



GenXHC: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data

Jim C. Huang^{1,*}, Quaid D. Morris¹, Timothy R. Hughes² and Brendan J. Frey¹

¹Probabilistic and Statistical Inference Group, Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada M5S 3G4 and ²Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada M5G 1L6

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Microarray designs containing millions to hundreds of millions of probes that tile entire genomes are currently being released. Within the next 2 months, our group will release a microarray data set containing over 12 000 000 microarray measurements taken from 37 mouse tissues. A problem that will become increasingly significant in the upcoming era of genome-wide exon-tiling microarray experiments is the removal of cross-hybridization noise. We present a probabilistic generative model for cross-hybridization in microarray data and a corresponding variational learning method for cross-hybridization compensation, GenXHC, that reduces cross-hybridization noise by taking into account multiple sources for each mRNA expression level measurement, as well as prior knowledge of hybridization similarities between the nucleotide sequences of microarray probes and their target cDNAs.

Results: The algorithm is applied to a subset of an exon-resolution genome-wide Agilent microarray data set for chromosome 16 of *Mus musculus* and is found to produce statistically significant reductions in cross-hybridization noise. The denoised data is found to produce enrichment in multiple gene ontology–biological process (GO–BP) functional groups. The algorithm is found to outperform robust multi-array analysis, another method for cross-hybridization compensation.

Contact: jim@psi.toronto.edu

INTRODUCTION

Oligonucleotide microarrays (Lockhart *et al.*, 1996; Hughes *et al.*, 2001) are quickly becoming the *de facto* standard for measuring transcript abundance. These microarrays generate measurements that are reproducible across different platforms between separate labs using independently generated tissues samples (Lee *et al.*, 2003). Oligonucleotide arrays have

been applied to the detection of *in silico*-predicted mRNAs (Zhang *et al.*, 2004; Sun *et al.*, 2004), and to quantification of the relative levels of different splice isoforms (Johnson *et al.*, 2003; Le *et al.*, 2004). An exciting new application of oligonucleotide arrays containing 10^5 – 10^7 probes is genome-wide high-resolution tiling assays (Bertone *et al.*, 2004; Shoemaker *et al.*, 2001; Frey *et al.*, submitted for publication, 2005), allowing the profiling of individual known and predicted transcripts to identify new non-coding RNAs, new gene models or to refine existing ones.

An important component in designing an oligonucleotide array is ensuring that each probe binds to its target with high specificity. The affinity of a probe for various transcripts can be estimated via the binding free energy, ΔG , between the probe and a corresponding subsequence of the transcripts. When a probe has high affinity to a non-target transcript, as in Figure 1, the abundance levels of the target transcript may be obscured by cross-hybridization. Algorithms have been developed (Li and Stormo, 2001) to design probe sets for sets of transcripts that minimize cross-hybridization with their non-target sequence.

However, in the case of high-density oligonucleotide microarrays, appropriate highly specific probes are not easily designed and cross-hybridization is unavoidable. This situation occurs most frequently when there are only a small number of feasible candidate oligos for some target transcripts. When highly specific probes cannot be designed, the effect of cross-hybridization on the probe intensity should nonetheless be predictable, especially if we are also measuring the abundance of the non-target transcript using a different probe. This abundance measurement, along with the binding affinity of the probe for the non-target sequence, can be used to determine how much of the probe intensity is due to cross-hybridization. Figure 2 shows an example of cross-hybridization compensation in microarray data, in which each probe can hybridize to multiple transcripts to various degrees.

*To whom correspondence should be addressed.

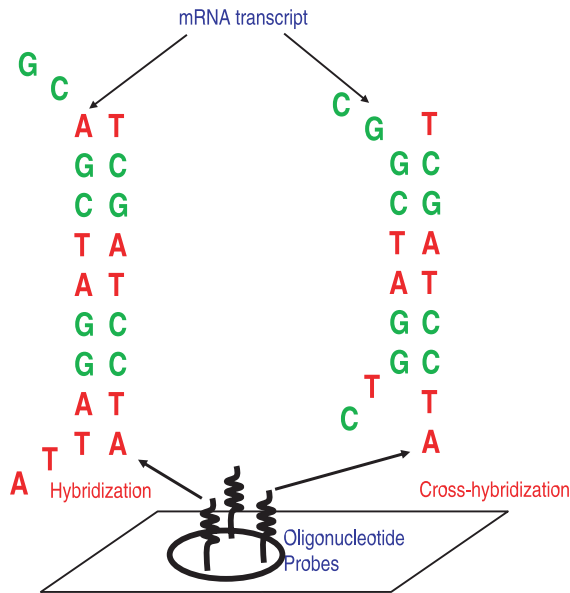


Fig. 1. Example of cross-hybridization on an oligonucleotide probe with two similar short nucleotide sequences: transcripts matching regions of the probe sequence are capable of hybridizing to the probe.

In this paper, we describe GenXHC, a probabilistic model for cross-hybridization compensation which estimates transcript abundances from probe intensities by accounting for potential cross-hybridization in high-density genome-wide microarray data. We use a BLAST (Altschul *et al.*, 1990) search against a database of transcripts to identify likely cross-hybridizing transcripts for each microarray probe. We use this information to fit a parameterized generative model of the observed probe intensities in terms of unobserved transcript abundances. Probabilistic inference in our generative model allows us to simultaneously estimate abundances of all targeted transcripts while removing the effects of cross-hybridization.

