Published online 14 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 3* ***e18***
doi:10.1093/nar/gkl1064

Free energy of DNA duplex formation on short oligonucleotide microarrays

Li Zhang[1,2,*], Chunlei Wu[1,2,3], Roberto Carta[4] and Haitao Zhao[1]

[1]Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 237, Houston, TX 77030, USA, [2]Program in Biomathematics and Biostatistics, The University of Texas Graduate School of Biomedical Sciences at Houston, 6767 Bertner Avenue, Houston, TX 77225, USA, [3]Genomic Institute of Novartis Research Foundation, 10675 John Jay Hopkins Dr, San Diego, CA 92121, USA and [4]Department of Statistics and Actuarial Sciences, University of Central Florida, Orlando, FL 32816, USA

Received August 7, 2006; Revised November 15, 2006; Accepted November 20, 2006

ABSTRACT

DNA/DNA duplex formation is the basic mechanism that is used in genome tiling arrays and SNP arrays manufactured by Affymetrix. However, detailed knowledge of the physical process is still lacking. In this study, we show a free energy analysis of DNA/DNA duplex formation these arrays based on the positional-dependent nearest-neighbor (PDNN) model, which was developed previously for describing DNA/RNA duplex formation on expression microarrays. Our results showed that the two ends of a probe contribute less to the stability of the duplexes and that there is a microarray surface effect on binding affinities. We also showed that free energy cost of a single mismatch depends on the bases adjacent to the mismatch site and obtained a comprehensive table of the cost of a single mismatch under all possible combination of adjacent bases. The mismatch costs were found to be correlated with those determined in aqueous solution. We further demonstrate that the DNA copy number estimated from the SNP array correlates negatively with the target length; this is presumably caused by inefficient PCR amplification for long fragments. These results provide important insights into the molecular mechanisms of microarray technology and have implications for microarray design and the interpretation of observed data.

INTRODUCTION

DNA microarrays have become a critical enabling technology in biological research (1–3). Initially, microarrays were designed only to assay gene expression values. However, recent advances have extended the application of the technology to SNP detection, DNA copy number measurements, alternative splicing analysis, DNA methylation characterization, gene structure discovery and genome resequencing (see reference 4 for review). The mechanisms of detection are different among these methods. However, concern over the quality of microarray data persists because the technology involves a complicated process that is difficult to control or comprehend (4). For Affymetrix GeneChip® arrays (Affymetrix, Santa Clara, CA), measurements of expression are obtained from DNA/RNA hybridization between the DNA probes and the target RNA, which are single-stranded and internally labeled with biotin molecules. But the new arrays designed for polymorphism detection (5,6) and gene discovery (7) use DNA/DNA hybridization, and the target DNA molecules are double-stranded and labeled at the 3′ ends rather than internally.

A number of physical models have been proposed to characterize DNA/RNA hybridization on expression arrays (8–14). Studies of these models have shown that the sensitivity and specificity of DNA/RNA hybridization depend on the sequences of the probes. No studies have been conducted for the new arrays using DNA/DNA hybridization. In this report, we investigate the free energy of binding on these new platforms based on the positional-dependent nearest-neighbor (PDNN) model (8). The basic idea of the PDNN model is that the binding free energy of a probe can be expressed as a weighted sum of its stacking energies (15,16), where the weights depend on the position along the probe:

$$\Delta G_{ij} = \sum_{k=1}^{24} \omega_k \, \varepsilon(b_k, b_{k+1}), \qquad \mathbf{1}$$

where ω_k is a weight factor that depends on the position of consecutive bases along the oligonucleotide, $b_1, b_2, \ldots, b_{25}$ denote the probe sequence, and ε (b_k, b_{k+1}) represents a stacking energy term.

