Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome

Caroline Bérard, UMR AgroParisTech/INRA MIA 518
Marie-Laure Martin-Magniette, UMR AgroParisTech/INRA MIA 518, URGV UMR INRA/CNRS/UEVE
Véronique Brunaud, URGV UMR INRA/CNRS/UEVE
Sébastien Aubourg, URGV UMR INRA/CNRS/UEVE
Stéphane Robin, UMR AgroParisTech/INRA MIA 518
Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome

Caroline Bérard, Marie-Laure Martin-Magniette, Véronique Brunaud, Sébastien Aubourg, and Stéphane Robin

Abstract

Tiling arrays make possible a large-scale exploration of the genome thanks to probes which cover the whole genome with very high density, up to 2,000,000 probes. Biological questions usually addressed are either the expression difference between two conditions or the detection of transcribed regions. In this work, we propose to consider both questions simultaneously as an unsupervised classification problem by modeling the joint distribution of the two conditions. In contrast to previous methods, we account for all available information on the probes as well as biological knowledge such as annotation and spatial dependence between probes. Since probes are not biologically relevant units, we propose a classification rule for non-connected regions covered by several probes. Applications to transcriptomic and ChIP-chip data of *Arabidopsis thaliana* obtained with a NimbleGen tiling array highlight the importance of a precise modeling and of the region classification. The "TAHMMAnnot" package is implemented in R and C and is freely available from CRAN.

KEYWORDS: bivariate Gaussian mixture, hidden Markov model, tiling arrays, unsupervised classification

Author Notes: The authors thank F. Roudier and V. Colot from IBENS for providing ChIP-chip data and for helpful discussions of biological interpretation, and S. Derozier and A. Lecharny from URGV for their help in bioinformatic analysis. The authors are grateful to S. Balzergue from URGV for providing transcriptomic tiling array data. This work was funded by MIA, GAP and MICA departments of INRA.
1 Introduction

For 15 years, the study of large-scale genomes has been possible thanks to DNA microarrays which originally had probes designed on genes. The tiling arrays now propose probes which cover the whole genome without a priori knowledge of structural annotation. The density is still increasing and companies now offer tiling arrays with 2 million probes. Thanks to technological advances and to the miniaturization of the support, tiling arrays have become a usual tool in biology laboratories. They make possible a large-scale exploration of the genome with a reasonable cost. Recently, Next Generation Sequencing (NGS) technology has revolutionized the domain because it directly produces nucleotide sequences. However, like any new technology, it remains expensive and suffers for now from uncontrolled technical biases (Oshlack, Robinson, and Young, 2010). NGS technology also raises new questions on read mapping or genome assembly. For all these reasons, tiling arrays, with a technology which is well controlled, remain widely used. They are a powerful tool for analyzing all kinds of experiments and are used in a wide range of studies such as DNA methylation, chromatin modification or transcription factor analysis with ChIP-chip experiments (Buck and Lieb, 2004), DNA copy number variation detection with CGH (Pinkel, Segraves, Sudar, Clark, Poole, Kowbel, Collins, Kuo, Chen, and Zhai, 1998, Snijders, Nowak, Segraves, Blackwood, and et al., 2001) and surveys of genomic transcriptional activities or transcript mapping with transcriptional experiments (Mockler, Chan, Sundaresan, and et al., 2005, Yamada, Lim, Dale, and et al., 2003, Hanada, Zhang, Borevitz, Li, and Shiu, 2007).

For comparative genomic hybridization, many different approaches exist for determining DNA copy number variations in CGH data such as segmentation (Hupé, Stransky, Thiery, Radvanyi, and Barillot, 2004, Picard, Robin, Lavielle, Vaisse, and Daudin, 2005) or Hidden Markov Models (Fridlyand, Snijders, Pinkel, Albertson, and Jain, 2004, Seifert, Banaei, Keilwagen, Mette, Houben, Roudier, Colot, Grosse, and Strickert, 2009).

Transcriptomic experiments may have one of two different purposes: the detection of transcribed regions or the study of gene expression across several conditions (also called differential analysis). Most methods previously developed for tiling array transcriptomic data deal with the first purpose. Among them, some methods are based on probe-by-probe statistical tests (e.g. Fisher test developed by Halasz, van Batenburg, Perusse, Hua, Lu, White, and Bussemaker (2006)) and others are based on segmentation methods such as Huber, Toedling, and Steinmetz (2006) or Zeller, Henz, Laubinger, Weigel, and Ratsch (2008) or HMM (Nicolas, Leduc, Robin, Rasmussen, Jarmer, and Bessières, 2009). The incorporation of annotation knowledge has also been proposed in a supervised framework (Du, Rozowsky, Korbel, Zhang, Royce, Schultz, Snyder, and Gerstein, 2006; Munch,
Gardner, Arctander, and Krogh, 2006). Surprisingly few methods are devoted to the study of gene expression profiles across samples based on tiling arrays. Some methods aggregate probes within regions and then apply hypothesis testing. The method gSAM (Ghosh, Hirsch, Sekinger, Kapranov, Struhl, and Gingeras, 2007) is an extension of SAM, which models the differential expression of a given region by a constant piece-wise function. In the TileMap method (Ji and Wong, 2005) each probe is used separately and a test statistic is proposed, based on a hierarchical empirical Bayes model.

For ChIP-chip experiments where the chromatin immunoprecipitation sample (ChIP) and the reference sample of genomic DNA are compared, the main goal is to detect regions enriched by ChIP. Johnson, Li, Meyer, Gottardo, Carroll, Brown, and Liu (2006) proposed a Model-based Analysis of Tiling arrays (MAT) algorithm dedicated to Affymetrix arrays. MAT models the baseline probe behavior based on probe sequence characteristics and genome copy number. Li, Meyer, and Liu (2005) proposed to model the behavior of each probe and a 2-state HMM is then used to estimate the enrichment probability at each probe location. In these two methods it is assumed that only a small proportion of probes is enriched by ChIP. This assumption is reasonable for ChIP-chip experiments dealing with transcription factors but not for histone modification or DNA methylation where a large enrichment is expected. Humburg, Bulger, and Stone (2008) have suggested a parameter estimation procedure for robust HMM analysis of chromatin structure where several long regions of interest are expected. ChIP-chip data can also be seen as one signal along the genome when using the log-ratio between the intensities of the ChIP and the reference samples. Analyses are then usually done using a sliding window (Cawley, Bekiranov, Ng, and et al., 2004) and statistic tests. Keles, van de Laan, Dudoit, and Cawley (2004) and He, Li, Zhou, Deng, Zhao, and Luo (2009) have respectively proposed the Welch t-statistic and the non parametric Wilcoxon rank-sum method.

