

Gene expression

Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays

Chunlei Wu^{1,2,3}, Haitao Zhao², Keith Baggerly^{2,3}, Roberto Carta⁴ and Li Zhang^{2,3,*}

¹Genomic Institute of Novartis Research Foundation, 10675 John Jay Hopkins Dr, San Diego, CA 92121,

²Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 237, Houston, TX 77030, ³Program in Biomathematics and Biostatistics, The University of Texas Graduate School of Biomedical Sciences at Houston, 6767 Bertner Avenue, Houston, TX 77225-0334 and ⁴Department of Statistics and Actuarial Sciences, University of Central Florida, Orlando, FL 32816, USA

Received on February 1, 2007; revised on April 22, 2007; accepted on May 11, 2007

Advance Access publication May 30, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In microarray experiments, probe design is critical to the specific and accurate measurement of target concentrations. Current designs select suitable probes through *in silico* scanning of transcriptome/genome based on first principles. However, due to lack of tools, the observed microarray data have not been used to assess the performance of individual probes to provide feedback to improve future designs.

Result: In this study, we describe a probe performance assessment method based on the concordance of the observed signals from probes that share common targets. Using this method, we found that probes containing multiple guanines in a row (G-stacks) have abnormal binding behavior compared with other probes, both in gene expression assays and genotyping assays using Affymetrix microarrays. These probes are less likely to covary with other probes that interrogate the same genes. Moreover, we found that these probes are much more likely to produce outliers when fitting the observed signals according to the positional dependent nearest neighbor model, which gives reasonable estimates of binding affinity for most other probes. These results suggest that probes containing G-stacks tend to have increased cross hybridization signals and reduced target-specific hybridization signals, presumably due to multiplex binding forming G-quartet structures. Our findings are expected to be useful in microarray design and data analysis.

Availability: URL: <http://odin.mdacc.tmc.edu/~zhangli/Perfect Match/> contains the computer program for calculating correlations of neighboring probes.

Contact: lzhangli@mdanderson.org

Supplementary information: *Bioinformatics* online or <http://odin.mdacc.tmc.edu/~zhangli/G-stack>

1 INTRODUCTION

Microarray technology has become widely used as a tool in biological research (Lander, 1999; Lockhart and Winzeler, 2000; Olson, 2004). A critical problem of this technology is ensuring that the signals observed from individual probes come from specific genes as designed, since thousands or tens of thousands of genes are measured simultaneously on an array. Typically, microarray probe design is based on a computational search of the transcriptome/genome, which considers uniqueness in the transcriptome/genome to avoid cross hybridization. The design also tries to limit the variation of binding affinity and melting temperature among different probes, and avoid secondary structure of target DNA (or RNA), which may interfere with binding (Li and Stormo, 2001; Matveeva *et al.*, 2003; Mei *et al.*, 2003; Rouillard *et al.*, 2003). However, because computational models have limited accuracy, actual probe performance varies. Thus, it would be highly desirable to utilize observed microarray data to optimize probe design and improve probe performance. However, this task is not trivial because the true content of the observed signals, which are mixtures of target-specific and cross hybridization, is not known. Through spike-in experiments, cross hybridization signals have been quantified on occasion (Wu *et al.*, 2005) and spike-in data have been used in array design (Mei *et al.*, 2003). However, such experiments are suitable only for a small set of probes because of experimental cost. Consequently, the bulk of observed microarray data have not been used to assess and improve probe design.

In this study, we propose a way to use the concordance of observed probe signals to evaluate probe performance. We studied data produced by short oligonucleotide arrays commercialized by Affymetrix, Inc. These arrays use *in situ* synthesized 25 mer DNA oligonucleotides as probes (Lockhart *et al.*, 1996). By design, multiple probes are used to target each gene to reduce cross hybridization effects. A group of probes targeted to the same gene is called a probe set. Ideally, the probes in a probe set should change concordantly as the target concentration varies between samples. However, a number

*To whom correspondence should be addressed.

of factors, such as random noise, cross hybridization and alternative splicing, can reduce the observed correlation between probes. We first searched for the probes in a probe set that were repeatedly found to be less concordant than their neighboring probes. We then searched for sequence motifs in these discordant probes to learn how to avoid such probes in array design. Such analysis led us to discover that probes that contain multiple guanines in a row (or G-stacks) display abnormal binding behavior compared with other probes. We show that probes that contain G-stacks are much less likely to covary with neighboring probes that interrogate the same genes.