*To whom correspondence should be addressed. Tel: +1 713 5634298; Fax: +1 713 5634243; Email: lzhangli@mdanderson.org

In this study, we apply this model to SNP arrays and genome tiling arrays. We focus on the quantitative characterization of stacking energy, the positional dependence and the cost of a single nucleotide mismatch. We show that because the probes are covalently attached to the microarray surface, that surface has a drastic impact on the observed microarray data. We further demonstrate that the DNA copy number estimated from the single nucleotide polymorphism (SNP) array is negatively correlated with the length of the target.

METHODS

Microarray data source and preprocessing

SNP data were obtained from the Affymetrix Web site at http://www.affymetrix.com. The data were generated from the Mapping50K_Xba240 array. This array is designed to interrogate 58960 SNP sites in the human genome. Typically, 40 probes are designed to interrogate a single SNP site. Half of the probe set consists of perfect match (PM) probes, and half consists of mismatch (MM) probes. There are two allele types associated with each SNP site. For each allele type, 10 PM probes are designed to match the sequence around the SNP site on both strands of DNA with the SNP site placed at five different positions near the center of the probe. Genotyping calls were performed using GDAS 3.0, developed by Affymetrix. To apply the PDNN model, we used PM probe signals that resulted from exact matches to the target DNA in homozygous sites according to Affymetrix genotyping calls. MM probe signals are not used in PDNN model fitting because binding on MM probes can potentially involve two mismatches, one for the allele type and the other manually designed at the 13th position on the MM probes. Data presented in this study were obtained from only one sample (sample name: NA06985_X_tH_B5_4000090). Results obtained from other samples were very similar, and hence are not shown.

We downloaded data (http://cgap.nci.nih.gov/Info/2002.1) on the genome tiling arrays from the supplementary materials published by Kapranov *et al*. (7) on the transcriptional activities in chromosomes 21 and 22. This array is designed to interrogate 362901 contiguous nucleotides of the DiGeorge syndrome minimal critical region (DGCR) on chromosome 22 (17). To use the PDNN model, it is necessary to compose an equivalence of 'probe sets', as in expression arrays. We assumed that probes sharing significant sequence overlaps would be able to bind the same fragments. We chose every consecutive 11 probes as a 'probe set' and applied the PDNN model to obtain the stacking energies and weight factors through a least-square fit via Monte Carlo simulations as previously described (8).

RESULTS

Stacking energy and positional weights

We applied the PDNN model to three different types of Affymetrix arrays: (i) the Mapping50k_Xba240 genotyping array designed to detect SNP in the human genome; (ii) the DGCR genome tiling array; and (iii) the HG-U133A expression array. Through optimizing the fit between the observed probe signals and the expected signals using the model, we obtained the nearest-neighbor stacking energies and positional-dependent weights (Figure 1). Note that the first two types of arrays use a DNA/DNA hybridization mechanism but the third type uses a DNA/RNA hybridization mechanism. Figure 1 shows that the overall patterns of the parameters are similar. From the stacking energies (Figure 1a), we saw that dinucleotide GC has favorable binding free energy but dinucleotide CG has unfavorable binding free energy. From the position-dependent weights (Figure 1b), we saw that the ends of the probes contribute less to the stability of binding than do the middle of the probes.

