2-3-2010

WAVELET BASED FUNCTIONAL MODELS FOR TRANSCRIPTOME ANALYSIS WITH TILING ARRAYS

Lieven Clement
Ghent University, Belgium, lieven.clement@gmail.com

Kristof DeBeuf
Ghent University, Belgium

Ciprian Crainiceanu
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Olivier Thas
Ghent University, Belgium

Marnik Vuylsteke
Ghent University, Belgium

See next page for additional authors

Authors
Lieven Clement, Kristof DeBeuf, Ciprian Crainiceanu, Olivier Thas, Marnik Vuylsteke, and Rafael Irizarry

Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

WAVELET BASED FUNCTIONAL MODELS FOR TRANSCRIPTOME ANALYSIS WITH TILING ARRAYS

By Lieven Clement[*], Kristof De Beuf[*], Ciprian Crainiceanu[‡], Olivier Thas[*], Marnik Vuylsteke[†], Rafael Irizarry[‡]

Ghent University[†], Johns Hopkins University[‡]*

For a better understanding of the biology of an organism a complete description is needed of all regions of the genome that are actively transcribed. Tiling arrays can be used for this purpose. Such arrays allow the discovery of novel transcripts and the assessment of differential expression between two or more experimental conditions such as genotype, treatment, tissue, etc. Much of the initial methodological efforts were designed for transcript discovery, while more recent developments also focus on differential expression. To our knowledge no methods for tiling arrays are described in the literature that can both assess transcript discovery and identify differentially expressed transcripts, simultaneously. The wavelet based functional model developed in this paper is designed to fill this methodological void. As opposed to existing methods, our statistical framework also permits a natural integration of preprocessing into the standard statistical analysis flow of tiling array data. We use Johnson transformations, which are based on cumulants, for computing false discovery rates (FDRs) and Bayesian credibility intervals for the estimates of the effect functions within the data space. A case study illustrates that our model is well suited for a simultaneous assessment of transcript discovery and differential expression, while remaining competitive with methods that perform only one of these tasks.

1. Introduction. In the last decade the genomes of many organisms have been entirely sequenced (e.g. Kim *et al.*, 2006). A detailed description of all genomic regions that are actively transcribed is needed for enhancing the knowledge of the organisms functioning and the regulation of its transcriptional networks. The complete set of these RNA transcripts is referred to as the transcriptome. It seems almost impossible to derive the entire transcriptome from the complete genome sequence alone. In addition, the transcriptome and the transcription level can vary considerably between different tissues and they typically depend on external environmental conditions (e.g. Halasz *et al.*, 2006). Thus, expression profiling has to be assessed experimentally. Genomic tiling arrays can provide an unbiased quantification

Keywords and phrases: tiling microarray, wavelets, adaptive regularization, transcript discovery, differential expression, genomics, *Arabidopsis thaliana*

of transcriptional activity (e.g. Bertone *et al.*, 2004). They are high-density microarrays that are designed without prior consultation of existing gene annotation. Their probes are roughly equally spaced along the genomic coordinate and span exonic, intronic and intergenic regions of the genome. Tiling array experiments thus enable the discovery of transcribed sequences and regulatory elements, which is not possible with classical microarrays that only contain probes in annotated regions.

1.1. *Motivation.* Methods for transcriptome analysis with tiling arrays initially focused on transcript discovery. They are often based on sliding windows and a thresholding criterion to identify transcriptionally active regions (TARs) (e.g. Bertone *et al.*, 2004; Kampa *et al.*, 2004; Royce *et al.*, 2005). Huber, Toedling and Steinmetz (2006), however, presented a structural change model (SCM) to provide the segmentation. After segmentation they used a threshold for partitioning the genome into transcribed and non-transcribed regions. Recent contributions also focus on the detection of differentially expressed TARs (e.g. Naouar *et al.*, 2009), enabling the biologists to identify genes that are differentially affected in their expression by two or more experimental conditions. A common approach for assessing differential expression with tiling arrays is to group probes in probesets by mapping tiling probes to a gene model. The construction of such probesets allows the detection of differentially expressed genes by using standard techniques for analysing the data of classical microarray experiments. This approach, however, does not allow the detection of differentially expressed transcripts in unannotated regions, because the unannotated probes cannot be grouped into probesets. Therefore, it does not use expression data of tiling array experiments to its full potential. To our knowledge no methods for tiling arrays are described in the literature that can both assess transcript discovery and identify differentially expressed transcripts, simultaneously. The development of such a method is our main goal. A second aim of this paper is to provide a statistical framework that permits the integration of preprocessing into the standard statistical analysis flow of tiling array data. The existing papers on tiling arrays typically require some preprocessing steps before applying a segmentation algorithm (e.g. Kampa *et al.*, 2004; Bertone *et al.*, 2004; Huber, Toedling and Steinmetz, 2006; Laubinger *et al.*, 2008; Naouar *et al.*, 2009). Huber, Toedling and Steinmetz (2006), for instance, introduced 1) a DNA reference based normalization procedure for background correction that accounts for the effect of probe affinities and 2) a between-array normalization. Although such preprocessing steps can have a large effect on the quality of the downstream analysis, it is common practice to ignore their

impact on the stochastic properties of the final statistical summaries. Within our functional data analysis framework, however, the preprocessing, transcript discovery and the identification of differentially expressed transcripts can be naturally integrated within the main analysis.

