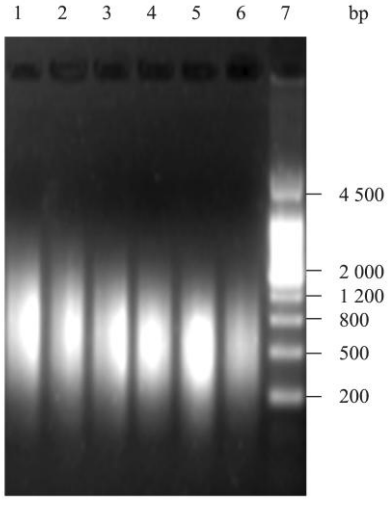


FIG. 1. DNA fragmentation with different sonication durations. A. DNA fragmentation with a short sonication duration. Each sonication lasted three seconds and then samples were cooled on ice for five seconds. DNA fragmentation with different sonication durations. before the next round of sonication. Lane 1, 3 s total sonication time; lane 2, 6 s; lane 3, 9 s; lane 4, 12 s; lane 5, 15 s; and lane 6, molecular marker. B. DNA fragmentation with a long sonication duration. Each sonication lasted five seconds and then samples were cooled on ice for ten seconds before the next round of sonication. Lane 1, 100 s total sonication time; lane 2, 120 s; lane 3, 140 s; lane 4, 160 s; lane 5, 180 s; lane 6, 200 s; and lane 7, molecular marker.

Microarray Hybridization Data

Hybridization results were highly reproducible based on correlation coefficients of the four datasets. Correlation coefficients of EV_1, EV_2, and EV_3 were above 0.95, whereas correlation coefficients between the first three datasets and the EV_4 were lower, with the values ranging from 0.87 to 0.92.Hierarchical clustering showed that signals in EV_4 differed significantly from the other hybridization signals. The only difference between the first three hybridizations and the fourth hybridization was the fragment size of the target DNA.

Relationship between Tm and Signal Intensity

Hybridization signals were sorted by the value of intensity; low and high intensity signals were selected for frequency analyses. For the 3 000 probes with lowest intensities, probe frequencies were significantly higher at 70 ℃-71.5 ℃ than at 72 ℃-74.5 ℃ (Fig. 3A). However, the differences of probe frequencies in the two temperature ranges were not significant for the 3 000 probes with highest intensities (Fig. 3B). We then analyzed the 300 lowest and highest intensity probes, finding that the frequencies for the lowest intensity probes were mainly distributed in the range from 70 ℃ to 72 ℃ and those for the 300 highest intensity probes in the range from 72 ℃ to 74 ℃ (Fig. 3C, D).

Estimation of Possible Synthesis Errors at Each Position of the Probe

There was no significant sequence bias found within probes that would imply errors in incorporation of a particular base resulting in low signal intensities (Fig. 4).

PCR Conformance for Hybridization

To determine whether probes that yielded low signal intensities might be false negatives, 46 PCR primer pairs were used to amplify ORFs that corresponded to low signal intensities. All were amplified, indicating false negative results in the microarray.

DISCUSSIONS

Microarray technology makes the study of thousands of genes simultaneously possible, but only a fraction of genes are differentially expressed and a relatively large portion of probes yield low signal intensities. Such low signal intensities may give rise to erroneous gene expression ratios or false negatives. A careful analysis of such signals before the subsequent analyses is essential. Techniques for determination of the microarray spot accuracy and for identification of the true signals have been suggested in the literature[8-11]. For the data analyses in this study, we first sorted the signal values from the lowest to the highest intensity. It is possible that the probes with the lowest signals might be spots where probe synthesis failed. We analyzed absence of 46 genes corresponding to some of those within the 300 lowest intensity spots by PCR. All were amplified. It is well known that the ORF sequences of *Yersinia pestis* are highly conserved, which suggests a systematic error in this microarray. To determine the cause of error, we first changed the conditions of DNA fragmentation. Results obtained when DNA fragments were shorter, 200-800 bp, rather than 200 bp to 2 kb, did not correlate well with our previous data (Table 1). Hierarchical clustering of the four datasets also showed that EV_4 was different from the other three datasets (Fig. 2), which suggested that DNA fragment size might be important for hybridization. Shorter fragments could probably

provide data with better quality, although further demonstration of this assumption would be needed. Thus optimization of sonication conditions was crucial for data reliability. Some sonicators, including the Hydroshear® (GeneMachines,CA) and the Bioruptor (Diagenode SA, Belgium), have a good