Additionally, we found that probes that contain G-stacks appear to have unexpected binding affinities. In our previous work (Zhang *et al.*, 2003), we developed the positional dependent nearest neighbor (PDNN) model, which gives reasonable estimates of the binding affinities of most probes on the arrays. In the PDNN model, probe binding affinity is formulated as a weighted sum of stacking free energies of neighboring base pairs in the double helix formed by the probe and its targeted mRNA transcript. The weights vary depending on the position of the base pairs along the probe, hence the naming of the PDNN model. We show that probes that contain G-stacks are abnormal because they tend to produce signals that are outliers far from the signals expected by the PDNN model. We also show that the abnormal behavior of such probes is not limited to data observed from gene expression assays since the probes produce outlier signals on genotyping assays (SNP detection) too. In the Discussion Section, we suggest a possible mechanism of the abnormal behavior of G-stacks.

2 METHODS

2.1 Sources of microarray data and processing

We obtained the gene expression data of Su *et al.* (Su *et al.*, 2004). This dataset includes 158 array images composed of 79 samples, each of which has two replicates hybridized on the human genome HG-U133A array. We discarded some of the samples because the correlation coefficients between some replicates appeared to be lower than those between others. Thus, we included 71 samples in our consequent analysis. Only PM probes were used; MM probes were discarded. We used the quantile normalization method (Bolstad *et al.*, 2003) to normalize the PM probe signals. The normalization process made the probe signal distribution the same for all the samples included in this study. To perform model fitting with the PDNN model, we used the software package PerfectMatch, available at <http://odin.mdacc.tmc.edu/~zhangli/PerfectMatch>.

We downloaded SNP (single nucleotide polymorphism) data from the Affymetrix, Inc. Web site (<http://www.affymetrix.com/support>). The array type is Mapping50k_xba (Matsuzaki *et al.*, 2004). To exclude probes that involve binding with mismatches, we used the following probe selection criteria: (1) the SNP type must be homozygous (i.e. AA or BB); (2) the allele type of the probe should match the SNP call according to the GDAS algorithm (Liu *et al.*, 2003) and (3) Probes with complementary sequences also exist on the array. In sum, 41 044 probes met these criteria, of which 515 contained GGGG in their probe sequences.

2.2 Differential correlation between probe neighbors

Let X , Y , Z be three consecutive probes in a probe set that targets a particular gene. Let X_i denote the signal of probe X on sample i , where $i = 1, \dots, n$, and n is the total number of samples. Similarly, let Y_i denote the signal of probe Y on sample i , and Z_i denote the signal of probe Z on sample i . We compute the correlation between these neighboring probes as

$$R_{\text{right}} = \frac{\sum Y_i Z_i - \sum Y_i \sum Z_i / n}{\sqrt{(\sum Y_i^2 - \sum Y_i \sum Y_i / n)(\sum Z_i^2 - \sum Z_i \sum Z_i / n)}}$$

$$R_{\text{left}} = \frac{\sum X_i Y_i - \sum X_i \sum Y_i / n}{\sqrt{(\sum X_i^2 - \sum X_i \sum X_i / n)(\sum Y_i^2 - \sum Y_i \sum Y_i / n)}}$$

$$R_n = \frac{\sum X_i Z_i - \sum X_i \sum Z_i / n}{\sqrt{(\sum X_i^2 - \sum X_i \sum X_i / n)(\sum Z_i^2 - \sum Z_i \sum Z_i / n)}}$$

$$D = R_n - (R_{\text{left}} + R_{\text{right}}) / 2$$

where D (the D score) is the differential correlation between neighboring probes with regard to probe Y . To perform these calculations, we use our software package PerfectMatch, available at <http://odin.mdacc.tmc.edu/~zhangli/PerfectMatch>, to compute R_{right} , R_{left} and R_n . The rationale of D score is that, when $D > 0$, it means that signals on probes X and Z are well correlated but signals on probe Y are discordant with them. There are two possible causes of $D > 0$ other than random noise. One is that Y is defective but both Z and X are performing well. The other possibility is that Y performs well but both X and Z are defective. And they are defective in the same way, so that signals on Z and X are well correlated but they are discordant with signals on Y . The second possibility is highly unlikely defective measurements are isolated events and they seldom behave concordantly. Hence, we use $D > 0$ as an indicator that probe Y performs worse than probes X and Z .