1.2. *Arabidopsis thaliana tiling array study.* The methods developed in this paper are illustrated on a dataset from a study that was conducted at the Flemish Institute of Biotechnology (VIB) Department of Plant Systems Biology, Ghent, Belgium. The study fits in the scope of a larger project that aims at increasing the knowledge of the role of E2F transcription factors in the regulation of the plant cell cycle and plant growth (Naouar *et al.*, 2009). E2Fs are conserved regulators of S phase-specific genes (Blais and Dynlacht, 2007). The *Arabidopsis thaliana* genome encodes three E2Fs (E2Fa, E2Fb and E2Fc; De Veylder, Beeckman and Inze, 2007), which are active in association with the dimerization partners DPa or DPb. A complete understanding of the role of the different E2F isoforms requires the comprehensive identification of their target genes. Within this context, Col-0 plants were used that are ectopically overproducing the heterodimer E2Fa-DPa (Naouar *et al.*, 2009). In the remainder of the paper these plants are referred to as the E2F-DPa$_{\mathrm{OE}}$ plants. Expression profiling was performed with Affymetrix GeneChip Arabidopsis Tiling 1.0R arrays. It is a single array with over 3.2 million perfect match and mismatch (PM/MM) probe pairs that are tiled across the complete non-repetitive *Arabidopsis thaliana* genome. Each probe consists of 25 bases. The center positions of the probes correspond to regions in the genome that are spaced on average 35 bases apart. Hence, the entire genome is tiled with non-overlapping probes with an average gap-width of 10 bases (Naouar *et al.*, 2009). In our study the 3 biological replicates for both wild type (WT) and E2F-DPa$_{\mathrm{OE}}$ strains are used; they correspond to the target preparation protocol number 3 (TPP3) in Naouar *et al.* (2009). The aim of the study was to quantify, compare and evaluate the expression and expression changes between wild type and E2F-DPa$_{\mathrm{OE}}$ plants. In addition wild type (WT; Col-0) genomic DNA was hybridized to a single tiling array for assessing the impact of the probe sensitivity.

Fig. 1 shows $\log_2$ transformed intensities obtained by the tiling array hybridizations as a function of the genomic coordinate. In the top panel of Fig. 1 the intensities from the DNA reference hybridization are depicted. All features should exhibit the same intensity, because the same copy number of genomic DNA is hybridized throughout the genomic coordinate. In practice, however, large differences in the measured intensities are observed. Although some of the variation can be explained by stochastic noise, the major part of

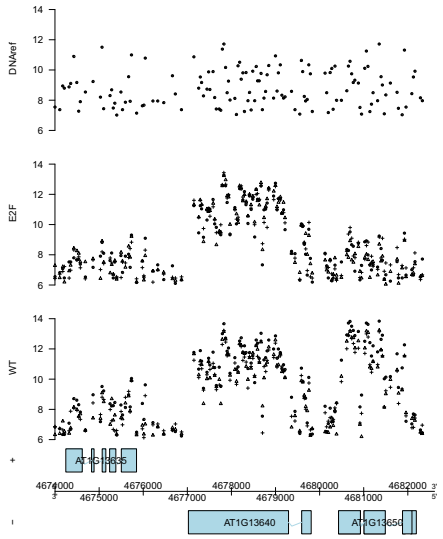

FIG 1. *Along-chromosome plot of array intensities from a reference DNA sample (DNAref), E2F-DPa$_{OE}$ plants (E2F-DPa$_{OE}$) and wild type plants (WT). The different replicates for WT and E2F are indicated by* •, + *and* △.

the variation is due to differences in probe affinity (e.g. Wu *et al.*, 2004). A similar pattern can be observed in the signals from the E2F-DPa$_{\mathrm{OE}}$ and WT plant hybridizations that are represented in the middle and bottom panel of Fig. 1, respectively. In contrast to classical microarrays, tiling arrays contain probes at intronic and intergenic regions. We also expect sudden jumps in the measured intensities due to differences in transcriptional activity among exonic regions, and, between exonic, intronic and intergenic regions.

1.3. *Outline.* Given the spiky and discontinuous nature of the data, wavelet based denoising seems very attractive. The use of wavelets allows an efficient regularization of the fixed and random effect functions without losing the ability to model local features. Morris and Carroll (2006) generalized wavelet regression to the case of multiple functions by introducing a Bayesian wavelet-based functional mixed model framework. Similar to Morris and Carroll (2006) we use a wavelet based functional model, but we develop a different approach for the estimation and regularization of the fixed effect functions. In particular, fast algorithms are developed for estimation and inference.

The paper is organized as follows. In Section 2, we introduce the wavelet

based functional model for transcriptome analysis and a parameter estimation procedure. We continue with an empirical Bayes false discovery rate procedure for identifying both expressed and differentially expressed regions in Section 3. In Section 4, our method is applied to the *Arabidopsis E2F-DPa*$_{OE}$ tiling expression data and its performance is compared to existing methods. Finally, we present conclusions and possible directions for further research in Section 5.

2. Wavelet based functional models for transcriptome analysis. We first present a basic functional model that can assess transcript discovery and differential expression simultaneously. Next, the basic model is extended to incorporate the information of DNA reference hybridizations. This extension enables us to account for differences in probe affinities by incorporating a kind of DNA reference normalization within the main analysis. The model is then transformed to the wavelet space in which an efficient regularization of the functional effects is accomplished.

2.1. *Functional Model.* Suppose that N_1 and N_2 tiling arrays are collected for two distinct experimental conditions C_1 and C_2, respectively. The expression functions $Y_i(t)$ are evaluated on an equally spaced grid, say $\boldsymbol{t} = (1, \ldots, T)$, corresponding to the genomic locations of the probes. We consider the functional model

$$Y_i(t) = \beta_1(t) + X_{1,i}\beta_2(t) + E_i(t), \tag{2.1}$$

with $i = 1, \ldots, (N_1+N_2)$, $Y_i(t)$ the $\log_2$ transformed probe intensity of probe t on array i, $X_{1,i}$ a dummy variable which is 1 for C_1 and -1 for C_2, $E_i(t)$ the zero-mean error term and the functions $\beta_1(t)$ and $\beta_2(t)$ are referred to as the mean and difference function, respectively. Note that for balanced designs the use of the (-1,1) coding allows an orthogonal estimation of both effect functions. After fitting the model, the estimated mean function, say $\hat{\beta}_1(t)$, can be used for transcript discovery. In particular, a segmentation can be performed by assessing in which genomic regions the mean intensity $\beta_1(t)$ exceeds a certain background level. The (-1,1) dummy coding implies that $2 \times \beta_2(t) = FC(t)$ enables inference on the $\log_2$-fold change between the two distinct experimental conditions. Model (2.1) will be referred to as the basic model. With the basic model, preprocessing might still be needed for background correction and normalization.