TABLE 1

Correlation Coefficients of the Four Hybridizations

Exp_ID	EV_1	EV_2	EV_3	EV_4
EV_1	1.000000	0.986775	0.951419	0.874462
EV_2	0.986775	1.000000	0.953125	0.868235
EV_3	0.951419	0.953125	1.000000	0.921876
EV_4	0.874462	0.868235	0.921876	1.000000

TABLE 2

Gene Location and PCR Primers for the Conformance of Microarray Results

No.	Gene Locus	Prime (sense)	Prime (anti-sense)	Product Length	PCR Results
1	NC_004088_334453_335229	5ATGATGAACCCGTTGGTC3;	5TTAGACCGAAATACGCG3	777	positive
2	NC_004088_1278040_1278441	5GTGGTTAAGATAAATAGGC3;	5CTATGGACATAGCTTTATATC3	402	positive
3	NC_004088_1373257_1375407	5TTGGCAACGACAAAACTGAACG3	5TTACAGCGAACAGGAAACGCAG3	2151	positive
4	NC_004088_1620799_1621269	5ATGATCAGTGGATCCTTGG3	5CTAGCTATTCAAAAGACTATAATG3	471	positive
5	NC_004088_1786372_1786584	5TTGATGATGGAAATCAGCATC3	5TTATGGACAGGCTCTGGCTT3	213	positive
6	NC_004088_1801819_1802319	5ATGCGTATGTCAACAACCACTG3	5TTATCGGGATTCGTTCGCTT3	501	positive
7	NC_004088_1881980_1882489	5ATGGATATGCTTTCAATATCAT3	5TTATTGCTTAAATTTAATCGAT3	510	positive
8	NC_004088_1992186_1992635	5ATGGACAAAATTGACGAAC3	5TTATATTTTTATCGGTCGAAC3	450	positive
9	NC_004088_2587160_2588500	5ATGGTTAACAGAATAAGCGAT3	5TTACGACGTAACTTTTTGAC3	1341	positive
10	NC_004088_2606479_2606793	5ATGATGAAAATATTGCTGTTAG3	5TTAGGGAGTTTTAGGTTTCG3	315	positive
11	NC_004088_2676774_2677196	5GTGACACCAATCTTTTTCCTTAAC3	5TCATAGCCTATCAGGGGGGGT3	423	positive
12	NC_004088_2678843_2679052	5GTGCCAGATGAAATCGAT3	5TTAACGTCTGCGTTTCTC3	210	positive
13	NC_004088_3221624_3221782	5TTGCGCTTCATCAATATAG3	5TCATTTATTGTCATCTAATGC3	159	positive
14	NC_004088_3874111_3874683	5TTGTTCAGGAAAAATAATACG3	5TTAGCGCCCGATTGTGAT3	573	positive
15	NC_004088_3884192_3885004	5GTGGGCGGTTATTTGAAAAT3	5TTAATGCTGCCTGTTTTTCC3	813	positive
16	NC_004088_4449163_4449468	5ATGGCTAAGCAATCAATGAAAG3	5TTACCAGCTAGCCTTCTTAAGG3	306	positive
17	NC_005810_238683_239630	5GTGTCTGATTCTTTGCGG3	5TTACACTAGTTGACTACCTGGT3	948	positive
18	NC_005810_1054716_1055045	5 ATGATGCCATCTGTGAAAGGTT3	5 TTACAGCTTTCTGATTTCAGTGAT3	330	positive
19	NC_005810_1183861_1184589	5 TTGATGATTTCCCTGAAGAATG3	5 TTAATGCAAGATTTTAGCCAGG3	729	positive
20	NC_005810_1829146_1830495	5 ATGCATCCTATTATCACGCG3	5 TCATTGTATTTTCACCAGAACG3	1350	positive
21	NC_005810_1917572_1918351	5 ATGCGCAGAACGCTCCTTAC3	5 TTATTGGGTCAGTTTTTGCGT3	780	positive
22	NC_005810_2866156_2867757	5GTGTCTTTACCGATTATGTCC3	5TCATAGTCCTGCTGTCCGT3	1602	positive