3 RESULTS

To evaluate the performance of a probe, we examined the correlation of its observed signals with those of its neighboring probes across many samples. A neighboring probe is one adjacent to the one under examination according to the ordering of the probes along the matching target gene sequence from the 5' end to the 3' end. (In the PDNN model, 'nearest neighboring' refers to the consecutive base pairs in the double helix formed in hybridization; here, 'neighboring probe' refers to the positions of the probes bound to the target gene.) Using probe level data from a previously published dataset (Su *et al.*, 2004), we computed the correlations of the probe signals between neighboring probes across the 71 samples. We then computed the differential correlation D score (see Methods Section) of each probe on the array to assess if each probe performed better ($D < 0$) or worse ($D > 0$) than its neighbors.

Simple observation of the correlations between neighboring probes led to the discovery that probes that contain G-stacks tend to have poorer performance than other probes. To search for probes with poor performance, we examined the probes that had correlations less than 0.5 with their left neighbors and right neighbors (see Methods Section), but whose neighbors had correlations greater than 0.85, i.e. probes that did not correlate well with their neighbors but whose neighbors correlated well

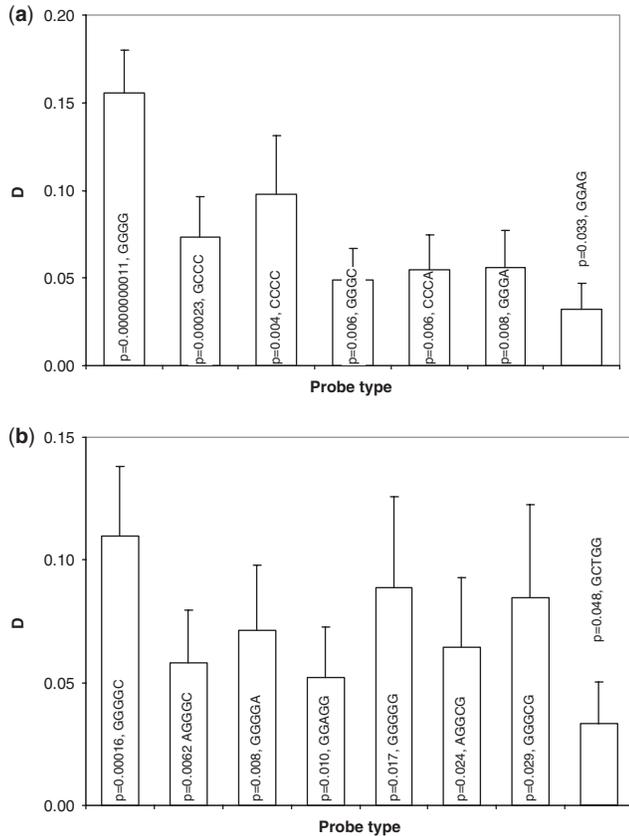


Fig. 1. Worst performing probes in gene expression assays. The vertical axis represents the average differential performance score in a group of probes. Each error bar shows the SD of the mean. A positive score means that the probe is less concordant than its neighboring probes. The probes are grouped according to (a) the central four bases from position 11 to 14 along the probe; or (b) central five bases from position 11 to 15. The P -values were estimated from t -test.

with each other. Of the 362 probes that met these criteria, 30% (120) contained GGGG in their sequences. Considering that only 6.8% of the 250 000 probes on the array have GGGG in their sequences, this association is clearly significant (P -value = 2×10^{-16} , χ^2 test) and suggests that the G-stack is not a desirable sequence motif for probe design.

To generally evaluate which sequence motifs may be poor choices for probe design, we stratified the probes according to their central bases from position 11 to 14. We examined the distribution of D scores in each group and found that G-rich probes were the worst performers. Figure 1a lists the most significant motifs that resulted from this analysis. On average, probes that contain GGGG had a D score of 0.16, which is significantly greater than 0 (P -value = 1.1×10^{-8} , t -test). Most motifs in Figure 1a are G-rich; the exception is CCCC. Similar results were obtained when the probes were stratified according to the central five bases from position 11 to 15 (Fig. 1b), from which the most significant motif was found to be GGGGC ($P=0.00016$, t -test).