Suppose that another N_0 arrays are hybridized to a DNA reference, which provides us with empirical evidence of the probe affinities. The information of the DNA reference hybridizations can be easily incorporated in the model.

Consider

$$Y_i(t) = \beta_0(t) + X_{1,i}\beta_1(t) + X_{2,i}\beta_2(t) + E_i(t), \tag{2.2}$$

with $i = 1, \ldots, N$, $N = N_0 + N_1 + N_2$, $\beta_0(t)$ a function that is related to the probe affinities derived from the DNA reference hybridizations, $\beta_1(t)$ the mean function that will be used for transcript discovery, $X_{1,i}$ a dummy variable which is 1 for the C_1 and C_2 arrays, and 0 for the reference DNA arrays, $\beta_2(t)$ the difference function, $X_{2,i}$ a dummy variable which is 1 for C_1 , -1 for C_2 and 0 for the reference DNA array. By including $\beta_0(t)$ in the model, the mean function $\beta_1(t)$ gets the interpretation of a $\log_2$-fold change with respect to the average intensities of the DNA reference hybridizations. Hence, DNA reference normalization is done automatically during the parameter estimation for Model (2.2). For balanced designs of the C_1 and C_2 arrays the use of the (-1,1,0) coding for $X_{2,i}$ implies an estimation orthogonality between $\beta_2(t)$ and the other two functions. In the remainder of the paper we focus on the more generic Model (2.2). The derivations below also hold for the basic Model (2.1).

Model (2.2) can be written in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}. \tag{2.3}$$

Here $\boldsymbol{Y}$ is a $N \times T$ matrix whose rows contain the $\log_2$ transformed intensities of one array observed on $\boldsymbol{t}$. $\boldsymbol{X}$ is a $N \times p$ design matrix of the covariates. Each row of the $p \times T$ matrix $\boldsymbol{B}$ contains one of the effect functions evaluated in $\boldsymbol{t}$. The rows of $\boldsymbol{E}$ contain the error processes evaluated in $\boldsymbol{t}$, corresponding to each of the observed tiling arrays. They are assumed to be i.i.d. $N(0, \sigma^2)$.

2.2. *Wavelet based functional model.* Functional models are commonly estimated by using basis functions. Morris and Carroll (2006) used a wavelet basis. Wavelets are well suited to deal with irregular functional data that are characterized by a high number of local features. In this paper we use the discrete wavelet transform (DWT) for projecting the data onto the wavelet space. This projection can be written as the matrix product $\boldsymbol{D} = \boldsymbol{Y}\boldsymbol{W}^T$, where $\boldsymbol{W}$ is an orthogonal DWT matrix. The rows of the matrix $\boldsymbol{D}$ contain the wavelet coefficients for each of the observed curves and they are double indexed by the location $k = 1, \ldots, K_j$ within the wavelet scale $j = 0, \ldots, J$. In practice, the projection from the data space onto the wavelet space is performed by a more efficient pyramid-based algorithm (e.g. Hastie, Tibshirani and Friedman, 2001).

The wavelet transform allows us to rewrite the model within the wavelet space by back-multiplying both sides of Model (2.3) with the DWT matrix

$\boldsymbol{W}^T$, resulting in

$$\boldsymbol{D} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}^*. \tag{2.4}$$

Hence, $\boldsymbol{B}^*$and $\boldsymbol{E}^*$ are the matrices whose rows contain the wavelet coefficients corresponding to the effect functions and the errors, respectively. Because the DWT is a linear projection, $\boldsymbol{E}^*$ is multivariate normal with mean zero and covariance matrix $\boldsymbol{S}^* = \boldsymbol{W}\boldsymbol{W}^T\sigma^2$. Moreover, the orthonormality of $\boldsymbol{W}$ implies that $\boldsymbol{S}^*$ is the diagonal matrix $\boldsymbol{I}\sigma^2$.

The wavelet transform concentrates most of the structure of the signal in relatively few large wavelet coefficients while distributing white noise equally over all wavelet coefficients. Denoising can thus be done by thresholding the smallest wavelet coefficients or shrinking them towards zero. One often makes the distinction between hard and soft thresholding of the wavelet coefficients. Hard thresholding sets all the coefficients below the threshold to zero and leaves the remaining coefficients unchanged. Soft tresholding sets the coefficients below the threshold to zero, but shrinks the remaining coefficients towards zero. Most of the thresholding rules can be linked to a regularization process using a penalty function (Antoniadis, 2007). The use of wavelet shrinkage allows a discontinuity-preserving denoising and typically consists of three steps:

1. Compute the wavelet coefficients of the noisy signal.
2. Modify the coefficients according to a certain rule.
3. Backtransform the modified coefficients to obtain the denoised signal.

2.3. *Estimation procedure.* Let $\beta_m^*(j,k)$, $m = 0,1,2$, indicate the element of $\boldsymbol{B}^*$ corresponding to scale j and location k, let $N(\mu,\sigma^2)$ denote the density function of a normal distribution with mean μ and variance σ^2 and let $MVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}$. A standard procedure in Bayesian wavelet regression exists in adaptive regularisation of the fixed effect functions by imposing a mixture prior on $\beta_m^*(j,k)$ of the form

$$\pi_m(j)N\left(0,\tau_m^2(j)\right) + \{1-\pi_m(j)\}\,\delta(0),$$

with $0 \leq \pi_m(j) \leq 1$, and $\delta(0)$ the density function of a point mass at zero. However, on biological grounds we can incorporate the following assumptions: (1) Fluctuations related to differences in probe affinity are always present, which is guaranteed by setting $\pi_0(j) = 1$. (2) Differential expression can only occur for exons that are expressed, which implies that $\beta_1^*(j,k)$ and $\beta_2^*(j,k)$ are both non-zero in differentially expressed regions. Therefore

