MixThres: mixture models to define a hybridization threshold in DNA microarray experiments

by

F. Picard, M.-L. Martin-Magniette, S. Gagnot, V. Brunaud, J. Aubert, A.-V. Gendrel, S. Robin, M. Caboche, A. Lecharny, V. Colot.

Research Report No. 20
October 2008

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

MixThres: mixture models to define a hybridization threshold in DNA microarray experiments

F. Picard[⋆,†,◦], M.-L. Martin-Magniette[†,‡,◦], S. Gagnot[‡], V. Brunaud[‡], J. Aubert[†], A.-V. Gendrel[†,∗], S. Robin[†], M. Caboche[‡], A. Lecharny[‡], V. Colot[‡,+].

[⋆] *Laboratoire Biométrie et Biologie Evolutive,*
UMR CNRS 5558-Univ. Lyon 1, F-69622, Villeurbanne, France,
[†] *UMR AgroParisTech/INRA MIA 518,*
16 rue C. Bernard, 75231 Paris Cedex 05, France.
[‡] *URGV UMR INRA 1165-CNRS 8114-UEVE,*
2 rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France.
[∗] *MRC Clinical Sciences Centre, Faculty of Medicine,*
Imperial College London, United Kingdom.
[+] *CNRS-ENS UMR 8186, Paris, France.*
[◦] *both authors contributed equally to this work.*

Contact:
`mlmartin@agroparistech.fr`

Abstract

Even if one of the major applications of two-color DNA microarray hybridizations is to detect differentially expressed genes using intensity log-ratios, single channel signals provide also useful information as absolute value measurements which allow the description of gene expression patterns. In this context, it becomes crucial to determine the set of probes that hybridize, that is for which the intensity signal is greater than a hybridization threshold to be fixed. Existing procedures are either an arbitrary thresholding or require the knowledge of a population of non-hybridized probes. In this work we present the MixThres method to determine an adaptive hybridization threshold from intensity levels of the complete set of probes hybridized on a chip. We define a hybridization threshold based on the histogram of the probe intensity values. Our procedure is divided into two steps. First the intensity distribution is estimated using mixture models. Second a hybridization threshold is defined from the components of the mixture. We validate our method on DNA tiling array and expression array data. We show that our method has a good reproducibility, its specificity is greater than 97 % and its precision of 88 %. The R package MixThres is available at `http://www.agroparistech.fr/mia/outil.html`

Introduction

It is well known that microarray data are corrupted by different sources of noise, one of them being the autofluorescence of the probes. When the goal of the experiment is to study differential expression between two conditions, no distinction needs to be made between non-differentially expressed and non-hybridized probes since the resulting intensity log-ratios are close to zero on average. Nevertheless, when the purpose of the experiment is to detect transcripts produced within one condition, the identification of hybridized probes becomes crucial. Throughout the paper a non-hybridized probe is a probe that has an intensity signal lower than the hybridization threshold. For these probes, it means that there is not enough signal resulting from hybridization with specific target, and not enough identity between the probe and the other targets to allow hybridization.

To the best of our knowledge, existing procedures are usually based on an estimation of a local background per spot from image acquisition softwares. In this context, an arbitrary threshold is defined for each individual probe. For example a probe can be declared above the background if red and green intensities are more than two standard deviations above background (1). Consequently, these methods are strongly dependent on the estimation of the background, which may be a poor measure of nonspecific fluorescence (2). Alternatives are proposed (3; 4), but their procedures require the knowledge of either positive and negative controls or a population of non-hybridized probes. In this work we propose to develop a statistical method to determine an adaptive hybridization threshold from intensity levels of the complete set of probes present on a chip. Our objective is also to develop a method which can be applied to any type of DNA microarray experiments.

We define a hybridization threshold based on the histogram of the logarithm (base 2) of the median intensities of the probes. Our procedure is divided into two steps. The first step consists in the estimation of the intensity distribution using mixture models. The second step is to define the threshold from the estimated density based on the components of the mixture. When mixture models are used, one needs to define the distribution of the components of the mixture as well as their number. Intensity histograms under study are defined on a finite space since the intensity signal varies between a lower bound defined as the value of the DNA autofluorescence and an upper bound defined as the saturation value. Moreover an important number of probes have a signal that is close to the lower bound which leads to a positively skewed histogram. We propose to use a Gaussian mixture model complemented with the introduction of a truncation parameter to model undirectly the dissymmetrical form. The number of components of the mixture is chosen using the BIC. We also propose to compare different models with or without truncation parameters. Finally in the second step of the procedure, the hybridization threshold is defined using conditional probabilities of membership to the different mixture components.