23	NC_008150_158592_159263	5ATGCCTTATGGTTATGCTTATG3	5TTACTTTTTGCTTTCATTACGG3	672	positive
24	NC_008150_1427652_1428440	5ATGTGCATCCCGCTGTGG3	5TCACCTTTCTGAAGTACTGGG3	789	positive
25	NC_008150_1467892_1468419	5ATGGATTCTTTATTATCTGC3	5TCAGTATCCTGCCATTT3	528	positive
26	NC_008150_1474396_1474839	5ATGCGTTGGTTTAGCGA3	5TTACACCCATCCACTTTGC3	444	positive
27	NC_008150_2543086_2545317	5ATGCGGCAAACAAGCATAGC3	5TCACGCTGTCCGCTCCAT3	2232	positive
28	NC_008149_353782_354975	5CATCCCATGATAAATACA3	5TAGCGATCAGTAGTGTCA3	1100	positive
29	NC_008149_635890_636471	5TTCATGGTTTTAATAGCT3	5ATGATTTCCACCTGATT3	497	positive
30	NC_008149_1287255_1289357	5AAAACGAGGCAGTGAAAAC3	5CTGGCTCCGCACGGCTACT3	2064	positive
31	NC_008149_1535679_1537958	5ATGGACGAACAATTGAAACAG3	5TCAAGGACGAAGAGTAATCG3	2280	positive
32	NC_008149_1655285_1656889	5GCAGTGTCAACAAACGAT3	5CACCAGATAACGGGCAA3	1549	positive
33	NC_008149_2225413_2225994	5ATGGTTGAGAAAAGTATTG3	5CGTCTAACCTGCTCATCC3	572	positive
34	NC_008149_2226092_2226349	5ATGAGTACGTCTGAATTACTC3	5TCATCCACGCGAGACTCT3	258	positive
35	NC_008149_2673205_2673462	5TTGGGCCGTTGGTGGAGATATAAG3	5TTATCCACGTCGATTCACCCCG3	258	positive
36	NC_008149_2953547_2954659	5TCAGCGCCGGTGGCTTG3	5CCACCGCGGCTGACGAT3	1080	positive
37	NC_008149_3165697_3166011	5ATGCAGCAAATGGTAATG3	5TTAATGGAATGGATAGTTTG3	315	positive
38	NC_008149_3183400_3184005	5GTTACTTGGCGCTTCAT3	5TTTAGATCAGTTGGCTCA3	582	positive
39	NC_008149_3188725_3190773	5ATGGTGACTTCAGACATCG3	5TTATTGATCGGGAAATGTC3	2049	positive
40	NC_008149_3268483_3268797	5ATGGCCAAAGCCCGTTTAC3	5TCATGCGGTAATCGATGCAAG3	315	positive
41	NC_008149_3546476_3547594	5CGTGGTACTTGATTGGT3	5GATTGCTTTGTGAGTTGT3	1065	positive
42	NC_008149_4305167_4306459	5ATGAGCAGCAATCAGATATTCCGC3	5TCATTCATTTACAGGCCCCGTTAC3	1293	positive
43	NC_003143_11016_11705	5ATGCAACTAAATACCCAACAGC3	5 TCAGTCTACCGTTTTTAATAGCAC3	690	positive
44	NC_003143_110732_111316	5 ATGCAAAGAGGCTGCTGTT3	5 TCATTGCACCTCATTGGTC3	585	positive
45	NC_003143_119972_120187	5 ATGAAACAAGGTATCCACCCT3	5 TTACTTCTTCGCACCCGGC3	216	positive
46	NC_003143_129479_130315	5 ATGACTCGAGTGATAGTAATC3	5 TTATTGTACGGTTCTCGC3	837	positive

reproducibility and can be standardized, but they are expensive and suitable for large-scale sequencing and microarray centers only. In most labs, a probe sonicator is used; therefore, a more detailed optimization and standardization protocol should be conducted to ensure the quality and reliability of a microarray experiment.

