http://www.jstage.jst.go.jp/browse/biophysics

BIOPHYSICS

Vol. 3, pp. 47–56 (2007)
doi: 10.2142/biophysics.3.47

Insight into the sequence specificity of a probe on an Affymetrix GeneChip by titration experiments using only one oligonucleotide

Shingo Suzuki[1], Chikara Furusawa[1,2], Naoaki Ono[2], Akiko Kashiwagi[1], Itaru Urabe[3] and Tetsuya Yomo[1,2,4]*

[1]*Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan*
[2]*Complex Systems Biology Project, ERATO, Japan Science and Technology Corporation, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan*
[3]*Department of Biotechnology, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan*
[4]*Graduate School of Frontier Biosciences, Osaka University, 1-3 Yamadaoka, Suita, Osaka 565-0871, Japan*

Received 13 November, 2006; accepted 20 July, 2007

High-density oligonucleotide arrays are powerful tools for the analysis of genome-wide expression of genes and for genome-wide screens of genetic variation in living organisms. One of the critical problems in high-density oligonucleotide arrays is how to identify the actual amounts of a transcript due to noise and cross-hybridization involved in the observed signal intensities. Although mismatch (MM) probes are spotted on Affymetrix GeneChips to evaluate the noise and cross-hybridization embedded in perfect match (PM) probes, the behavior of probe-level signal intensities remains unclear. In the present study, we hybridized only one complement 25-mer oligonucleotide to characterize the behavior of duplex formation between target and probe in the complete absence of cross-hybridization. Titration experiments using only one oligonucleotide demonstrated that a substantial amount of intact target was hybridized not only to the PM but also the MM probe and that duplex formation between intact target and MM probe was efficiently reduced by increasing the stringency of hybridization conditions and shortening probe length. In addition, we discuss the correlation between potential for secondary structure of target oligonucleotide and hybridization intensity. These findings will be useful for the development of genome-wide analysis of gene expression and genetic variations by optimization of hybridization and probe conditions.

Key words: microarray, cross-hybridization, probe length, hybridization, probe-level signal

Corresponding author: Tetsuya Yomo, Department of Bioinformatics Engineering, Graduate School of Informaiton Science and Technology, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan.
e-mail: yomo@ist.osaka-u.ac.jp

The Affymetrix GeneChip system, which is one of the major types of DNA microarray, is a powerful tool for the analysis of genome-wide expression of genes[1] and for genome-wide screens of genetic variation and disease-causing mutations[2,3]. One of the greatest problems for all DNA microarray platforms is cross-hybridization because it adds background intensity, which is not related to the true amount of a transcript. Therefore, effective methods of decreasing and/or controlling cross-hybridization are required for accurate microarray assays. The Affymetrix GeneChip system makes use of two types of probe: a perfect match (PM) probe and a corresponding mismatch (MM) probe to estimate noise and cross-hybridization. The terminology for the Affymetrix GeneChip system is omitted in this paper, because it is available in a number of earlier reports[1–5]. Although MM probes are spotted onto GeneChips to evaluate noise and cross-hybridization, GeneChip analyses have shown that a number of MM probes possess greater fluorescence intensity than their cognate PM probes[6,7]. This indi-

cates that the use of MM probes for assessment of nonspecific binding is unreliable. Therefore, several algorithms for analysis using only the PM signal intensities have been developed[6,8,9]. However, it is difficult to account for all of the noise and cross-hybridization using only the PM signal intensity. Recently, several models to account for hybridization on microarrays have been developed based on published data sets of transcriptome analyses[10–15]. However, these models were based on transcriptome analysis that included a background cDNA or cRNA sample in which the concentrations of the various targets were unknown and the sequences were different. The difficulty in using both PM and MM probe signal intensities is mainly because it is not clear what the PM and MM actually do in the GeneChip system, although they were designed with well-defined theoretical expectations. Given this situation, the behavior of probe-level signal intensities was analyzed empirically using well-defined RNAs[16,17]. However, the signal intensities derived from these microarray analyses using well-defined RNAs may involve cross-hybridization, although the amounts of cross-hybridization may be less than in transcriptome analyses. Optimization of probe design and appropriate analysis algorithms require improved understanding of the hybridization behavior between target and probe oligonucleotides under completely controlled conditions.

In the present study, we investigated the hybridization behaviors of PM and MM probes, which play significant roles in the sensitivity, specificity, and reliability of detection as well as for accurate quantification of the abundance of target transcripts. In addition, to reduce cross-hybridization on the microarray, it was important to use probe oligonucleotides to discriminate target and non-target of highly similar sequence. Therefore, we focused on the ratios of the signal intensities of PM to those of cognate MM probes. The PM/MM ratio is an index of specificity[12] and is especially important in re-sequencing experiments in screening of genetic variation. To avoid the effects of noise and cross-hybridization, we hybridized only a single target 25-mer oligonucleotide complementary to the sequence of the spiked probe. This experimental approach using only one target oligonucleotide enables us to identify the absolute signal intensities of probes clearly, even if the concentration of the applied target oligonucleotide is very low. The sequence specificity of probes depends on the hybridization conditions, such as hybridization temperature, duration of hybridization, and washing conditions. Therefore, we evaluated the PM/MM ratios under various hybridization conditions as described above. Our results indicate that a substantial amount of intact target hybridized not only to the PM but also to the MM probe and increases in hybridization stringency and shortening probe length reduced the duplex formation between intact target and MM probe efficiently as compared to the PM probe. In addition, we discuss the correlation between the potential for secondary structure formation of the target oligonucleotide and absolute signal intensity.