When two samples are compared, most methods rely on the log-ratios. But this can mask the multimodality of the data due to the dimension reduction. To overcome this problem, Martin-Magniette, Mary-Huard, Bérard, and Robin (2008) have argued that is worth working directly with the two measurements of each probe. For ChIP-chip data, they have proposed to model the distribution of the ChIP sample conditionally to the reference sample by a mixture of two linear regressions. The same idea was used by Johannes, Wardenaar, Colomé-Tatché, and et al. (2010) to directly compare two ChIP samples. Assuming that the samples play a symmetric role, they introduced a bidimensional mixture model of four components with constraints on the mean parameters to study the differential enrichment.

In this article, we focus on the modeling of the joint distribution with an unsupervised classification point of view to study the difference between two ChIP
or transcriptomic samples. Comparing the two samples requires distinguishing four different biologically interpretable groups of probes: a group with similar behavior in both samples, a group with higher intensity in the first sample than in the second sample, a symmetric group with higher intensity in the second sample and a last group with low intensity in both samples which can be viewed as noise, corresponding to the non-transcribed regions (cf Figure 1).

Figure 1: Schematic explanation of the 4 groups to consider when comparing two samples. Example of transcriptomic data.

A parametric classification method based on multivariate mixture models permits a direct comparison of two (or more) samples. This differs from a log-ratio based study which does not distinguish the group of identical behavior from the noise group. Our method simultaneously analyses the expression difference between two conditions and detects hybridized regions. In contrast to previous methods, we consider all the available information: the intensity of the two signals, the position of the probe along the genome and its structural annotation. The position of the probe is important because there is a signal dependence between adjacent probes due to the high resolution of tiling arrays. Structural annotation informs us about the location of the probes in intergenic, exonic or intronic regions (see Figure 2, screen capture of FLAGdb++ (Samson, Brunaud, Duchêne, De Oliveira, Caboche, Lecharny, and Aubourg, 2004)). This must be accounted for as, in a transcriptomic experiment, probes annotated as exonic are more likely to be hybridized whereas intergenic or intronic (non-coding) probes should be mainly in the noise group.
We use a 4-state heterogenous hidden Markov model with bidimensional Gaussian emission densities to gather all this information. Finally since genome annotation is an on-going process with possible errors, we will discuss the relevance of its use for each specific application.

Figure 2: Example of genome annotation. Yellow squares: probes; blue arrows: known genes. The arrows correspond to exons and the fine lines between arrows correspond to introns.

Most methods provide probe-by-probe results. As for the classification purpose, the HMM provides an answer for each probe, via the posterior probabilities. However probes are not relevant units from a biological point of view. Although HMM are widely used for classification problems in genomic data, the classification of biologically interesting regions is not a common practice. To get a result by region, the most commonly used method is a sliding window approach where the probe signals are merged \textit{a priori}. Another method proposed by Li et al. (2005) is to define a region as at least two probes with positive log-odds enrichment values in the ChIP sample and at least one probe with a log-odds enrichment value lower than $-15$ in the control sample. But these methods do not deal with regions covered by several non-adjacent probes, such as genes with exons and introns. We propose a new solution deriving a posterior probability for a region given \textit{a priori} with arbitrary structure (such as a non-connected region) and also a procedure of gene classification which allows us to quickly get a list of differentially expressed genes. This calculation clearly improves the classification of regions compared to the results derived from the classification of the probes.

The article is organized as follows. The statistical model is described in Section 2.1. The inference is given in Section 2.2. Section 2.3 describes the classification method for a probe and a gene. In Section 3, we discuss the different sub-models and the method is illustrated on NimbleGen tiling arrays for transcriptomic and ChIP-chip data of the plant \textit{Arabidopsis thaliana}. We also perform a simulation study in Section 3.3 to compare our approach with three existing methods. The main conclusions and some possible extensions are discussed in Section 4. The method is implemented in R and C and is freely available from CRAN, with the “TAHMMAnnot” package.

DOI: 10.2202/1544-6115.1692
2 Methods

We propose a non-supervised classification model to compare the intensities of the two samples hybridized on the array for each probe. It accounts for all available information for each probe: the two intensities to be compared, the position of the probe along the chromosome and the current annotation of the probe (for example exonic, intronic, intergenic, transposable element, etc). As in Johannes et al. (2010), our method does not deal with the usual log-ratio but rather considers the two intensities separately and the joint signal of the two samples is modeled.

2.1 Model

For probe \( t \), we denote

- \( X_t = (X_{t1}, X_{t2}) \) the log-intensities for both samples,
- \( C_t \in \{1, ..., P\} \) the annotation category,
- \( Z_t \in \{1, ..., K\} \) the unknown status.

In our case, \( K = 4 \), Groups 1 and 2 will refer respectively to ‘noise’ and ‘identical’ probes, whereas Groups 3 and 4 will refer to differentially hybridized probes. To account for the dependence between adjacent probes, we assume that the process \( \{Z_t\} \) is a first order Markov chain with heterogenous transition \( \pi^p \) depending on the annotation category:

\[
P(Z_t = l | Z_{t-1} = k, C_t = p) = \pi^p_{kl}
\]

We then assume that the \( \{X_t\} \) are independent conditionally to the \( \{Z_t\} \) with distribution

\[
(X_t | Z_t = k) \sim \mathcal{N}(\mu_k, \Sigma_k).
\]

The parameters \( \mu_k \) and \( \Sigma_k \) of the Gaussian distribution do not depend on the annotation category.

If there is no spatial dependence, the \( \{Z_t\} \) are independent and distributed according to a multinomial of parameter \( \pi^p_k \) which corresponds to the proportion of probes from group \( k \) in annotation category \( p \). If there is no annotation and no spatial dependence, the model comes down to a mixture model with four components. All these sub-models are discussed in Section 3. We focus on the model with \( K = 4 \) groups, but the models with \( K = 2 \) and \( K = 3 \) are also possible if a differential expression does not exist in one (or two) directions. The constraint of the orientation of the first two components remains unchanged. If there is no biological information about the number of groups, the choice between a model with 2, 3 or 4 groups can be made with a selection criterion such as BIC.
2.2 Inference

We use the parametrization proposed by Banfield and Raftery (1993) which enables us to characterize geometric properties of the Gaussian density (volume, shape, orientation). This parametrization considers the eigenvalue decomposition of the variance matrix of the group $k$:

$$\Sigma_k = D_k \Lambda_k D_k'.$$

The matrix $\Lambda_k$ describes both the volume and the shape of the ellipse associated with the Gaussian distribution. The matrix $D_k$ describes the orientation of this ellipse. A similar decomposition of variance matrix is studied by Celeux and Govaert (1995) in the Gaussian mixture context and is implemented in the Mixmod software (Biernacki, Celeux, Echenim, Govaert, and Langrognet, 2007) and in the Mclust R package (Fraley and Raftery, 2006, revised 2010). In their approach, each term of the decomposition is either equal in all groups or specific to each group.