Two approaches are used to validate the method. The first one consists in using expression data obtained with a DNA tiling microarray that covers the whole chromosome 4 of *Arabidopsis thaliana*. The probes of this array cover genic as well as intergenic regions and when hybridizing labeled mRNAs only, we expect that probes corresponding to intergenic regions should not hybridize. We show with these dataset that our method has a good reproducibility and a specificity of 98%. The second approach consists in applying the method

to transcriptome data produced on a DNA microarray and to confirm significant signal by RT-PCR approaches. With this second approach, we estimate the method precision at 88%.

The objectif of this paper is to provide both a detailed description (histogram modelling, parameter estimation, threshold definition) and a validation of the methodology. A R package named MixThres is also available on our web site. Our methodology has already been applied in three different biological projects published in biological journals, where the methodology cannot be detailed. The first application concerned ChIP-chip data that were analyzed with truncated Gaussian mixture models to determine enriched probes (5). The authors have studied the chromatin factor TERMINAL FLOWER 2/LIKE HETEROCHROMATIN PROTEIN 1 (TFL2/LHP1) and have shown that TFL2/LHP1 associates with hundreds of small domains, almost all of which correspond to genes located within euchromatin. The aim of the second application was to improve gene annotation of *Arabidopsis thaliana* at the structural and functional levels by studying 522 samples hybridized on CATMA microarrays (6). The hybridization threshold of these 522 hybridized samples allowed them to identify 465 novel genes. In the last application the authors built a repertoire of approximately 20 000 *Arabidopsis thaliana* promoter regions, amplified by PCR and printed on glass arrays to get a promoter microarray, used then for ChIP-chip experiment (7). Truncated Gaussian mixture models was used to model the immunoprecipited sample (IP sample) of the histone acetyltransferase GCN5. The authors have shown that GCN5 associated with 40 % of the tested promoters.

Materials and Methods

Data

In this article, the signal under study comes from a two-color microarray experiment which is used to perform single-channel analysis. In this context, we define the signal as the logarithm (base 2) of the intensity of one of the two channels. As it is the case in differential analysis, a normalization procedure is required to remove technical biases. To our knowledge the only single-channel normalization procedure consists in a within-array correction followed by a possible between-array correction (8). The within-array normalization is a redistribution of the lowess correction on each channel, and the possible between-array normalization is used to force signals coming from different arrays to share a common distribution. We use their within-array correction to define our data: let $R_{ig}^{\texttt{raw}}$ and $G_{ig}^{\texttt{raw}}$ denote the raw logarithms (base 2) of the red and green channel median intensities for array i and probe g ($g = 1, \dots G$). After the within-array normalization, the normalized signals R_{ig} and G_{ig} are defined by

$$R_{ig} = R_{ig}^{\texttt{raw}} - \frac{1}{2}c(A_{ig}), \qquad G_{ig} = G_{ig}^{\texttt{raw}} + \frac{1}{2}c(A_{ig}), \tag{1}$$

where $c(A_{ig})$ is the lowess correction (9). We could work from these data, but we prefer to work with a dye-swap to control dye biases. Consequently Y_g the signal intensity of gene g is defined such that: $Y_g = (R_{1g}+G_{2g})/2$ or $Y_g = (G_{1g}+R_{2g})/2$ depending on the experimental design. Since we assume that the distributions of the intensity signal of the two arrays of a dye-swap are close by definition, we do not perform a between-array normalization.

The range of intensities measured by the Genepix scanner for one probe on one array is between 0 and 16. Since probes present some autofluorescence, the signal is generally

greater than the lower bound, say ℓ and lower than an upper bound, say u. In practice the lower bound ℓ is close to 5. As for the upper bound u, it is about 16 and may be greater due to the reconstruction of the normalized signal using Equations (1). It does not mean there was saturation as it is the case for the raw data. A second major characteristics of the signal intensity is that an important number of probes have a signal which is close to the lower bound. For the intensity histogram, it leads to a dissymmetrical form with a left peak as shown in Figure 1.

Our objective is to define a hybridization threshold from the intensity histogram. For this purpose, we propose to estimate the distribution of the intensity signal by using a mixture model and to define a hybridization threshold based on conditional probabilities of belonging to each component of the mixture for each probe.