THE PROBLEM OF CROSS-HYBRIDIZATION IN MICROARRAY DATA

Although it is known that cross-hybridization can be a significant problem in microarray data analysis (Wren *et al.*, 2002), we demonstrate here the effect of cross-hybridization in a new genome-wide tiling dataset based on over 1 million probes (Frey *et al.*, submitted for publication). Figure 3 shows the relative frequencies of pairwise Pearson correlation coefficients (excluding self correlation) computed between reciprocal-best-match pairs and randomly matched pairs of normalized expression profile measurements in a subset of *Mus musculus* chromosome 16 gene expression data. Figure 3 shows that 33% of measured expression profiles that exhibit sequence similarity tend to exhibit large Pearson correlation values ($r > 0.95$) due to the effect of cross-hybridization, whereas

only 2% of randomly matched expression profiles exhibit such large pairwise correlation values. This indicates that the amount of cross-hybridization noise is significant, as the effect of cross-hybridization noise increases the correlation between measured gene expression profiles.

For a given set of cross-hybridizing transcripts, there can be variation in each transcript's ability to hybridize to a given probe sequence. One important factor that determines the potential for hybridization between a transcript and a probe is the magnitude of their pairwise stacking hybridization free energy, ΔG (SantaLucia *et al.*, 1998). This quantity is computed by summing (i.e. stacking) the free energies associated with each base pair. Different transcript-probe pairs will have free energy quantities that vary as functions of sequence length and base composition. Note that although the number of base pairs between two sequences determines their potential for hybridization, the proportion of the probe's intensity due to cross-hybridization depends on the relative abundance of the target and non-target transcripts at the time of hybridization. These quantities are often unknown and must be estimated. To accommodate for this missing information, we require the use of probabilistic models for cross-hybridization that account for uncertainties in the measurements.

PREVIOUS WORK IN CROSS-HYBRIDIZATION COMPENSATION

Previous work in cross-hybridization compensation using probabilistic models include Affymetrix' MAS 5.0 algorithm, robust multi-array analysis (RMA) and GeneChip RMA (GC-RMA), all of which have been applied to Affymetrix' GeneChip microarray technology (Irizarry *et al.*, 2003; Wu and Irizarry, 2004; Wu *et al.*, 2004). In GeneChips, a pair of probes for each target transcript is provided in the form of the perfect match (PM) probe and its corresponding mismatch (MM) probe, which provides a measure of cross-hybridization noise for the PM probe via imperfect complementarity to the PM target. The measurements from the MM probes provide calibration information that can be taken advantage of to perform compensation by subtracting them from their PM probe's measurement. MAS 5.0 performs cross-hybridization compensation by subtracting the MM measurements from their corresponding PM values and computing a robust average of the PM-MM values as an expression measure. However, adjusting for cross-hybridization using a PM-MM measure has been shown to yield estimates with large variance (Irizarry *et al.*, 2003; Wu and Irizarry, 2004; Wu *et al.*, 2004). The RMA method addresses the issue by using a global adjustment procedure based on a probabilistic linear model of the data that ignores the measurements from the MM probes: this has been shown to reduce variance in compensation estimates, but at the cost of increased bias at low expression levels. The GC-RMA algorithm expands on RMA by combining a probabilistic model with a physical model that uses the additional

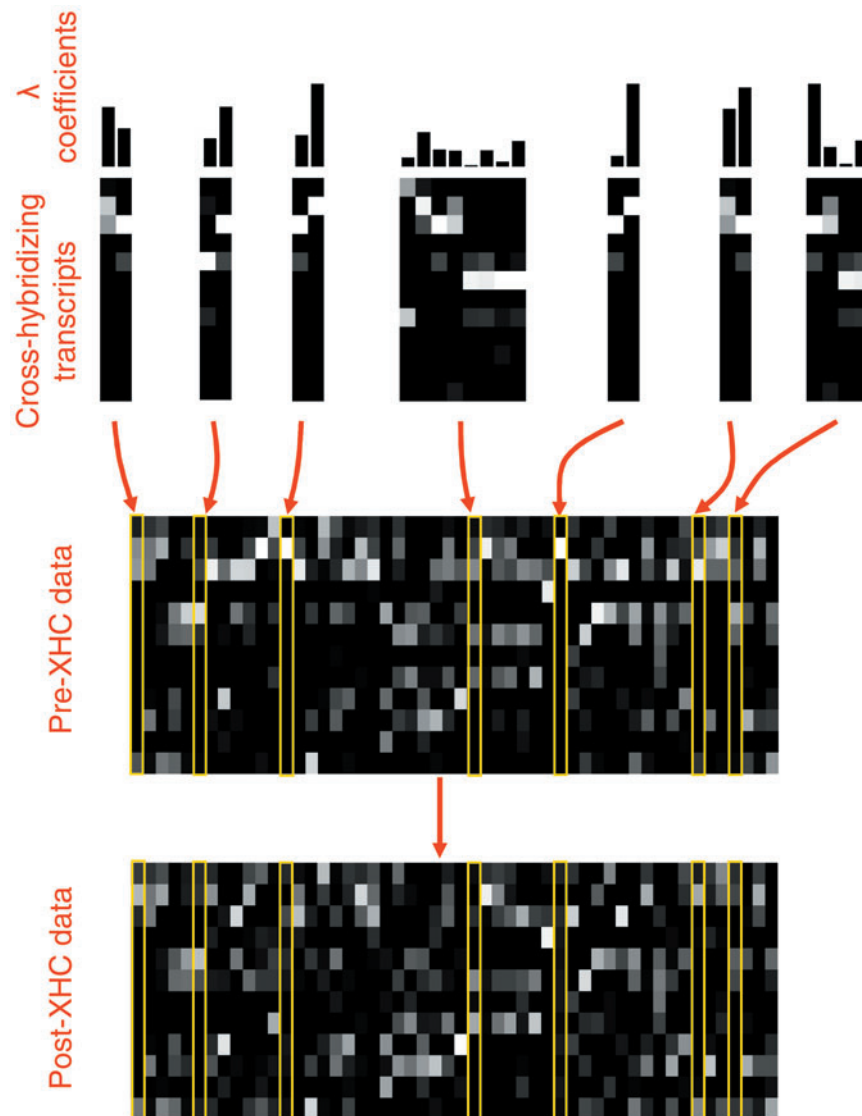


Fig. 2. A flow diagram of the generative process for cross-hybridization using sample measured expression profiles in chromosome 16 *M.musculus* gene expression data. Expression profiles are arranged by column, with measurements across the 12 tissue pools arranged along rows. White indicates high expression. Each probe can hybridize to multiple transcripts and thus measures components of expression from transcripts other than its target. The measured expression levels can be used in tandem with knowledge of probe-transcript hybridization constraints to infer expression levels for the transcripts.