In our case we need an intermediate modeling. By definition groups 1 and 2 should have the same orientation (see Figure 1), which implies that $D_1 = D_2$. Furthermore the dispersion around the main axis is expected to be similar in all groups, which amounts to fixing the second eigenvalue of $\Sigma_k$ for all groups. This can be summarized as

$$\Sigma_k = D_k \Lambda_k D_k', \quad \text{for } k = 1, \ldots, 4;$$

$$D_1 = D_2;$$

$$\Lambda_k = \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \quad \text{with } u_{1k} > u_2, \quad \text{for } k = 1, \ldots, 4.$$

The parameters $\{\pi_p\}, \{\mu_k\}, \{D_k\}, \{u_{1k}\}$ and $u_2$ are estimated using the EM algorithm. The E-step is achieved with the Forward-Backward algorithm (Baum (1972), Rabiner (1989)). This model requires a specific M-step to satisfy the prescribed constraints on the variance matrices (see Appendix for formulas). These constraints cannot be satisfied with Mixmod or Mclust. In Johannes et al. (2010) the constraints are related to the means which assumes a strong symmetry in the distribution of the data.

2.3 Posterior probabilities for a region

The posterior probability for each probe to belong to each group

$$\tau_{ik,X} = P(Z_t = k|X,C), \quad \text{where } X = \{X_t\},$$

is obtained as a by-product of the Forward-Backward algorithm and can be used for probe classification. Nevertheless, the probe may not be the relevant biological entity and we would rather look at the status of a region such as a gene or a
transposable element. We propose the analogous calculation of the posterior probabilities for a region. Many quantities could be calculated but our criterion is the natural classification criterion for a region in the context of HMM. It is a generalization of the posterior probabilities, which allows us to locally restore the hidden path. A region is declared in the group $k$ if the hidden path remains in the $k$-th state throughout this region.

We define a region as a set of probes that can be decomposed into sub-regions of adjacent probes. As a reference to the gene structure and without loss of generality, we will refer to these sub-regions as ‘exons’ and to the spaces between them as ‘introns’. In eukaryotic genes, exons correspond to coding regions that are spliced together in the transcript to become the mRNA, after removal of introns, which are not expressed. We define the posterior probability for such a region $g$ to belong to group $k$ as the probability for all its probes to belong to group $k$:

$$Q_{gk} = P(\forall t \in g, Z_t = k | X, C) \quad (1)$$

A region is covered by several probes and our definition considers the case of a homogeneous region, which is when all probes have the same status. We compute this probability for a gene $g$ with $Q$ exons (and $Q - 1$ introns). We denote $e_q$ the position of the first probe of exon $q$ and $i_q$ the position of the first probe of intron $q$; thus $i_q - 1$ refers to the last probe of exon $q - 1$. As convention, we denote $i_Q$ the position of the first probe after the end of the gene. We also denote $X_u^v = \{X_t\}_{u \leq t \leq v}$.

We get

$$Q_{gk} = P(\forall t \in g, Z_t = k | X, C)$$

$$= P(Z_{e_1} = k | X_{e_1}^{i_1}) \times \left( \prod_{t = e_1 + 1}^{i_1 - 1} A_{k,t} \right) \times \prod_{q = 2}^{Q - 1} \left( B_{k,q} \times \prod_{t = e_q + 1}^{i_q - 1} A_{k,t} \right) \times B_{k,Q} \times \left( \prod_{t = e_Q + 1}^{i_Q - 2} A_{k,t} \right) \times P(Z_{i_Q - 1} = k | Z_{i_Q - 2} = k, X_{i_Q - 1}^{n_Q})$$

where $A_{k,s} = P(Z_s = k | Z_{s-1} = k, X_s, C)$,

$B_{k,q} = P(Z_{e_q} = k | Z_{i_q - 1} = k, X_{i_q - 1}^{e_q}, C)$, with $C = \{C_t\}$.

All these terms can be calculated with the Forward recursion of the Forward / Backward algorithm. Note that the sum of the $Q_{gk}$ for $k \in \{1, \ldots, 4\}$ is not equal to one, as all probes from the same gene may not have the same status. Changing the list of exons associated to a gene allows us to account for alternative splicing or to exclude the last exon for which the expression level could be lower due to the labeling protocol (Nicolas et al., 2009).
3 Applications

We now illustrate the use of the proposed modeling on both ChIP-chip and transcriptomic data. All experiments have been carried out on a two-color NimbleGen array of about 700,000 probes designed to insure a constant hybridization temperature. For each dataset two biological replicates are available, for which hybridizations are performed in dye-swap. The normalization step is done by averaging on the dye-swap the two signals of each technical replicate to remove the gene-specific dye bias (Mary-Huard, Picard, and Robin, 2006). Analyses are performed per chromosome on the normalized data. Then we present a simulation study to compare our approach with 3 existing methods.

3.1 ChIP-chip dataset

We analyse the data from a histone modification (H3K9me2) study in *Arabidopsis thaliana* for a wildtype and a mutant (polIV). We directly compare the ChIP samples of the wildtype and the mutant to study their difference in methylation.

The methylation mainly affects transposable elements but also large adjacent regions (Humburg et al., 2008). Therefore the enriched probes are expected to be found both in the transposable elements and in wide neighboring regions. As the methylation does not affect a specific annotation category, the standard annotation information is not useful to detect enriched probes. This suggests using an HMM model without the annotation knowledge.

The histone methylation under study is known to be weakly present in the genome and the mutant is known to have a loss of methylation compared to the wildtype (Bernatavichute, Zhang, Cokus, Pellegrini, and Jacobsen, 2008). We find consistent results as shown by the estimated proportions in each group given by our model: 43% noise, 21% identical, 22% loss in mutant, 14% gain in mutant.