the prior on the $[\beta_m^*(j,k)]$ can be written as a mixture of multivariate normal distributions:

$$\begin{aligned}(2.5)\quad \begin{bmatrix} \beta_0^*(j,k) \\ \beta_1^*(j,k) \\ \beta_2^*(j,k) \end{bmatrix} \sim \{1-\pi_1(j)-\pi_2(j)\} MVN\left(\mathbf{0}, \begin{bmatrix} \tau_0^2(j) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) + \\ \pi_1(j) MVN\left(\mathbf{0}, \begin{bmatrix} \tau_0^2(j) & 0 & 0 \\ 0 & \tau_1^2(j) & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) + \\ \pi_2(j) MVN\left(\mathbf{0}, \begin{bmatrix} \tau_0^2(j) & 0 & 0 \\ 0 & \tau_1^2(j) & 0 \\ 0 & 0 & \tau_2^2(j) \end{bmatrix}\right).\end{aligned}$$

Within the wavelet space, Model (2.2) can be written as

$$(2.6)\quad \begin{cases} D_i(j,k) & = & \beta_0^*(j,k) + X_{1,i}\beta_1^*(j,k) + X_{2,i}\beta_2^*(j,k) + E_i^*(j,k), \\ \epsilon_i^*(j,k) & i.i.d. & N\left(0,\sigma^2\right), \end{cases}$$

where $j = 0,\ldots,J$, $k = 1,\ldots,K_j$ and the $\beta_m^*(j,k)$ are distributed as in (2.5).

The resulting posterior distributions of $[\beta_m^*(j,k)]$, given the observed values of $D(j,k)$, are again i.i.d. for each (j,k) and they are given by

$$\begin{aligned}(2.7)\quad \left[\begin{array}{c} \beta_0^*(j,k) \\ \beta_1^*(j,k) \\ \beta_2^*(j,k) \end{array}\middle| \boldsymbol{D}(j,k)\right] \sim \omega_0(j,k) MVN\left(\begin{bmatrix} \hat{\beta}_{0,0}^*(j,k) \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_0^*(j)\right) \\ + \omega_1(j,k) MVN\left(\begin{bmatrix} \hat{\beta}_{0,1}^*(j,k) \\ \hat{\beta}_{1,1}^*(j,k) \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_1^*(j)\right) \\ + \omega_2(j,k) MVN\left(\begin{bmatrix} \hat{\beta}_{0,2}^*(j,k) \\ \hat{\beta}_{1,2}^*(j,k) \\ \hat{\beta}_{2,2}^*(j,k) \end{bmatrix}, \boldsymbol{\Sigma}_2^*(j)\right).\end{aligned}$$

The analytical expressions for the $\hat{\beta}_{m,m'}^*(j,k)$, the variance covariance matrices $\boldsymbol{\Sigma}_{m'}^*(j,k)$ and the $\omega_{m'}(j,k)$ are given in Appendix A. The estimation orthogonality between $\beta_2^*(j,k)$ and the remaining effect functions, however, implies that $\hat{\beta}_{s,1}^*(j,k) = \hat{\beta}_{s,2}^*(j,k)$, with $s = 1,2$. This allows us to obtain

the following marginal posterior distributions of $\beta_1^*(j,k)$ and $\beta_2^*(j,k)$:

$$\begin{aligned}(2.8)\ \beta_1^*(j,k)|\boldsymbol{D}(j,k) &\sim \{\omega_1(j,k)+\omega_2(j,k)\}\, N\left(\hat{\beta}_1^*(j,k), \sigma^2_{\hat{\beta}_1}\right) + \\ &\quad \{1-\omega_1(j,k)-\omega_2(j,k)\}\, \delta(0),\\ (2.9)\ \beta_2^*(j,k)|\boldsymbol{D}(j,k) &\sim \omega_2(j,k) N\left(\hat{\beta}_2^*(j,k), \sigma^2_{\hat{\beta}_2}\right) + \\ &\quad \{1-\omega_2(j,k)\}\, \delta(0).\end{aligned}$$

For the model to be fully specified, we still have to define the hyperparameters $\pi_m(j)$ and $\tau_m^2(j)$. Extending the expressions in Abramovich, Sapatinas and Silverman (1998) to the functional model framework, we assume the hyperparameters of the prior model to be of the form

$$\begin{aligned}&\tau_m^2(j) = c_m \sigma^2 2^{-\alpha_m j},\\ &\pi_0(j) = 1-\pi_1(j)-\pi_2(j),\\ &\pi_1(j) = \min(1-\pi_2(j), q_1 2^{-\phi_1 j}),\\ &\pi_2(j) = \min(1, q_2 2^{-\phi_2 j}),\end{aligned} \tag{2.10}$$

where $m = 0, 1, 2$, and, c_m, q_1, q_2, α_m, ϕ_1 and ϕ_2 are non-negative constants. Abramovich, Sapatinas and Silverman (1998) have shown that 0.5 and 1 are robust choices for α_m and ϕ_m, respectively. We have chosen to impose the same amount of shrinkage to all non-zero wavelet coefficients by setting $c_0 = c_1 = c_2 = c$. The differences in smoothness between the effect functions will thus only be influenced by their corresponding prior probabilities π_m.

When the noise level σ is unknown, it can be robustly estimated by the median absolute deviation of the wavelet coefficients at the finest level, divided by 0.6745 (e.g. Abramovich, Sapatinas and Silverman, 1998). The remaining hyperparameters can be estimated by empirical Bayes using direct maximum marginal likelihood. The marginal density function for each $\boldsymbol{D}(j,k)$ is given in Equation (A.1). It is straightforward to obtain the marginal likelihood, because all wavelet coefficients between the different locations and different scales are assumed to be independent. However, the marginal density function is a mixture of multivariate normal distributions. It is therefore not feasible to come up with an analytical solution for maximum marginal likelihood estimators so that the marginal loglikelihood has to be optimized numerically.