MM measurements to perform cross-hybridization compensation. GC-RMA has been shown to reduce this bias, without much of an increase in variance with respect to RMA.

One concern with the GC-RMA model is that the MM probes are modeled as measuring only noise, though they have been shown to measure some amount of signal from their corresponding PM probe's target (Wu *et al.*, 2004). It should be possible to estimate each probe's target transcript abundance by modeling probe measurements as containing both signal and noise. This would allow for double the number of signal-measuring probes on a microarray, which can be a

significant savings in practice as microarray technology scales to higher and higher probe densities. This is the motivation for the GenXHC model, which uses latent variables and knowledge of the transcript population to estimate transcript abundances directly from measured expression data, without the need for additional calibration information. The problem of cross-hybridization compensation then becomes one of inference in which the maximum-likelihood settings for the latent variables correspond to the desired gene expression profiles, subject to an explicit sparsity constraint on probe-transcript interactions.

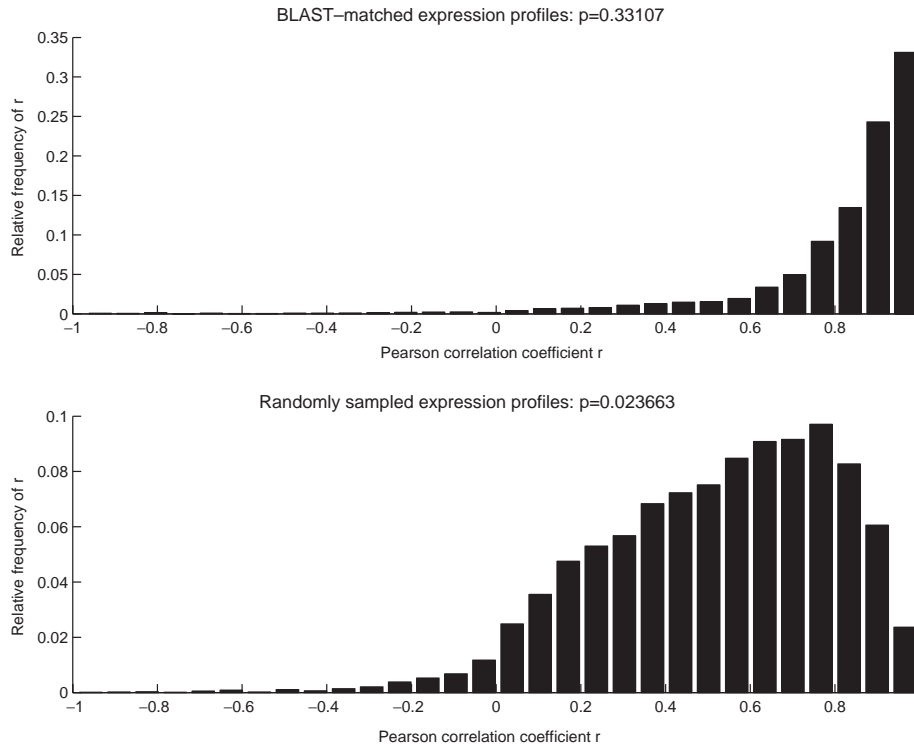


Fig. 3. Distribution of pairwise Pearson correlation coefficients for sample normalized *M.musculus* chromosome 16 gene expression data: cross-hybridization tends to occur between microarray probes with similar sequence.

A PROBABILISTIC GENERATIVE MODEL FOR CROSS-HYBRIDIZATION

We propose a generative model for cross-hybridization that can be viewed as a large-scale factor analysis model which includes constraints on probe-transcript hybridization interactions inferred by probe sequence-similarity. The model, while incorporating such prior constraints, makes very few additional assumptions based on prior knowledge about dependencies between model variables in order to avoid biasing the inferred data. Standard techniques for inference in factor analysis models cannot be applied to the problem of cross-hybridization compensation due to the size of our model. In the next section, we formulate a sparsity-constrained factor analysis model for cross-hybridization. We then describe an efficient algorithm for inference that iteratively estimates the latent variables and model parameters from the observed expression data.

Factor analysis

A probabilistic factor analysis model represents an N -dimensional observed data point, \mathbf{x} , as an affine mapping of another point, \mathbf{z} , that lies in an M -dimensional subspace, where $M < N$. More precisely, the factor analysis model represents the observed data \mathbf{x} as a linear combination of lower-dimensional latent factors \mathbf{z} , with weights given by the elements of the factor loading matrix $\mathbf{\Lambda}$ plus additive Gaussian

noise, \mathbf{v} , such that

$$\mathbf{x} = \mathbf{\Lambda}\mathbf{z} + \mathbf{v}. \quad (1)$$

This model can be also formulated probabilistically as:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = N(\mathbf{\Lambda}\mathbf{z}, \mathbf{\Psi})N(0, \mathbf{I}), \quad (2)$$

where the conditional mean is $E[\mathbf{x}|\mathbf{z}] = \mathbf{\Lambda}\mathbf{z}$, $\mathbf{\Lambda}$ is the $N \times M$ factor loading matrix with elements λ_{ij} and $\mathbf{\Psi} = \text{diag}(\psi_1^2, \psi_2^2, \dots, \psi_N^2)$ is the covariance matrix of the noise on the observed N -dimensional data, \mathbf{x} .