The studied histone modification is also known as a heterochromatin mark. Most regions covered by H3K9me2 are adjacent and cover several megabases in pericentromeric regions or in interstitial heterochromatin regions (a tightly packed form of chromatin) as the knob of chromosome 4, but there are also smaller regions (islands of heterochromatin) located in euchromatin (a lightly packed form of chromatin) and covering mainly transposable elements (Bernatavichute et al., 2008). The results obtained using our method corroborate this information: 91.3% of probes in heterochromatin are methylated whereas only 49.5% of probes in euchromatin are methylated. In heterochromatin, 82% of probes have identical behaviour between wildtype and mutant whereas only 9.5% of probes are identical in euchromatin. Moreover 56% of methylated probes cover transposable elements or a 500 base-pair (bp) surrounding region.
The transition probabilities provide insights about the length of regions from each group through mean sojourn time. The average size of the binding sites is 14.3 probes (corresponding to 3289 bps) for the identical group, 4.5 probes (1035 bps) for the group with lost in mutant, 3.7 probes (851 bps) for the group enriched in mutant, and 7.7 probes (1771 bps) for the noise group. These calculations show that impoverished or enriched regions are three times smaller than regions with identical behaviour between wildtype and mutant. Moreover, the transposable elements are 2 to 3 times smaller in the euchromatin compared to the heterochromatin. This suggests that most of the methylation losses of the mutant occur in transposable elements from the euchromatin. The transposable element META1 (located between positions 5326458 and 5331580 on chromosome 4) is known to have a loss of methylation in the mutant. The regulatory region of META1 is located at the beginning of the transposable element with small RNAs which are involved in the methylation process. Our method declared the first half of the probes covering META1 (near the start position) in the group where methylation is lost. The other probes are declared identically methylated between the two samples. This example shows the advantage of the high resolution of the tiling array.

**Comparison with the models of Johannes et al. (2010)** As in Section 2.1, Groups 1 and 2 refer respectively to ‘noise’ and ‘identical’ probes, whereas Groups 3 and 4 correspond to differentially enriched probes. Johannes et al. (2010) proposed two mixture models of 4 bidimensional Gaussian distributions with constraints on the mean parameters. The first model is a full-switching model (Model 2) where the component means are constrained as follows: \( \tilde{\mu}_1 = (\mu_1, \mu_1), \tilde{\mu}_2 = (\mu_2, \mu_2), \tilde{\mu}_3 = (\mu_2, \mu_1), \tilde{\mu}_4 = (\mu_1, \mu_2) \) and the covariance matrices of Groups 3 and 4 are equal (\( \Sigma_3 = \Sigma_4 \)). The flexible-switching model (so-called Model 3) is the full-switching model (Model 2) with less restrictive constraints on \( \tilde{\mu}_3 = (\mu_4, \mu_3) \) and \( \tilde{\mu}_4 = (\mu_3, \mu_4) \). We compared the HMM model with their models on the H3K9me2 dataset. Model 2 leads to a smaller proportion of differentially enriched regions (7.8% lost in mutant and 1.2% gain in mutant, see Figure 3) than the HMM (22% and 14% respectively). The transposable element META1 that is declared differentially enriched with our model (see above) is found in the identical group according to their Model 2. The classification of Model 3 seems to be unsuitable for probes with similar intensities between 8 and 10 where more probes are expected to be declared in the identical group (see Figure 3). By comparing the results of the HMM with the results given by our simplest model (mixture model, without dependence), we note that the good definition of the differentially enriched probes is a consequence of the constraints on the variance matrices we put in the model. In fact, the added advantage of HMM is only to provide blurred boundaries between groups.
(due to the dependence assumption). In conclusion, it seems that the independence assumption, the symmetrical constraints on the means, and the equal variances for the differentially enriched probes lead to a model which is too simple to analyze such data. These two models also do not fit well the transcriptional dataset defined in Section 3.2 (results not shown).

Figure 3: Classification comparison between the two models of Johannes et al. (2010) and the mixture model and HMM on H3K9me2 dataset. top left: full-switching model (Model 2), top right: flexible-switching model (Model 3), bottom left: mixture model, bottom right: HMM.

3.2 Transcriptional dataset

We now study the gene differential expression between the leaf and the seed 10 days after pollination of the plant Arabidopsis thaliana. First we compare the 4 sub-models, second we present the results by gene, then we consider the detection of new transcribed regions. In the results, over-expressed (under-expressed) refers to probes with a higher (smaller) signal in the seed.
Table 1: Fit of the 4 models. $\mathcal{M}_1 = \text{mixture}$, $\mathcal{M}_2 = \text{HMM}$, $\mathcal{M}_3 = \text{mixture + annotation}$, $\mathcal{M}_4 = \text{HMM + annotation}$.

|                | $\mathcal{M}_1$ | $\mathcal{M}_2$ | $\mathcal{M}_3$ | $\mathcal{M}_4$ |
|----------------|-----------------|-----------------|-----------------|-----------------|
| number of parameters | 19              | 31              | 25              | 61              |
| $-2 \log$-likelihood | 406249          | 371309          | 373283          | 356617          |
| BIC            | 406469          | 371668          | 373573          | 357323          |
| ICL            | 436197          | 412706          | 399986          | 398272          |

3.2.1 Comparison of the 4 models

The model presented in Section 2.1 uses all available information and is referred to as model $\mathcal{M}_4$. For the annotation, $P = 3$ categories are considered: intergenic, intron and exon, and only exonic RNA is expected to be found in the sample. Model $\mathcal{M}_4$ can be simplified if either the structural annotation (Model $\mathcal{M}_2$) or spatial dependence between probes (Model $\mathcal{M}_3$) is not taken into account. Model $\mathcal{M}_1$ is the simplest model with neither annotation nor spatial dependence. It comes down to an independent mixture model. The constraints on the variance matrices detailed in Section 2.2 are kept in all models. Table 1 presents the fit of the four models for chromosome 4. First we note that combining the HMM with the annotation information leads to a real improvement in terms of likelihood. This can be seen when comparing models $\mathcal{M}_1$ and $\mathcal{M}_3$ or models $\mathcal{M}_2$ and $\mathcal{M}_4$. In both cases, the best BIC is obtained when adding the annotation in the model. The full model $\mathcal{M}_4$ achieves the best BIC criterion, suggesting that all available information should be taken into account. Biernacki, Celeux, and Govaert (2000) proposed an alternative selection criterion named ICL dedicated to classification purposes. In contrast to BIC which aims at finding the best fitting of the data distribution, the purpose of the ICL criterion is to assess the number of mixture components that leads to the most reliable clustering (with small entropy of the posterior distribution). It is a penalised criterion based on the integrated likelihood which corresponds to the BIC penalised by the entropy. ICL has been established in the independent mixture context but Celeux and Durand (2008) have shown in a simulation study that it seems to have the same behaviour in the HMM context. According to ICL, model $\mathcal{M}_4$ is also chosen, so we use model $\mathcal{M}_4$ to compare the two transcriptomic samples.