Modelling the intensity signal distribution

The use of mixture of distributions appears natural to model the intensity histogram. Each component of the mixture can be interpreted in terms of clusters of probes with similar signal intensities and the component of interest is the one with the highest mean intensity, corresponding to hybridized probes. When using mixture models, the choice of the form of the distributions is crucial. As a preliminary analysis, many usual distributions were tested including non-symmetrical distributions such as Lognormal, Gamma or Weibull distributions, but none systematically fitted the left peak of the empirical distribution (data not shown). In order to model this asymmetry, we introduce truncation parameters $\ell = \min_g(y_g)$ and $u = \max_g(y_g)$, and we build mixture models for truncated Gaussian distributions. The expression of a Gaussian density truncated at ℓ and u is easily derived from the density of a non-truncated Gaussian. It equals

$$g_\ell^u(y;\theta) = \frac{f(y;\theta)}{F(u;\theta) - F(\ell;\theta)} \mathbb{I}\{\ell \leq y \leq u\},$$

where $f(\bullet;\theta), F(\bullet;\theta)$ represent the density and cumulative distribution functions of a non-truncated Gaussian of parameter $\theta = (\mu, \sigma^2)$. The introduction of truncation parameters allows us to re-weight the densities on the support $[\ell, u]$. Now if we consider a mixture model of K truncated Gaussians and we denote p_k the proportion of the k-th component in the mixture, the density of the data is defined by:

$$g_\ell^u(y;p,\theta) = \sum_{k=1}^{K} p_k g_\ell^u(y;\theta_k).$$

The vectors of parameters of the mixture model are $\theta = (\theta_1, \ldots, \theta_K)$ and $p = (p_1, \ldots, p_K)$ with $\sum_{k=1}^K p_k = 1$. The key element of this model is that weights $(F(u;\theta_k) - F(\ell;\theta_k))$ will be more important for components with a mean close to the truncation bounds. This strategy is used to model the left peak observed on real intensity histograms.

An EM algorithm to estimate the parameters

When using truncated Gaussian distributions, it appears that the empirical estimators of the mean and of the variance are biased due to truncation. Then a fixed-point algorithm

can be used for correction (10), using the fact that the maximum likelihood estimators are equal to the moment estimators when ℓ and u are known. In the context of a mixture of truncated Gaussian distributions, it is also crucial to perform this correction. We propose thus to modify the traditional EM algorithm (11) by including a fixed-point algorithm in the M-step.

Let Z_{kg} be a hidden random variable equal to 1 if gene g belongs to component k and 0 otherwise. By definition conditional probabilities τ_{kg} equal $\Pr\{Z_{kg} = 1|Y_g = y_g\}$ and are computed in the E-step of the algorithm. Let $p^{(h)}$ and $\theta^{(h)}$ the values of the parameters at iteration (h), then at iteration $(h+1)$ of the E-step conditional probabilities are equal to :

$$\tau_{kg}^{(h+1)} = \frac{p_k^{(h)} g_\ell^u(y_g; \theta_k^{(h)})}{\sum_{l=1}^K p_l^{(h)} g_\ell^u(y_g; \theta_l^{(h)})}.$$

The M-step is a maximisation step where the new values of parameter p and θ are computed. In our version the M-step is divided into two steps: in the M_1-step, the proportions of the components in the mixture and the empirical estimators of the mean and the variance are computed and then in the M_2-step these estimators are corrected, leading to the unbiased estimators of θ. More precisely at iteration $(h+1)$ in the M_1-step, we compute $m_k^{(h+1)}, (s_k^2)^{(h+1)}$ and $p_k^{(h+1)}$ defined by:

$$\begin{aligned}
m_k^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)} y_g}{\sum_{g=1}^G \tau_{kg}^{(h+1)}}, \\
(s_k^2)^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)} (y_g - m_k^{(h+1)})^2}{\sum_{g=1}^G \tau_{kg}^{(h+1)}}, \\
p_k^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)}}{G}.
\end{aligned}$$

The estimators of m_k and s_k^2 are biased estimators of μ_k and σ_k^2, which are corrected in the M_2-step using a fixed-point algorithm. Denoting (j) the j^{th} iteration of the fixed-point algorithm, and setting $(\sigma_k^2)^{(0)} = (s_k^2)^{(h+1)}$, $\mu_k^{(0)} = m_k^{(h+1)}$, and $\theta_k^{(j)} = (\mu_k^{(j)}, (\sigma_k^2)^{(j)})$, we get the following algorithm:

$$\begin{aligned}
\mu_k^{(j+1)} &= m_k^{(h+1)} - A_k^{(j)} (\sigma_k^2)^{(j)}, \\
(\sigma_k^2)^{(j+1)} &= (s_k^2)^{(h+1)} \times \left\{ 1 + B_k^{(j)} + (\sigma_k^2)^{(j)} (A_k^2)^{(j)} \right\}^{-1}.
\end{aligned}$$

with

$$\begin{aligned}
A_k^{(j)} &= \frac{f(\ell; \theta_k^{(j)}) - f(u; \theta_k^{(j)})}{F(u; \theta_k^{(j)}) - F(\ell; \theta_k^{(j)})}, \\
B_k^{(j)} &= \frac{(\ell - \mu_k^{(j)}) f(\ell; \theta_k^{(j)}) - (u - \mu_k^{(j)}) f(u; \theta_k^{(j)})}{F(u; \theta_k^{(j)}) - F(\ell; \theta_k^{(j)})}.
\end{aligned}$$

In MixThres package we use the ε-accelerated version of the EM algorithm (12). The theoretical convergence of this EM algorithm is proved in (13).