Cross-hybridization as a factor analysis model with sparsity constraints

The cross-hybridization problem can be formulated in terms of a factor analysis model with a sparsity structure on the weights λ_{ij} as follows: let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$ denote a matrix in which the t^{th} column denotes the noisy measurements across all N probes in the microarray for conditions $t = 1, \dots, T$, with x_{it} denoting the amount of mRNA transcript measured by probe i in condition t . Let $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T]$ denote the matrix of latent transcript levels to be inferred; the t^{th} column of \mathbf{Z} , therefore, denotes the latent mRNA expression levels across all M probed transcripts for condition t , with z_{jt} denoting the amount of mRNA transcript j and condition t . Let $\mathbf{\Lambda}$ denote the hybridization matrix that maps the set of latent transcript levels \mathbf{Z} to the set of measured profiles \mathbf{X} , with λ_{ij} denoting the hybridization coefficient associated with probe i and

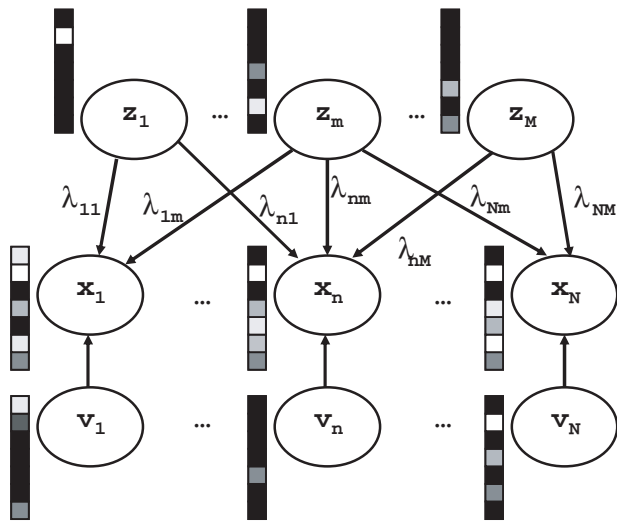


Fig. 4. Probabilistic graphical model representation for cross-hybridization: each observed expression profile \mathbf{x}_n is a weighted sum of latent expression profiles $\mathbf{z}_1, \dots, \mathbf{z}_M$ plus Gaussian noise \mathbf{v}_n .

transcript j . Thus, each measured expression profile can be expressed as a linear combination of a small number of latent expression profiles with additive Gaussian noise, as depicted in Figure 4. The problem of cross-hybridization compensation is, to infer the optimal settings for the latent expression profiles \mathbf{Z} , the hybridization matrix $\mathbf{\Lambda}$ and the noise variance model parameter Ψ , using noisy expression measurements \mathbf{X} .

Specifying a sparsity structure for the hybridization matrix

In the above model for cross-hybridization, the matrix $\mathbf{\Lambda}$ is treated as a model parameter and is, not given a probability distribution which accounts for uncertainty in the values for the hybridization coefficients λ_{ij} . We wish to introduce strong prior information about both the sparsity structure of the set of hybridization interactions between the microarray probes and the set of target transcripts, but also allow for uncertainty in the values of the hybridization coefficients, λ_{ij} . This necessitates a Bayesian factor model in which the factor loading matrix, $\mathbf{\Lambda}$, is given a prior distribution $p(\mathbf{\Lambda})$ which encodes our prior knowledge of the problem. We model the elements of the factor loading matrix $\mathbf{\Lambda}$ as being statistically independent, allowing the prior $p(\mathbf{\Lambda})$ to be factorized into a product of local distributions, $p(\lambda_{ij})$:

$$p(\mathbf{\Lambda}) = \prod_{(i,j)} p(\lambda_{ij}). \quad (3)$$

Not all of the probed transcripts can hybridize to any given microarray probe (i.e. many of the λ_{ij} coefficients are 0). By using local sequence-alignment methods such as BLAST, we can establish constraints on the possible hybridization interactions within a given set of probes and their target transcripts by

thresholding matching sequences by their BLAST E -values. This is approximately equivalent to thresholding matching sequences by their stacking hybridization free energy ΔG , as a strongly-bound pair of nucleotide sequences will have a larger number of nucleotide base pairs and hence, a lower BLAST E -value: only probe-transcript pairs with E -values below the threshold are allowed to have non-zero hybridization coefficients. Similarly, paired sequences with a small number of base pairs will be weakly paired and hence will be unlikely to make a contribution to cross-hybridization noise (Hughes *et al.*, 2001).

Note that the sparsity structure of the hybridization matrix, $\mathbf{\Lambda}$, could be partially constrained according to probe-transcript free energy parameters ΔG ; a simple approach would be to set the non-zero elements of $\mathbf{\Lambda}$ to some parametric form of the free energy, and then solve for the latent transcript levels \mathbf{Z} that explain the measurements for the given $\mathbf{\Lambda}$. However, we have found empirically that the free energy alone does not determine the hybridization strength between a probe and a transcript. The probabilistic model above can account for this ambiguity of the hybridization coefficients due to incomplete information; instead of setting the non-zero coefficients deterministically, the generative model represents uncertainty in the values of these non-zero coefficients so as to maximize the probability of observing the given data.

Let S denote the set of non-zero elements in the matrix $\mathbf{\Lambda}$. We set $p(\lambda_{ij}) = 1 \quad \forall (i, j) \notin S$: the distribution $p(\mathbf{\Lambda})$, factors as

$$p(\mathbf{\Lambda}) = \prod_{(i,j) \in S} p(\lambda_{ij}). \quad (4)$$

We impose standard normal priors $p(z_j)$ and $p(\lambda_{ij})$ on the latent expression levels and the hybridization coefficients. The probabilistic model for cross-hybridization is thus represented as a joint distribution over the observed and latent expression profiles, \mathbf{X} and \mathbf{Z} , and the hybridization matrix $\mathbf{\Lambda}$, such that

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}, \mathbf{\Lambda}) &= N(\mathbf{\Lambda}\mathbf{z}, \Psi) = \prod_i N\left(\sum_{j:(i,j) \in S} \lambda_{ij}z_j, \psi_i^2\right), \\ p(\mathbf{z}) &= \prod_j p(z_j) = \prod_j N(0, 1), \\ p(\mathbf{\Lambda}) &= \prod_{(i,j) \in S} p(\lambda_{ij}) = \prod_{(i,j) \in S} N(0, 1), \\ p(\mathbf{X}, \mathbf{Z}, \mathbf{\Lambda}) &= p(\mathbf{\Lambda}) \prod_t p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{\Lambda}) p(\mathbf{z}_t), \end{aligned} \quad (5)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$ and $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T]$ are the measured and latent expression measurements which are indexed by tissue. The above model treats expression levels as being statistically independent across the T conditions: each condition corresponds to a sample drawn independently from the probability model. We, therefore, make no assumption

in our model about patterns of covariance between particular microarray conditions so as not to bias the inferred data.

