Published in final edited form as:

Comput Biol Chem. 2007 April ; 31(2): 92–98.

Using probe secondary structure information to enhance Affymetrix GeneChip background estimates

Raad Z. Gharaibeh, Anthony A. Fodor, and Cynthia J. Gibas*

Bioinformatics Research Center, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, North Carolina 28223, USA.

Abstract

High-density short oligonucleotide microarrays are a primary research tool for assessing global gene expression. Background noise on microarrays comprises a significant portion of the measured raw data. A number of statistical techniques have been developed to correct for this background noise. Here, we demonstrate that probe minimum folding energy and structure can be used to enhance a previously existing model for background noise correction. We estimate that probe secondary structure accounts for up to 3% of all variation on Affymetrix microarrays.

Keywords

DNA Microarray; Probe secondary structure; Background correction

1. Introduction

Microarray technology holds the promise of capturing global gene expression by providing global molecular snapshots of the cell's transcriptional machinery products (Lockhart et al., 1996). The ultimate goal of gene expression microarrays is to measure the abundance of each known transcript in the sample under investigation. The abundance is inferred from the signal generated by each probe as a result of a hybridization reaction with a labeled target (transcript). However, this signal includes *background noise* that not only measures the target abundance, but also non-specific binding and autofluorescence of the chip surface.

In the Affymetrix GeneChip system, each transcript's abundance is measured by a set of 11-20 probe pairs. Each pair is composed of a perfect match probe (PM), which exactly complements a region on the transcript, and a mismatch probe (MM), which is identical to the PM probe except at the 13th base, where the reverse complement nucleotide is introduced. MM probes were originally introduced by Affymetrix to measure background noise. However, it has been shown by many groups that MM contain significant amount of PM signal and are therefore unreliable as estimators of background noise (Chudin et al., 2001; Forman et al., 1998; Irizarry et al., 2003; Naef et al., 2003). A true estimate of background noise would improve the quality of Affymetrix GeneChip data.

Inconsistency of the signal generated from each probe is a common phenomenon in GeneChip microarray experiments (Li and Wong, 2001; Nielsen et al., 2005). The differences in the signal produced can be attributed to many sources: optical noise, cross-hybridization, dye-related

*Corresponding author. Tel.: +1 704 687 8378; fax: +1 704 687 6610. E-mail address: cgibas@uncc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

contributions and probe sequence composition. Many algorithms have been developed to attempt to correct for these inconsistencies (Irizarry et al., 2006; Wu and Irizarry, 2005; Zhang et al., 2003). In particular, it has been found that probe sequence composition can significantly affect the intensity of the signal generated from that probe, independent of the concentration of its target. A number of groups have suggested models where the background intensity of probes could be estimated based on their sequence composition (Naef and Magnasco, 2003; Zhang et al., 2003).

The process of nucleic acid hybridization in solution has been well studied and models such as the nearest-neighbor model provide a robust description of hybridization thermodynamics (SantaLucia and Hicks, 2004). Probe-target hybridization on the microarray surface, however, does not follow the solution analogue, and the nearest-neighbor parameters that describe solution hybridization appear to be different than those for microarrays (Zhang et al., 2003). On-chip DNA hybridization is likely to be complicated by the geometric constraints of having one strand (i.e. probe) attached to the surface of the chip (Shchepinov et al., 1997). In addition, many other factors like probe and target secondary structure, effective reaction volume, electrostatics, diffusion and surface effects, reaction thermodynamics and kinetics, competitive binding effects, hybridization buffer composition and probe-probe interactions are believed to affect microarray DNA hybridization (Lima et al., 1992; Southern et al., 1999).

In this study, we examine the effect of predicted probe secondary structure on background hybridization in Affymetrix microarrays. Although microarray probes are attached to the surface of the chip, they are dynamic molecules that, depending on their sequence composition, can fold onto themselves into stable secondary structure. Such stable secondary structure has the potential to interfere with probe-target hybridization (Lima et al., 1992). Consequently, the signal obtained from such probes may not reflect the actual transcript concentrations. It has been shown, for example, that a stable secondary structure motif in a 20-mer probe dramatically decreases the final signal obtained to a point where the probe is considered insensitive to its intended target (Anthony et al., 2003).