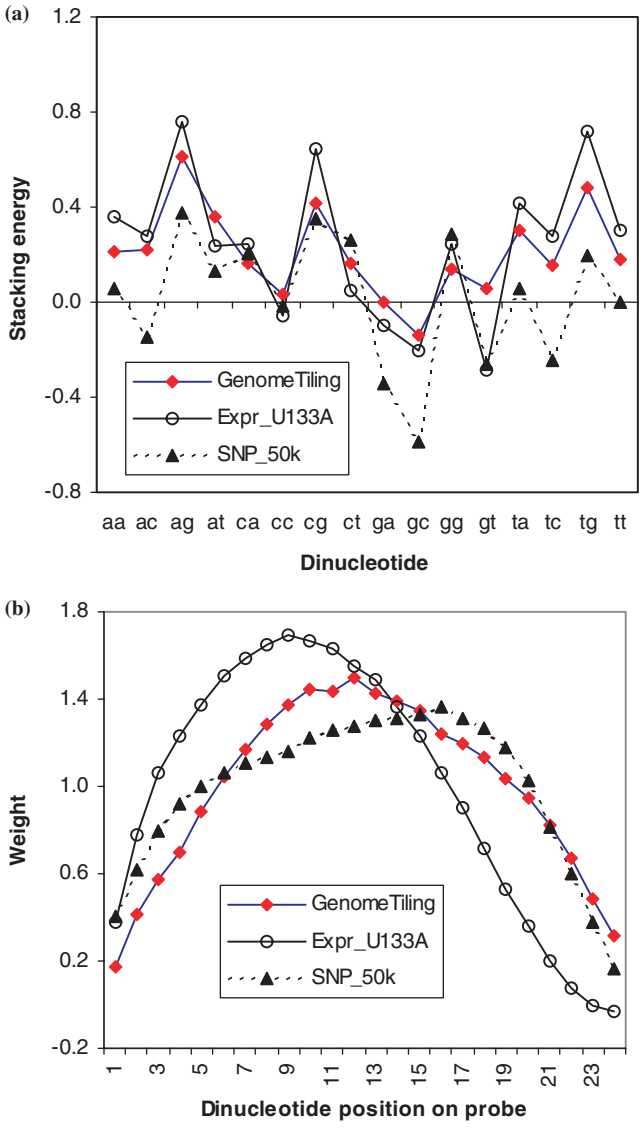

Figure 1. Free energy parameters obtained from different arrays. (**a**) Nearest-neighbor stacking energies; (**b**) Weight factors.

As shown in Figure 2, the model-fitted values agree well with the observed signals. Based on our observations, the occasional misfits tend to happen at exon boundaries (details not shown); this is to be expected because the probes in the probe set could not bind the same transcripts.

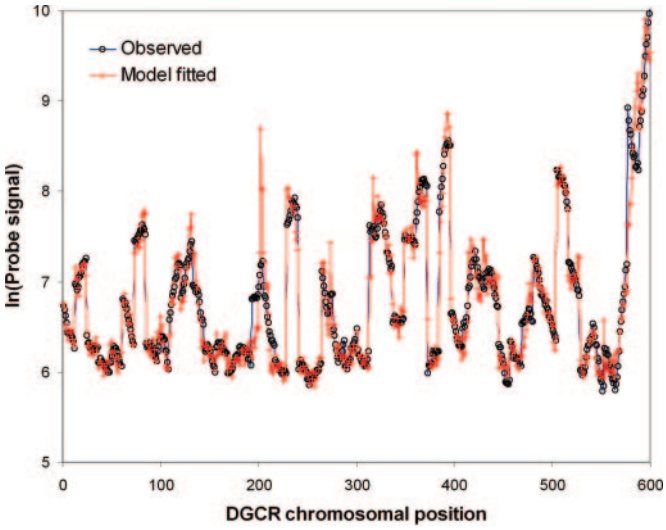

Figure 2. Fitting probe signals on the DGCR array with the PDNN model. Every consecutive set of 25 probes was considered a probe set for application of the PDNN model.