Materials and methods

Preparation of biotin-labeled target oligonucleotides

This study was carried out with a GeneChip Test3 Array (Affymetrix, Santa Clara, CA) designed for evaluating target quality and labeling efficiency. This array contains spiked control probes corresponding to five *Bacillus subtilis* genes involved in amino acid biosynthesis, *dap*, *lys*, *phe*, *thr*, and *trp*, along with commonly expressed genes from various organisms, including mammals, plants, and eubacteria. These spiked control probes are derived from 5′, middle, and 3′ portions of the genes. The sequences of 25-mer target oligonucleotides, Dap5-11 (5′-CCGAGCGCAAAATTTGGCGC GATGA-3′), DapM-19 (5′-CATCATCACTGTGGGCGCC AAAAGC-3′), and Dap3-02 (5′-ATATGCGGGCTGCTTC AGCTGCTTC-3′), are complementary to the spiked control probes, AFFX-DapX-5_at No. 11, AFFX-DapX-M_at No. 19, and AFFX-DapX-3_at No. 02, respectively. These target oligonucleotides all have the same GC content (56%). The 19-mer oligonucleotide target, Dap3-02-6nt (5′-ATATGCG GGCTGCTTCAGC-3′), is six bases shorter than Dap3-02 from the 3′ end.

Methods for target preparation, described in the earlier version of the Expression Analysis Technical Manual (Affymetrix, 2001), were followed. Briefly, aliquots of 100 pmol of synthetic oligonucleotide target were labeled at the 3′ end with biotinylated ddUTP (Biotin-16-2′,3′-dideoxyuridine-5′-triphosphate) using a BioArray Terminal Labeling Kit with Biotin-ddUTP (Enzo, Farmingdale, NY) in accordance with the manufacturer's instructions.

Preparation of biotin-labeled background of prokaryotic transcripts

For all experiments that included a background cDNA sample, aliquots of 10 μg of *Escherichia coli* total RNA were used. Briefly, *E. coli* K-12 strain W3110 was grown overnight with shaking at 37°C in 5 ml of liquid Luria-Bertani medium. To maintain logarithmic growth, the overnight cultures were diluted to an optical density at 600 nm (OD_{600}) of 0.05 into 5 ml of fresh liquid Luria-Bertani medium. Then, cultures were grown with shaking at 37°C to an OD_{600} of 0.8. Cells were harvested by centrifugation and stored at −80°C prior to RNA extraction. Total RNA was isolated and purified from cells using an RNeasy mini kit with on-column DNA digestion (Qiagen, Hilden, Germany) in accordance with the manufacturer's instructions. For preparation of cDNA background samples, standard methods for cDNA synthesis, fragmentation, and end-terminus biotin labeling were carried out in accordance with Affymetrix protocols.

Array hybridization, washing, staining, scanning, and data analysis

Hybridization, washing, staining, and scanning were carried out according to the earlier version of the Expression Analysis Technical Manual (Affymetrix, 2001). Briefly, the labeled target oligonucleotide was diluted in hybridization cocktail containing 1×manufacturer's recommended buffer (100 mM MES, 1 M NaCl, 20 mM EDTA, and 0.01% Tween-20), 50 pM B2 Control Oligo, 0.1 mg/mL herring sperm DNA, and 0.5 mg/mL BSA, in tenfold dilutions such that the labeled oligonucleotide target would yield final concentrations of 1.4 nM to 140 aM. In experiments that included cDNA background, aliquots of 1.2 μg of labeled cDNA were added to the hybridization cocktail. The labeled and diluted target oligonucleotide samples with or without background cDNA were hybridized to GeneChip Test3 Arrays at 45°C for 16 h in a Hybridization Oven 640 (Affymetrix) set at 60 rpm under standard conditions. In experiments for evaluation of hybridization conditions, hybridization temperature and time were varied as described in each section. After hybridization, washing and staining procedures were carried out with the Fluidics Station 450 using Micro_1v1_450 fluidics script (Affymetrix) under standard conditions. In experiments for evaluation of washing conditions, the arrays were washed and stained using Flex_Micro_1v1_450 fluidics script, in which the number of stringent wash cycles was increased from 8 (default) to 30. Following washing and staining procedures, the arrays were scanned using a GeneChip Scanner 3000 (Affymetrix). All GeneChip experiments were performed in duplicate using two different biotin-labeled target oligonucleotides prepared separately for each sample. Absolute signal intensities of all probes in all samples were generated using GCOS 1.0 software (Affymetrix). Duplicate measurements were averaged to obtain a single absolute signal intensity for each target.

Results

Absolute probe signal intensities

To characterize the absolute signal intensities of perfect match (PM) and mismatch (MM) probes precisely, target oligonucleotides labeled at the 3′ end with biotin were hybridized to the GeneChip Test3 Array (Affymetrix) with and without cDNA background generated from *Escherichia coli* total RNA under standard hybridization conditions. We chose three target oligonucleotides, Dap5-11, DapM-19, and Dap3-02, which were complementary to the spiked control probes, AFFX-DapX-5_at No. 11, AFFX-DapX-M_at No. 19, and AFFX-DapX-3_at No. 02, respectively. These target oligonucleotides had the same GC content of 56% to exclude the effect of GC content-dependent hybridization strength. Recently, hybridization models based on the nearest neighbor model[18] were reported[10,12,13], and we predicted Gibbs free energy of target oligonucleotides. The nearest neighbor model predicted that the potentials for duplex formation, ΔG, of oligonucleotide targets, Dap5-11, DapM-19, and Dap3-02, were −31.3, −29.7, and −30.0 kcal/mmol, respectively. Figures 1A and 1B show the signal intensities and ratios of signal intensity of PM to that of cognate MM of AFFX-DapX-5_at No. 11, AFFX-DapX-M_at No. 19, and AFFX-DapX-3_at No. 02 probe pairs as a function of target oligonucleotide concentration with and without cDNA background. Although all three target oligonucleotides with cDNA background were applied to the Test3 Array at once, only one target oligonucleotide was hybridized to the array without cDNA background to avoid the effects of cross-hybridization completely. Both with and without cDNA background, the line plots of intensity *vs.* target concentration showed the typical sigmoidal shape encountered in chemical kinetics and the signal intensity was saturated at a target concentration of 140 pM. However, the detection limit depended heavily on the abundance of the background cDNA. Although the detection limit was at a concentration of *ca.* 1.4 fM without background cDNA, addition of background cDNA caused a shift in the detection limit to *ca.* 1.4 pM. It follows that linearity was changed from the target oligonucleotide concentration range of 14 fM–14 pM to 1.4 pM–14 pM due to addition of cDNA background, depending on the variety of the target sequences. These results indicated that cDNA background conceals the behavior of target-probe hybridization at less than or equal to a target concentration of 140 fM. Particularly in the case of Dap3-02, substantial signal intensities derived from background cDNA were observed, although the three target oligonucleotides used in the present study had the same GC content of 56% and similar potentials for duplex formation. Moreover, in the case of Dap5-11, signal intensities of MM probes were stronger than those of PM probes in the target oligonucleotide concentration range of 1.4 fM–140 fM. Our findings further supported the importance of evaluation of the background signal embedded in PM probes. Consequently, we focused on analysis without cDNA background to characterize the hybridization behavior of PM and MM probes in the low concentrations of target oligonucleotides. It is worth noting that signal intensities of MM probes increased with increasing target concentration despite the complete absence of cross-hybridization (Fig. 1B). This result clearly indicated that a substantial amount of intact target was hybridized not only to the PM but also to the MM probe, even if the target oligonucleotides were applied at low concentrations. In Figure 1B, the horizontal position of the curves reflects differences in target-binding strength for hybridization among them, although the target oligonucleotides had the same GC content and similar potentials for duplex formation. Although there was a small difference in the potentials for duplex formation of oligonucleotide targets, the order of predicted free energy does not seem to reflect the observed signal intensity. Possible reason for this discrepancy is addressed in *Discussion*. The bar graph shows the ratios of signal intensities of PM to those of