Microarray probes are usually screened for the presence of stable secondary structure either by a simple base complementarity check or using more sophisticated and time consuming energy minimization algorithms (Markham and Zuker, 2005). The base complementarity check is more routinely used for its simplicity and speed. Discrepancies between methods do exist, and there are no guidelines that determine which method is preferable (Koehler and Peyret, 2005). It is therefore likely that, despite these screening procedures, a significant amount of secondary structure is present in probes in microarray experiments.

Here we propose that the background noise of each probe can be modeled as a function of its sequence composition and its minimum folding energy and secondary structure. By incorporating probe secondary structure information into a previously described model of background concentration (Naef and Magnasco, 2003), we improved the fit of that model to microarray data by 1-3% with minimal addition of significant free parameters.

2. Methods

2.1. Data sets

Seven data sets were used in this study (Table 1): the human genome U133 Latin Square data set (http://www.affymetrix.com/support/technical/sample_data/datasets.affx), theChoe control data set (Choe et al., 2005), a Leukemia data set (Armstrong et al., 2002), a Malaria PM only data set (Le Roch et al., 2003), an Etoposide response data set (Fodor et al., 2006), a BK potassium channel knockout data set (Meredith et al., 2006; Pyott et al., 2006) and an alternative splicing PM only tiling microarray data set (Sugnet et al., 2006).

2.2. System and software

All the computational work was done on a 73-node Apple cluster. Each node is a dual 2.7 GHz PowePC G5 with 2GB RAM running Mac OSX 10.4.

Secondary structure prediction was done using the hybrid-min-ss program of the UNAFold-2.5 software package (Markham and Zuker, 2005). All probes were folded as single DNA strands at 45 °C and 1.0 M sodium concentration. All other options were set to the program defaults.

Simple linear model fitting and *p*-value calculations were done using R linear model function (lm) (<http://www.r-project.org>). The Naef and Magnasco model (Naef and Magnasco, 2003) and the position-dependent secondary-structure attenuated affinity model were implemented in Perl. All Perl code is available upon request.

3. Results

3.1. Simple linear models

The signal intensity generated from each probe can be modeled as:

$$I_j = O_j + N_j + S_j \quad (1)$$

Where I is the raw intensity value of probe j , O is the optical noise, N is the background noise of non-specific binding, and S is the signal generated from specific binding between probe j and its intended target (Wu and Irizarry, 2005). In this paper, we do not model the signal and none of our models therefore contain terms for S . Since the S term, which we are ignoring in our models, is significantly higher in the PM probes than the MM probes, each probe type was modeled separately.

Controlling the GC content of the probe is one of the basic principles of microarray probe design. A probe with high GC content tends to hybridize better and to form a stable duplex with both target and non-target sequences. A simple linear model that relates probe intensity to GC content can be written as follows:

$$I_j = B0 + B1\langle GC \rangle_j + \epsilon_j \quad (2)$$

where I is the raw intensity of probe j , $\langle GC \rangle_j$ is the number of GC nucleotides in probe j (which is a number between 0 and 25), $B0$ and $B1$ are free parameters and ϵ_j is an error term. The model explains a modest amount of the overall intensity when applied to the Latin Square data set; $R^2 \approx 0.02$ for PM and 0.12 for MM (Fig. 1). The model explains more of the MM probes intensity because most of the signal obtained from MM probes is background noise. MM intensity is therefore more independent of the concentration of the target gene.

We wondered, compared to the GC content, how much of the background noise probe secondary structure would explain when put into a simple linear model. The free energy of probe secondary structure formation (ΔG_{ss}) is an indicator of the stability of secondary structure in which the probe folds on itself. The more stable the secondary structure, the less a probe will be able to hybridize to its target or non-target sequences. As a result, one would expect to observe a low signal from such probe.

How much of all probe variance can be explained directly by secondary structure predictions? A simple linear model is:

$$I_j = B0 + B1\langle \Delta G_{ss} \rangle_j + \epsilon_j \quad (3)$$

where $\langle \Delta G_{ss} \rangle_j$ is probe j minimum folding energy in Kcal/mol.