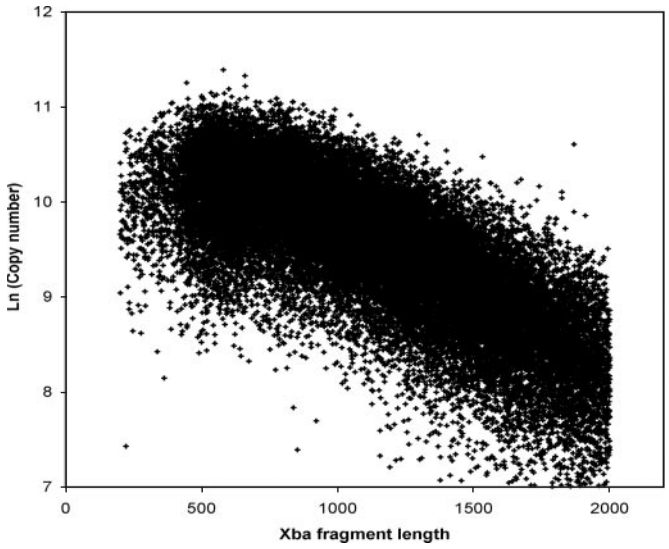

Figure 3. DNA copy number is negatively correlated to target length. The DNA copy numbers were extracted from SNP sites that have homozygous calls. The *y*-axis is presented on a natural logarithm scale.

DNA copy numbers are related to the target lengths

DNA copy numbers in a cell are expected to be the same, with the exception of the sex chromosomes and chromosomal aberrations. Therefore, one might expect similar target concentrations to be estimated from SNP arrays after correcting for the binding affinity of a probe. Using the PDNN model, we extracted the DNA copy numbers from homozygous SNP sites. In Figure 3, the values of the DNA copy numbers are plotted against the target lengths (the length information was obtained from the Affymetrix Web site). There is an apparent negative correlation ($r = -0.75$), which means that long target fragments are not well represented on the microarray.

Free energy cost of a single mismatch on SNP arrays

We observed that the mismatch discrimination is stronger in DNA/DNA duplexes than in DNA/RNA duplexes. On the 50K_Xba240 arrays, we observed that only 0.5% of a total of 413930 probe pairs had fewer PM signals than MM signals (PM < MM), where PM and MM stand for the probe signals on the perfect match and mismatch probes, respectively. This is in sharp contrast to our observations of the expression arrays, in which 30% of the probe pairs had PM < MM.

To evaluate the free energy cost of a single mismatch in DNA/DNA duplexes on 50K-Xba240 arrays, we collected PM probe signals associated with homozygous calls. For an SNP site with an SNP call of AA, we expect the PM probes designed to interrogate the A type to have exact matches and the PM probes designed to interrogate the B type to have single mismatches. For such a pair of PM_A and PM_B probes, we define the free energy cost of a single mismatch to be

$$\Delta G = k_B T^* [\ln(PM_A - c) - \ln(PM_B - c')], \quad \mathbf{2}$$

where c and c' are cross hybridization signals on the PM_A and PM_B probes, respectively; T is the temperature of the hybridization experiment; and k_B is the Boltzmann constant. The cross hybridization signals can be ignored when the PM_A and PM_B signals are strong. We observed that the value of $\ln(PM_A/PM_B)$ is largely independent of the magnitude of the probe signals, indicating that the effect of cross hybridization is small for most probes. Therefore, we simply used the approximation:

$$\Delta G = \ln(PM_A) - \ln(PM_B), \quad \mathbf{3}$$

ignoring the effects of cross hybridization and expressing ΔG in $k_B T$ units. We stratified the types of single mismatches according to the following four types of nucleotides: the nucleotide at the mismatch point on the PM_A probe; its immediate left and right neighboring nucleotides; and the mismatched nucleotide of the target bound on the PM_B probe. Such stratification produced 192 categories. We computed the average values of ΔG in the 192 categories (see Supplementary Table S1, in the Supplementary information). The overall average of ΔG was found to be ~1.34. The most costly mismatch is found in $\Delta \boldsymbol{G}$ (TGT_p/ACA_t) − $\Delta \mathbf{G}$ (TAT_p/ACA_t) = 1.78 ± 0.01 $k_B T$, where the subscripts p and t denote probe and target, respectively. The least costly mismatch is found in $\Delta \boldsymbol{G}$ (ATC_p/GAT_t) − $\Delta \boldsymbol{G}$ (AGC_p/GAT_t) = 0.81 ± 0.02 $k_B T$.