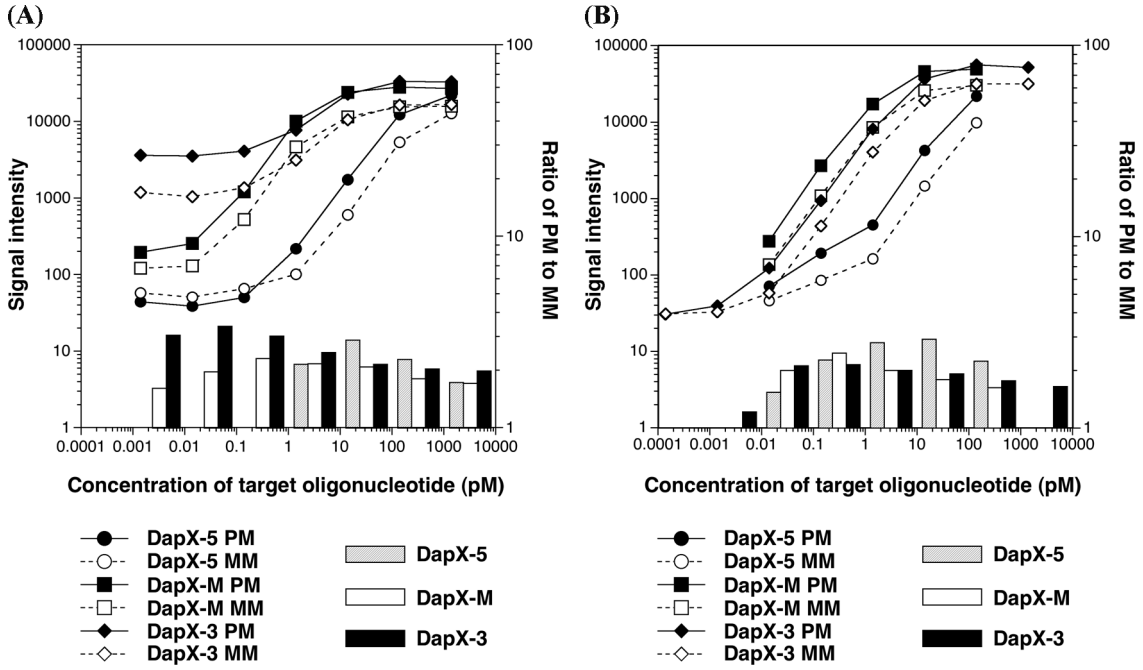


Figure 1 Absolute signal intensities of perfect match (PM) and mismatch (MM) probes, such as AFFX-DapX-5_at No. 11, AFFX-DapX-M_at No. 19, and AFFX-DapX-3_at No. 02, which were identified by independent titration assay of hybridization with DNA target oligos, Dap5-11, DapM-19, and Dap3-02, respectively. Bar graphs show the ratios of signal intensity of PM to those of cognate MM probes. The average signal intensities determined using GCOS 1.0 software were derived from two replicate GeneChip analyses. (A) PM, MM signals, and PM/MM ratios with cDNA background. (B) PM, MM signals, and PM/MM ratios without cDNA background.