2.4. *Wavelet thresholding.* In the wavelet domain the coefficient-wise posterior median corresponds to a point estimate of the posterior distribution under a family of loss functions that are equivalent to the use of L1 norms on the function and its derivatives. Within a Bayesian wavelet regression context, the posterior median thus acts as a true thresholding

rule because it sets wavelet coefficients effectively to zero when the posterior probability ω_m is small. Hence it accommodates for inhomogeneous functions, which is one of the aims of wavelet regression (Johnstone and Silverman, 2005). In their paper one can also find the formula for calculating the posterior median corresponding to the Gaussian multiple shrinkage prior.

3. Empirical Bayes inference for tiling array data. Here we describe an inference procedure for tiling array experiments using an approximate empirical Bayes FDR procedure. The FDR procedure relies on the posterior distributions of the effect functions from the wavelet based functional model. In particular, the mean function $\beta_1(t)$ is used for transcript discovery and the $\log_2$ fold change $FC(t) = 2 \times \beta_2(t)$ for assessing differential expression. The FDR procedures that are presented here are based on the work of Morris *et al.* (2008). However, we avoid the use of computationally intensive Bayesian MCMC methods. In Section 3.1 we adopt their method within our wavelet based functional model context for tiling array data and Section 3.2 presents of a procedure to approximate the posterior distributions of the effect functions in the data space. Finally we present a procedure for controlling the global FDR in Section 3.3.

3.1. *Local FDR procedure.* In tiling microarrays experiments it is often very intuitive for experimenters to identify the set of differentially expressed regions as all regions that exhibit a fold change above a threshold, say δ_{FC}. Until recently the use of such thresholds was lacking statistical rigour. Morris *et al.* (2008), however, provided a procedure for wavelet based functional mixed models that flags regions significantly exceeding a δ_{FC}-fold change between treatment groups while controlling the expected Bayesian FDR at the desired level α. For our model, they would define the local FDR estimate that $FC(t)$ exceeds a threshold $\log_2(\delta_{FC})$ at a certain genomic location t by

$$FDR_2(t) = Pr\left\{|2 \times \beta_2(t)| \leq \log_2(\delta_{FC}) | \boldsymbol{Y}\right\}. \tag{3.1}$$

With the basic Model (2.1) we can also infer on transcript discovery by using the mean function $\beta_1(t)$, for which we first have to set a threshold δ_{TD} for the background intensity. The use of a global background threshold, however, seems to be less interesting as high background values might occur at probes with high sensitivity. A classical background correction step prior to the main analysis may solve this issue. In this paper we avoid the use of *ad hoc* preprocessing procedures because it is hard to account correctly for these steps within the main analysis. Therefore, we use Model (2.2) that

incorporates the data of reference DNA hybridizations. Remember that the function $\beta_1(t)$ then gets the interpretation of a $\log_2$-fold change with respect to the DNA reference hybridization. As before, the biologists can provide an appropriate threshold, say δ_{TD}. Regions will be flagged as discoveries provided that they exceed a certain fold change compared to the reference DNA hybridizations. The local FDR can be defined as

$$FDR_1(t) = Pr\left\{\beta_1(t) \leq \log_2(\delta_{TD}) | \boldsymbol{Y}\right\}. \tag{3.2}$$

3.2. *Approximation of the posterior distributions of the effect functions in the data space.* Morris *et al.* (2008) would estimate the FDRs in Equations (3.1) and (3.2) by using MCMC samples from the posterior distributions of $\beta_1(t)$ and $\beta_2(t)$. We apply an approximate empirical Bayes method to estimate the marginal posterior distributions of $\beta_1(t)$ and $2 \times \beta_2(t)$. However, these distributions are intractable since they involve linear combinations of mixture distributions. Within the setting of wavelet based scatter plot smoothing, Barber, Nason and Silverman (2002), approximated the posterior distribution of the smoother by a suitable parametric distribution. In particular, they used Johnson curves. At each location t there exists precisely one Johnson curve with the same first four cumulants, say $\kappa_1(X), \ldots, \kappa_4(X)$, as the posterior distribution of the smoother. Note, that the first four cumulants also have a direct interpretation: $\kappa_1(X)$ and $\kappa_2(X)$ are the mean and variance of X, respectively, $\kappa_3(X)/\kappa_2^{3/2}(X)$ is the skewness and $\kappa_4(X)/\kappa_2^2(X)+3$ is the kurtosis. Johnson curves fall into three categories,

1. the log normal case (SL), $z = \gamma + \delta \log(x - \zeta)$ with $\zeta < x$,
2. the unbounded case (SU), $z = \gamma + \delta \sinh^{-1}\{(x - \zeta)/\lambda\}$, and
3. the bounded case (SB), $z = \gamma + \delta \log\{(x - \zeta)/(\zeta + \lambda - x)\}$, with $\zeta < x < \zeta + \lambda$,

in which z has a standard normal distribution and x is the Johnson variable. The Johnson curves are a rich family of distributions which provide good approximations of the tails of the distribution (Barber, Nason and Silverman, 2002).

Barber, Nason and Silverman (2002) approximated the first four cumulants of the posterior distribution of a wavelet based scatterplot smoother at each location t. Here, we will provide an analytical solution for the first four cumulants of the posterior distributions of the effect functions within the data space. The backtransformation from the wavelet space to the data space is a linear transformation and within the wavelet space the effect functions at each (j, k) are assumed to be independent. The cumulants of the effect functions in the original space can therefore be easily acquired by using

the following standard properties of cumulants,

$$\kappa_r\left(\sum_i \phi_i Z_i\right) = \sum_i \phi_i^r \kappa_r\left(Z_i\right), \tag{3.3}$$

where ϕ_i represent constants and the Z_i are independently distributed random variables. Once the cumulants are known within the wavelet space, they are readily available in the original space by applying modified versions of the inverse discrete wavelet transform (IDWT) using (3.3). Within the wavelet space the marginal posterior distribution of each effect function is a mixture of a point mass at zero and a normal distribution. For such a mixture distribution analytical expressions for the cumulants can be calculated. The first 4 cumulants are given in Appendix B. They are used to fit Johnson curves, which enable the calculation of the FDRs defined in Equations (3.1)-(3.2) and provide pointwise credibility intervals around the effect functions.