To further examine the cause of poor performance of probes that contain G-stacks, we compared the observed signals

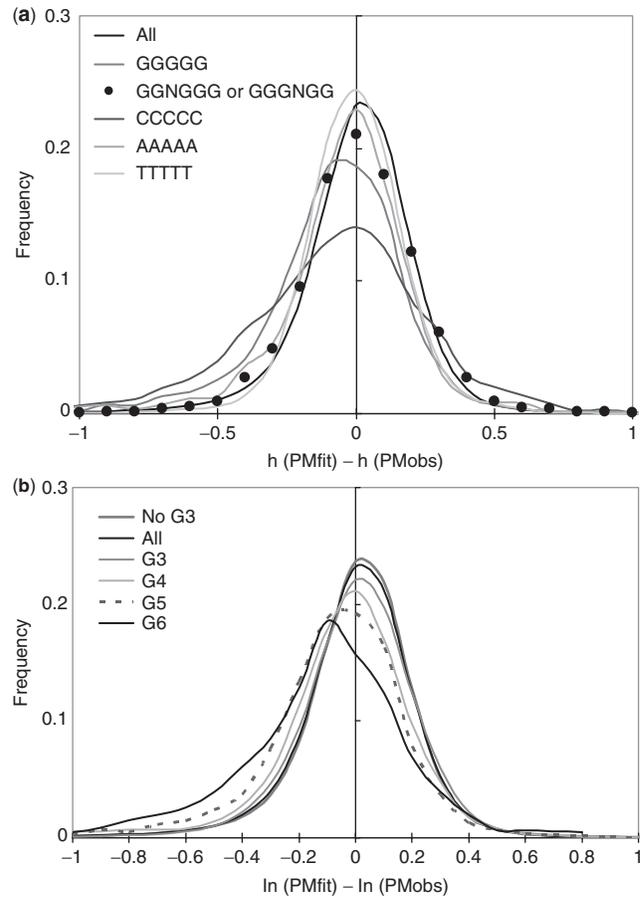


Fig. 2. Distributions of residuals. All distribution curves were normalized to have area of 1. The residuals were obtained from model fitting on one array HG-U133A [data from (Su *et al.*, 2004)], using PDNN model (Zhang *et al.*, 2003). (a) The black line includes all probes and the red includes 3538 probes containing GGGGG in their sequences; the blue for 2134 CCCCC probes; the green for 2028 AAAAA probes; the brown for 7213 TTTT probes and the black dots for 4863 probes containing GGNGGG or GGGNGG, where N is a nucleotide other than G. (b) Probes were stratified by the G-stack length except that the red line includes probes that do not contain GGG and the black line includes all probes. G3, G4, G5 and G6 represent probes with G-stack length of 3, 4, 5 and 6, respectively.

(PMobs) on these probes with the model fitted values (PMfit) according to PDNN model (Zhang *et al.*, 2003). From the distribution of residuals [defined as $\ln(\text{PMfit}) - \ln(\text{PMobs})$], we saw heavier tails from probes that contain GGGGG or CCCCC, compared with those from all probes (Fig. 2a). This implies that CCCCC and GGGGG probes tend to create more outliers. In contrast, probes that contain TTTT or AAAAA demonstrated behavior similar to the group that included all probes. Interestingly, when the G-stack is interrupted, as in probes with GGNGGG or GGGNGG, where N is a base other than G, the probes behave rather normally (black dots in Fig. 2a). It means that it is the G-stack rather than the individual Gs that causes the poor performance.

From Figure 2a, we can also see that the observed signals from probes containing GGGGG tend to be greater than

expected from PDNN model as the residual distribution curve is tilted to the left. From probes that contain GGGGG in their sequences, we found 218 probes had $\ln(\text{PMfit}) - \ln(\text{PMobs}) < -0.5$, but only 39 probes had $\ln(\text{PMfit}) - \ln(\text{PMobs}) > 0.5$. Interestingly, we found that the former group of probes is associated with low gene expression values, but the latter group is associated with high gene expression values. Using the signals from probes that are in the same probe sets as the probes that contain G-stacks, we estimated the gene expression values according to PDNN model. For the probe sets (genes) associated with the 218 probes, the average gene expression value $\pm \text{SD} = 5.85 \pm 0.12$ (values presented on natural logarithm scale), while for the probe sets associated with the 39 probe, the average gene expression value $\pm \text{SD} = 6.41 \pm 0.45$ (on natural logarithm scale as well). This difference is statistically significant (P -value = 3×10^{-13} , t -test). These results indicate that probes that contain G-stacks tend to get extra signals when target concentration is low but miss signals when target concentration is high. We also examined CCCCC probes in detail to look for the same pattern. We found 226 probes with $\ln(\text{PMfit}) - \ln(\text{PMobs}) < -0.5$, the associated average gene expression $\pm \text{SD} = 5.86 \pm 0.30$. We also found 80 probes with $\ln(\text{PMfit}) - \ln(\text{PMobs}) > 0.5$, the associated average gene expression $\pm \text{SD} = 5.95 \pm 0.33$. Thus, CCCC probes also tend to have higher than expected signals, but there is no significant association with the gene expression values as that observed in GGGG probes. These results suggest that GGGG probes and CCCC probes may have different mechanisms that lead to their poor performance on the microarrays.