If we apply this simple linear model to the Latin Square data set, we find a very low r -squared values; $R^2 < 10^{-4}$ (Fig. 1). However, the p -value of the null hypothesis that the $B1$ parameter is equal to zero is rejected with high confidence (Fig. 1). These data suggest that there is a statistically significant influence of ΔG_{ss} on the observed intensity, although this relationship does not explain very much of the overall intensity on the array.

One may argue that the low r -squared values in Eq. 3 are due to the fact that ΔG_{ss} value does not reflect the size of the secondary structure motif found in that probe and the number of free bases available for hybridization. The program hybrid-min-ss reports for the most stable secondary structure of a probe whether a given nucleotide is involved in secondary structure formation or not. We can define a value, S_L , which is the longest stretch of nucleotides that are not involved in secondary structure formation (for example, $S_L=10$ in Fig. 2). To investigate the relationship between the longest free string of bases of the folded probe (S_L) and the observed intensity, we can again apply a simple linear model:

$$I_j = B0 + B1 \langle S_L \rangle_j + \epsilon_j \quad (4)$$

where $\langle S_L \rangle_j$ is the longest free string of bases in probe j based on its minimum energy structure.

This model also has a very low r -squared values when applied to the Latin Square data set; $R^2 < 10^{-3}$ (Fig. 1), only slightly higher than Eq. 3 suggesting little direct effects of probe S_L on the observed intensity. However, the p -value of the null hypothesis that the $B1$ parameter is equal to zero is also rejected with high confidence (Fig. 1).

We see that GC content can explain a modest amount of overall intensity. Models based on secondary structure explain much less of the intensity data, although they are still highly statistically significant.

3.2. Position-dependent secondary-structure attenuated affinity model

Since the three simple linear models (Eq. 2-4) all hold significant relationships with the observed intensity (Fig. 1), we wanted to combine them into one model that takes into account the base composition, ΔG_{ss} and S_L of the probe. We found that a simple linear combination of GC, ΔG_{ss} and S_L did not significantly improve on the power of the individual models (data not shown). We reasoned that a model that is aware of each probe's base position and involvement in the overall secondary structure of the probe would outperform models that ignore this information.

The model of Naef and Magnasco (Naef and Magnasco, 2003) provides a starting point that meets our requirement for individual base information. In this model, probe background is modeled based on sequence composition:

$$\ln \langle B/M \rangle = \sum_{k=1}^{25} \sum_{l \in (A, T, C, G)} S_{lk} A_{lk} \quad (5)$$

where B is the raw probe intensity, M is the median intensity of the array, l is nucleotide index, k is the position of l along the probe, S a Boolean variable equals to 1 if the probe sequence has l at k and zero otherwise, and A is the per-site per letter affinity. To clarify, consider the probe shown in Fig. 2, then Eq. 5 will read:

$$\begin{aligned}
\ln \langle B/M \rangle = & (S_{1G} \times A_{1G}) + (S_{1A} \times A_{1A}) + (S_{1T} \times A_{1T}) + (S_{1C} \times A_{1C}) + \\
& (S_{2G} \times A_{2G}) + (S_{2A} \times A_{2A}) + (S_{2T} \times A_{2T}) + (S_{2C} \times A_{2C}) + \\
& (S_{3G} \times A_{3G}) + (S_{3A} \times A_{3A}) + (S_{3T} \times A_{3T}) + (S_{3C} \times A_{3C}) + \\
& \dots \dots \dots \\
& + (S_{25G} \times A_{25G}) + (S_{25A} \times A_{25A}) + (S_{25T} \times A_{25T}) + (S_{25C} \times A_{25C}) \\
\ln \langle B/M \rangle = & A_{1C} + A_{2G} + A_{3A} + \dots + A_{25C}
\end{aligned}$$

Eq. 5 is a simple model that has four free parameters for each probe base (100 free parameters for a 25-base probe). The values of these 100 free parameters are generated by linear least squares fit (Naef and Magnasco, 2003). Given the large number of probes on each chip (about half a million for the human genome U133 chip, for example) over-fitting is not a concern.

