



<http://www.psi.toronto.edu>

Iterated Conditional Modes for Cross-Hybridization Compensation in DNA Microarray Data

Jim C. Huang, Quaid D. Morris, Brendan J. Frey

October 06, 2004

PSI TR 2004-031

Iterated Conditional Modes for Cross-Hybridization Compensation in DNA Microarray Data

Jim C. Huang, Quaid D. Morris, Brendan J. Frey
University of Toronto

1 Introduction

The principal technology that is driving the rush of knowledge in genomic functional prediction is the DNA microarray. Whereas in the past, biological assays would be performed on a discrete "one gene, one experiment" basis, microarray technology has allowed biologists to perform thousands of assays on a single miniature chip. DNA microarrays consist of an array of probes, each of which measures the approximate amount of mRNA transcript produced from a target gene. In such microarray experiments, the functions of an organism's genes are inferred by examining how these genes vary in their measured mRNA expression level across several tissue pools, such as brain, liver or skin.

One of the important problems in DNA microarray analysis and gene functional prediction is that the data generated from microarray experiments is typically prone to many types of noise from various sources. A particular class of noise that has received little attention so far is the biological effect of cross-hybridization on microarray data ([1]). Indeed, many genes belong to homologous families and therefore tend to exhibit a large similarity in sequence. As a result, probes on a microarray will measure not only the mRNA level of their target gene, but also the unwanted mRNA levels of other, similar non-specific genes. We thus propose a signal model to represent the effect of cross-hybridization in microarray data, in which the expression profiles measured by probes are to be explained by taking a weighted linear sum of a relatively small number of latent gene expression profile variables. To compensate for the effect of cross-hybridization, these hidden gene expression profiles would be iteratively inferred from the data via a method based on a Factor Analysis model using the Iterated Conditional Modes inference algorithm to perform cross-hybridization compensation (XHC). Microarray data in which this cross-hybridization effect has been removed will allow for increased specificity and sensitivity in predicting the function of genes from microarray data produced from microarray experiments.

2 Methods

2.1 Signal Model for Cross-Hybridization Compensation (XHC)

Let N denote the number of probes on the array. We model each probe x_n , $n = 1, 2, \dots, N$ in the array as a set of T -dimensional zero-mean observed data vectors. Let $X = [x_1, x_2, \dots, x_T]$ denote a matrix in which the t^{th} column denotes the measurement across all N probes in the microarray for tissue pool t , $t = 0, 1, \dots, T$. The set of measurements x_t is assumed to be a set of i.i.d. Gaussian random vectors. We assume a *factor analysis* model for the microarray data in which the set of N observed expression profile vectors X can be explained using a smaller set of $M < N$ *latent gene expression profiles*. Let $Z = [z_1, z_2, \dots, z_T]$ denote the matrix of latent gene expression profiles to be inferred; the t^{th} column of Z therefore denotes the latent mRNA expression levels across all M probed genes for tissue pool t . According to the factor analysis model, each probe on the microarray is capable of hybridizing to multiple genes and conversely, each gene can have multiple probes that can hybridize to it. Thus, the observed mRNA expression level x_{nt} for probe n in tissue pool t can be represented as a weighted sum of latent gene mRNA expression levels z_{mt} for tissue pool t plus some random additive Gaussian noise. Thus, we can represent the relationship between the observed and latent expression levels as

$$\begin{aligned} x|z &\sim \mathcal{N}(\Lambda z, \sigma^2 I) \\ z &\sim \delta(z - z_t) \end{aligned} \quad (1)$$

where $x \in \mathbb{R}^N$, $z \in \mathbb{R}^M$ and $\Lambda \in \mathbb{R}^{(N \times M)}$ is the *factor loading matrix* that couples the observed expression profiles to the latent gene expression profiles. For the purposes of the XHC problem, Λ is referred to as the *hybridization matrix*.

We make the assumption that Λ is sparse: we assume that each probe can only hybridize to a maximum of 2 genes. More precisely, we assume that probe n in the array hybridizes to its intended target, or *primary* gene with coefficient $\lambda_1^{(n)} = 1$, as well as cross-hybridizes to a single *secondary* gene other than its intended target with a non-negative coefficient $\lambda_2^{(n)} < 1$. To determine primary and secondary genes for a given probe, the BLAST program [3] is used to perform pairwise alignments between each probe and all matching genes and ranks each match according to sequence similarity. The primary and secondary genes for a given probe are thus determined by taking the top 2 ranking genes based on sequence similarity.