We now turn to the problem of inferring the optimal settings of the latent expression profiles, \mathbf{Z} , and the hybridization matrix, $\mathbf{\Lambda}$. This requires us to compute the posterior distribution $p(\mathbf{z}_t, \mathbf{\Lambda} | \mathbf{x}_t), \forall t = 1, \dots, T$. For the model in (5), this distribution is intractable to compute analytically. In order to perform inference, we must resort to approximate methods that make use of a surrogate distribution $q(\mathbf{z}_t, \mathbf{\Lambda})$ over the latent variables that approximates $p(\mathbf{z}_t, \mathbf{\Lambda} | \mathbf{x}_t)$. Two popular methods for approximate inference are Markov-chain Monte Carlo (MCMC) methods and variational methods. The variational method for approximate inference has been chosen in this case due to concerns with the convergence and mixing properties of MCMC methods and the more deterministic flavor of the variational method, where there is a guarantee on convergence to an optimal solution.

VARIATIONAL LEARNING FOR CROSS-HYBRIDIZATION COMPENSATION

In variational learning (Neal and Hinton, 1998; Jaakkola and Jordan, 2000) with latent variables \mathbf{H} and observed variables \mathbf{V} , the exact posterior $p(\mathbf{H} | \mathbf{V})$ is approximated by a distribution $q(\mathbf{H}; \phi)$ parameterized by a set of variational parameters ϕ . Thus, inference consists of iteratively improving the fit of $q(\mathbf{H}; \phi)$ to $p(\mathbf{H} | \mathbf{V})$ with respect to these parameters. This is measured by the relative entropy, $D(q || p)$, between $q(\mathbf{H}; \phi)$ and the model $p(\mathbf{H}, \mathbf{V})$ of the latent and observed variables; this can be written as

$$\begin{aligned} D(q || p) &= \int_{\mathbf{H}} q(\mathbf{H}; \phi) \log \frac{q(\mathbf{H}; \phi)}{p(\mathbf{H}, \mathbf{V})} d\mathbf{H} \\ &= \int_{\mathbf{\Lambda}} \int_{\mathbf{Z}} q(\mathbf{Z}, \mathbf{\Lambda}) \log \frac{q(\mathbf{Z}, \mathbf{\Lambda})}{p(\mathbf{X}, \mathbf{Z}, \mathbf{\Lambda})} d\mathbf{Z} d\mathbf{\Lambda}, \end{aligned} \quad (6)$$

where \mathbf{X} , \mathbf{Z} and $\mathbf{\Lambda}$ have been substituted as the observed and latent variables, \mathbf{V} and \mathbf{H} , for the model in (5).

The approximating distribution can be further simplified via a mean-field decomposition of all latent variables, where the complex joint distribution $q(\mathbf{z}_t, \mathbf{\Lambda})$ can be decomposed as the product of simpler local distributions over each latent variable, such that

$$q(\mathbf{z}_t, \mathbf{\Lambda}) = \prod_{(i,j) \in S} q(\lambda_{ij}) \prod_j q(z_{jt}). \quad (7)$$

Now, let the marginal distributions $q(z_{jt})$ and $q(\lambda_{ij})$ be Gaussian:

$$q(z_{jt}) = N(\mu_{jt}, \sigma_j^2), \quad (8)$$

$$q(\lambda_{ij}) = N(l_{ij}, \phi_{ij}^2), \quad (9)$$

where l_{ij} , ϕ_{ij}^2 , μ_{jt} and σ_j^2 are variational parameters that are used to perform approximate inference.

Learning and inference in the model specified in (5) is accomplished via the variational EM algorithm. Denoting the sets of variational and model parameters by ϕ and θ , the algorithm alternates between minimizing $D(q || p)$ with respect to the set of variational parameters ϕ (the variational E-step), and then minimizing $D(q || p)$ with respect to the model parameters θ (the variational M-step). The variational EM algorithm continues to alternate between the E-Step and the M-Step until convergence, at which point the inferred values for μ_{jt} correspond to the latent expression profiles of interest. For the model of (5) and the mean-field variational approximation, $D(q || p)$ can be simplified to

$$\begin{aligned} D(q || p) &= -\frac{1}{2} \left(\sum_{j,t} (\log \sigma_j^2 - \sigma_j^2 - \mu_{jt}^2) \right. \\ &\quad + \sum_{(i,j) \in S} (\log \phi_{ij}^2 - \phi_{ij}^2 - l_{ij}^2) \\ &\quad - \sum_{i,t} \left(\log \psi_i^2 + \frac{(x_{it} - \sum_{j:(i,j) \in S} l_{ij} \mu_{jt})^2}{\psi_i^2} \right) \\ &\quad \left. + \sum_{i,t} \frac{\sum_{j:(i,j) \in S} l_{ij}^2 \sigma_j^2 + \phi_{ij}^2 (\mu_{jt}^2 + \sigma_j^2)}{\psi_i^2} \right) + K, \end{aligned} \quad (10)$$

where K does not depend on the variational and model parameters.