In our approach, we add the continuous variable θ to reflect the involvement of the probe nucleotides in secondary structure formation. The model now is written as:

$$\ln \langle B/M \rangle = \sum_{k=1}^{25} \sum_{l \in \{A, T, C, G\}} \theta_{lk} S_{lk} A_{lk} \quad (6)$$

The θ term reflects the degree to which an individual probe base participates in secondary structure formation. In our model, it is represented by any value between 0 and 1. There are a large number of ways in which values for θ could be generated. We made the following simplifying assumptions. We begin by considering nucleotides that are not involved in secondary structure formation. In cases where a probe's $\Delta G_{ss} > 0$ Kcal/mol we can set θ for all bases within that probe to 1. Likewise, when a base within a probe is not involved in secondary structure hydrogen bonding (yellow ovals in Fig 2, for example), we can set θ to 1 for that base. To calculate θ for the remaining bases, we set a $\Delta G_{ss\text{-cutoff}}$ value, below which θ will be constant ($\theta = tb$). We assumed that the relationship between θ and ΔG_{ss} is linear in the region where ΔG_{ss} is between $\Delta G_{ss\text{-cutoff}}$ and 0 (Fig. 3). From the assumption of linearity, we can derive a slope and an intercept to yield:

$$\theta = \frac{tb - 1}{\Delta G_{ss\text{-cutoff}}} \langle \Delta G_{ss} \rangle_j + 1 \quad (7)$$

This equation has two unknown parameters $\Delta G_{ss\text{-cutoff}}$ and tb . To find the best values for these parameters, we tested the effects of changing $\Delta G_{ss\text{-cutoff}}$ and tb on the performance of the model (Eq. 6) on a single chip from the Latin Square data set. We found that the best performance of the model was obtained at $\Delta G_{ss\text{-cutoff}} = -3.6$ Kcal/mol and a $tb = 0.35$ (Fig. 4).

To summarize, we define our *position-dependent secondary-structure attenuated affinity model* (PSAA) as Eq. 6, where B is the raw probe intensity, M is the median intensity of the array, l is letter index, k is the position of l along the probe, A is the per-site per letter affinity, S a Boolean variable equal to 1 if the probe sequence has l at k and zero otherwise, and θ is:

- a. 1, if the probe $\Delta G_{ss} > 0$ Kcal/mol.
- b. 1, if the probe $\Delta G_{ss} \leq 0$ Kcal/mol and l is not involved in secondary structure hydrogen bonding.
- c. $\left(\begin{array}{ll} 0.35, & \text{if } \Delta G_{ss} \leq -3.6 \\ \frac{0.65}{3.6} \langle \Delta G_{ss} \rangle_j + 1, & \text{if } -3.6 < \Delta G_{ss} \leq 0 \end{array} \right)$ and l is involved in secondary structure hydrogen bonding.

Here, the involvement of each probe base in secondary structure hydrogen bonding is based on its minimum energy structure.

When we consider the folded probe presented in Fig. 2, Eq. 6 reads:

$$\begin{aligned} \ln \langle B/M \rangle = & \\ \theta & ((S_{1G} \times A_{1G}) + (S_{1A} \times A_{1A}) + (S_{1T} \times A_{1T}) + (S_{1C} \times A_{1C})) + \\ \theta & ((S_{2G} \times A_{2G}) + (S_{2A} \times A_{2A}) + (S_{2T} \times A_{2T}) + (S_{2C} \times A_{2C})) + \\ \theta & ((S_{3G} \times A_{3G}) + (S_{3A} \times A_{3A}) + (S_{3T} \times A_{3T}) + (S_{3C} \times A_{3C})) + \\ & \dots \dots \dots \\ + \theta & ((S_{25G} \times A_{25G}) + (S_{25A} \times A_{25A}) + (S_{25T} \times A_{25T}) + (S_{25C} \times A_{25C})) \\ \ln \langle B/M \rangle = & A_{1C} + A_{2G} + 0.64A_{3A} + \dots + A_{25C} \end{aligned}$$

The model defined in Eq. 6 was fitted to all the data sets (Table 1). The fitting was done on the PM and MM probes separately. Table 1 shows a comparison between the native Naef and Magnasco model (Naef and Magnasco, 2003) and our position-dependent secondary-structure attenuated affinity model. We see that including probe secondary structure information improved the fit of the native Naef and Magnasco model (Naef and Magnasco, 2003) by 1-3%, depending on the chip and probe type. Note that all the models (Eq. 2,3,4,5 and 6) perform better on the MM probes due to the higher background noise present in the MM signal.