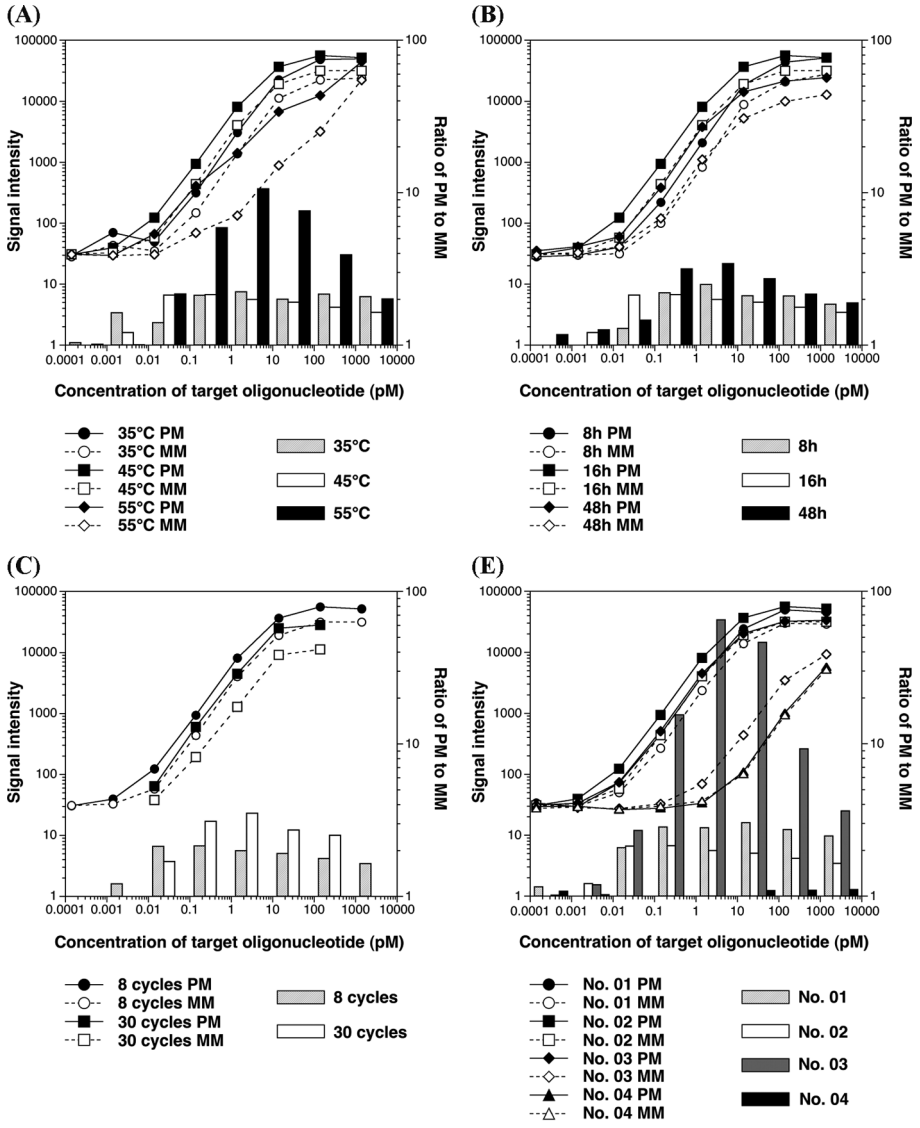
cognate MM probes. The PM/MM ratio is an index of specificity and is especially important in re-sequencing experiments in screening of genetic variation. Although the ratios of signal intensity of PM to that of cognate MM were almost 2, there was little variation. It is likely that the largest PM/MM ratio was correlated with target-binding strength. Briefly, the PM/MM ratio of DapM-19, which shows high target-binding strength, was the largest at the comparatively low target concentration of 140 fM, while that of Dap5-11, which shows low target-binding strength, was largest at the high target concentration at 14 pM.

The MM probe is used to quantify the background noise and cross-hybridization embedded within the signal intensity of the PM probe. Typical algorithms that use GeneChip data to infer quantitative transcript expression levels begin by subtracting the intensity of the MM probe signal from that of the cognate PM probe (with adjustments to the MM value if MM > PM)[19]. Our results indicated that substantial amounts of intact target were hybridized to the MM probe, even if the target oligonucleotides were applied at low concentrations. In addition, the observation that the intensities of MM probes were close to those of PM even without a chemical background implies that the reversal of signal intensities between PM and MM probes is very easy.

Effects of hybridization and washing conditions

To characterize the basic behavior of target-probe hybridization, different hybridization temperatures, durations of hybridization, and number of stringent washing cycles were tested. The results were obtained by titration assay of hybridization with Dap3-02 target oligonucleotide, which was representative of the performance of other target oligonucleotides, without cDNA background to avoid the effects of cross-hybridization completely. Figure 2A shows the effects of different hybridization temperatures, such as 35°C, 45°C (standard conditions), and 55°C, for 16 h on absolute signal intensities and PM/MM ratios of AFFX-DapX-3_at No. 02 probe pairs as a function of target oligonucleotide concentration. Unexpectedly, hybridization at 35°C led to a decrease in signal intensities in the target concentration range of 14 fM to 14 pM in comparison to that at 45°C. This result suggests that it is insufficient for 16 h of hybridization to be equilibrated at 35°C. In the case of hybridization at 55°C, a significant increase in PM/MM ratio was observed in the target oligonucleotide concentration range of 140 fM to 140 pM. However, increasing the hybridization temperature to 55°C increased the PM/MM ratio but also decreased the overall signal intensities as expected. In addition, increasing the hybridization temperature to 65°C destroyed the arrays by melting the adhesive fixing the quartz wafer on the cartridge (data not shown).

Figure 2B shows the effects of different durations of hybridization, such as 8 h, 16 h (standard conditions), and 48 h, at 45°C on signal intensities that were increased by hybridization with target oligonucleotide, Dap3-02, as a function of target oligonucleotide concentration. Hybridiza-



(D)

```
Target  5'---AAGAAGAAGCAGCTGAAGCAGCCCGCATATTAGGCGCAGATA---3'

No.01   5'  AAGAAGAAGCAGCTGAAGCAGCCCGC  3'                21
             **********T***********
No.02       5'  GAAGCAGCTGAAGCAGCCCGCATAT  3'             25
                ***********T**************
No.03           5'  GCTGAAGCAGCCCGCATATTAGGCG  3'         19
                    **************T*******
No.04               5'  GCAGCCCGCATATTAGGCGCAGATA  3'  13
                        ************T
Dap3-02     3'  CTTCGTCGACTTCGTCGGGCGTATA  5'
```

Figure 2 Effects of variation of hybridization, wash, and probe conditions on signal intensities identified by titration assay of hybridization with target oligonucleotide, Dap3-02, without cDNA background. The line plots show absolute signal intensities of PM and MM probes. The bar graphs show ratios of signal intensity of PM to that of cognate MM. The average signal intensities determined using GCOS 1.0 software were derived from two replicate GeneChip analyses except for the experiment regarding wash conditions. The signal intensities under stringent wash conditions were derived from a single GeneChip analysis. (A) Hybridization temperature. (B) Duration of hybridization. (C) Number of stringent wash cycles. (D) Setting of adjacent probe pairs of AFFX-DapX-3_at No. 02, such as AFFX-DapX-3_at No. 01, No. 03, and No. 04, and the possible duplexes formed between target oligonucleotide, Dap3-02, and the adjacent probes. The constitutive part of the sense strand of the spiked probe, *Dap*, is shown as the target sequence. The middle four sequences are adjacent probe pairs and the flip-flop position is underlined. The possible base pairs hybridizing with target oligonucleotide, Dap3-02, are indicated by asterisks and the numbers beside the sequences indicate the numbers of hybridizing base pairs. The sequence of the target oligonucleotide, Dap3-02, is shown at the bottom. The gray nucleotides in the 3′ end of Dap3-02 indicate the excised six bases in the target oligonucleotide, Dap3-02-6nt, from which results are shown in Figures 3 and 4B. (E) Absolute signal intensities of adjacent probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04.

tion for 8 h caused a decrease in signal intensity in the target concentration range of 14 fM to 14 pM analogous to the decrease in hybridization temperature. This result also indicated that it is insufficient for 8 h of hybridization to be equilibrated at 45°C. Similar to the increase in hybridization temperature, prolongation of the duration of hybridization to 48 h increased PM/MM ratio in the target oligonucleotide concentration range of 140 fM to 14 pM. However, the increase in PM/MM ratio was much larger under conditions of high temperature than with a long duration of hybridization. Unexpectedly, hybridization for 48 h led to a decrease in signal intensities over the whole range of target concentrations tested. These results, which were observed under conditions of low temperature and short hybridization period, suggested that for adequate hybridization it would be necessary to incubate arrays at 45°C for 16 h.