Minimizing $D(q || p)$ with respect to the variational and model parameters by taking derivatives and setting to zero to find stationary points yields the variational updates for GenXHC:

Variational E-step:

$$\begin{aligned} \mu_{jt} &\leftarrow \frac{\sum_{i:(i,j) \in S} (l_{ij} / \psi_i^2) (x_{it} - \sum_{k \neq j} l_{ik} \mu_{kt})}{1 + \sum_{i:(i,j) \in S} (l_{ij} / \psi_i^2) (l_{ij}^2 + \phi_{ij}^2)}, \\ \sigma_j^2 &\leftarrow \frac{1}{1 + \sum_{i:(i,j) \in S} (l_{ij} / \psi_i^2) (l_{ij}^2 + \phi_{ij}^2)}, \\ l_{ij} &\leftarrow \frac{\sum_t \mu_{jt} (x_{it} - \sum_{k \neq j} l_{ik} \mu_{kt})}{\psi_i^2 + \sum_t (\mu_{jt}^2 + \sigma_j^2)}, \\ \phi_{ij}^2 &\leftarrow \frac{\psi_i^2}{\psi_i^2 + \sum_t (\mu_{jt}^2 + \sigma_j^2)}, \end{aligned} \quad (11)$$

Variational M-step:

$$\begin{aligned} \psi_i^2 &\leftarrow \frac{1}{T} \sum_t \left(\left(x_{it} - \sum_{j:(i,j) \in S} l_{ij} \mu_{jt} \right)^2 \right. \\ &\quad \left. + \sum_{j:(i,j) \in S} (l_{ij}^2 + \phi_{ij}^2) \sigma_j^2 + \sum_{j:(i,j) \in S} \mu_{jt}^2 \phi_{ij}^2 \right) \end{aligned} \quad (12)$$

RESULTS

GenXHC for *M.musculus* microarray data

Our algorithm was applied to a subset of a genome-wide *M.musculus* microarray data set (Frey *et al.*, submitted for publication). The subset consisted of 26 486 probes that were designed for putative exons in chromosome 16. These putative exons were predicted from the Repeat-masked mouse draft genome (Build 28) using five different exon-prediction programs. One 60-mer oligonucleotide probe was selected for each putative exon, such that its cross-hybridization potential to the full set of microarray probes was minimal. Array designs were submitted to Agilent Technologies (Palo Alto, CA) for array production. The resulting scanned microarray images were then quantitated using GenePix (Axon Instruments); complex noise structures in the microarray images were then removed via a spatial detrending algorithm (Shai *et al.*, 2003), and each set of images representing measurements over 12 tissue pools was calibrated with the VSN algorithm (Huber *et al.*, 2002) using a set of 100 reference genes that were represented on each microarray.

The sparsity structure on the hybridization matrix, \mathbf{A} , was determined by running a probe-to-RefSeq cDNA pairwise sequence alignment using BLAST with an E -value cutoff of 1×10^{-2} . We set the E -value cutoff to be permissive enough to identify all transcript-probe pairings which could possibly contribute to cross-hybridization noise with a sufficient number of base-pairings (Hughes *et al.*, 2001). A list of potential cross-hybridizing cDNAs was built in this fashion for each microarray probe. The resulting dataset was further reduced by removing transcripts that only had a single matching probe sequence (as they would not contribute to cross-hybridization) and their corresponding probes. The dataset obtained in this fashion had $N = 9904$ probe measurements and $M = 2965$ latent transcripts confirmed by the RefSeq cDNA database.

Before running the variational EM algorithm on the microarray dataset, the observed expression levels were normalized to the range $[0, 1]$ by subtracting the minimum and dividing by the maximum expression value for each probe. The algorithm was then run for 30 iterations.

We note that the model proposed in (5) allows both the hybridization coefficients and the latent expression levels to be negative. Such negative values, however, do not correspond to meaningful mRNA transcript quantities. In order to avoid negative values, the algorithm implements a simple heuristic whereby negative values are set to zero at the end of each iteration of the variational EM algorithm.

Reconstruction error using denoised profiles

To evaluate the inferred expression profiles, a permutation test was applied to the measured data as follows: the variational learning method is trained on 2 expression datasets, one of which is the original dataset, and the other which is a dataset

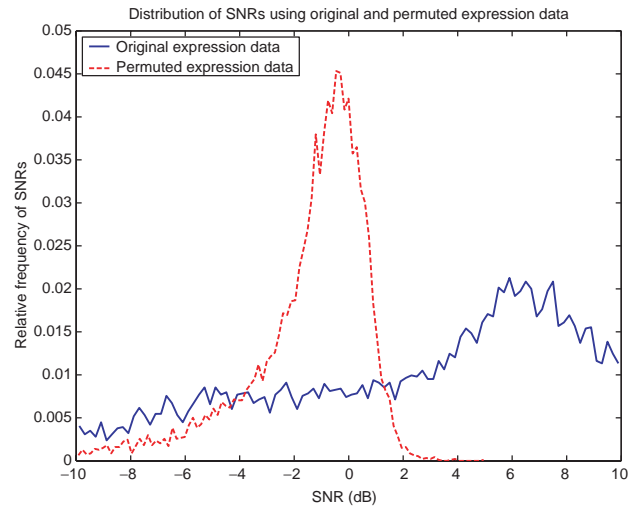


Fig. 5. Decrease in reconstruction error, measured as SNR, using original and randomly permuted data for learning: taking into account hybridization constraints allows us to better explain noisy measurements.

with the probe order randomly permuted. This permutation alters the set of possible hybridization interactions between the microarray probes and the set of cross-hybridizing transcripts. We then define the signal-to-noise ratio (SNR), as a measure of reconstruction error:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_t \|\hat{\mathbf{x}}_t\|^2}{\sum_t \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2} \right), \quad (13)$$

where $\|\cdot\|$ is the Euclidean norm of a vector, $\hat{\mathbf{x}}_t = \mathbf{A}\mathbf{z}_t$ is the estimate of the observed expression profile, \mathbf{x}_t , according to the probabilistic model outlined in (5). Thus, one expects the SNR figure to be larger for data inferred from the original dataset due to the fact that the original dataset has incorporated explicit hybridization constraints between probes and transcripts, whereas the sparsity constraint for the permuted data has no biological meaning.