Model choice

The modified EM algorithm allows us to estimate the parameters of the mixture model for a given number of components and for given truncation parameters. To fit best the histogram, we consider a collection of mixture models of untruncated, left, right and left-right truncated Gaussian distributions for which the number of components varies between 1 and K_{max}. Then we choose the best model which minimizes the BIC. Let us denote $m_K\{\ell, u\}$ a mixture model with K components with left-right truncation bounds ℓ and u respectively, the BIC is defined such that:

$$BIC(m_K\{\ell, u\}) = -2\log \mathcal{L}(Y; p, \theta | m_K\{\ell, u\}) + \log(G) \times (3K - 1),$$

where $\mathcal{L}(Y; \theta | m_K\{\ell, u\})$ denotes the likelihood of the mixture model $m_K\{\ell, u\}$.

Definition of the hybridization threshold

An ideal situation would be to select a mixture model with two components, one component for the non-hybridized population and one for the hybridized population. But in practice this situation does not occur because the reality is more complex leading to a signal distribution which is not bimodal. So, once the best model has been selected using the BIC, the components are ordered according to their mean value and the number of components is denoted $\widehat{K}$. Then our aim is to define a hybridization threshold to distinguish non-significant hybridization signal from significant hybridization signals. What we know is that the component $\widehat{K}$ with the highest mean is composed of hybridized probes, but nothing can be inferred for other components, since there is still an ambiguity between truly hybridized probes with low intensity and non-hybridized probes. One classical method to cluster probes is the *maximum a posteriori* rule (MAP): probe g belongs to component k_g^* if $k_g^* = \underset{s}{\text{Argmax}} \{\hat{\tau}_{sg}\}$. This procedure defines a natural threshold

$$T_{MAP} = \min \left\{ y_g \mid k_g^* = \widehat{K} \right\}.$$

Nevertheless, this procedure is very conservative in practice. This is why we propose the following threshold, $T(\varepsilon)$ above which a probe is declared as being hybridized:

$$T(\varepsilon) = \max \left\{ y_g \mid k_g^* < \widehat{K} \; \exists s \in \{1, \ldots, k_g^* - 1\}, \hat{\tau}_{sg} \geq \varepsilon \right\}.$$

In other words, intensity values are ranked by descending order, for each probe g k_g^* is determined using the MAP rule. Then if k_g^* differs from $\widehat{K}$, we calculate the conditional probability of belonging to each component with lower mean than $\mu_{k_g^*}$. The threshold $T(\varepsilon)$ is then defined as the first intensity value such that one of the calculated conditional probabilities is greater than ε. The performance of this procedure is assessed in the following with $\varepsilon = 10^{-4}$.

Results

We apply this method on two datasets. The first one consists of data obtained using a DNA tiling microarray of the entire known sequence of *Arabidopsis thaliana* chromosome

4 (19 Mb). The probes of this array cover genic as well as intergenic regions. The advantage of this array is that when hybridizing labeled mRNAs only, we expect that probes corresponding to intergenic regions should not hybridize. This dataset, named tiling array dataset, is used to estimate the specificity of our method. It is available on the ftp site of CATdb (`ftp://urgv.evry.inra.fr/CATdb/`). The second dataset consists of transcriptome experiments on CATMA microarray, available in the database CATdb (14). This second dataset, names CATMA dataset, has already been used to determine genes missed by the official annotation (6). This second dataset allows us to estimate the method precision. For both microarrays, all information relative to the probe sets is available on the integrative database FLAGdb++ (15).

Materials

The tiling array dataset contains results from 8 hybridizations corresponding to four biological samples in dye-swap. For each sample, polyA+ RNA was extracted from flower buds and open flowers harvested a fixed time over a one-week period from approximately 200 Arabidopsis wild-type plants of ecotype Columbia. Plants were grown under long-day conditions (16hrs white light, 22 degrees Celsius/ 8 hrs darkness, 18 degrees Celsius). Reverse transcription and cDNA labeling were performed as previously described (16). The DNA tiling microarray is described in detail elsewhere (5) and its accession number in ArrayExpress is A-MEXP-602. Briefly, this array contains ∼21000 sequential ∼1kb fragments that cover the 19 Mb Arabidopsis chromosome 4 sequence as well as a few regions located on the other four Arabidopsis chromosomes.