As expected, intergenic probes mostly belong to the noise group (84%) and few belong to expressed groups: 9% in the under-expressed group and 6% in the over-expressed. Intronic probes display a similar, although different, repartition: 60% noise, 7% identical, 24% under-expressed and 9% over-expressed (cf Section 3.2.3 for discussion about expressed probes in intergenic and intronic regions). As expected, most exonic probes (78%) belong to the expressed groups: 41% identi-
cal, 23% under-expressed and 14% over-expressed. The transition matrices for the intronic and intergenic categories are very similar (not shown): whatever the status of probe $t$, probe $t+1$ has a 70% to 95% chance of being noise. This is different for the exonic probes where the transition matrix has high probabilities on the diagonal meaning that probe $t+1$ has high probability (80% to 90%) of having the same status as probe $t$. All these results seem to be coherent with what is expected for transcriptomic data.

### 3.2.2 Gene classification

We now consider the classification of each gene. To this end, we compute the posterior probability $Q_{gk,X}$ defined in Equation (1). The advantage of our classification criterion is to offer the possibility not to classify the gene if it is too heterogeneous in terms of probe status covering the gene. In fact we propose to classify the genes via a two-step procedure. The probability for a gene to be homogeneous whatever the status is $\sum_k Q_{gk,X}$. We first verify whether the gene has homogeneous status by considering a ratio similar to a Bayes factor: $\sum_k Q_{gk,X} / \sum_k Q_{gk}$, where $Q_{gk} = P(\forall t \in g, Z_t = k | C)$ is the non-conditional version of $Q_{gk,X}$, which is for a gene

$$Q_{gk} = m_k^E \times (\pi_k^E)^{\sum Q_q=1 (i_q-e_q)-1} \times \prod_{q=1}^{Q-1} [((\pi_q)_{e_q+1-i_q})_{kk}],$$

where superscripts $E$ and $I$ refer to the exonic and intronic categories, respectively.

As its computation involves a product of probabilities with as many terms as the number of exonic probes in the gene, $Q_{gk,X}$ goes to zero for long genes. The ratio with $Q_{gk}$ does not correct this effect; therefore we apply an additional linear correction on the log-ratio with respect to the length of exons and the number of exons in the gene. We define this corrected log-ratio as a unistatus value which is a tool for decision support. If the homogeneous assumption seems verified, the second step is to calculate the conditional posterior probability $Q_{gk,X} / \sum_l Q_{gl,X}$ to assign the gene to the group $k$ for which this posterior probability is the highest.

We found 80% of genes which have a unistatus value higher than 0 (corresponding to 22528 genes). Among these 22528 genes, 11900 are declared identically expressed in the seed and in the leaf, 3632 are declared under-expressed in the seed and 2667 are declared over-expressed in the seed. It is difficult to interpret biologically these results given the fact that the functional annotation is still unclear. Available tools are databases such as Genevestigator (Zimmermann, Hirsch-Hoffmann, Hennig, and Gruissem, 2004) which makes possible the visualization of gene expression across thousands of experimental conditions through data from Affymetrix microarrays. To illustrate how our results confirm the actual knowledge
in biology, we identified 96 genes linked to seed using Flagdb++ which summarizes all available information on annotation. Among them, 70 have a probe on Affymetrix microarray and are so represented in Genevestigator. Only 8 genes are known for sure to be specifically expressed in the mature silique which is the experimental condition under study hybridized on the tiling array. For the other 62 genes, their expression is not located specifically in the mature silique but also in other development stage of seed. Among these 8 genes, 7 have a unistatus value higher than 0 and are declared over-expressed in the seed with our calculation. For the others, the expression is not clearly located in seed, which makes it difficult the comparison.

### 3.2.3 Detection of new transcripts

Although our model is built for the comparison of two samples, it also allows the detection of previously unknown transcription sites thanks to the high resolution of the tiling array. To this aim, the model without annotation seems more suitable, since we are bringing it into question. A lot of regions with expressed probes are found in intergenic regions: 1328 small regions with 2 or 3 consecutive expressed probes, 185 regions with 4 or 5 consecutive expressed probes and 90 regions with more than 5 consecutive probes (including 25 regions with more than 10 consecutive probes). For the 90 regions with more than 5 consecutive probes, we check with other annotation information such as Expressed Sequence TAG (EST) or genes predicted by the Eugene software (Schiex, Moisan, and Rouzé, 2001) which are not yet in the official TAIR annotation. We found 39 regions matching with annotation like small RNA, rRNA, tRNA, including 12 regions corresponding to a coding sequence defined in Eugene and 10 corresponding to transcriptional units recently annotated due to the presence of EST. Figure 4 (from FLAGdb++ (Samson et al., 2004)) shows examples of results for two annotated genes and also for two expressed regions which correspond to EST and Eugene genes. Moreover the obtained results show many other interesting things, such as surprisingly many transcriptions in the introns in 5’UTR (40% of intronic probes declared expressed in Section 3.2.1). This seems to be consistent with a recent article of Cenik, Derti, Mellor, Berriz, and Roth (2010) assuming a functional role of 5’UTR short introns.

### 3.3 Simulation study

We performed a simulation study to compare our approach with 3 existing methods: ChIPO Tile (Buck, Nobel, and Lieb, 2005), the method of Johannes et al. (2010) and the one of Seifert et al. (2009). As there is no annotation in the simulation datasets,
Figure 4: Presentation of results in Flagdb++. Circles represent EST, blue arrows are official TAIR annotation genes and violet arrows represent Eugene genes. The small rectangles are probes colored according to their status: grey if not expressed, black if identically expressed and red or green if differentially expressed. On the first 2 lines, there are 2 expressed genes covered by a majority of probes with the same status except the intronic probes which are reasonably declared as not expressed. The last 2 lines present expressed probes where there are no official genes but the signal coincides with EST and/or Eugene genes.

we applied our model $\mathcal{M}_2$ corresponding to the simple HMM with constraints on the variance matrices.

3.3.1 Design

We generated two datasets of size $n = 90\,000$ for which we are interested in retrieving the differentially expressed probes. The first two datasets are simulated with a latent variable $Z$ being a first order Markov chain taking its values in \{1, ..., 4\}. The transition matrix $\pi$ and the stationary distribution $m$ have been adjusted on real datasets described below. To overcome the Gaussian assumption, the observations $X$ were sampled according to an empirical distribution in each of the four groups of a real dataset. More precisely, the observations are sampled from real datasets in order to be similar to realistic tiling array data. The resampling is done using the posterior probabilities as weight for each probe. Two real datasets have been chosen with different expected proportions of differentially expressed probes. The first dataset, presented in Section 3.1, concerns the study of a histone mark (H3K9me2) in Arabidopsis thaliana for a wildtype and a mutant (polIV). In this dataset, about 30% of probes are expected to be differentially expressed. The second dataset is a publicly available ChIP-chip dataset coming from Penterman, Zilberman, Huh, Ballinger, Henikoff, and R.L. (2007) which compares the methylation profile of a
wild-type *Arabidopsis* plant to that of a triple mutant. It leads to low proportions of differentially expressed probes.