Figure 2C shows the effects of increasing the number of stringent washing cycles on signal intensities that were increased by hybridization with target oligonucleotide, Dap3-02, as a function of target oligonucleotide concentration. The standard protocol for the GeneChip Test3 Array includes 8 cycles of washing of arrays in stringent washing buffer at 50°C. Thirty cycles of stringent washing increased the PM/MM ratio due to effective dissociation of target oligonucleotides binding to MM probes. It follows that increasing the number of cycles of stringent washing decreases the overall signal intensity.

Absolute signal intensities of adjacent probe pairs.

On the GeneChip Test3 Array, spiked control genes are represented by 20 different probe pairs. Several probe pairs have overlaps with some adjacent probe pairs. Figure 2D shows the setting of adjacent probe pairs of AFFX-DapX-3_at No. 02, such as AFFX-DapX-3_at No. 01, No. 03, and No. 04, and the possible duplexes formed between target oligonucleotide, Dap3-02, and the adjacent probes. These probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04, have overlaps of stretches of 21, 19, and 13 nucleotides, respectively, with AFFX-DapX-3_at No. 02 probe pairs. This setting of probes with overlaps enables us to analyze the behavior of duplex formation of partial hybridization. In principle, it is possible to identify absolute signal intensities of adjacent PM and MM probe pairs with poor hybridization specificity and sensitivity based on the thermodynamic properties of hybridization. Figure 2E shows the absolute signal intensities of adjacent probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04, which were increased on titration assay of hybridization with Dap3-02 target oligonucleotide without cDNA background. Equivalent results were obtained for the two other target oligonucleotides studied (data not shown). In the present study, signals of probes with overlaps of stretches of less than 12 nucleotides were not observed (data not shown). Although probes with long overlaps yield better signal intensity than those with short overlaps, the signal intensities of AFFX-DapX-3_at No. 01

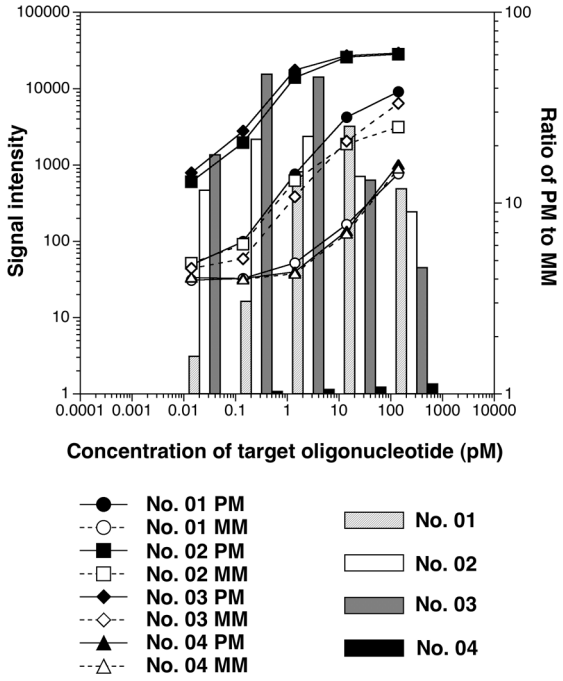


Figure 3 Effects of shortening target length on absolute signal intensities without cDNA background. The line plots show absolute signal intensities of PM and MM probes, which were increased by titration assay of hybridization with target oligonucleotide, Dap3-02-6nt. The line plots show absolute signal intensities of PM and MM probes. The bar graphs show ratios of signal intensity of PM to those of cognate MM probes. The average signal intensities determined using GCOS 1.0 software were derived from two replicate GeneChip analyses.

and No. 03 PM probes were slightly lower than that of AFFX-DapX-3_at No. 02. It is worth noting that slightly shorter probe pairs (*i.e.*, AFFX-DapX-3_at No. 01 and No. 03) showed significantly large PM/MM ratios under standard hybridization conditions. Particularly, the PM/MM ratio of AFFX-DapX-3_at No. 03 was 30-fold greater than that of No. 02. These results indicate that use of a slightly shorter probe would be able to reduce the number of MM probes showing greater fluorescence intensity than the cognate PM probes under our standard conditions.

Next, to confirm that the number of base pairs involved in duplex formation was important for increasing the PM/MM ratio, the 19-mer oligonucleotide target, Dap3-02-6nt, which was six bases shorter than Dap3-02 from the 3′ end, was hybridized to the GeneChip Test3 Array at 45°C for 16 h. Figure 3 shows the absolute signal intensities of adjacent probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04, that were increased by the titration assay of hybridization with 19-mer Dap3-02-6nt target oligonucleotide. The bar graph shows increases in PM/MM ratios in AFFX-DapX-3_at No. 01 and No. 02 probe pairs. Although the number of base pairs involved in duplex formation decreased, the PM signal intensities of AFFX-DapX-3_at No. 02 increased in the target concentration range of 14 fM to 1.4 pM, in contrast to our expectations.