The CATMA dataset contains microarray data were extracted from the CATdb database developped (14). All samples were hybridized on a same array type, named CATMA for Complete Arabidopsis Thaliana MicroArray. CATMA microarray is a generic array which contains 24 576 Gene Specific Tag, small ORF and also probes designed in the chloroplastic et mitochondrial genomes. The CATMA dataset has already been analyzed with our method to identify hundreds of novel functional genes in the *Arabidopsis* genome (6). The interest of this analysis is an experimental validation allowing us an estimation of the method precision.

Data pre-processing

For both datasets, the raw data comprise the logarithm base 2 of median feature pixel intensities at wavelength 635 nm (red) and 532 nm (green). No background was subtracted. The array-by-array normalization is performed to remove systematic biases. First, we exclude spots that are considered badly formed features. Then we perform a global intensity-dependent normalization using the lowess procedure (9). Finally, for each block, the log-ratio median calculated over the values for the entire block is subtracted from each individual log-ratio value.

For the tiling array dataset, as explained in the experimental design and the Materials section, we focus on the expression data of wild-type plants of Columbia ecotype since the *Arabidopsis thaliana* annotation and the microarray probes are based on this ecotype. For each biological sample, we define the intensity signal as the average on the dye-swap, since the correlations between the two normalized signals obtained from a dye-swap varies

between 0.92 and 0.98. This indicates a high reproducibility of the microarray data between technical replicates. Informations about the data are given in Table 1.

For both datasets, to estimate the intensity histogram, a collection of mixture model of Gaussian untruncated, left-truncated, right-truncated or left-right truncated is considered. For each family of Gaussian the number of components of the mixture model varies between 1 and K_{max}=5. Consequently the model collection consists in 20 models (4 families of Gaussian times 5 different number of components). The hybridization threshold is calculated following the procedure described in Methods. The results of the tiling-array dataset are summarized in Table 2. Figures 1, 2, 3 show an example of an intensity signal histogram and its estimation, as well as the components of the selected mixture model and the estimated hybridization threshold. For the CATMA dataset, we refer to the paper of (6) for the results and their interpretation.

Reproducibility of MixThres

For each biological sample and each probe of the tiling array dataset, we calculate a hybridization index which equals 1 if the signal intensity is higher than the hybridization threshold and 0 otherwise. Since four biological samples are available, we obtain for each probe a hybridization score between 0 and 4. Out of the 21602 probes, 4681 are declared hybridized four times (score=4) and 13681 are never declared hybridized (score=0), thus 85 % of the results are coherent between the four biological samples. Moreover, the threshold as well as the percentage of probes declared hybridized vary between the four biological samples reasonably(Table 2). This shows that our method provides reproducible results. Consequently from now, we present results based on the hybridization score. A probe is declared hybridized if the hybridization score is at least 3. We choose this definition since a score of 3 means that the probe is declared for the two biological replicates and for one of the technical replicates. According to this rule, 27,3 % among the 21602 probes are declared hybridized: 4681 have a score of 4 and 1218 a score of 3. At first sight this percentage of 27,3 % can be considered low, nevertheless since targets are labeled mRNA it must be interpreted with respect to the percentage of tiles covering genes, which is about 73 %.

Specificity of the method estimated with the first dataset

By definition the specificity is the probability to declare a probe not hybridized rightly. To estimate it we focus on the set of 5724 probes which cover intergenic regions.

Among the 5724 probes, 143 are declared hybridized. Consequently the specificity is estimated at at least 97,5 %. We analyze the 143 potential false positives in detail to find an explanation of the hybridization signal. To do so we perform a bioinformatic analysis. We first perform a blast of the sequences of the 143 probes against the whole genome (e-value $< 10^{-10}$) to find hits where at least 85 nucleotides are identical. The number of hits is greater than 5 for 67 probes among the 143, indicating possible cross-hybridization. Secondly we investigate the quality of the microarray probes, which were built from PCR products. During a validation step of the microarray, a PCR product quality index, available in the FLAGdb++ database, was attributed to each probe and for 10 probes this index indicates that either the size of the PCR product associated to the probe is wrong or it gives a multiple band product. So the nature of these 10 probes is unknown and they should have

been removed from the set of probes which cover inter-genomic regions. Third thanks to the database FLAGdb++, we find that 13 probes cover at least an Expressed Sequence Tag (EST), and 4 probes have associated Massively Parallel Signature Sequencing data (MPSS data). Briefly, Ests are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene and MPSS data are short sequence signatures from a defined position within an mRNA. The output of MPSS is similar to SAGE but the method of obtaining the data is different. We refer to the data help of FLAGdb++ for more complete explanations. A similar expertise was performed for the 5724 probes and it allows us to refine the set of truly non-hybridized probes from 5664 to 3701 probes, which gives a new estimation of the specificity at 98.7 % (49 false positives among 3701). Similar estimate of the specificity is found when the procedure is applied to a seedling experiment (data not shown).