We analysed the synthetic datasets with Model $\mathcal{M}_2$ and with three other methods.

- ChIPOTle is a method dedicated to peak-finding in classical ChIP-chip experiments. Therefore it only provides two populations. It uses a sliding window approach based on the log-ratio. The window size and step parameters have to be tuned. With default parameters, ChIPOTle detects only one peak. We put window=200 and step=50, which seems to be a good combination giving a reasonable number of peaks for each simulated dataset. We used the absolute value of the log-ratio to mimic situations usually analysed.

- Seifert et al. (2009) proposed a three-state HMM that models the log-ratios of the two intensities. It requires the incorporation of *a priori* knowledge using prior distributions. The choice of the priors is not easy and those given by default do not provide three populations. Hence we modified them. We put startDistribution = (0.1,0.7,0.2), means = (-1,0,0,1), stds = (0.3,1,0.5), scaleOfAprioriMeans = (0.1,1,1.75) and shapeOfStandardDeviations = (20;1;100).

- Johannes et al. (2010) proposed two mixture models of four bidimensional Gaussian distributions with constraints on the mean parameters for the simultaneous analysis of two samples. These two models are described more precisely in Section 3.1.

In the simulated datasets, we are looking for the four groups defined in Section 1. The classification is done with the MAP rule for our model $\mathcal{M}_2$ and the model of Johannes et al. (2010). However, ChIPOTle and the method of Seifert et al. (2009) do not provide four groups: they merge the noise and the identical groups. The method of Seifert et al. (2009) applies the Viterbi algorithm to determine the most probable state sequence and give a classification of probes into three groups. For ChIPOTle, the differentially expressed probes are deduced from the detected peaks and there are only two groups. To compare our method with the ones of Seifert et al. (2009) and ChIPOTle, we summed the posterior probabilities of the noise and identical groups to obtain a classification into three groups, and also those of the over-expressed and under-expressed groups to obtain a classification into two groups.

The methods are compared using the classification results in terms of sensitivity, specificity and False Discovery Rate (FDR) for a given group $k$. The sensitivity is defined as: $\frac{T_P}{T_P + F_N}$, the specificity is defined as $\frac{TN}{TN + FP}$, and the FDR is defined as $\frac{FP}{TP + FP}$, where:
We hence focus on the probes assigned as differentially expressed. Both sensitivity and specificity are expected to be large whereas FDR is hoped to be small.

### 3.3.2 Results

The results are presented in Tables 2 and 3. The peaks detected with the ChIPOTle method only represent between 31% and 35% of differentially expressed probes for the two datasets. The flexible-switching model (Model 3) of Johannes et al. (2010) provides better results than the full-switching model (Model 2); therefore we focus on Model 3. In the first dataset, Model 3 of Johannes et al. (2010) and the method of Seifert et al. (2009) have similar behavior. They respectively find 85% or 82% of under-expressed probes and 63% or 72% of over-expressed probes. In the second dataset where few probes are differentially expressed, the two methods behave differently. The method of Seifert et al. (2009) has difficulty finding the over-expressed group (only 55% of detected probes). On the contrary, Model 3 of Johannes et al. (2010) finds 100% of differentially expressed probes but detects a lot of false positives in return (FDR=99%). About 40 000 probes with similar intensities between 8 and 10 are declared differentially expressed whereas they are expected to be declared in the identical or in the noise group. In the two simulated datasets, our model $M_2$ identifies more than 83% of differentially expressed probes, with fewer false positives (FDR between 2 and 13%). Among the four methods, Model $M_2$ provided the best triplets of sensitivity, specificity and FDR whatever the proportion of differentially expressed probes in the dataset.

In order to overcome the Markovian dependency assumption, we also simulated two other datasets where the hidden path $Z$ was sampled in a 4-state jump process with Markovian between-state transitions and Negative Binomial sojourn times (instead of Geometric). Similar results are obtained. The triplets of sensitivity, specificity and FDR obtained with Model $M_2$ are (92,99,4) and (91,99,5) for the two differentially expressed groups in the first dataset respectively, and (99,100,3) and (85,100,16) in the second dataset. This ensures that our model is not too dependent on the Markovian assumption.

\[
\begin{align*}
TP_k &= \sum_t I(\hat{Z}_t=k) I(Z_t=k) \\
FN_k &= \sum_t I(\hat{Z}_t=k) I(Z_t\neq k) \\
FP_k &= \sum_t I(\hat{Z}_t\neq k) I(Z_t=k) \\
TN_k &= \sum_t I(\hat{Z}_t\neq k) I(Z_t\neq k)
\end{align*}
\]
Table 2: Dataset derived from H3K9me2 with a large proportion of differentially expressed probes. The notation corresponds to the order sensitivity, specificity, FDR in %.

|                  | noise | identical | under-expressed | over-expressed |
|------------------|-------|-----------|-----------------|---------------|
| Z                | 38135 | 19527     | 19413           | 12925         |
| ChIPOTle         | 99, 35, 27 | 35, 99, 6 |                 |               |
| $\mathcal{M}_2$ with 2 groups | 97, 94, 4 | 94, 94, 5 |                 |               |
| Seifert et al.   | 92, 79, 11 | 82, 95, 18 | 72, 98, 13 |               |
| $\mathcal{M}_2$ with 3 groups | 98, 93, 4 | 92, 99, 6 | 90, 99, 6 |               |
| Johannes et al. M3 | 94, 90, 12 | 82, 100, 0.1 | 63, 99, 11 |               |
| TAHMMAnnot $\mathcal{M}_2$ | 96, 96, 5 | 99, 100, 1 | 92, 98, 6 | 90, 99, 6 |

Table 3: Dataset derived from Penterman et al. (2007) with a small proportion of differentially expressed probes. The notation corresponds to the order sensitivity, specificity, FDR, in %.

|                  | noise | identical | under-expressed | over-expressed |
|------------------|-------|-----------|-----------------|---------------|
| Z                | 45300 | 43401     | 782             | 517           |
| ChIPOTle         | 100, 31, 1 | 31, 100, 0 |                 |               |
| $\mathcal{M}_2$ with 2 groups | 100, 91, 0.1 | 91, 100, 6 |                 |               |
| Seifert et al.   | 98, 99, 0  | 100, 99, 48 | 55, 98, 84 |               |
| $\mathcal{M}_2$ with 3 groups | 100, 91, 0.1 | 97, 100, 2 | 82, 100, 12 |               |
| Johannes et al. M3 | 23, 96, 13 | 74, 79, 12 | 100, 100, 31 | 100, 56, 99 |
| TAHMMAnnot $\mathcal{M}_2$ | 85, 84, 16 | 83, 85, 16 | 97, 100, 2 | 83, 100, 13 |