Thus, we seek to iteratively estimate the sparse hybridization matrix Λ and the set of latent gene expression profiles Z over all T tissue pools by maximizing the log-likelihood $\mathcal{L}(\Lambda|X)$ of the observed data X under the assumption of uniform random sensor noise level across all probes. In order to perform this optimization and infer the hidden profiles, we will use the batch form of a greedy, Iterated Conditional Modes (ICM) algorithm.

2.2 Iterated Conditional Modes Algorithm

Under the assumptions outlined above, the log-likelihood $\mathcal{L}(\Lambda|X)$ of the observed data X can be written as

$$\begin{aligned}
\mathcal{L}(\Lambda|X) &= \log \prod_{t=1}^T p(x_t) = \sum_{t=1}^T \log p(x_t) = \sum_{t=1}^T \log \int_z p(x_t, z) dz \\
&= \sum_{t=1}^T \log \int_z p(x_t|z)p(z) dz = \sum_{t=1}^T \log \int_z p(x_t|z)\delta(z - z_t) dz \\
&= \sum_{t=1}^T \log p(x_t|z_t) \\
&= \log p(X, Z) = \mathcal{L}_c(\Lambda|X, Z)
\end{aligned} \tag{2}$$

where $\mathcal{L}_c(\Lambda|X, Z) = \log p(X, Z)$ is the *complete log-likelihood* given the latent variables Z . The *Iterated Conditional Modes Algorithm* thus consists in iteratively optimizing $\mathcal{L}_c(\Lambda|X, Z)$ by alternating between the following 2 optimization steps:

$$Z^* = \arg \max_Z \mathcal{L}_c(\Lambda|X, Z) \tag{3}$$

$$\Lambda^* = \arg \max_{\Lambda} \mathcal{L}_c(\Lambda|X, Z)$$

Thus, the algorithm alternates between finding the set of latent variables that maximizes $\mathcal{L}_c(\Lambda|X, Z)$ and then obtaining the optimum value for Λ .

2.3 Factor Analysis using the Iterated Conditional Modes algorithm

Substituting the conditional distribution $p(x|z) = \mathcal{N}(x, \sigma^2 I)$ into the expression for $\mathcal{L}_c(\Lambda|X, Z)$ yields (omitting constant terms):

$$\mathcal{L}_c(\Lambda|X, Z) = \sum_{t=1}^T \log p(x_t|z_t) \tag{4}$$

$$= -\frac{1}{2\sigma^2} \sum_{t=1}^T (x_t - \Lambda z_t)^T (x_t - \Lambda z_t) \tag{5}$$

$$= -\frac{1}{2\sigma^2} \sum_{t=1}^T \sum_{n=1}^N (x_{tn} - z_{tm(n,1)} - \lambda z_{tm(n,2)})^2 \tag{6}$$

where $m(n, 1)$ and $m(n, 2)$ denote the gene indices corresponding to the primary and secondary gene respectively for probe n . Optimizing the above equations by taking derivatives of $\mathcal{L}_c(\Lambda|X, Z)$ with respect to the hidden profiles Z and the free $\lambda_2^{(n)}$ parameters and setting them to zero yields the following ICM update equations for the latent profiles and the $\lambda_2^{(n)}$ parameters:

$$z_{tk} = \frac{\sum_{n:m(n,1)=k} (x_{tn} - \lambda_2^{(n)} z_{tm(n,2)}) + \sum_{n:m(n,2)=k} \lambda_2^{(n)} (x_{tn} - z_{tm(n,1)})}{M_1^{(k)} + \sum_{n:m(n,2)=k} (\lambda_2^{(n)})^2} \tag{7}$$

and

$$\lambda_2^{(n)} = \frac{\sum_t z_{tm(n,2)}(x_{tn} - z_{tm(n,1)})}{\sum_t z_{tm(n,2)}^2} \quad (8)$$

where $M_1^{(k)}$ is the number of array probes that can hybridize to gene k with a coefficient of 1. Thus, we iterate between equations (7) and (8) until convergence. The latent variable and parameter estimates z_{tn} and $\lambda_2^{(n)}$ are therefore the pointwise *maximum likelihood* values corresponding to the maxima of the joint distribution $p(X, Z)$.