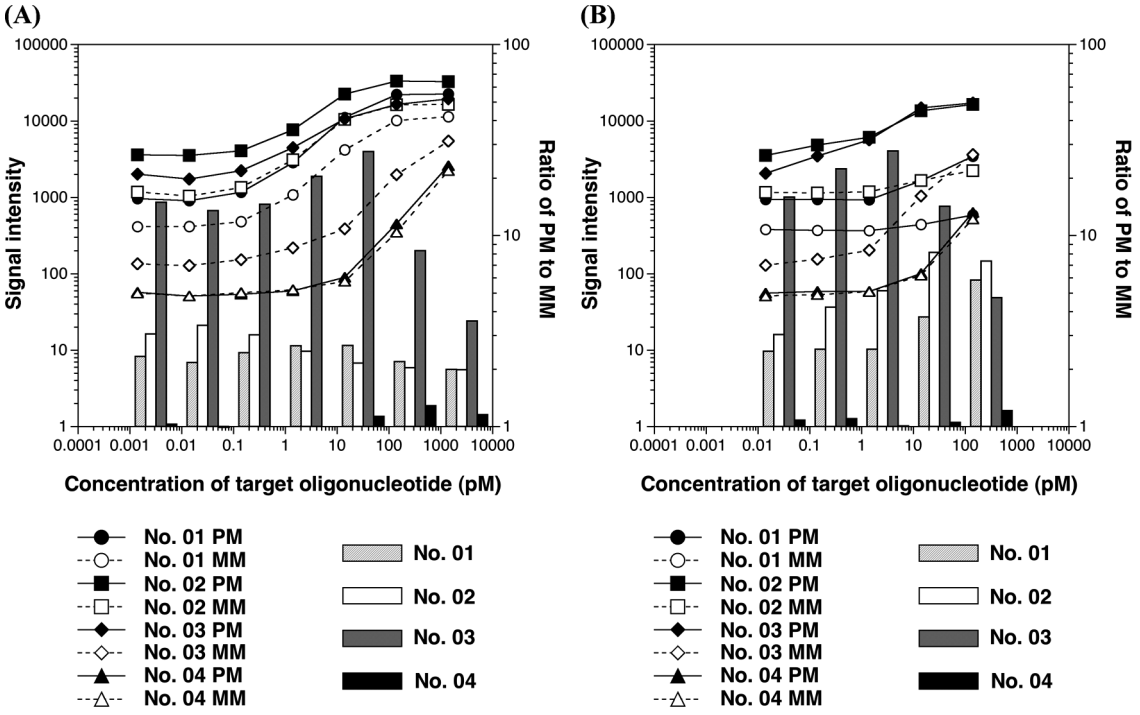


Figure 4 Absolute signal intensities of adjacent probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04 with cDNA background. (A) PM, MM signals, and PM/MM ratios, which were demonstrated by the titration assay of hybridization with target oligonucleotide Dap3-02, with cDNA background. The line plots show absolute signal intensities of PM and MM probes. The bar graphs show ratios of signal intensity of PM to those of cognate MM probes. The average signal intensities determined using GCOS 1.0 software were derived from two replicate GeneChip analyses. (B) Effects of shortening target length on absolute signal intensities with cDNA background. The line plots show absolute signal intensities of PM and MM probes, which were increased by titration assay of hybridization with target oligonucleotide, Dap3-02-6nt. The bar graphs show the ratios of signal intensity of PM to those of cognate MM probes. The average signal intensities were determined as described above.

Effects of cDNA background on absolute signal intensities of adjacent probe pairs

Our findings that optimization of probe length and target length had a marked impact on specificity of duplex formation in the complete absence of cross-hybridization. These conditions were very different from the standard conditions of genome-wide analyses of gene expression and genetic variation. To validate our findings under standard hybridization conditions, we hybridized the target oligonucleotide in the presence of the complex cDNA background. Figures 4A and 4B show the absolute signal intensities of adjacent probe pairs, AFFX-DapX-3_at No. 01, No. 03, and No. 04, that were increased by titration assay of hybridization with target oligonucleotides Dap3-02 and Dap3-02-6nt, respectively, with the cDNA background. As shown in Figure 1A, the cDNA background concealed the behavior of target-probe hybridization at less than or equal to a target concentration of 140 fM, and we therefore focused on the signal intensities obtained from the titration experiments above the target concentration of 1.4 pM. In the case of both Dap3-02 and Dap3-02-6nt, similar results were obtained in the presence of cDNA background. However, the addition of cDNA background reduced the PM/MM ratios slightly, although the absolute signal intensities of both PM and MM probes were decreased. It is expected that the addition of cDNA background would increase the absolute signal intensities, and the absolute signal intensities at the low target concentrations increased markedly. These results indicated that the effective target concentrations were reduced by target-target interactions between the target oligonucleotide and cDNA background in the hybridization solution.

Discussion

The results of the present study revealed the behavior of probe-level signal intensities with and without complex cDNA background. Whereas earlier studies demonstrated the effects of hybridization conditions and probe length[20], the experimental design of the present study was significantly different from those of earlier studies to characterize the behavior of probe-level signal intensities of widely used Affymetrix GeneChips. Affymetrix provided the publicly available results of experiments to characterize the probe-level behavior of hybridization (see www.affymetrix.com/support/technical/sample_data/datasets.affx). However, as these experimental studies were performed using well-defined RNAs spiked into transcriptome, a certain amount of cross-hybridization was inevitable. In such analyses, quantification of a small difference in signal intensity between PM and MM probes can be difficult due to the presence of

cross-hybridization. In the present study, the use of artificially synthesized oligonucleotides only allowed us to quantify the absolute signal intensity without the effect of cross-hybridization, especially when the target concentration is low.

Although the three target oligonucleotides used in the present study had the same GC content of 56% and similar potentials for duplex formation, various absolute signal intensities were observed (Fig. 1B). This may be explained by secondary structure formation of the targets[21–23]. From this viewpoint, we predicted the possible intramolecular structures of three targets by Mfold analysis[24]. The target oligonucleotides were assumed to be linear and the ionic conditions were set at 1 M sodium ions. The potential for secondary structure, ΔG, of oligonucleotide targets, Dap5-11, DapM-19, and Dap3-02, were −3.0, 0.3, and −2.2 kcal/mmol, respectively. That is, the target strands of Dap5-11 can form stable folded structures, which will obstruct duplex formation and reduce signal intensity. This result suggests that the intramolecular structure would reduce the affinity of the target to specific probes and it is very important for prediction of target concentration to evaluate not only hybridization energy between target and probe but also the potential for secondary structure of probe and/or target. Recently, the effects of secondary structure of target and probe have also attracted a great deal of attention[21–23]. However, secondary structure has not been considered in software applications for gene expression analysis in Affymetrix GeneChips.