Precision of the method estimated with the second dataset

Another quantity of interest is the sensitivity defined by the probability to declare rightly a probe hybridized. Unfortunately it is impossible to determine a set of probes which must hybridize and this is why we do not carry out this study. In contrast we are able to estimate the precision which is the proportion of probes declared righly hybridized amongst all probes declared hybridized with the CATMA dataset. Indeed in the analysis of the 522 hybridized samples, for all probes declared hybridized at least once, an experimental validation was performed to validate this result. Among the 465 new genes found hybridized at least one times in the 522 hybridized samples, the hybridization evidence was confirmed by RT-PCR approaches for 88%, thus the method precision is estimated at 88%.

Discussion

We propose a method to identify hybridized probes using mixture distributions. We provide a definition of a hybridization threshold to identify hybridized probes. From the modelling point of view, considering the truncation leads to a better fit of the left peak of intensity histograms. On the studied datasets we observe that the first smallest value of BIC is associated with a truncated Gaussian mixture model and the second smallest value with the untruncated Gaussian mixture with the same number of components. Interestingly both hybridization thresholds derived from these two models are equal. This indicates that the threshold value is well-defined and does not depend on the truncation parameters. We suggest users to fix K_{max} to 5, this number of component is usually sufficient to model correctly the intensity histogram.

The hybridization threshold depends on parameter ε, which has been set at 10^{-4} in this study. Consequently the specificity and sensitivity depend on ε since a greater value will lead to a decrease (increase) in specificity (sensitivity). In order to assess the role of ε, we have tested several values from 10^{-3} and 10^{-6} and estimations of the specificity and the precision are the same. We emphasize that our method focuses on the identification of hybridized probes. Consequently the user needs to be careful regarding the interpretation of the hybridization threshold. When the signal is higher than the threshold, we showed that the corresponding probe is hybridized (see the section Specificity). However when the signal is lower than the threshold, the corresponding probe is not necessarily non-hybridized.

We apply the method on normalized data from dye-swaps, but it can also be applied on raw data or on normalized data without dye-swap. In these cases the user must be aware that technical biases could exist and alter the results. At present our model does not consider biological replicates. When available, we propose to check for reproducibility using a hybridization score as shown in Reproducibility section. Note that the generalization of a mixture model taking several samples into account is straightforward if the number of components is independent of the sample.

Funding

A.-V. Gendrel was supported by a graduate studentship from the French Ministry of Research. V. Colot was supported by a grant from the EU Network of Excellence The Epigenome.

References

[1] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[2] C.S. Brown, P.C. Goodwin, and P.K. Sorger. Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Science of the United States of America.*, 98(16):8944–8949, 2001.

[3] M Bilban, LK Buehler, S Head, G Desoye, and V Quaranta. Defining signal thresholds in dna microarrays: exemplary application for invasive cancer. *BMC Genomics*, 3(1):19, 2002.

[4] V. Stolc, Z. Gauhar, C. Mason, G. Halasz, MF. van Batenburg, SA. Rifkin, S. Hua, T. Herreman, W. Tongprasit, PE. Barbano, HJ. Bussemaker, and KP. White. A gene expression map for the euchromatic genome of drosophila melanogaster. *Science*, 306:655–660, 2004.

[5] F. Turck, F. Roudier, S. Farrona, M.-L. Martin-Magniette, E. Guillaume, N. Buisine, S. Gagnot, R.A. Martienssen, G. Coupland, and V. Colot. Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet*, 3(6):e86, Jun 2007.

[6] S. Aubourg, M.-L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, PE. Jullien, M. Ingouff, V. Thareau, T. Schiex, A. Lecharny, and JP. Renou. Analysis of catma transcriptome data identifies hundreds of novel functional genes and improves gene models in the arabidopsis genome. *BMC Genomics*, 8:401, 2007.

[7] M. Benhamed, M.-L. Martin-Magniette, L. Taconnat, F. Bitton, C. Servet, R. De Clercq, B. De Meyer, C. Buysschaert, S. Rombauts, R. Villarroel, S. Aubourg, J. Beynon, R.P. Bhalerao, G. Coupland, W. Gruissem, F.L.H. Menke, B. Weisshaar, J.-P. Renou, D.-X. Zhou, and P. Hilson. Genome-scale arabidopsis promoter array identifies targets of the histone acetyltransferase GCN5. *Plant Journal*, 2008.

[8] Y. H. Yang and N. Thorne. Single channel normalisation for cDNA microarray data. *IMS Lecture Notes– Monograph Series*, 40:403–418, 2003.

[9] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[10] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. (2nd Edition)*. Wiley Series in Probability and Statistics. John Wiley & Sons. N.Y., 1994.