To study the effects of G-stack length, we stratified the probes according to the length of consecutive Gs in their sequences and examined the distribution of the residuals. As Figure 2b shows, the residual distribution starts to show deviation from normal probes only when the G-stack length is more than 3. When the G-stack length is 6, the deviation becomes quite obvious.

We found that the unusual binding behavior of probes that contain G-stack is not limited to gene expression assays. We examined data produced from genotyping arrays for SNP detection (Kennedy *et al.*, 2003). The measurement mechanism on this type of arrays differs from that of gene expression arrays because the target molecules used in genotyping assays are double-stranded, end-labeled DNA molecules as opposed to the single-stranded, internally labeled RNA molecules used in gene expression assays. For simplicity, we collected probes signals that involved no mismatches based on genotype calls determined by the GDAS algorithm (Liu *et al.*, 2003) (see Methods Section for details). Because the target molecules are double stranded, both sense and antisense sequences are adopted to design the probes. Consequently, a pair of probes with sequences complementary to each other should bind to the same target molecules. Because the same double helix forms for each probe in the probe pair upon binding to the targets, we expect probes with complementary sequences to have similar binding affinity. Therefore, we used the ratio of observed signals between complementary sequences (cPM/PM) to examine the binding affinity of the probes.

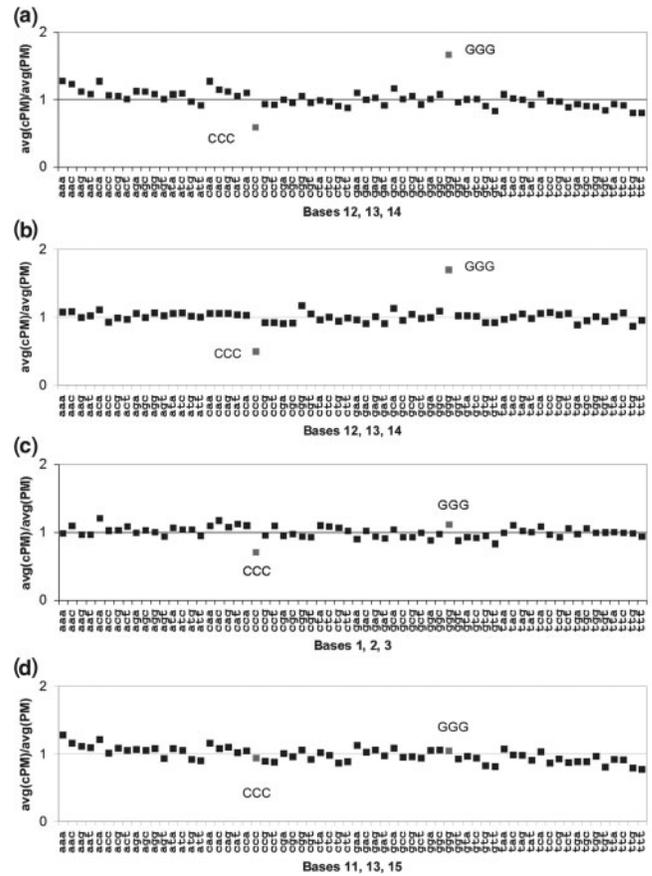


Fig. 3. Signal ratio between of complementary probe pairs. Two probes that have reverse complementary sequences to each other are called complementary probe pair. The vertical axes in the figures represent the average of complementary probe signals (cPM) divided by the average of PM probe signals. (a) The probes were stratified according to the central three bases on the sense probe. All 41044 probes on Mapping 50k-Xba met homozygous criteria were included; (b) the probes included are a subset of that in (a) by requiring that the number of As and Ts are equal in a probe sequence; (c) the probes were the same as that in (a) but the probes stratified according to the first three bases on a probe and (d) the probes were the same as that in (a) but the probes stratified according to the bases 11, 13 and 15 on a probe.