2.4 Detecting Cross-Hybridization in DNA Microarray Data

To gauge the impact of cross-hybridization on the observed expression data, we performed a statistical test in which each observed expression profile vector was normalized to have unit norm. Taking the oligonucleotide sequences of the probes on the array and the corresponding sequences of their target genes, we performed a BLAST search using default parameters to find the possible probe-to-gene mappings. We then computed the Pearson correlation coefficient between randomly-paired and BLAST-matched expression profiles. Figure (1) shows the resulting distribution of correlation coefficients. We can see that approximately 33% of the BLAST-matched profiles show high correlation whereas only 2% of the randomly-paired profiles have high correlation. The considerably higher proportion of highly-correlated probe pairs in the BLAST-matched set suggests that the effect of cross-hybridization is noticeably present in DNA microarray data in the form of a large proportion of highly-correlated expression profiles. This is consistent with our signal model in which observed mRNA expression levels can be expressed as a linear combination of the expression levels of matching genes.

2.5 Preprocessing of Expression Data

We pre-processed the data by removing genes that exhibited a large variance in their corresponding probes' measured expression levels. Indeed, by plotting the distribution of the variance in expression levels of probes (Figure (3)) across the set of genes, we see a bimodal distribution in which there is a set of genes for which probes have a low variance in expression level and another set of genes for which probes have high variance. Given that we model the random sensor noise level of all probes as being equal to σ^2 , the measured profiles with high variance (defined as having a log-variance higher than a certain threshold $\log(\sigma_0)$) in expression level for a given gene were considered to be particularly noisy measurements that were subsequently removed from the training set to avoid inferring noisy gene expression profiles that allow for overfitting of the observed data.

The possible probe-to-gene sequence mappings were determined via BLAST with the default settings between the probe oligonucleotide sequences and the sequences of the targeted genes. The output of the search is a ranked list of matching gene nucleotide sequences for each probe sequence, each of which is ranked by its number of mismatches to the probe. The number of mismatches is a good measure of the hybridization binding potential between a probe and its target gene according to the thermodynamic properties of hybridization [2].

Each gene was assumed to have at least one *primary probe* associated with it: that is, a probe which hybridizes to that gene with a unity hybridization coefficient. We removed any genes that hybridized to less than 3 probes on the array; this was done to remove gene profiles that would be uninformative for the purposes of inferring the hidden gene expression profiles based on the effect of cross-hybridization.

2.6 Results

We used *Mus Musculus* gene expression data [4] measured over 12 tissue pools to train the ICM algorithm. The observed expression levels were normalized to the range 0 to 1 before running the algorithm. We tested the algorithm for two cases:

1. A probe hybridizes to a primary and secondary gene according to its BLAST matches;
2. A probe hybridizes to its primary probe according to its best BLAST match and to a randomly-selected secondary gene;

The objective is thus to verify that the ICM algorithm is not simply fitting coefficients to noisy gene expression profiles which are not biologically meaningful. In addition, we ran the ICM algorithm in the cases that each probe hybridizes to only 1 gene, and then 2 genes. This allows us to verify that the use of 2 gene expression profiles to explain an observed expression profile allows for a significant reduction in reconstruction error than if only 1 gene profile is used.

We thus use the Signal-to-Quantization Noise Ratio (SQNR) over the entire array as a measure of how much error is incurred in estimating the observed expression profiles using the inferred gene profiles. We define the microarray SQNR (in dB) as follows:

$$SQNR_{array} = 10 \log \left(\frac{\sum_t \|x_t\|^2}{\sum_t \|x_t - \hat{x}_t\|^2} \right) \quad (9)$$

where $\hat{x}_t = \Lambda z_t$ is the estimate of the observed expression profile x_t given Λ and z_t .

As can be seen in Table (1), an appreciable microarray SQNR gain with respect to both the case in which a single gene is used to represent a probe's measured profile and the case in which the second gene for a probe is randomly assigned. Figure (2) shows the cumulative distribution plot of the resulting SQNRs. As can be seen, a higher proportion of probes have high SQNR figures with XHC using the pre-processed data than the proportion obtained from optimizing on the full measured data set or a subset thereof that is thresholded at the 33rd percentile in expression level.

Figure (4) shows a sample probe and the corresponding target gene profiles that have been inferred by our algorithm. In this instance, the high expression level in tissue pools 10 and 12 have been mapped separately to the probe's primary and secondary target genes, demonstrating that the observed expression profile can be explained by the cross-hybridization of the probe's sequence with both its primary and secondary genes.