MM probes were introduced as nonspecific hybridization controls, based on the idea that the true signal would be proportional to the difference between PM and MM signal intensities. Whereas previous studies using well-defined RNAs implied that intact target would also hybridize to MM probes if the targets were applied at high concentration[17], our GeneChip analysis clearly indicated that a substantial amount of intact target was hybridized to the MM probe not only at high target concentration but also at low target concentration. This may be one of the causes of the reversal of signal intensities between PM and MM probes in especially low target concentrations. This reversal could prevent the typical algorithms using the intensities of MM probe from being sensitive to high-precision assessment of expression levels for genes. In fact, there have been several reports of analysis algorithms using only PM signal intensities[6,8,9]. One of the most critical problems in the GeneChip system is how to deal with cross-hybridization, which produces spurious data. That is, for identification of the true amounts of a transcript, it is important to subtract the noise and cross-hybridization involved in the observed signal intensities of the PM probe on GeneChip analysis[25]. Recently, several models to account for hybridization on microarrays have been developed based on published data sets of transcriptome analyses[6–15]. Unlike the present study, these models were based on transcriptome analysis that included a background cDNA or cRNA sample in which the concentrations of the various targets were unknown and their sequences were different. This complex background conceals the behavior of target-probe hybridization and prevents us from understanding the detailed thermodynamic properties of hybridization on microarrays. Therefore, to characterize the probe-level behavior of hybridization in detail, it is necessary to avoid the effects of cross-hybridization on the signal intensities of probes. It appears that our data are insufficient to improve the hybridization model, because we analyzed only four target oligonucleotides. Further experimental calibration would be required to improve the accuracy of the hybridization models on microarrays. To further improve the hybridization models, we are currently performing titration experiments to examine the hybridization properties of at least 100 target oligonucleotides.

We also found that changes in hybridization and stringent washing conditions affected the PM/MM ratio. The PM/MM ratio can represent specificity in GeneChip analysis and it is especially important for screening of genetic variation. It also appears that an increase in PM/MM ratio can reduce the number of MM probes that have greater fluorescence intensity than their cognate PM probes. Therefore, the increase in PM/MM ratio also leads to improvement of the analysis of genome-wide expression of genes in GeneChip systems. Figures 2A, 2B, and 2C show negative correlations between signal intensities and PM/MM ratio as a function of hybridization stringency. In fact, the updated hybridization solutions included 7.8% dimethylsulfoxide, as described in the Expression Analysis Technical Manual (Affymetrix, 2004). With regard to changes in hybridization conditions, the decrease in signal intensities due to prolongation of the duration of hybridization to 48 h is puzzling (Fig. 2B). This phenomenon was not observed in the previous study[20]. The decrease in the overall signal intensities for 48 h could be explained by breakage of the coating and/or linker moiety by long incubation because it is recommended that the GeneChip should be stored at 4°C.

An important observation in the present study is the increase in PM/MM ratio with decreases in probe and/or target length. This phenomenon was confirmed not only in the complete absence but also in the presence of complex cDNA background. This was particularly clear with probe or target sequences 19 nucleotides in length (Fig. 2E, 3, 4A, and 4B). These results also suggest that not only probe length but also flip-flop position may affect duplex formation between target oligonucleotide and probe, because the signal intensity of the AFFX-DapX-3 No. 03 MM probe was very similar to that of the AFFX-DapX-3 No. 04 probe pair (Fig. 2E). This finding indicates that the behavior of duplex formation between the target and AFFX-DapX-3 No. 03 MM probe is similar to that of the AFFX-DapX-3 No. 04 probe pair, although the predicted numbers of hybridizing base pairs of AFFX-DapX-3 No. 03 MM, AFFX-DapX-3 No. 04 PM, and MM probe are 18, 13, and 12,

respectively. In comparison with the AFFX-DapX-3_at No. 04 probe pair, the difference in signal intensities between PM and MM was very small because the effect of a mismatch at the end of a hybridizing target may be small. In the case of the AFFX-DapX-3 No. 03 MM probe, it was suggested that only the center 12 nt, complementary to the 5′ end of the probe sequence, could contribute to duplex formation. That is, the repelling force of the mismatched base pair could prevent the 6-nt sequence at the 5′ end of the target oligonucleotide, complementary to the 3′ end adjacent to the center nucleotide of the probe, from undergoing duplex formation. However, this effect of the repelling force of the mismatched base pair was not observed with hybridization of the 19-mer oligonucleotide target (Fig. 3). Shortening of the target oligonucleotide changed the behavior of duplex formation between the target and AFFX-DapX-3 No. 03 MM probe. This may have been because the 6-nt sequence at the 3′ end of the target, which was predicted not to hybridize, also affected duplex formation of the 6-nt sequence of 5′ end of the target oligonucleotide, complementary to the 3′ end adjacent to the center nucleotide of the probe. Shortening of the target oligonucleotide also changed the behavior of duplex formation between the target and AFFX-DapX-3_at No. 02 PM probe. Although the number of base pairs involved in duplex formation decreased, the PM signal intensities of AFFX-DapX-3_at No. 02 increased in the target concentration range of 14 fM to 1.4 pM (Fig. 2E and 3). This result suggests that shortening the target length decreases the molecules that cross-hybridize to other probes. Similar results were observed for AFFX-DapX-3_at No. 03. It is unusual for targets to be shorter than probes and part of the probe oligonucleotide hybridize to the specific target because it is recommended that the target samples are fragmented to 50–200 bases in length. However, these observations suggest that the fragmentation pattern has a significant influence on the accuracy of GeneChip analysis. Although this is consistent with the previous report[26], the underlying mechanisms are still unclear. Further studies, such as hybridization of only short targets in a fixed length isolated from complex fragments by high-performance liquid chromatography or polyacrylamide gel electrophoresis, are necessary to understand the hybridization behaviors of PM and MM probes under standard conditions of genome-wide analyses of gene expression and genetic variation.