We compared the values shown in Supplementary Table S1 with those obtained from studies conducted in aqueous solution, and observed that the two data sets are moderately correlated ($R^2 = 0.35$, Figure 4). We found that the outliers were involved in the disruption of GGG/CCC binding. When they are excluded, $R^2 = 0.46$. It has been suggested that poly G-stacks incur multiplex binding (10); our observation, which may explain these outliers.

The effect of the microarray surface on binding

Because the target samples are double-stranded DNA fragments on SNP arrays and genome tiling arrays, one might

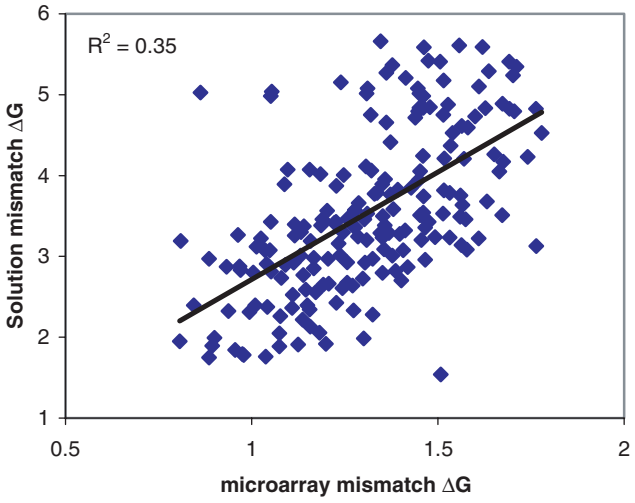

Figure 4. Correlation between free energy costs of single mismatches obtained from SNP arrays and those from solution studies. Values (in kcal/mol) from solution studies were adapted from reference (16). The values (in K_BT) from microarray studies were obtained from Supplementary Table S1. The R^2 value becomes 0.46 when the six outmost outliers are excluded.

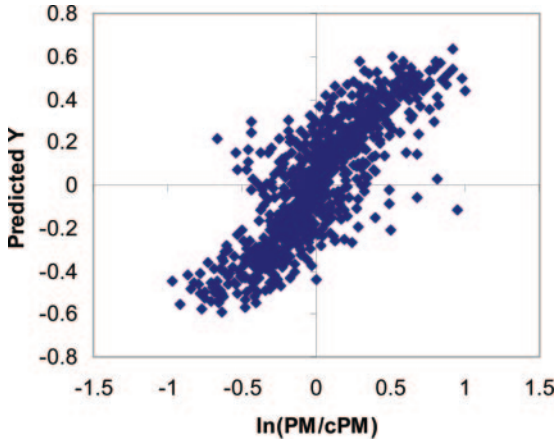

Figure 5. Modeling the surface effects on probe binding. The *x*-axis represents the logged signal ratio between the probes with complementary sequences. The *y*-axis represents the model-fitted values, which depend on the numbers of As, Ts, Cs and Gs in the probes.

expect the probes with complementary sequences (i.e. antisense to each other) to yield equal amounts of signals. Upon binding to their targets, the probes with complementary sequences form the same DNA/DNA double helices; consequently, one might expect the binding affinity of these probes to be the same.

However, data observed from the arrays sharply contradict these expectations. We reason that the discrepancy may be caused by interactions between the DNA/DNA duplexes and the surface of the microarray. For DNA hybridization in aqueous solution, the roles of probes and targets are reciprocally symmetrical so that probes and targets are interchangeable. This symmetry is broken for hybridization on the microarrays because the probes are covalently bounded to the surface while the targets can roam free in solution. Since the targets and the microarray surface are likely charged, there is a repulsion force between them. The strength of the repulsion depends on the sequence of the targets. Consequently, the costs of free energy to bring different single-stranded targets to the microarray surface are different. For a pair of probes with complementary sequences, because their corresponding targets are chemically different, they can have different binding affinities due to interaction with the microarray surface. We can see this effect in the stacking energies in Figure 1a; for example, dinucleotides CA and TG have different stacking energies.