3 Discussion

The resulting gain in microarray SQNR with respect to the case in which a single gene is assigned to a probe suggests that performing XHC produces an increase in SQNR by increasing the precision with which the observed data can be estimated by taking into account multiple sources of hybridization for each probe. In addition, the significant gain in SQNR with respect to the case in which each probe's secondary gene is randomly assigned further enhances the biological validity and significance of the inferred gene expression profiles.

Figure (2) suggests that the full expression data set contains many noisy and outlier probe measurements which can be used to infer noisy gene expression profiles that are not biologically significant and that can overfit the observed data. Thus, pre-processing of the data to remove outliers is especially crucial given that ICM is particularly sensitive to outliers in the training data.

4 Conclusion

The problem of cross-hybridization in DNA microarray data is a problem that has received relatively little attention. We have shown that cross-hybridization indeed has a significant impact on a non-negligible proportion of microarray measurements. We have defined a signal model for cross-hybridization, and we have subsequently proposed a method for compensating for this effect via a factor analysis approach using Iterated Conditional Modes. We have shown using a relatively simple model to account for multiple sources of hybridization in the mRNA levels measured by microarray probes allows for an increase in the array SQNR. Possible improvements to the current approach include:

1. Allowing for more than 2 hybridizing genes per probe and allowing the number of cross-hybridizing genes to be randomly distributed;
2. Allowing for a probability distribution to model uncertainty in the hidden gene profiles (e.g.: as in factor analysis)
3. Introduce a prior distribution for modeling the sparse structure of the factor loading matrix Λ ;
4. Allowing for a more robust estimation of gene expression profiles with respect to noisy expression data and outliers;

We are currently in the process of implementing the above improvements to allow for more robust XHC that is less dependent and sensitive to data pre-processing and noisy measurements. Based on the preliminary results shown here, we believe that such an improved algorithm can provide a dramatic increase in the accuracy and reliability of microarray expression data as well as increased sensitivity and specificity in the functional prediction of genes.

References

- [1] J.D. Wren, A. Kulkarni, J. Joslin, R.A. Butow, and H.R. Garner. "Cross-Hybridization on PCR-Spotted Microarrays," *IEEE Engineering in Medicine and Biology*, Vol. 21, pp. 71-75, March 2002.
- [2] J.J. SantaLucia, H.T. Allawi, and P.A. Seneviratne. "Improved Nearest-Neighbor Parameters for predicting DNA duplex stability," *Biochemistry*, Vol. 35, pp. 3555-3562, 1998.
- [3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. "Basic local alignment search tool," *J Mol Biol* 215(3):403-10, 1990.
- [4] B.J. Frey, N. Mohammad, W. Zhang, Q.D. Morris, M.D. Robinson, R. Chang, O. Shai, S. Mnaimneh, Q. Pan, J. Rossant, J. Aubin, B.J. Blencowe, and T.R. Hughes. "Full-genome exon profiling in mus musculus " (*in preparation*).

5 Figures and Tables

	$\Delta SQNR$
With respect to a single gene per probe	0.90059 dB
With respect to a randomly-assigned 2 nd gene	1.2259 dB