For microarray data analysis, it is important to avoid cross-hybridization of highly similar sequences. In the present study, we demonstrated that an increase in hybridization stringency or use of a 19-mer probe or target reduced duplex formation of intact target and MM probe in the complete absence of cross-hybridization. Our findings will be useful for the design of high-precision GeneChip analysis methods. In addition, our findings will also assist in the improvement of detection methods of SNPs based on microarray technology because the contribution of a single-base mismatch to signal intensity is enhanced as the probe length is shortened. It is expected that the detection sensitivity of shorter probes will be lower[27] and much larger numbers of shorter probes are needed for accurate transcriptional profiling as compared with longer probes. However, cross-hybridization decreases as probe length is reduced[28]. Therefore, shorter probes work well in genome-wide expression analysis of genes if multiple probes are used per gene[29]. Recently, tiling arrays have been used for several types of study, such as non-biased transcriptome analysis, novel gene discovery, analysis of alternative splicing, mapping of regulatory DNA motifs using chromatin immunoprecipitation, and whole-genome DNA methylation analysis[30–32]. As tiling arrays are made with probes tiled throughout the genome without any reference to coding or non-coding regions, shorter probes should also work well in such analyses. Shorter probes may raise concerns about reduction of sequence specificity to the complex target. To estimate sequence specificity, we calculated the redundancy of sequences throughout the *E. coli* genome sequence in lengths of 14-mer to 25-mer. Although the redundancy of sequences increases significantly when the probe is too short, 19-mer probes seemed to be of sufficient length. For example, more than 97% of sequences are unique if the length was longer than 18-mer, and most of the remaining non-unique sequences likely came from repeating sequences, which should be removed from the analysis.

In the present study, we demonstrated the behavior of probe-level signal intensity in the complete absence of cross-hybridization. Our findings suggest that optimization of probe conditions, such as probe length and mismatch position, has a greater impact on specificity than increases in stringency of hybridization conditions, such as hybridization temperature, duration of hybridization, and number of washing cycles. Moreover, a clear relationship was identified between potential for secondary structure formation and absolute signal intensity. The data presented here will assist in probe design for microarray analyses for genome-wide gene expression and screening of genetic variation and in improvement of the hybridization model on microarrays.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research (A) (no. 15207020) from the Japan Society for the Promotion of Science, “The 21st Century Center of Excellence Program”, “Special Coordination Funds for Promoting Science and Technology: Yuragi Project” and “Global COE (Centers of Excellence) Program” of the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

1. Noordewier, M. O. & Warren, P. V. Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol.* **19**, 412–415 (2001).

2. Hacia, J. G. Resequencing and mutational analysis using oligonucleotide microarrays. *Nat. Genet.* **21**, 42–47 (1999).
3. Mir, K. U. & Southern, E. M.: Sequence variation in genes and genomic DNA: methods for large-scale analysis. *Annu. Rev. Genomics Hum. Genet.* **1**, 329–360 (2000).
4. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24 (1999).
5. Affymetrix Microarray Suite 5.0 User's Guide (2001).
6. Zhou, Y. & Abagyan, R. Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics* **3**, 3 (2002).
7. Naef, F., Lim, D. A., Patil, N. & Magnasco, M. DNA hybridization to mismatched templates: a chip study. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **65**, 040902 (2002).
8. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. & Speed, T. P. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
9. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
10. Zhang, L., Miles, M. F. & Aldape, K. D. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* **21**, 818–821 (2003).
11. Hekstra, D., Taussig, A. R., Magnasco, M. & Naef, F. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.* **31**, 1962–1968 (2003).
12. Binder, H. & Preibisch, S. Specific and non specific hybridization of oligonucleotide probes on microarrays. *Biophys. J.* **89**, 337–352 (2005).
13. Wu, C., Carta, R. & Zhang, L. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.* **33**, e84 (2005).
14. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).
15. Naef, F. & Magnasco, M. O. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **68**, 011906 (2003).
16. Relógio, A., Schwager, C., Richter, A., Ansorge, W. & Valcárcel, J. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* **30**, e51 (2002).
17. Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K. & Kreder, D. E. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* **3**, RESEARCH0005.1–RESEARCH0005.10 (2002).
18. SantaLucia, J. Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95**, 1460–1465 (1998).
19. Affymetrix Statistical Algorithms Reference Guide (2001).
20. Forman, J. E., Walton, I. D., Stern, D., Rava, R. P. & Trulson, M. O. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesized oligonucleotide Arrays. *ACS Symp. Ser.* **682**, 206–28 (1998).
21. Karaman, M. W., Groshen, S., Lee, C. C., Pike, B. L. & Hacia, J. G. Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. *Nucleic Acids Res.* **33**, e33 (2005).
22. Ratushna, V. G., Weller, J. W. & Gibas, C. J. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* **6**, 31(2005).
23. Koehler, R. T. & Peyret, N. Effects of DNA secondary structure on oligonucleotide probe binding efficiency. *Comput. Biol. Chem.* **29**, 393–397 (2005).
24. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
25. Irizarry, R. A., Wu, Z. & Jaffee, H. A. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**, 789–794 (2006).
26. Zhang, Y., Price, B. D., Tetradis, S., Chakrabarti, S., Maulik, G. & Makrigiorgos, G. M. Reproducible and inexpensive probe preparation for oligonucleotide arrays. *Nucleic Acids Res.* **29**, E66-6 (2001).
27. He, Z., Wu, L., Fields, M. W. & Zhou, J. Use of microarrays with different probe sizes for monitoring gene expression. *Appl. Environ. Microbiol.* **71**, 5154–5162 (2005).
28. He, Z., Wu, L., Li, X., Fields, M. W. & Zhou, J. Empirical establishment of oligonucleotide probe design criteria. *Appl. Environ. Microbiol.* **71**, 3753–3760 (2005).
29. Chou, C. C., Chen, C. H., Lee, T. T. & Peck, K. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res.* **32**, e99 (2004).
30. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**, 93–102 (2005).
31. Mockler, T. C. & Ecker, J. R. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1–15 (2005).
32. Royce, T. E., Rozowsky, J. S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M. & Gerstein, M. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **21**, 466–475 (2005).