Figure 5 shows a histogram of the SNR figures computed for each of the 9904 probes; as can be seen, the SNR is on average much higher on the original dataset than on the permuted set. By taking into account the set of hybridization constraints between microarray probes and gene transcripts, we have achieved a lower reconstruction error with respect to the case where the sparsity structure of the hybridization matrix is randomized. This result demonstrates that the GenXHC algorithm has not simply overfit the observed data using the available hidden variables and parameters of the probabilistic model.

Gene Ontology–biological process (GO–BP) enrichment using denoised expression data

We then sought to verify whether removing cross-hybridization noise from a set of gene expression profiles

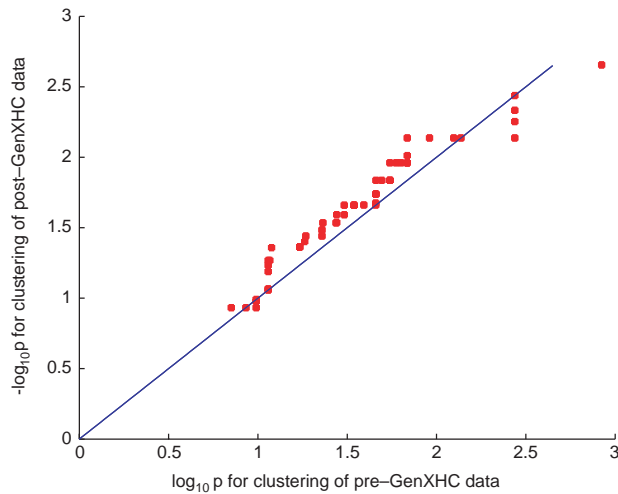


Fig. 6. Comparison of functional group enrichment for noisy data and data denoised using GenXHC: taking into account cross-hybridization noise produces enrichment for most of the functional groups.

would produce enrichment for a given set of functional groups associated to a set of GO–BP functional annotations. Because co-expression of genes predicts co-functionality (Tavazoie *et al.*, 1999), with all else being equal, a higher enrichment indicates improved measurements of transcript levels.

We used GO annotations from the Mouse Genome Database (<http://www.informatics.jax.org>; MGD, 2005) to associate each gene to a set of biological functions. Of these GO–BP categories, we selected GO–BP categories that were represented by at least five cDNAs, for a total of 328 GO–BP categories. We constructed two datasets for comparison: the first consisted of the averaged probe measurements for each transcript and the other consisted of the inferred expression data. We performed agglomerative hierarchical clustering (using the pairwise Pearson correlation between two profiles as a similarity metric) on both datasets. Each dataset was then clustered into 120 functional groups, and a hypergeometric p -value was then computed for each functional group in both datasets.

Figure 6 shows a comparison of enrichment for each of the 120 functional groups: the increased enrichment with respect to noisy pre-XHC data across multiple functional groups suggests that the amount of noise present in the expression data has been reduced by GenXHC.

Comparison of GenXHC to RMA

We then sought to compare the performance of GenXHC to RMA on our dataset via the functional enrichments they produce, using the same GO–BP enrichment test above. We implemented RMA background subtraction using maximum-likelihood estimates of the model parameters for the RMA linear model (Irizarry *et al.*, 2003). Figure 7 compares the performance of GenXHC and RMA, as well as the performance

of RMA on un-normalized noisy expression data. Though RMA improves functional enrichment with respect to noisy data, GenXHC produces equal or greater enrichment with respect to RMA across the majority of the functional groups. This is probably due to the fact that unlike RMA, which only models the statistical properties of cross-hybridization noise, GenXHC models the explicit sparse structure of the set of probe-transcript interactions and is able to take into account specific probe-transcript cross-hybridization effects.

DISCUSSION

In the example of cross-hybridization compensation shown in Figure 2, GenXHC has managed to remove noisy expression levels in the measured gene expression data that are due to cross-hybridizing transcripts. However, for many probes, the target transcript for which they were designed has been assigned a hybridization coefficient of 0. This can be explained by considering that certain probes are particularly noisy and, therefore, incur high reconstruction error for using the optimal inferred variable settings. The algorithm, therefore, sets the optimal values of the hybridization coefficients for these probes to 0, as setting them to greater values would increase the error. Also, GenXHC produces statistically significant improvements on average, but with a broad variance in SNR improvements (as evidenced in Fig. 5). This may correspond to the presence of many outlier expression profiles in the training set that are poorly explained by the model parameters, and hence contribute to the algorithm finding poor solutions. One possible remedy to this is to include an outlier model that accounts for microarray probes which measure low expression levels, under the assumption that these may in fact correspond to measurement noise and experimental error.

It is also worth noting that our probabilistic model only uses information on binding energy to identify likely cross-hybridization targets for each probe. It does not take into account other physical effects, opting instead to capture such effects using the uncertainty in the model variables. The binding energy value between a probe and a transcript, as well as the probe effect (Li and Wong, 2001) are, nevertheless, important determinants of probe intensity, and we are therefore currently investigating the use of such information to set the priors on hybridization coefficients to account for such physical effects.