Probe design also affects the data reliability as has been reported for other platforms, such as the Affymetrix system[12]. Probe design generally relies on flexible bioinformatics programs with parameters like probe length, GC content, and melting temperature set by the user. The software finds probes matching with these criteria in the target gene sequences and generates a list of potential probes. In the case of the whole genome of a bacterial pathogen, there are thousands of genes, adding difficulty to probe design. Often the *T*m range is set broadly to allow all genes to be targeted; this may lead to some false results[13]. Although it is difficult to make all probes hybridize within a narrow temperature range in design, the error generated should be modeled.

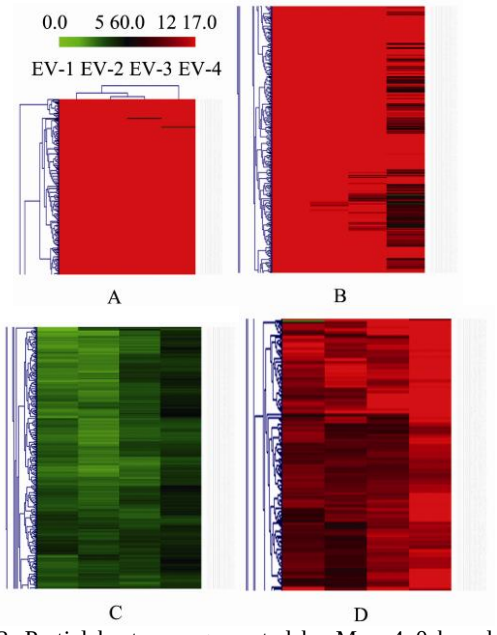
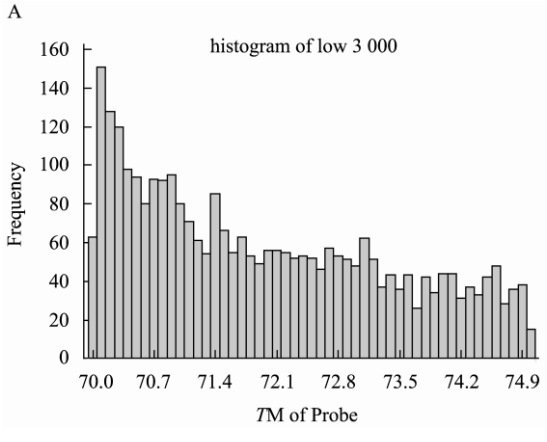


FIG. 2. Partial heatmaps generated by Mev_4_0 based on the four datasets.Lane 1. EV_2, Lane 2. EV_1, Lane 3. EV_3, Lane 4. EV_4. A. The clustering of the four datasets. Panels B, C, and D show that some signals in EV_4 were significantly weakened or enhanced compared with the other data sets.