3.3. Gains in performance can not be trivially explained by additional free parameters

We note that there are two distinct kinds of free parameters in our model. The 100 free parameters from the original Naef and Magnasco model (Eq. 5) are calculated for each chip by linear least squares fit. We have added two free parameters in Eq. 6, $\Delta G_{ss-cutoff}$ and tb . These parameters were determined from one of the Latin Square data set chips from the curves shown in Fig. 4 and were held constant for all the data sets in this paper. Given that our fits contain between 173,262 and 496,468 data points (Table 1), it seems unlikely that the improvements in performance could be explained by the addition of the free parameters $\Delta G_{ss-cutoff}$ and tb . Nonetheless, to further rule out this possibility, we refolded the Latin Square data set probes with either a completely random sequence (generated with an equal probability of A, C, G and T) or a shuffled sequence. Then we fed Eq. 6 the original probe sequence (i.e. the right l at k) along with the new ΔG_{ss} and the new minimum energy structure that resulted from folding the random or shuffled probe sequence. For the random sequence case, the performance of the original Naef and Magnasco model (Naef and Magnasco, 2003) was severely degraded (Fig. 5). For the shuffled sequences, the probe's base composition is not affected, but the position of each base has been changed due to the shuffling process. For the shuffled sequences, the fit of the model dropped down to that of the original Naef and Magnasco model (Naef and Magnasco, 2003). These results on shuffled and random sequence show that the presence of the two additional free parameters $\Delta G_{ss-cutoff}$ and tb cannot by themselves explain the improved performance over the original Naef and Magnasco model (Naef and Magnasco, 2003). This strongly supports our argument that the gain in the r-squared values of our model came from including probe secondary structure information and do not arise trivially from the addition of free parameters.

4. Discussion

In the absence of a clear understanding of the microarray hybridization mechanisms and the frequent use of probes that fold into stable secondary structure under the hybridization conditions on microarrays, a model is needed to explain or approximate the effects of such behavior on microarray signal. Using simple linear models, we saw a modest relationship

($R^2 < 10^{-3}$) between probe intensity and its ΔG_{ss} or S_L . We propose as a more powerful alternative to two parameter linear models, a modification of the Naef and Magnasco model (Naef and Magnasco, 2003) to include probe secondary structure effects on the background intensity. Our model works by equating an increase in secondary structure with a decreased contribution to a linear least square fit. If a particular base is involved in secondary structure hydrogen bonding (Fig. 2), we assign it a low θ score depending on the overall ΔG_{ss} of the probe (Eq. 7). Consequently, this base contribution is attenuated in the Naef and Magnasco model (Naef and Magnasco, 2003). Consider, for example, the third adenine base in the folded probe presented in Fig. 2, in the Naef and Magnasco model (Naef and Magnasco, 2003) its contribution to the brightness is A_{3A} . Based on the predictions of hybrid-min-ss, this base is involved in secondary structure hydrogen bonding and we therefore expect a reduced contribution to the intensity caused by background binding. We therefore attenuate its contribution to the brightness by θ , and its contribution now is $0.64A_{3A}$ instead of A_{3A} . Results attenuated by θ have more power than the original Naef and Magnasco model (Naef and Magnasco, 2003) over a wide range of Affymetrix data sets (Table 1).

The secondary structure information used here is based on the minimum folding energy (ΔG_{ss}) and the minimum energy structure, as predicted by an energy minimization algorithm (Markham and Zuker, 2005) that uses the nearest-neighbor parameters (SantaLucia, 1998) to predict secondary structure of single-stranded DNA molecules in solution. In the absence of clear understanding of the effects of the geometric constraints of attaching one end of the DNA probe to the chip surface on its secondary structure, the nearest-neighbor parameters represent a reasonable approximation for microarray (Held et al., 2003). We are also fully aware that single-stranded DNA molecules are highly dynamic and each molecule is likely to exist in an ensemble of structures. Based on that, predicting the minimum folding energy (ΔG_{ss}) and the minimum energy structure for any single-stranded DNA molecule can be different when using different prediction algorithms, even when the same folding conditions are used. The results presented here are based on the minimum folding energy (ΔG_{ss}) and the minimum energy structure calculated using UNAFold (Markham and Zuker, 2005). It has been shown that the differences in the predicted minimum folding energy (ΔG_{ss}) and the minimum energy structure between different prediction algorithms are small (Ding et al., 2004; Ratushna et al., 2005). Consequently, we would expect similar results no matter which of the currently popular secondary structure prediction algorithms were used.