[11] A. Dempster, Laird N., and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[12] M. Kuroda and M. Sakakihara. Accelerating the convergence of the em algorithm using the vector ϵ algorithm. *Computational Statistics and Data Analysis*, 51:1546–1561, 2006.

[13] M. Wang, M. Kuroda, M. Sakakihara, and Z. Geng. Acceleration of the em algorithm using the vector epsilon algorithm. *Computational Statistics*, 23:469–486, 2008.

[14] S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J.-P. Renou, A. Lecharny, and V. Brunaud. CATdb: a public access to arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36(Database-Issue):986–990, 2008.

[15] F. Samson, V. Brunaud, S. Duchene, Y. De Oliveira, M. Caboche, A. Lecharny, and S. Aubourg. FLAGdb++: a database for the functional analysis of the arabidopsis genome. *Nucleic Acids Res.*, 32:D347–50, 2004.

[16] Z. Lippman, AV. Gendrel, M. Black, MW. Vaughn, N. Dedhia, WR. McCombie, K. Lavine, V. Mittal, B. May, KD. Kasschau, JC. Carrington, RW. Doerge, V. Colot, and R. Martienssen. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430:471–476, 2004.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 2 about here.]

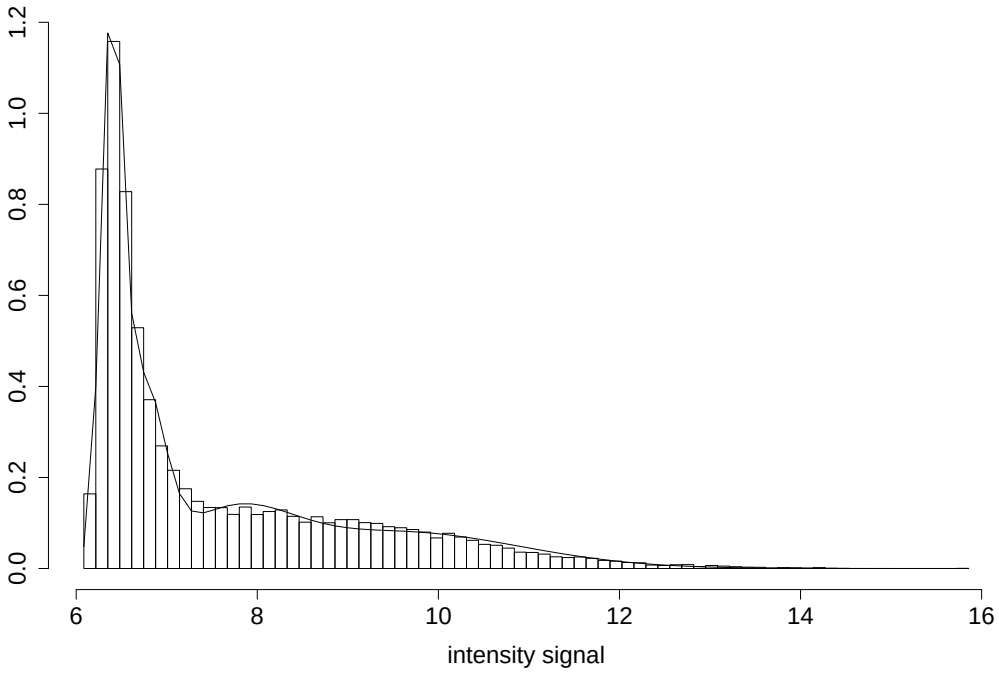

Figure 1: Intensity histogram of the second biological sample with estimated density (mixture of 4 no-truncated Gaussian distributions).