Again, we found that probes that contain G-stacks appear to be outliers in terms of cPM/PM ratios. Figure 3a shows the average cPM/PM ratios for probes stratified by the central three bases on the PM probes. Probes that contain GGG at the center of the probe sequence have much lower signals than their complementary probes, which have CCC at the center of the probe sequences (the average ratio is ~ 1.7). Similar results were obtained when the probes were stratified according to the central four bases (data not shown).

From our previous study, we found that the assumption that complementary probes ought to have the same binding affinity does not hold exactly (Zhang *et al.*, 2007). A possible cause is interaction between target molecules and the microarray surface, which is not equivalent for complementary probes. We performed regression analysis of cPM/PM ratios in terms

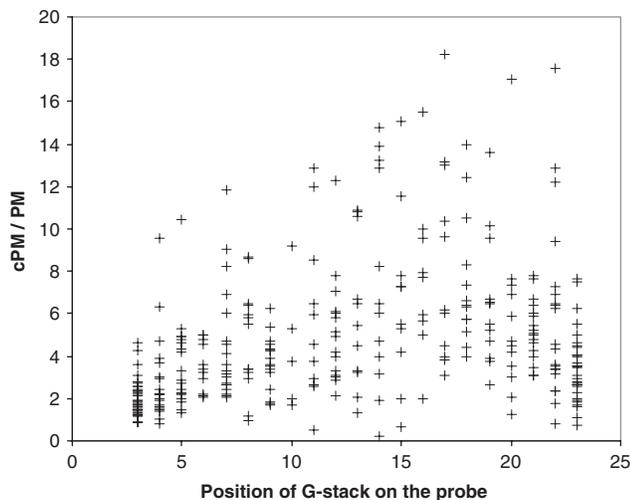


Fig. 4. Positional effects of G-stacks. The vertical axis shows the ratio of probe signals between complementary probes (cPM/PM). Only PM probes that contain GGGGG are included with the horizontal axis showing the midpoint position of GGGGG on the PM probe. The left side is the 5' end of the PM probe; the right side is the 3' end of the PM probe. The ratios appear to decrease on both ends of the probe. All probe signals shown here were assumed to involve no mismatches.

of A, T, C, G composition of the probes. We have found that the cPM/PM ratio depends to some extent on the number of As minus the number of Ts in the probe sequence (Zhang *et al.*, 2007). Consequently, we examined probes with equal number of As and Ts in their sequences (Fig. 3b). For probes in Figure 3b, the surface effects are supposed to be similar for PM and cPM probes. Interestingly, with these probes, cPM/PM ratio is close to 1 mostly, and the GGG probes as a group of outliers become even more striking. This result suggests that when the surface effect is corrected for, the abnormality of probes containing G-stacks is more prominent. Furthermore, to find out if it is the G-stacks or the individual Gs that lead to the abnormal cPM/PM ratios, we stratified the probes according to the bases 11, 13 and 15 instead of the central three bases (Fig. 3d). In Figure 3d, the probes with three Gs at these bases did not result in abnormal cPM/PM ratios. Thus, similar to that found in the gene expression arrays, G-stacks seemed to be the cause rather than the individual Gs.

We found that the effects of G-stacks seem to depend on the position on the probe. When the probes were stratified according to the first three bases (i.e. the 5' end. The 3' end of the probe is tethered to the microarray surface.) instead of the central three bases, the contrast between GGG probes and CCC probes diminished (Fig. 3c). In Figure 4, we show all probes that have GGGGG in the sequences. It is striking to note that 98% of the 324 cases shown in this figure, have cPM/PM ratios greater than 1. The cPM/PM ratio appears to be smaller when GGGGG is at either ends of a probe. These results are consistent with existing models (Held *et al.*, 2006; Mei *et al.*, 2003; Zhang *et al.*, 2003), which find that the ends of the probes contribute less to binding affinity.