4 Discussion

Tiling array is a powerful technology which requires adapted statistical methods to deal with the large quantity and the variability of data. We focus on the comparison of two samples from transcriptomic or ChIP-chip experiments and also on the detection of transcribed regions by directly modeling the joint distribution of the two sample intensities. We consider all the available information from the probes: the intensity of the two signals, the dependence between neighboring probes and the structural annotation. The annotation knowledge is very useful information with the aim of classification because of the intrinsic difference between exonic or intergenic probes. This method can be used for ChIP-chip or transcriptomic data whenever there are two conditions to compare. Both one-color and two-color tiling arrays can be analysed. This method could be adapted for the comparison of $d > 2$ samples. The number of parameters would be linear in $d$ and quadratic in $K$ but
still very small compared to the number of observations. However, the definition of generic geometrical constraints in dimension $d$ is not straightforward and would need to be adapted to the experimental design. A simulation study highlighted the performance of our approach and applications on *Arabidopsis thaliana* tiling array show the ability of the model to interpret the data and provide a new insight on gene expression or gene expression control as well as new biological hypotheses. *Arabidopsis thaliana* is a model plant with a very well-known genome annotation but for many organisms the annotation is not available or unreliable. That is why the sub-models are also useful. The model without annotation allows us not to be limited by the quality of the available annotation and this model is also useful to detect new genes to improve the current official annotation.

This work also raises the question of classification. The results are given by probe and by region. We compute a posterior probability by region and we propose a procedure for region classification. The most common regions are the genes which are non-connected regions, but any other region can be defined. It would be interesting to control the False Discovery Rate, *i.e.* the expected proportion of misclassifications, in the case of having 4 groups and under the dependence hypothesis, and also for the results given by region. Moreover it is clear that the assumption of a normal distribution for the emission distribution may not be realistic. We are now working on a model where the emission distributions are themselves mixtures, in order to get more flexible distributions and therefore a better fit to real data.

### Appendix: Computation of the estimates of $D$ and $\Lambda$.

Recall that $X_t = (X_{t1}, X_{t2})$ are the log-intensities for both samples, $t$ varies from 1 to $n$, where $n$ is the total number of observations.

Let $\bar{X}_k = \frac{\sum_{t=1}^{n} \tau_{tk} X_t}{n_k}$, where $n_k = \sum_{t=1}^{n} \tau_{tk}$.

Let $W_k = \sum_{t=1}^{n} \tau_{tk} (X_t - \bar{X}_k) (X_t - \bar{X}_k)'$ be a matrix like $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$ and $\Lambda_k = \begin{pmatrix} u_{1k} & 0 \\ 0 & u_{2k} \end{pmatrix}$, with $u_{1k} > u_{2k}$, for $k = 1, \ldots, 4$. The maximum likelihood estimator of the orientation matrix $D$ identical for the first two components with the same orientation is in the form of $\begin{pmatrix} \hat{d} & -\sqrt{1 - \hat{d}^2} \\ \sqrt{1 - \hat{d}^2} & \hat{d} \end{pmatrix}$, where $\hat{d}$ is the minimum of the function:

$$f(\hat{d}) = \sum_{k=1}^{2} \left\{ \frac{d^2 w_{1k} + 2w_{2k} \sqrt{1 - \hat{d}^2} + w_{4k}(1 - \hat{d}^2)}{u_{1k}} + \frac{d^2 w_{4k} + 2w_{2k} \sqrt{1 - \hat{d}^2} + w_{1k}(1 - \hat{d}^2)}{u_{2}} \right\}$$
The estimator of $\hat{d}$ is defined by:
\[
\hat{d}^2 - \frac{1}{2} = \pm \frac{N_{1,4}}{2 \left[ \{N_{1,4}\}^2 + 4 \{N_2\}^2 \right]^{1/2}}, \text{ with } \hat{d} > 0,
\]
where
\[
N_{1,4} = \sum_{k=1}^2 (w_{1k} - w_{4k})(u_2 - u_{1k})/u_1ku_2 \quad \text{and} \quad N_2 = \sum_{k=1}^2 (w_{2k})(u_2 - u_{1k})/u_1ku_2.
\]

Let $B_k$ be a matrix defined by $B_k = D_k'W_kD_k$ like
\[
\begin{pmatrix}
  b_{1k} & b_{3k} \\
  b_{4k} & b_{2k}
\end{pmatrix}
\]

The maximum likelihood estimator of $\Lambda_k$ is in the form of
\[
\begin{pmatrix}
  \hat{u}_{1k} & 0 \\
  0 & \hat{u}_2
\end{pmatrix}, \quad \text{where}
\]
\[
\begin{align*}
\hat{u}_{1k} &= b_{1k}/n_k \\
\hat{u}_2 &= \frac{4}{\kappa=1} b_{2k}/n.
\end{align*}
\]

References

Banfield, J. and A. Raftery (1993): “Model-based gaussian and non-gaussian clustering.” *Biometrics*, 49, 803–821.

Baum, L. (1972): “An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes.” *Inequalities*, 3, 1–8.

Bernatavichute, Y., X. Zhang, S. Cokus, M. Pellegrini, and S. Jacobsen (2008): “Genome-wide association of histone h3 lysine nine methylation with chg dna methylation in arabidopsis thaliana.” *PLoS ONE*, 3(9):e3156.

Biernacki, C., G. Celeux, A. Echenim, G. Govaert, and F. Langrognet (2007): “Le logiciel mixmod d’analyse de mélange pour la classification et l’analyse discriminante.” *La Revue de Modulad*, 35, 25–44.

Biernacki, C., G. Celeux, and G. Govaert (2000): “Assessing a mixture model for clustering with the integrated completed likelihood.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Buck, M. and J. Lieb (2004): “Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments,” *Genomics*, 83, 349–360.

Buck, M., A. Nobel, and J. Lieb (2005): “Chipotle: a user-friendly tool for the analysis of chip-chip data.” *Genome Biol.*, 6(11).
Cawley, S., S. Bekiranov, H. Ng, and et al. (2004): “Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas,” Cell, 116, 499–509.

Celeux, G. and J. Durand (2008): “Selecting hidden markov model state number with cross-validated likelihood,” Computational Statistics, 23(4), 541–564.