Table 1: Increases in SQNR due to XHC

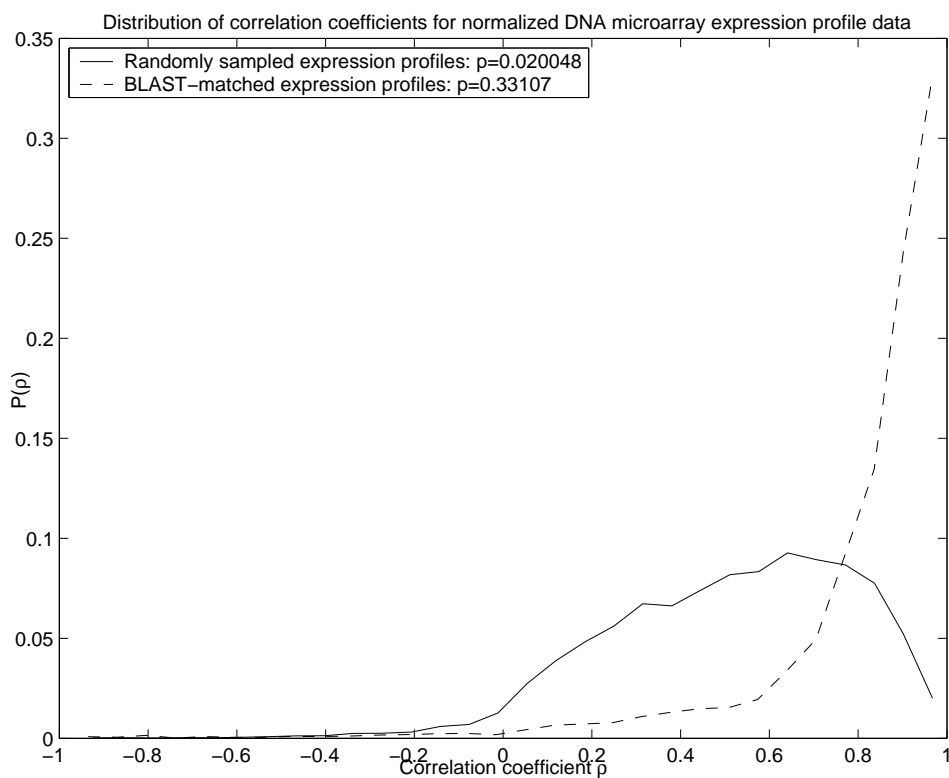


Figure 1: Distribution of Pearson correlation coefficients for normalized DNA microarray data