To model the effects of the microarray surface, we collected pairs of probe signals (PM and cPM) with complementary sequences on 50K_Xba240 arrays. We discarded probes involved in heterozygous calls or no calls and kept only the PM probes so that the binding of the targets would involve no mismatches. Based on the composition of the nucleotides in the PM probes, we stratified the probe pairs and computed the average value of Y = ln(PM/cPM) in each category (categories with less than 10 pairs were discarded). Here, PM and cPM stand for a pair of probe signals. The sequences of PMs and cPMs are complementary to each other. We then used a linear regression model to find the dependence of $Y = \ln(\mathrm{PM}/\mathrm{cPM})$ on the composition of the nucleotides of the PM probe. The regression results were as follows:

$$\begin{aligned}\hat{Y} &= -0.17 + 0.050N_{\mathrm{T}} - 0.047N_{\mathrm{A}} + 0.020N_{\mathrm{G}} \\ &\approx -0.17 + 0.05(N_{\mathrm{T}} - N_{\mathrm{A}} + 0.4N_{\mathrm{G}}),\end{aligned} \quad \mathbf{4}$$

where N_{T}, N_{A} and N_{G} represent the number of Ts, As and Gs in the PM probe, respectively. As shown in Figure 5, the observed Y and the model-fitted $\hat{Y}$ are well correlated. ($R^2 = 0.67$, $N = 875$). Again, we found that the outliers may be associated with probes having poly G-stacks. When the probes with GGG or CCC in their sequences were discarded, the R^2 value changed to 0.73.

DISCUSSION

We have presented a quantitative analysis of the DNA/DNA duplex formation on short oligonucleotide microarrays. We showed the sequence dependence of the probe binding affinities using the PDNN model. We did not address the effects of secondary structures, which have been shown to affect microarray measurements (18). In terms of stacking energy and positional dependence, we found that DNA/DNA duplex formation is very similar to DNA/RNA duplex formation (Figure 1). However, in terms of the free energy cost of mismatches, we found that DNA/DNA duplexes are less tolerant, as probe pairs with MM < PM occurred much less frequently in the SNP arrays compared with their occurrence in the gene expression arrays. We also examined how neighboring bases affect a single mismatch on SNP arrays (Supplementary Table S1). The values in Supplementary Table S1 are correlated to values determined in aqueous solution, but the magnitude of the values is much smaller for the microarrays.

SNP arrays also have been used to detect aberrations in DNA copy number (19,20). However, estimates of DNA copy number appear to be very noisy, and a better understanding of the observed signals is needed. We found that the probe signals depend on the binding affinity as well as

the length of the target. In processing SNP array samples (6), restriction enzyme Xba is used to excise chromosomes, and the resulting fragments are ligated to a common primer for PCR amplification. The efficiency of PCR amplification is not the same for all nucleotide fragments, and long fragments tend to be poorly amplified, which explains why there is a negative correlation between the DNA copy number and the target length.

Another important finding of this study is that there must be a strong surface effect in short oligonucleotide microarray hybridization. This is consistent with findings in other studies that used experiments (21) and theoretical calculations (22) to demonstrate such surface effects. Our conclusion is based on the contrast in signals found between probes with complementary sequences. Another potential cause of the signal contrast is the biotin-labeling of DNA fragments. However, because the arrays we analyzed have the biotin molecules attached to the target DNA fragments at the 3′ ends and the target fragments are typically longer than 100 bases, the biotin molecule is far away from the binding sites. Thus, it is unlikely that the biotin molecules interfere with binding. Note that for gene expression arrays using DNA fragments that are internally labeled with biotin, it is difficult to determine whether it is the biotin molecule or the microarray surface that interferes with the binding.