3.3. *Global empirical Bayes FDR procedure.* Equations (3.1) and (3.2) present so-called local FDR bounds. They control the FDR on probe level. As a consequence the global FDR of a set of locations that have a local FDR below α is too conservative. Morris *et al.* (2008) introduced a method for controlling the global FDR at the desired level α. The principle is to flag the set of locations $\psi_l = \{t_l : FDR_l(t_l) < \varphi\alpha\}$ as significant regions for the factor l. The threshold $\varphi\alpha$ has to be chosen so that $N(\psi_l)^{-1}\sum_{t_l\in\psi_l} FDR_l(t_l) < \alpha$ with $N(\psi_l)$ the cardinality of the set ψ_l. When simultaneous inference on p functions is required, either a common threshold $\varphi\alpha$ can be used or a set, say $\{\varphi_1\alpha, \ldots, \varphi_p\alpha\}$, of separate thresholds has to be proposed for controlling the simultaneous FDR at the α-level.

4. Results and Discussion. In this section we apply our procedure to the *Arabidopsis E2F-DPa$_{OE}$* dataset and compare our method with the methods of Kampa *et al.* (2004) and Huber, Toedling and Steinmetz (2006).

4.1. *Example: the Arabidopsis E2F-DPa$_{OE}$ tiling experiment.* The tiling data are obtained by hybridizing $N_0 = 1$ array to a DNA reference, $N_1 = 3$ arrays for the E2F-DPa$_{\mathrm{OE}}$ plants and $N_2 = 3$ arrays for the WT plants. We first remap the PM probes to the *Arabidopsis thaliana* genome annotation TAIR 8. Next, we select all PM/MM probe pairs with a PM probe that only maps to a unique location in the genome and a MM probe that has no match in the genome. Although no MM data are included in our model, we choose this approach for enabling a comparison of our method with the one

of Kampa *et al.* (2004) that makes use of MM probes. Finally, the tiling array data are filtered to remove probes for which the data show a low variance. We believe that this can be indicative for "dead" probes. In particular, probes with a $\log_2$-intensity below 7 in the DNA reference hybridisation show a reduced variability in the hybridisation of the plant material and are removed from the analysis.

All analyses are based on Model (2.2). In the matrix notation of Equation (2.3), the design matrix $\boldsymbol{X}$ equals

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{bmatrix}^T . \tag{4.1}$$

The Haar basis is used for the wavelet transformation. We expect level shifts among expressed exons and between expressed and non-expressed regions. Within an exonic region, the same mRNA concentration is present for all probes. In theory this should result into a constant fold change between 1) E2F-DPa$_{\text{OE}}$ and WT plants, and 2) the mean intensity with respect to the DNA reference hybridization. We perform the DWT down to $J = 10$. Both FDR procedures for transcript discovery and for the detection of differential expression rely on a threshold value that is driven by the biological problem at hand. This eventually leads to results that are both statistically significant and biologically relevant. The biologists involved in this study, consider a fold change between the E2F-DPa$_{\text{OE}}$ and WT arrays to be relevant as soon as it exceeds $\delta_{FC} = 1.5$. They also propose to set the threshold for the mean function $\beta_1(t)$ at $\delta_{TD} = 1/2$. The average intensities of the probes in the discovered transcripts are therefore larger than one half of the corresponding mean reference DNA signal.

We wish to control the total global FDR at 5%, i.e. $\alpha = 0.05$. According to Section 3.3 this can be done by using a set of separate thresholds for transcript discovery and differential expression. Here, both the FDRs for transcript discovery and differential expression analysis are controlled at 2.5%.

With our method we find 53821 transcribed regions and 1553 differentially expressed regions. Of these discovered TARs and differentially expressed TARs, 1356 and 40 TARs do not overlap with existing annotation, respectively. They can thus be considered as new discoveries that have to be biologically validated. A more detailed overview of the results for each chromosome is given in Table 1.

Fig. 2 shows two genomic regions on chromosome 1. The 3 top panels of each figure consist of the raw $\log_2$ DNA, E2F-DPa$_{\text{OE}}$ and WT intensi-

TABLE 1
Transcript discovery and differential expression for all chromosomes of Arabidopsis thaliana in the E2F-DPa$_{OE}$ experiment.

	Transcript discovery		Differential expression	
Chromosome	Detected	Non-annotated	Detected	Non-annotated
1	14088	336	405	11
2	8041	246	240	5
3	10648	241	295	12
4	8364	222	269	5
5	12680	311	344	7
1–5	53821	1356	1553	40

ties. In the middle panel, the genomic coordinate and the annotation are displayed. The bottom panels show the posterior medians (black lines) of $\beta_1(t)$ and $FC(t) = 2 \times \beta_2(t)$ along with 95% credibility intervals (light blue lines). The black boxes indicate the discovered transcripts and the shaded regions represent TARs that are significantly differentially expressed above the threshold δ_{FC}, controlling the global FDR at the 5% level.

From the left panel in Fig. 2 it can be seen that the genes AT1G13635 and AT1G13640 are transcribed in both strains, and that gene AT1G13650 is downregulated in the E2F-DPa$_{\mathrm{OE}}$ plants. In the right panel gene AT1G43580 is transcribed to the same level in both strains and an unannotated region is discovered which is upregulated in E2F-DPa$_{\mathrm{OE}}$.

4.2. *Comparison with existing methods.* The results of our model for transcript discovery are compared with two commonly used methods. First, the method of Kampa *et al.* (2004) which is based on the calculation of the pseudomedian within a sliding window. If the pseudomedian of all the probes within the window exceeds a certain threshold value, then the center probe of the window is called transcribed. Second, the method of Huber, Toedling and Steinmetz (2006) which uses a structural change model for performing the segmentation of the genomic expression profile. A segment is thus called transcribed if the mean intensity of the probes in the segment is larger than some background expression value, which is deduced from the observed intensities of non-annotated regions.