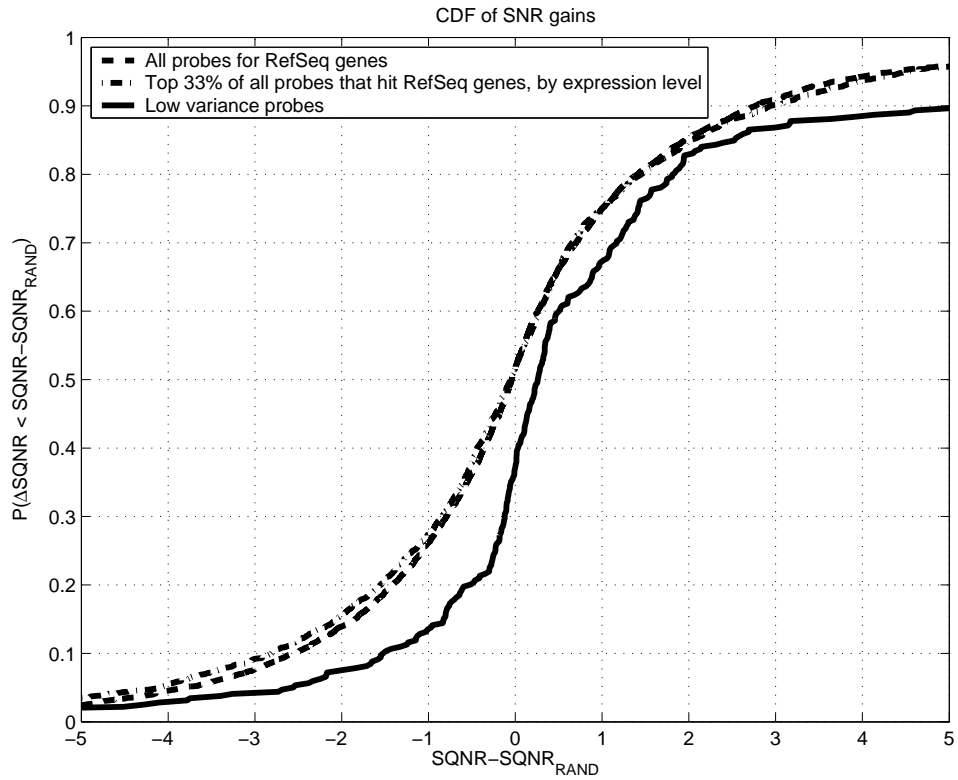


Figure 2: Cumulative distribution functions of resulting SQNRs

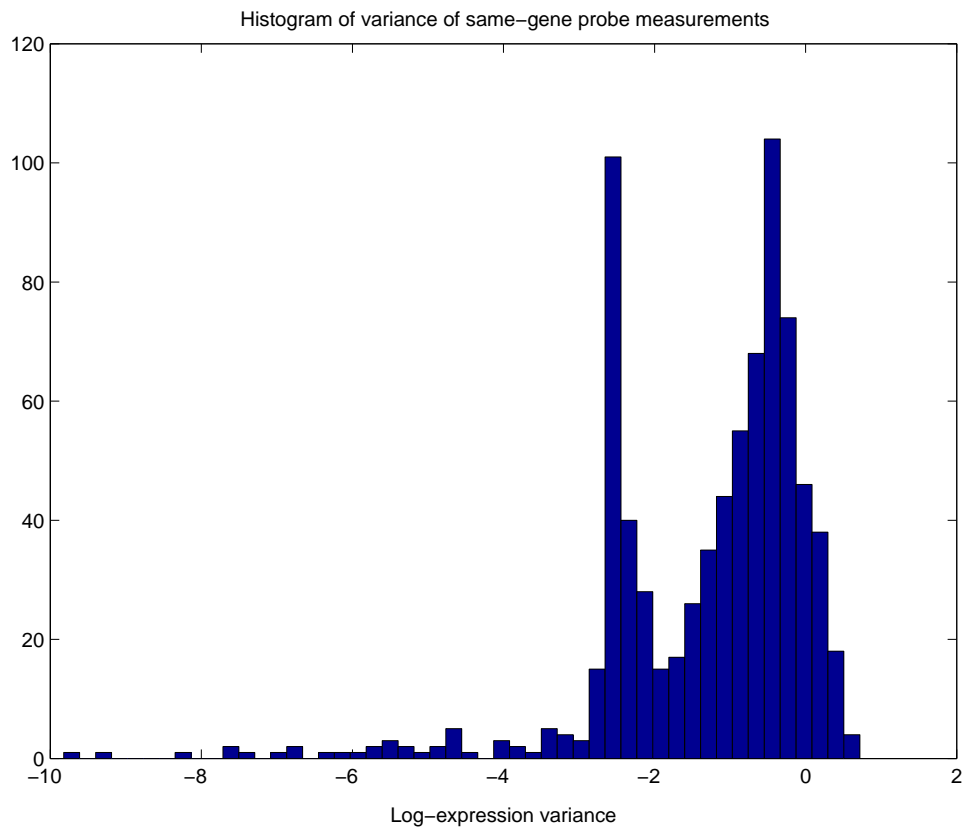


Figure 3: Distribution of variances of same-gene probe measurements

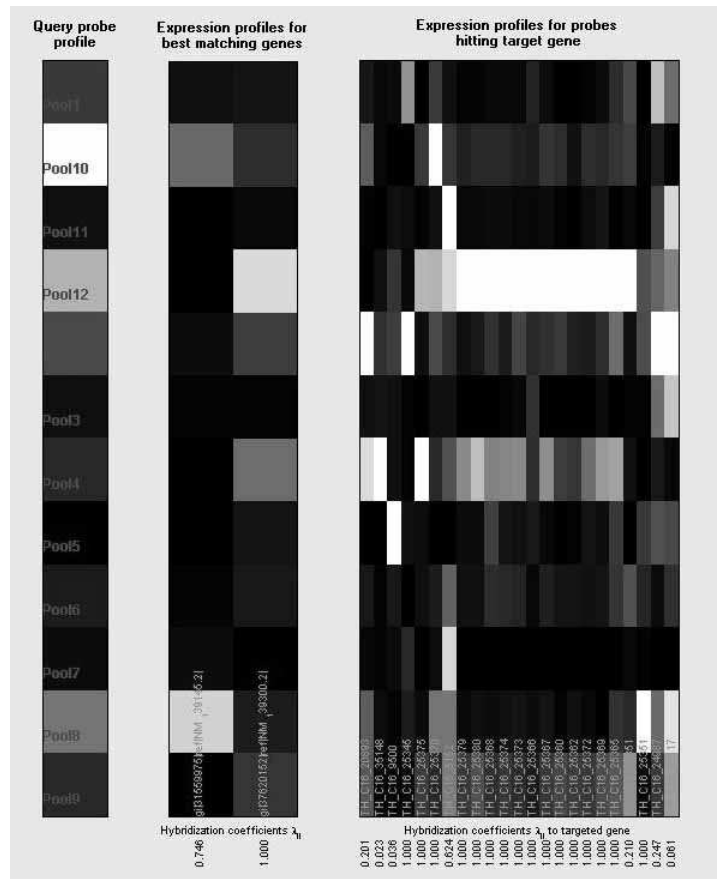


Figure 4: Sample probe and inferred profiles from XHC