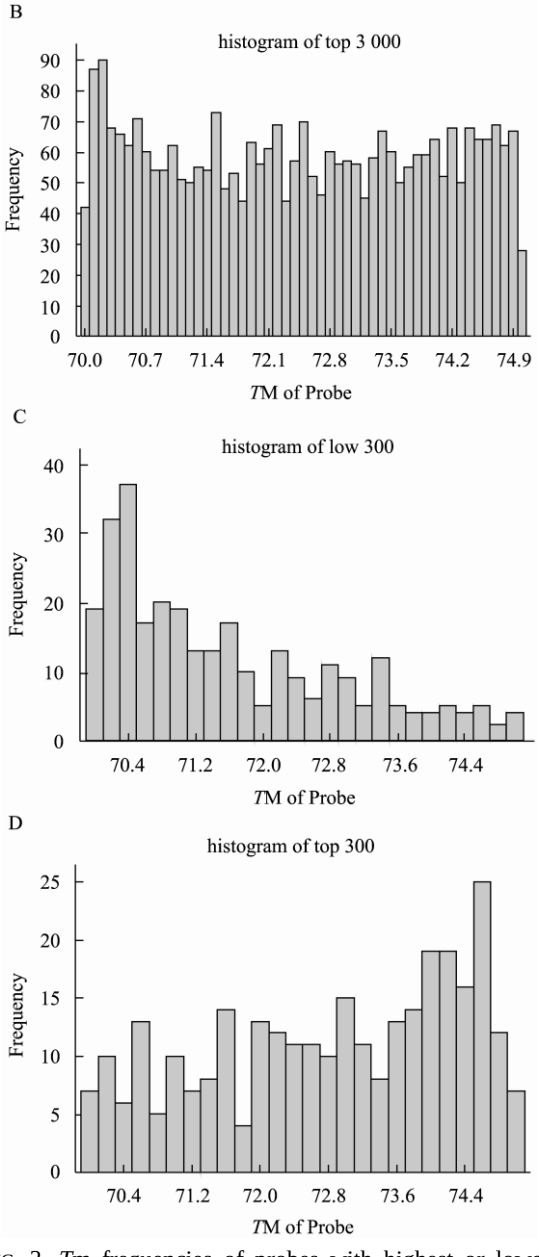


FIG. 3. *T*m frequencies of probes with highest or lowest signals. A, *T*m frequencies of the lowest 3 000 probes with signals ranging from 50 to 288.5. B, *T*m frequencies of 3 000 probes with signals ranging from 915.5 to 65 535. C, *T*m frequencies of the 300 probes with lowest signals ranging from 50 to 98.5. D, *T*m frequencies of the 300 probes with highest intensity signals ranging from 11 865 to 65 535.

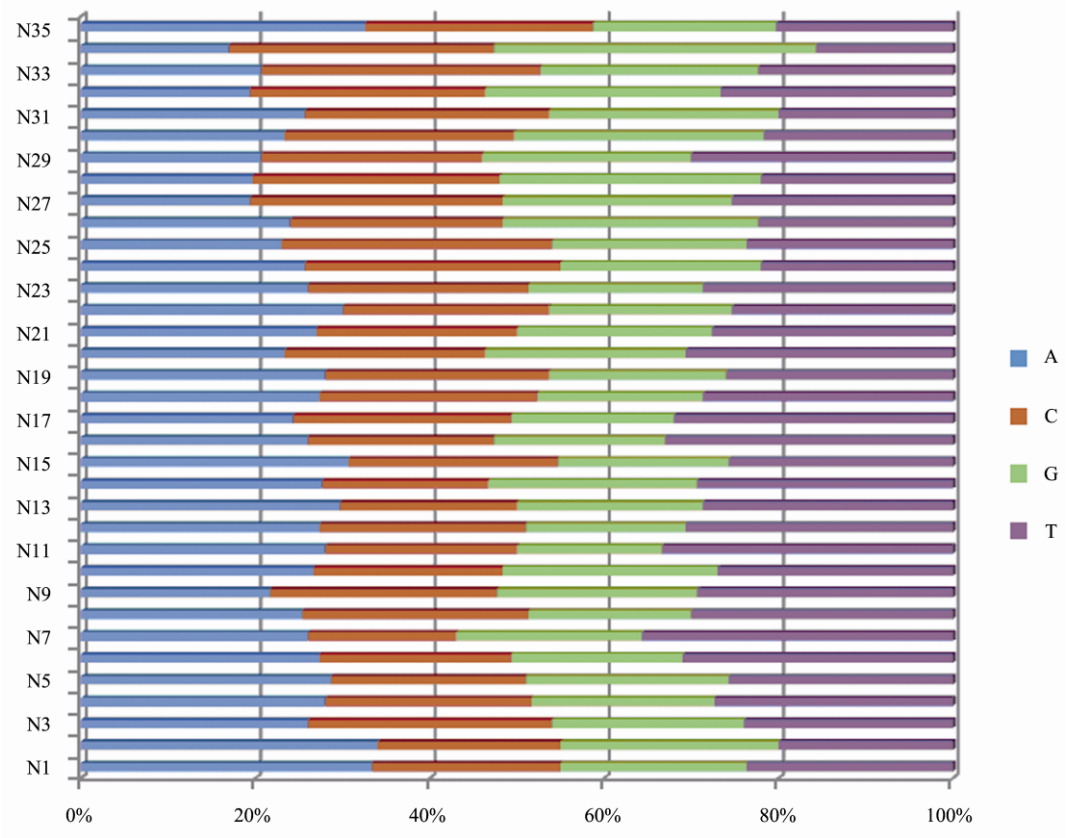


FIG. 4. The proportion of A, C, G, and T at each position within the 35 nt of the 300 probes with lowest signal intensities (ranging from 50 to 98.5).