We next compare the different methods in terms of sensitivity and positive predictive value (PPV) at nucleotide-level. The sensitivity is defined as the number of nucleotides in the detected TARs that overlap with annotated regions (true positives) divided by the total number of nucleotides in annotated regions (sum of true positives and false negatives). The PPV is defined as the number of nucleotides in the detected TARs that overlap with annotated regions (true positives) divided by the total number of nu-

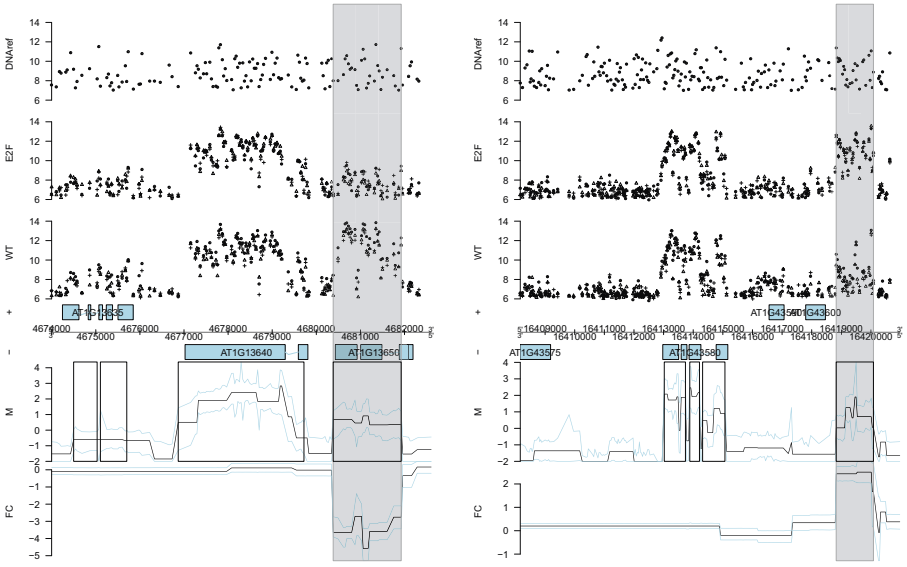

FIG 2. *Along-chromosome plot of an annotated region that is downregulated in E2F plants (left panel) and an unannotated region which is upregulated in E2F plants (right panel). Array intensities from a reference DNA hybridization (DNAref), E2F-DPa$_{OE}$ plants (E2F-DPa$_{OE}$) and wild type plants (WT), mean model $\beta_1(t)$ (M), fold change $FC(t) = 2 \times \beta_2(t)$ (FC). 95% credibility intervals are indicated with light blue lines, discovered transcripts are indicated by black boxes and differentially expressed regions are indicated by shaded regions. The different replicates for WT and E2F are indicated by $\bullet$, $+$ and $\triangle$*

cleotides in the detected TARs (sum of true and false positives). Annotation data are based on the TAIR 8 release. Furthermore, the computation times of the different methods are considered; they are measured on a 2.2 GHz Dual Core AMD Opteron® Processor 275 GNU/Linux server system with 16 GB RAM.

The results for chromosome 1 of the *Arabidopsis* data are summarized in Table 2. The existing methods are applied using parameter values and threshold values similar to the ones reported in the papers in which the methods have been introduced: Kampa *et al.* (2004), Huber, Toedling and Steinmetz (2006), and David *et al.* (2006). The results for the wavelet based method are obtained only by inferring on the mean model. From Table 2 we see that the wavelet based method gives the highest PPV. However, this is at the cost of a sensitivity that is lower than the other two methods. Furthermore, it is clear from Table 2 that the wavelet based method performs well in terms of numerical speed. It clearly outperforms Huber's method and it is only marginally slower than Kampa's method, which was implemented

TABLE 2
Comparison of our wavelet based method with existing methods for transcript discovery (chromosome 1).

Method	Sensitivity	PPV	Computation time
Wavelet	0.477	0.961	10 min
Kampa	0.551	0.929	2 min
Huber	0.575	0.930	1 h 50 min

in C by an efficient algorithm for calculating pseudomedians (Royce, Carriero and Gerstein, 2007).

A graph of the results for the region with genomic coordinates 782228 to 798823 on chromosome 1 is shown in Fig. 3. In the top panel, the TAIR 8 annotation is displayed. In the bottom panels, the average $\log_2$ transformed intensities over the six arrays are displayed (black dots) along with the transcripts discovered by the different methods (black boxes). Transcribed annotated regions with fairly long exons (e.g. AT1G03220 and AT1G03230) are in general well detected as one long TAR, by both our wavelet based method and the two existing methods. If the annotated regions contain many introns that are spliced out (e.g. AT1G03240 to AT1G03260), the wavelet based and the Kampa method seem to mimic the exonic structure better, while the Huber method usually detects this region as only one or just a few longer TARs. These findings are consistent with the mean length (Kampa: 534 bp, wavelet based: 619 bp, Huber: 1751 bp) and the total number (Kampa: 19631, wavelet based: 14088 , Huber: 6249) of the detected TARs in chromosome 1 for the different methods.

The sensitivity and PPV results presented in Table 2 depend heavily on the threshold value chosen in the analysis. Thus, they still lack some valuable information on the sensitivity-PPV trade-off for the three methods, which can not be derived from a single threshold. Therefore, we also inspect the nucleotide-level sensitivity and PPV for other threshold values. Moreover, Fig. 3 suggests that different results might be attained if one considers the TAIR 8 annotation of the entire genes and pseudo-genes (exonic + intronic regions) as compared to the calculation that only involves the annotated exonic regions. Both cases are examined and the results are presented in Fig. 4.