Celeux, G. and G. Govaert (1995): “Gaussian parsimonious clustering models,” Pattern Recognition, 28, 781–793.

Cenik, C., A. Derti, J. Mellor, G. Berriz, and F. Roth (2010): “Genome-wide functional analysis of human 5’untranslated region introns.” Genome Biology, 11:R29.

Du, J., J. Rozowsky, J. Korbel, Z. Zhang, T. Royce, M. Schultz, M. Snyder, and M. Gerstein (2006): “A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chip-chip experiments: Systematically incorporating validated biological knowledge.” Bioinformatics, 22(24), 3016–3024.

Fraley, C. and A. Raftery (2006, revised 2010): “Mclust version 3 for r: Normal mixture modeling and model-based clustering.” Technical Report no. 504, Department of Statistics, University of Washington.

Fridlyand, J., A. Snijders, D. Pinkel, D. Albertson, and A. Jain (2004): “Hidden markov models approach to the analysis of array cgh data.” J Multivariate Analysis, 90, 132–153.

Ghosh, S., H. Hirsch, E. Sekinger, P. Kapranov, K. Struhl, and T. Gingeras (2007): “Differential analysis for high density tiling microarray data.” BMC Bioinformatics, 8:359.

Halasz, G., M. van Batenburg, J. Perusse, S. Hua, X. Lu, K. White, and H. Busse-maker (2006): “Detecting transcriptionally active regions using genomic tiling arrays.” Genome Biology, 7.

Hanada, K., X. Zhang, J. Borevitz, W. Li, and S. Shiu (2007): “A large number of novel coding small open reading frames in the intergenic regions of the arabidopsis thaliana genome are transcribed and/or under purifying selection.” Genome Research, 17, 632–640.

He, K., X. Li, J. Zhou, X. Deng, H. Zhao, and J. Luo (2009): “Ntap: for nimblegen tiling array chip-chip data analysis.” Bioinformatics, 25, 1838–1840.

Huber, W., J. Toedling, and L. Steinmetz (2006): “Transcript mapping with high-density oligonucleotide tiling arrays.” Bioinformatics, 22(6), 1963–1970.

Humbug, P., D. Bulger, and G. Stone (2008): “Parameter estimation for robust hmm analysis of chip-chip data.” BMC Bioinformatics, 9:343.

Hupé, P., N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot (2004): “Analysis of array cgh data: from signal ratio to gain and loss of dna regions,” Bioinformatics, 20(18), 3413–3422.

DOI: 10.2202/1544-6115.1692
Ji, H. and W. Wong (2005): “Tilemap: create chromosomal map of tiling array hybridizations.” *Bioinformatics*, 21, 3629–3636.

Johannes, F., R. Wardenaar, M. Colomé-Tatché, and *et al.* (2010): “Comparing genome-wide chromatin profiles using chip-chip or chip-seq.” *Bioinformatics*, 26, 1000–1006.

Johnson, W., W. Li, C. Meyer, R. Gottardo, J. Carroll, M. Brown, and X. Liu (2006): “Model-based analysis of tiling-arrays for chip-chip,” *PNAS*, 103, 12457–12462.

Keles, S., M. van de Laan, S. Dudoit, and S. Cawley (2004): “Multiple testing methods for chip-chip high density oligonucleotide array data.” *University of California Berkeley Division of Biostatistics Working Paper Series*, 147.

Li, W., A. Meyer, and X. Liu (2005): “A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences.” *Bioinformatics*, 211, 274–282.

Martin-Magniette, M., T. Mary-Huard, C. Bérard, and S. Robin (2008): “Chipmix: mixture model of regressions for two-color chip-chip analysis,” *Bioinformatics*, 24:i181-i186.

Mary-Huard, T., F. Picard, and S. Robin (2006): *Introduction to Statistical Methods for Microarray Data Analysis, Mathematical and Computational Methods in Biology*.

Mockler, T., S. Chan, A. Sundaresan, and *et al.* (2005): “Applications of dna tiling arrays for whole genome analysis.” *Genomics*, 85, 1–15.

Munch, K., P. Gardner, P. Arctander, and A. Krogh (2006): “A hidden markov model approach for determining expression from genomic tiling micro arrays.” *BMC Bioinformatics*, 7:239.

Nicolas, P., A. Leduc, S. Robin, S. Rasmussen, H. Jarmer, and P. Bessières (2009): “Transcriptional landscape estimation from tiling array data using a model of signal shift and drift.” *Bioinformatics*, 25(18), 2341–2347.

Oshlack, A., M. Robinson, and M. Young (2010): “From rna-seq reads to differential expression results.” *Genome Biology*, 11:220.

Penterman, J., D. Zilberman, J. Huh, T. Ballinger, S. Henikoff, and F. R.L. (2007): “Dna demethylation in the arabidopsis genome.” *PNAS*, 104, 6752–6757.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J. Daudin (2005): “A statistical approach for array cgh data analysis,” *BMC Bioinformatics*, 6:27.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, and Y. Zhai (1998): “High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays.” *Nat.Genet.*, 20, 207–211.

Rabiner, L. (1989): “A tutorial on hidden markov models and selected applications in speech recognition.” *Proc. IEEE*, 77, 257–286.
Samson, F., V. Brunaud, S. Duchêne, Y. De Oliveira, M. Caboche, A. Lecharny, and S. Aubourg (2004): “Flagdb++: a database for the functional analysis of the arabidopsis genome.” *Nucleic Acids Res.*, Jan 1;32 Database issue: D347-50.

Schiex, T., A. Moisan, and P. Rouzé (2001): “Eugene: An eucaryotic gene finder that combines several sources of evidence.” *Computational Biology*, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, 111–125.

Seifert, M., A. Banaei, J. Keilwagen, M. Mette, A. Houben, F. Roudier, V. Colot, I. Grosse, and M. Strickert (2009): “Array-based genome comparison of arabidopsis ecotypes using hidden markov models.” in *Biosignals*.

Snijders, A., N. Nowak, R. Segraves, S. Blackwood, and *et al.* (2001): “Assembly of microarrays for genome-wide measurement of dna copy number.” *Nat.Genet.*, 29, 263–264.

Yamada, K., J. Lim, J. Dale, and *et al.* (2003): “Empirical analysis of transcriptional activity in the arabidopsis genome.” *Science*, 302, 842–846.

Zeller, G., S. Henz, S. Laubinger, D. Weigel, and G. Rätsch (2008): “Transcript normalization and segmentation of tiling array data.” *Pacific Symposium on Bio-computing*, 12, 527–538.

Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem (2004): “Genevestigator. arabidopsis microarray database and analysis toolbox.” *Plant Physiology*, 136(1), 2621–2632.