For our CombiMatrix array, we assessed the relationship between *T*m and signal intensity. There was a different distribution of *T*m in the lowest 300 and highest 300 probes. Signals were weaker with probes that hybridized at temperatures lower than 72 ℃ than with probes that hybridized in a higher temperature range. In this study, the hybridization temperature was set to 50 ℃, the default temperature. A temperature of 50℃ might be too stringent for the probes with *T*ms lower than 72 ℃ and this may have resulted in false negative hybridizations. Consistent *T*ms of probes (a narrow range of 72 ℃-74 ℃) and a lower hybridization temperature (48 ℃) may result in fewer false negative results.

We also made an assessment of the accuracy of electrochemical synthesis used in CombiMatrix CustomArray fabrication. Possible errors may arise from improper control of electrode activation, incorrect deprotection at some positions, or misuse of virtual flasks, which may lead to poor incorporation of A,G, C, or T at certain positions. We examined the percentages of the four bases at each position in the 300 probes with lowest signals and found that there was no bias toward A, G, C, or T in the sequences (Fig. 4).

Through the analysis of the data generated from a CombiMatrix comparative genomic hybridization (CGH) array, we found that all probe *T*ms should be within a narrow range, *T*m and hybridization temperatures should be balanced, and proper target genomic DNA fragments might be in the range from 200 to 800 bp or in a shorter range. These parameters all affected hybridization signal intensity, data reliability, and subsequent bioinformatics analyses and thus should be taken into account in order to obtain accurate and reliable data from *in situ* synthesized microarrays.

REFERENCES

1. Wilkes T, Laux H, Foy C A (2007). Microarray data quality – review of current developments. *OMICS: A Journal of Integrative Biology* **11**(1), 1-13.
2. Eads B, Cash A, Bogart K, Costello J, *et al.* (2006). Troubleshooting microarray hybridizations. *Methods in Enzymology* **411**, 34-49.
3. Copois V, Bibeau F, Bascoul-Mollevi C, *et al.* (2007). Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. *Journal of Biotechnology* **127**(4), 549-559.
4. Archer K J, Dumur C I, Joel S E, *et al.* (2006). Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models. *Biostatistics* **7**(2),198-212.

5. Jones L, Goldstein D R, Hughes G, *et al*. (2006). Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data. *BMC Bioinformatics* **7**, 211.
6. Reimer M, Weinstein J N (2005). Quality assessment of microarrays: visualization of spatial artifacts and quantitation ofregional biases. *BMC Bioinformatics* **6**,166.
7 Li Y, Dai E, Cui Y, *et al*. (2008). Different region analysis for genotyping Yersinia pestis isolates from China. *PLoS One* **3**, e2166.
8. Burgoon L D, Eckel-Passow J E, Gennings C, *et al*. (2005). Protocols for the assurance of microarray data quality and process control. *Nucleic Acids Research* **33(19)**,e172.
9. Bilban M, Buehler L, Head S, *et al*. (2002). Defining signalthresholds in DNA microarrays: exemplary application for invasive cancer. *BMC Genomics* **3**, 19.
10. Tran P H, Peiffer D A, Shin Y, *et al*. (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res* **30**, e54.
11 Heber S, Sick B (2006). Quality assessment of Affymetrix GeneChip data. *OMICS: A Journal of Integrative Biology* **10(3)**, 358-368.
12. Gautier L, Cope L, Bolstad B M, *et al*. (2004). affy: analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**(3),307-315.
13. Bonner T I, Brenner D T, Neufeld B R, *et al*. (1973). Reduction in the rate of DNA reassociation by sequence divergence. *J Mol Biol* **81**, 123-135.

(*Received March* 12, 2010 *Accepted September* 18, 2010)