Taken together, our results provide important insights into the molecular mechanisms of microarray technologies. Such insights can help facilitate future advancements of this important technology. For example, our findings may underscore the value of using biochemical means (i.e. ionic strength or pH) to reduce the surface effects on microarrays to enhance the sensitivity of the measurements. Such insights also have implications for the interpretation of data obtained from microarray technologies. For instance, because long target fragments (>1000 bp) are poorly represented in SNP arrays, the DNA copy numbers estimated from the corresponding probes are less reliable. Thus, down-weighting the corresponding probes in the DNA copy number estimates may improve the analysis of the observed data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful for funding from The University of Texas M. D. Anderson Cancer Center through an institutional research grant, and thank LeeAnn Chastain for editorial assistance. Funding to pay the Open Access publication charges for this article was provided by The faculty incentive funds to L.Z. from MD Andersen Cancer Center.

Conflict of interest statement. None declared.

REFERENCES

1. Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.
2. Chittur,S.V. (2004) DNA microarrays: tools for the 21st Century. *Comb. Chem. High Throughput Screen* **7**, 531–537.
3. Stamatoyannopoulos,J.A. (2004) The genomics of gene expression. *Genomics*, **84**, 449–457.
4. Draghici,S., Khatri,P., Eklund,A.C. and Szallasi,Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*., **22**, 101–109.
5. Dong,S., Wang,E., Hsie,L., Cao,Y., Chen,X. and Gingeras,T.R. (2001) Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res*., **11**, 1418–1424.
6. Kennedy,G.C., Matsuzaki,H., Dong,S., Liu,W.M., Huang,J., Liu,G., Su,X., Cao,M., Chen,W., Zhang,J. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol*., **21**, 1233–1237.
7. Kapranov,P., Cawley,S.E., Drenkow,J., Bekiranov,S., Strausberg,S.L., Fodor,S.P. and Gingeras,T.R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.
8. Zhang,L., Miles,M.L. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*., **21**, 818–821.
9. Hekstra,D., Taussig,A.R., Magnasco,M. and Naef,F. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res*., **31**, 1962–1968.
10. Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.M., Lu,G., Fang,J., Liu,W.M., Ryder,T. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.
11. Halperin,A., Buhot,A. and Zhulina,E.B. (2004) Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophys. J*., **86**, 718–730.
12. Vainrub,A. and Pettitt,B.M. (2004) Theoretical aspects of genomic variation screening using DNA microarrays. *Biopolymers*, **73**, 614–620.
13. Held,G.A., Grinstein,G. and Tu,T. (2006) Relationship between gene expression and observed intensities in DNA microarrays—a modeling study. *Nucleic Acids Res*., **34**, e70.
14. Held,G.A., Grinstein,G. and Tu,T. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. zNatl Acad. Sci. USA*, **100**, 7575–7580.
15. Sugimoto,N., Nakano,S., Katoh,MN., Matsumura,A., Nakamuta,H., Ohmichi,T., Yoneyama,M. and Sasaki,M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211–11216.
16. SantaLucia,J., Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
17. Budarf,M.L., Collins,J., Gong,W., Roe,B., Wang,Z., Bailey,L.C., Sellinger,B., Michaud,D., Driscoll,D.A. and Emanuel,B.S. (1995) Cloning a balanced translocation associated with DiGeorge syndrome and identification of a disrupted candidate gene. *Nature Genet*., **10**, 269–278.
18. Lane,S., Evermann,J., Loge,F. and Call,D.R. (2004) Amplicon secondary structure prevents target hybridization to oligonucleotide microarrays. *Biosens. Bioelectron*., **20**, 728–735.
19. Huang,J., Wei,W., Chen,J., Zhang,J., Liu,G., Di,X., Mei,R., Ishikawa,S., Aburatani,H., Jones,K.W. *et al*. (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*, **7**, 83.
20. Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*., **65**, 6071–6079.
21. Heaton,R.J., Peterson,A.W. and Georgiadis,R.W. (2001) Electrostatic surface plasmon resonance: direct electric field-induced hybridization and denaturation in monolayer nucleic acid films and label-free discrimination of base mismatches. *Proc. Natl Acad. Sci. USA*, **98**, 3701–3704.
22. Vainrub,A. and Pettitt,P.M. (2002) Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., **66**, 041905.