CONCLUSION

The problem of removing cross-hybridization noise will become increasingly significant in the upcoming era of genome-wide exon-tiling microarray experiments. We have developed GenXHC, a probabilistic model for cross-hybridization compensation in high-density, genome-wide microarray data. The model is able to take into account explicit hybridization constraints between microarray probes and

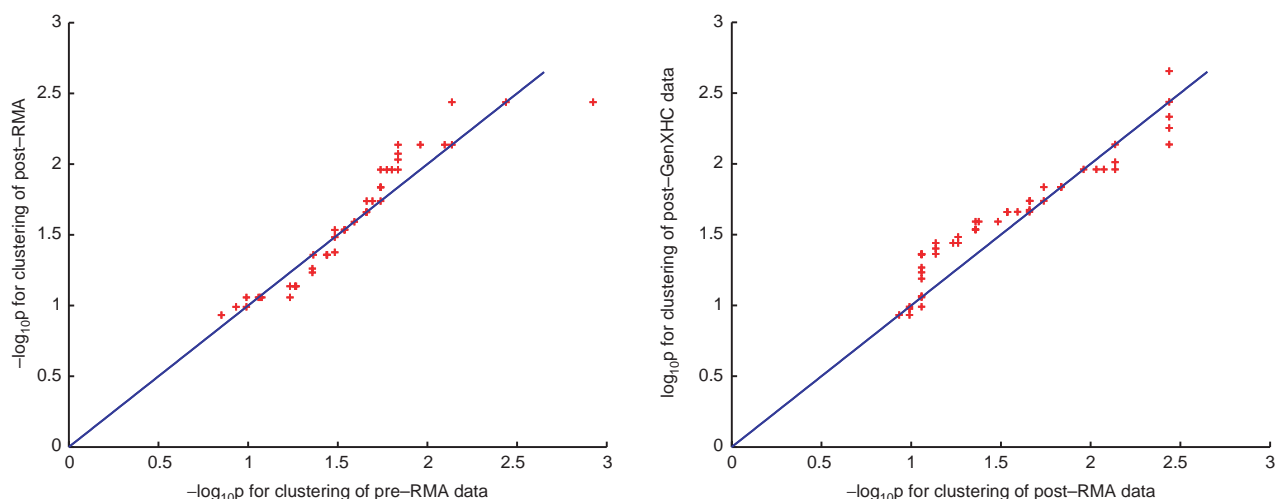


Fig. 7. Comparison of functional group enrichment using GenXHC and RMA: taking into account the explicit sparsity inherent to the set of probe-transcript hybridization interactions produces enrichment.

gene transcripts, outperforming the popular RMA method for cross-hybridization compensation. The algorithm was applied to a subset of a genome-wide *M.musculus* exon-tiling microarray dataset and was shown to produce a significant reduction in cross-hybridization noise, albeit with large variance on the range of improvements in terms of SNR. The inferred gene expression data was also shown to produce enrichment in many GO-BP functional groups. We are investigating the addition of an outlier model that can account for noisy data such that this variance in reconstruction errors can be reduced. As microarrays scale to higher densities, we believe that algorithms which can accurately compensate for cross-hybridization will play an important role in future large-scale microarray assays.

ACKNOWLEDGEMENTS

J.C.H. was supported by a NSERC Canada Graduate Scholarship. J.C.H. and Q.D.M. were supported by a CIHR Net grant. Q.D.M. was supported by an NSERC PDF. B.J.F. was supported by a Premier's Research Excellence Award and a gift from Microsoft Corporation.

REFERENCES

- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
- Frey,B.J., Mohammad,N., Morris,Q.D., Zhang,W., Robinson,M.D., Mnaimneh,S., Shai,O., Chang,R., Pan,Q., Laurin,N. *et al.* Genome-wide analysis of mouse transcription using exon-resolution microarrays and factor graphs. Submitted for publication.
- Frey,B.J., Morris,Q.D., Robinson,M. and Hughes,T.R. (2005) Finding novel transcripts in high-resolution genome-wide microarray data using the GenRate model. *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology (RECOMB)*, May 2005.
- Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics*, **18**, S96–S104.
- Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
- Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U., Speed,T.P. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
- Jaakkola,T. and Jordan,M.I. (2000) Bayesian parameter estimation via variational methods. *Stat. Comput.*, **10**, 25–37.
- Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
- Le,K., Mitsouras,K., Roy,M., Wang,Q., Xu,Q., Nelson,S.F. and Lee,C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res.*, **32**, e180.
- Lee,J.K., Bussey,K.J., Gwady,F.G., Reinhold,W., Riddick,G., Pelletier,S.L., Nishizuka,S., Szakacs,G., Annereau,J.P., Shankavaram,U. *et al.* (2003) Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol.*, **4**, R82.

- Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
- Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 98–106.
- Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Mouse Genome Database (MGD) (2005) Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine.
- Neal,R.M. and Hinton,G.E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan,M.I. (ed.), *Learning in Graphical Models*. Kluwer Academic Publishers, Dordrecht, pp. 355–368.
- SantaLucia,J.J., Allawi,H.T. and Seneviratne,P.A. (1998) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
- Shai,O., Morris,Q. and Frey,B.J. (2003) Spatial bias removal in microarray images. *University of Toronto Technical Report PSITR-2003-21*.
- Shoemaker,D.D., Schadt,E.E., Armour,C.D., He,Y.D., Garrett-Engle,P., McDonagh,P.D., Loerch,P.M., Leonardson,A., Lum,P.Y., Cavet,G. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.
- Sun,Y., Koo,S., White,N., Peralta,E., Esau,C., Dean,N.M. and Perera,R.J. (2004) Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Res.*, **32**, e188.
- Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 213–215.
- Wren,J.D., Kulkarni,A., Joslin,J., Butow,R.A. and Garner,H.R. (2002) Cross-hybridization on PCR-spotted microarrays. *IEEE Eng. Med. Biol.*, **21**, 71–75.
- Wu,Z. and Irizarry,R.A. (2004) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology (RECOMB)*, March, pp. 98–106.
- Wu,Z., Irizarry,R.A., Gentleman,R., Murillo,F.M., Spencer,F. *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.
- Zhang,W., Morris,Q.D., Chang,R., Shai,O., Bakowski,M.A., Mitsakakis,N., Mohammad,N., Robinson,M.D., Zingibl,R., Somogyi,E. *et al.* (2004) The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21–43.