4 DISCUSSION

We have developed a method to evaluate probe performance according to concordance of probe signals between neighboring probes in the same probe sets. It should be noted that the method is only applicable for comparing large groups of probes. If we look at only three consecutive probes, it is not clear which probe signals are closer to the true expression values, although the correlation between two of them may be higher than that of the other pairing. Only in large groups can we expect probes that are well correlated with their neighboring probes to be more trustworthy than those that are not well correlated with their neighbors. In this study, we searched for sequence motifs that are associated with poor performance and found that probes that contain G-stacks tend to be poorly correlated with other probes. Of the 250 000 probes on the HG-U133A array, 16 743 contain GGGG in their sequences; of those, 3538 contain GGGGG in their sequences. These probes provide ample sample size to determine the statistical significance of our results. The abnormal behavior of the probes containing G-stack seemed to be general on Affymetrix microarrays, as we observed that the probes containing G-stacks also had discordant signals (See Fig. S1 in Supplementary Material) with other probes from a different dataset, which used a denser probe design (the array type is HG-U133 Plus 2.0).

There are multiple causes of poor correlation between probes. Among the 362 probes with the highest D scores, 1/3 of them contained G-stacks. The causes of the remaining 2/3 of the probes are not clear. We examined one of such probes in detail. Its target gene is tyrosine phosphatase, non-receptor type 6. The probe's sequence is 'CCTATCCCCAGCCATGAAGAATGC'. The probe's signal is discordant ($r \sim 0.2$) with other probes in the probe set (206687_s_at). If this bad probe is removed from the probe set, the correlations between other probes are around 0.8. From residual analysis using PDNN model, we found the bad probe had signals that were 3 times higher than that expected from the model fitted values. Interestingly, we also found 51 probes on the HG-U133A array that had a fragment of the bad probe, 'CCCCAGC', in their sequences. Most of these 51 probes have $D > 0$ (mean = 0.09; SD = 0.2; P -value = 0.003). These results highly suggest that CCCCAGC is a magnet for attracting cross hybridization.

In general, the possible causes of high D -scores are random noise (Naef *et al.*, 2002), alternative splicing and cross hybridization, saturation (Naef *et al.*, 2003), target-target or probe-probe interaction (Forman *et al.*, 1998), degradation of target samples (Auer *et al.*, 2003) and secondary structure formed by targets and probes (Mir and Southern, 1999; Shchepinov *et al.*, 1997). Use of incorrect gene sequencing in probe design also could lead to uncorrelated probe signals (Dai *et al.*, 2005; Sliwerska *et al.*, 2006). Figure 1 suggested the probes with C-stacks may also result in poor performance. It may be interesting to explore further in the remaining 2/3 of the probes for common patterns. But regardless of its causes, poor correlation is always an undesirable trait in probe performance because the desired behavior is that the signal linearly responds to the target concentration without

interference from other factors. Therefore, linear correlation between neighboring probes appears to be a reasonable index to reflect probe performance.

Why are probes that contain G-stacks problematic on microarrays? Nucleotides rich in Gs are known to form quadruplex bundles involving G-quartets (Dapic *et al.*, 2003; Keniry, 2000; Mergny *et al.*, 2005), but their role in microarrays is not widely recognized. On microarrays, probes containing G-stacks may form quadruplex bundles with target molecules. Because the probes are immobilized on the Affymetrix arrays, it is not possible for them to form the quadruplexes among themselves. The target molecules, on the other hand, may form quadruplex among themselves in solution. The target molecules may quadruplexes among themselves. Mei *et al.* (Mei *et al.*, 2003) suggested that probes that contain GGGG in their sequences may invoke quadruplex binding, but did not determine if GGGG sequences harm or help probe performance. Consequently, probes manufactured by Affymetrix, Inc. still contain G-stacks. The longest G-stack in a probe on the HG-U133A array is nine guanines.

The G-quartet quadruplex hypothesis may not be the only explanation to our results. To form stable G-quartets in solution, the guanines need not to be contiguous (Dapic *et al.*, 2003). However, our results show that probes with GGGNGG sequences behaved very differently from probes with GGGGG sequences. It is not clear why GGGNGG sequences would not form quadruplexes on the microarrays. Besides quadruplex formation, difficulties in synthesizing the probes containing G-stacks may also be a cause of poor performance. Our current study only analyzed data collected from Affymetrix microarrays. It would be interesting to see if the same phenomena can be observed on microarrays using other techniques. Apparently, future experiments are needed to reveal how quadruplex formation may hinder microarray hybridization.