The results presented in this work suggest that, on average, 1-3% of all the intensities on Affymetrix GeneChip microarrays can be explained by probe secondary structure independent of any target information. Given that not all the probes form stable secondary structure (50% of the human genome U133 Latin Square data set probes, for example have predicted $\Delta G_{ss} > 0$), the 1-3% enhancement over the original model is quite satisfactory, and represent a step forward in understanding the factors that affect the on-chip hybridization process.

The current design of GeneChip microarrays devotes half of the chip to MM probes. The sole purpose of these probes is to estimate the background noise portion present in the PM signal to enhance the chip ability to detect differently expressed genes. Advances in the ability to correctly estimate background noise on Affymetrix GeneChip microarrays based on probe sequence information may in the future eliminate the need of MM probes on these arrays offering more space to interrogate more genes on the same array.

Acknowledgements

This research was supported in part by NIH 1R01GM072619-01 (C.J.G.) and by the UNC-Charlotte GASP program (R.Z.G.). Cel files for the splicing microarray data set were generously provided by Manny Ares.

References

- Anthony RM, Schuitema AR, Chan AB, Boender PJ, Klatser PR, Oskam L. Effect of secondary structure on single nucleotide polymorphism detection with a porous microarray matrix; implications for probe selection. *Biotechniques* 2003;34:1082–1089. [PubMed: 12765035]
- Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 2002;30:41–47. [PubMed: 11731795]
- Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* 2005;6:R16. [PubMed: 15693945]
- Chudin E, Walker R, Kosaka A, Wu S, Rabert D, Chang T, Kreder D. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip(R) arrays. *Genome Biol* 2001;3:research0005.0001–research0005.0010. [PubMed: 11806828]
- Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. *Nucl Acids Res* 2004;32:W135–141. [PubMed: 15215366]
- Fodor AA, Tickle TL, Richardson C. Towards the uniform distribution of null p-values on microarrays. *Genome Biol*. 2006Submitted
- Forman JE, Walton ID, Stern D, Rava RP, Trulson MO. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesized oligonucleotide arrays. *ACS Symp Ser* 1998;682:206–228.
- Held GA, Grinstein G, Tu Y. Modeling of DNA microarray data by using physical properties of hybridization. *Proc Natl Acad Sci U S A* 2003;100:7575–7580. [PubMed: 12808153]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat* 2003;4:249–264.
- Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006;22:789–794. [PubMed: 16410320]
- Koehler RT, Peyret N. Effects of DNA secondary structure on oligonucleotide probe binding efficiency. *Comput Biol Chem* 2005;29:393–397. [PubMed: 16290040]
- Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 2003;301:1503–1508. [PubMed: 12893887]
- Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 2001;98:31–36. [PubMed: 11134512]
- Lima WF, Monia BP, Ecker DJ, Freier SM. Implication of RNA structure on antisense oligonucleotide hybridization kinetics. *Biochemistry* 1992;31:12055–12061. [PubMed: 1280997]
- Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675–1680. [PubMed: 9634850]
- Markham NR, Zuker M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* 2005;33:W577–581. [PubMed: 15980540]
- Meredith AL, Wiler SW, Miller BH, Takahashi JS, Fodor AA, Ruby NF, Aldrich RW. BK calcium-activated potassium channels regulate circadian behavioral rhythms and pacemaker output. *Nat Neurosci* 2006;9:1041–1049. [PubMed: 16845385]
- Naef F, Magnasco MO. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003;68:011906. [PubMed: 12935175]
- Naef F, Socci ND, Magnasco M. A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics* 2003;19:178–184. [PubMed: 12538237]
- Nielsen HB, Gautier L, Knudsen S. Implementation of a gene expression index calculation method based on the PDNN model. *Bioinformatics* 2005;21:687–688. [PubMed: 15509605]
- Pyott SJ, Meredith AL, Fodor AA, Vázquez AE, Yamoah EN, Aldrich RW. Cochlear function in mice lacking the BK channel α , β 1, or β 4 subunits. *J Biol Chem*. 2006In Press