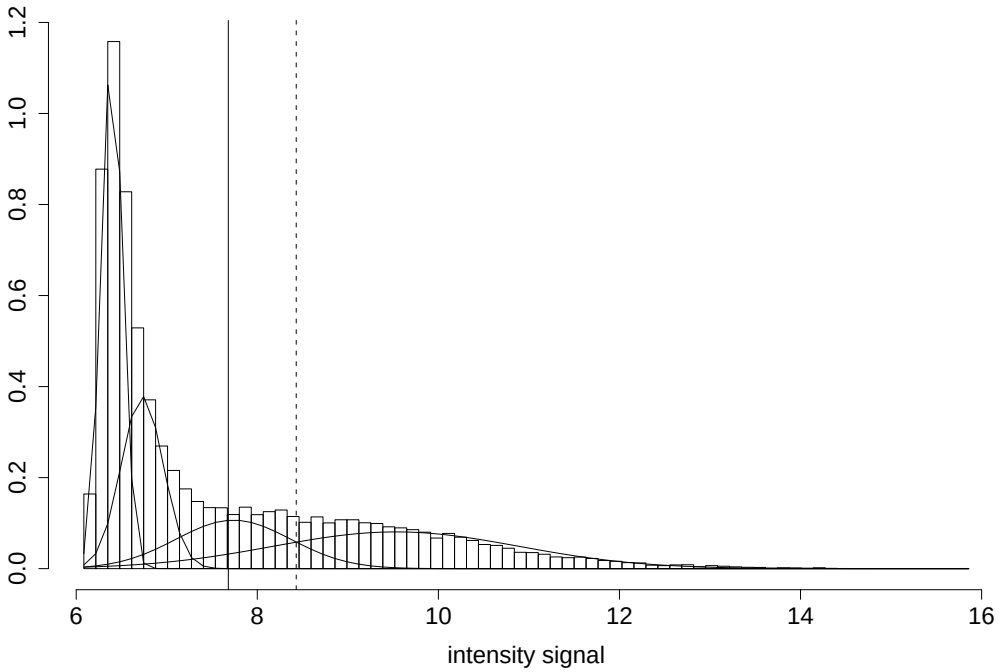

Figure 2: Intensity histogram of the second biological sample with the 4 components of the selected model. Plain vertical line indicates the hybridization threshold $T(\varepsilon) = 7.68$ with $\varepsilon = 10^{-4}$. Dot vertical line indicates the MAP threshold $T_{MAP} = 8.43$.

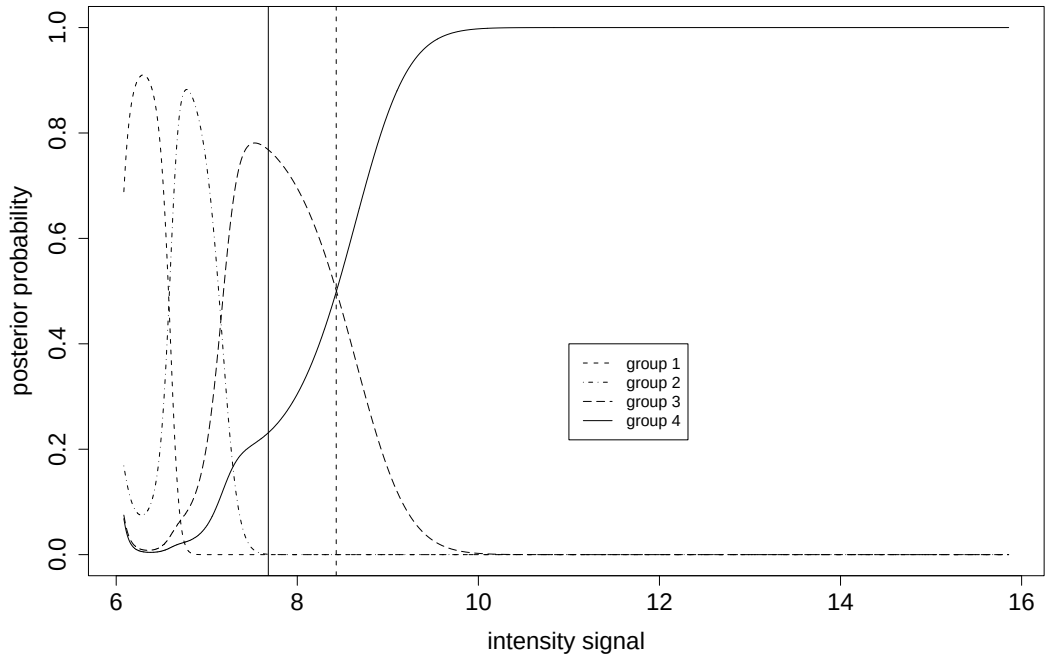

Figure 3: Conditional probabilities for the second biological sample according to the intensity signal. Plain vertical line for the hybridization threshold $T(\varepsilon)$ with $\varepsilon = 10^{-4}$. Dot vertical line for T_{MAP}.

Table 1: Data summary for the *Arabidopsis thaliana* experiment.

Biological sample	Technical sample	Normalized signal range
1	117.Red	[5.50;15.55]
1	116.Green	[5.76;15.91]
2	142.Red	[6.00;16.15]
2	143.Green	[5.86;15.56]
3	118.Red	[6.09;15.09]
3	119.Green	[6.26;15.57]
4	134.Red	[6.20;15.39]
4	135.Green	[6.24;15.56]

Table 2: Results of the hybridization study for each dye-swap. The second column is the correlation between the 2 arrays of each dye-swap.

Biological sample	Corr.	Selected model	Threshold	% of hybridized probes
1	0.92	5 right trunc. Gauss.	8.55	24.2 %
2	0.97	4 no trunc. Gauss.	7.68	35.4%
3	0.96	5 right trunc. Gauss.	9.23	27.1%
4	0.98	5 right trunc. Gauss.	9.47	30.2%