Based on our analysis, we assert that probes that contain G-stacks perform poorly on microarrays, because G-stacks tend to increase cross hybridization and reduce target-specific hybridization. This poor performance is not likely to be caused by saturation because it can apparently happen at low target concentration. Probe and target molecules that contain G-stacks could form intra- and/or inter-molecular G-quartets. When target concentration is zero or low, the probe signal is dominated by cross hybridization, so contributions from off-target, G-rich molecules bound to probes with G-stacks could be identifiable. As the target concentration increases, the content of gene specific hybridization in the probe signal increases so that the effects of cross hybridization are less obvious. At very high target concentrations, the availability of the target molecules may be reduced by target-target interactions. Target molecules with C-stacks can cross hybridize to molecules with G-stacks forming duplexes. Alternatively, molecules with C-stacks can also form i-motifs. These interactions hinder hybridization so that fewer than expected targets are accessible on the microarray surface. This mechanism can explain our residual analysis results (Fig. 2). It may also explain the data observed from genotyping assays, in which the target molecules are nearly always present, so that probes with G-stacks generally have reduced signals (Fig. 4). Note that for hybridization in aqueous solution, the

roles of probes and targets are symmetrical so that we expect cPM/PM ratio to be one. However, for hybridization on the microarrays, because the probes are immobilized, some probe-target interactions, such quadruplex formation, are prohibited. Thus, when the roles of probes and targets are switched, all types of molecular interactions cannot be symmetrically switched. Therefore, cPM/PM can be significantly different from one, which was observed in our data.

The fact that probes that contain G-stacks tend to have abnormal signals both on gene expression assays and genotyping assays strongly suggests that they should be avoided in probe design. In commonly used methods for microarray data analysis (Hubbell *et al.*, 2002; Irizarry *et al.*, 2003; Li and Wong, 2001; Zhang *et al.*, 2003), the effects of outliers are suppressed because of the use of robust estimators. Consequently, the effects of probes that contain G-stacks have limited scope. However, the existing algorithms cannot reliably detect the outliers and remove their effects. Therefore, removing probes that have poor performances in probe design is a cleaner, more efficient solution to the problem.

ACKNOWLEDGEMENTS

We thank Margaret Newell for editing the manuscript and support provided by M. D. Anderson Cancer Center start-up fund and MDACC Institutional Research Grant to L.Z.

Conflict of Interest: none declared.

REFERENCES

- Auer, H. *et al.* (2003) Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat. Genet.*, **35**, 292–293.
- Bolstad, B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
- Dai, M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.
- Dapic, V. *et al.* (2003) Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Res.*, **31**, 2097–2107.
- Forman, J.E. *et al.* (1998) Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesized oligonucleotide arrays. In *Molecular Modeling of Nucleic Acids*. American Chemical Society, Washington, DC, pp. 206–228.
- Held, G.A. *et al.* (2006) Relationship between gene expression and observed intensities in DNA microarrays – a modeling study. *Nucleic Acids Res.*, **34**, e70.
- Hubbell, E. *et al.* (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
- Irizarry, R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
- Keniry, M.A. (2000) Quadruplex structures in nucleic acids. *Biopolymers*, **56**, 123–146.
- Kennedy, G.C. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233–1237.
- Lander, E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.
- Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
- Li, F. and Stormo, G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
- Liu, W.M. *et al.* (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397–2403.
- Lockhart, D.J. and Winzler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

- Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Matsuzaki,H. *et al.* (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.
- Matveeva,O.V. *et al.* (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res.*, **31**, 4211–4217.
- Mei,R. *et al.* (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.
- Mergny,J.L. *et al.* (2005) Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.*, **33**, 81–94.
- Mir,K.U. and Southern,E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.*, **17**, 788–792.
- Naef,F. *et al.* (2002) Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.*, **3**.
- Naef,F. *et al.* (2003) A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics*, **19**, 178–184.
- Olson,J.A. Jr (2004) Application of microarray profiling to clinical trials in cancer. *Surgery*, **136**, 519–523.
- Rouillard,J.M. *et al.* (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
- Shchepinov,M.S. *et al.* (1997) Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res.*, **25**, 1155–1161.
- Sliwerska,E. *et al.* (2006) SNPs on Chips: The hidden genetic code in expression arrays. *Biol. Psychiatry*.
- Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
- Wu,F.X. *et al.* (2005) Dynamic model-based clustering for time-course gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 821–836.
- Zhang,L. *et al.* (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.
- Zhang,L. *et al.* (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.*, **35**, e18.