- Ratushna VG, Weller JW, Gibas CJ. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* 2005;6:31. [PubMed: 15755320]
- SantaLucia J Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 1998;95:1460–1465. [PubMed: 9465037]
- SantaLucia J Jr, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 2004;33:415–440. [PubMed: 15139820]
- Shchepinov MS, Case-Green SC, Southern EM. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res* 1997;25:1155–1161. [PubMed: 9092624]
- Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. *Nat Genet* 1999;21:5–9. [PubMed: 9915493]
- Sugnet CW, Srinivasan K, Clark TA, Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D, Ares M. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol* 2006;2:e4. [PubMed: 16424921]
- Wu Z, Irizarry RA. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol* 2005;12:882–893. [PubMed: 16108723]
- Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol* 2003;21:818–821. [PubMed: 12794640]

Fig 1.
 R^2 distribution for the simple linear models (Eq.s 2-4) for all the U133 Latin Square chips. The null hypothesis that the BI parameter is equal to zero is rejected with high confidence ($P < 10^{-4}$) for all the models.

Fig 2. Folded probe showing its sequence, minimum folding energy (ΔG_{ss}) and minimum energy structure. The longest free string of bases of the folded probe (S_L) is shown in gray box; bases involved in hydrogen bonding are shown in green ovals.

Fig 3.

A model for the relationship between ΔG_{ss} and θ for the bases involved in secondary structure formation.

Fig 4. Effects of changing the values of $\Delta G_{ss\text{-cutoff}}$ and tb on the performance of the position-dependent secondary-structure attenuated affinity model using the human genome U133 Latin Square Experiment 2 Replicate 1 PM probes.

Fig 5.

Effects of changing tb or $\Delta G_{ss-cutoff}$ on the performance of the PSAA model. Effects of changing (A) the values of tb while holding $\Delta G_{ss-cutoff} = -3.6$ or (B) the values of $\Delta G_{ss-cutoff}$ while holding $tb = 0.35$ on the performance of PSAA: the position-dependent secondary-structure attenuated affinity model (Eq. 6). Data shown are for the human genome U133 Latin Square Experiment 2 Replicate 1 PM probes. NM: Naef and Magnasco (Naef and Magnasco, 2003) model (Eq. 5). The suffixes (-SH) and (-RD) indicates the R^2 after generating the minimum folding energy (ΔG_{ss}) and the minimum energy structure from shuffled and random sequences, respectively (see section 3.3 for explanation).

R² of Naef and Magnasco model (Naef and Magnasco, 2003) (NM) and the position-dependent secondary-structure attenuated affinity model (PPSA) for the seven data sets used in this study. Results presented as average R² ± SD.

Table 1

Data set	<i>nd</i> ^a	<i>np</i> ^b	NM	PSAA ^c
Latin Square	42	248152	0.17±0.009	0.184±0.010
Choe (Choe et al., 2005)	6	248152	0.40±0.009	0.416±0.009
		195994	0.20±0.022	0.216±0.025
Leukemia (Armstrong et al., 2002)	72	195994	0.46±0.017	0.49±0.017
		201800	0.49±0.063	0.51±0.062
Etoposide response (Fodor et al., 2006)	60	201800	0.60±0.036	0.61±0.035
		496468	0.05±0.040	0.06±0.040
BK knockout (Meredith et al., 2006;Pyott et al., 2006)	20	496468	0.11±0.062	0.12±0.062
		496468	0.09±0.035	0.10±0.036
Splicing microarray (Sugnet et al., 2006)	75	496468	0.29±0.050	0.30±0.049
		505916	0.30±0.062	0.31±0.063
Malaria (Le Roch et al., 2003)	17	173262	0.36±0.043	0.38±0.043

^a*nd*: number of chips.

^b*np*: number of probes.

^cThe differences in R² between NM and PSAA are all statistically significant (P < 10⁻³) using paired one-sided Wilcoxon test.