When the TAIR 8 annotation of the genes and pseudo-genes is used, the method of Huber provides the best results in general. Usually, the main concern in transcript discovery analysis is to keep the proportion of false positives at a reasonable level in order to attain a high enough PPV. In the range of high PPV the three methods are very competitive. When the introns are omitted from the calculation of the sensitivity and the PPV, the wavelet

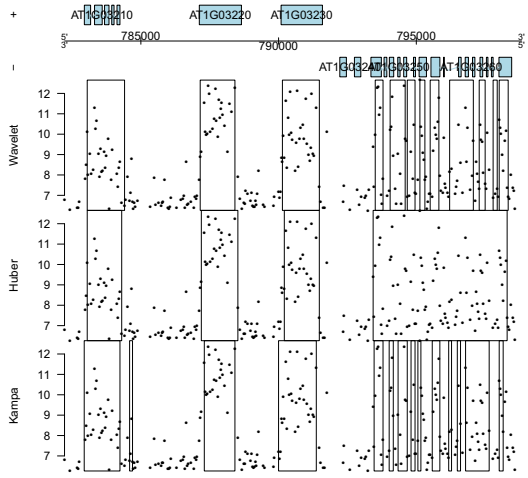

FIG 3. *Comparison of transcript discovery between wavelet based, Huber and Kampa method (chromosome 1: bp 782228-798823).*

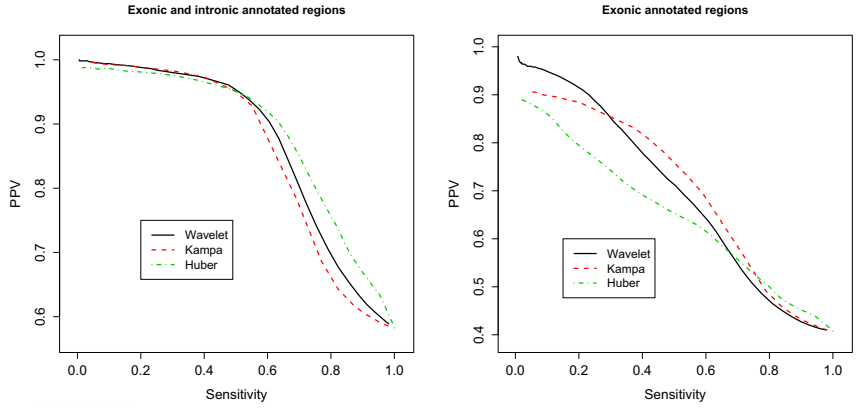

FIG 4. *Nucleotide-level sensitivity vs. PPV for the wavelet based, Huber and Kampa method based on the complete TAIR 8 annotation of chromosome 1 (left) and based on the TAIR 8 annotation of chromosome 1 without introns (right).*

based and the Kampa method perform better than the Huber method; this confirms the findings of Fig. 3. When the proportion of false positives has to be kept low, the wavelet based method outperforms Kampa's method as well.

In the literature only few contributions are available that assess differential expression by using tiling arrays. Most of them first construct probe-

sets for known annotation, which allows them to use conventional methods for assessing differential expression that were developed for the analysis of classical microarray experiments. This approach, however, does not allow for the discovery of novel transcripts that are differentially expressed. Our wavelet based method enables the discovery of both expressed and differentially expressed transcripts in annotated as well as in unannotated regions simultaneously and it remains very competitive with existing methods for transcript discovery. Hence, the method uses tiling array data to its full potential.

5. Conclusions and further research. We have presented a new method for transcriptome analysis using tiling arrays. In contrast to existing methods, it can asses transcript discovery and identify differentially expressed TARs, simultaneously. Meanwhile, our method remains competitive with the existing methods that perform only one of these tasks. Moreover, it also improves upon them in the sense that preprocessing is incorporated within the main analysis flow and it performs well in terms of numerical speed.

Extending our approach to other members of the exponential family may be important for other applications. In particular, the Poisson case is interesting, because that would allow us to use the model for analysing expression data generated by next generation sequencers such as Solexa, 454 and SOLiD®.

APPENDIX A: POSTERIOR DISTRIBUTIONS

The marginal distribution of the data in the wavelet space after integrating out the functional effects is given by

$$
\begin{aligned}
\text{(A.1)} \quad f\left(\boldsymbol{D}(j,k)\right) &= \left\{1-\pi_1(j)-\pi_2(j)\right\} g_0\left(\boldsymbol{D}(j,k)\right) + \\
&\qquad \pi_1(j) g_1\left(\boldsymbol{D}(j,k)\right) + \pi_2(j) g_2\left(\boldsymbol{D}(j,k)\right) \\
\text{(A.2)} \quad g_0\left(\boldsymbol{D}(j,k)\right) &= MVN(\mathbf{0}, \boldsymbol{V}_0(j)\sigma^2) \\
\text{(A.3)} \quad g_1\left(\boldsymbol{D}(j,k)\right) &= MVN(\mathbf{0}, \boldsymbol{V}_1(j)\sigma^2) \\
\text{(A.4)} \quad g_2\left(\boldsymbol{D}(j,k)\right) &= MVN(\mathbf{0}, \boldsymbol{V}_2(j)\sigma^2),
\end{aligned}
$$

with

$$
\begin{aligned}
\boldsymbol{V}_0 &= \boldsymbol{I} + \boldsymbol{X}_0 \boldsymbol{X}_0^T \rho(j), \\
\text{(A.5)} \quad \boldsymbol{V}_1 &= \boldsymbol{I} + \left\{ \boldsymbol{X}_0 \boldsymbol{X}_0^T + \boldsymbol{X}_1 \boldsymbol{X}_1^T \right\} \rho(j) \\
\boldsymbol{V}_2 &= \boldsymbol{I} + \left\{ \boldsymbol{X}_0 \boldsymbol{X}_0^T + \boldsymbol{X}_1 \boldsymbol{X}_1^T + \boldsymbol{X}_2 \boldsymbol{X}_2^T \right\} \rho(j),
\end{aligned}
$$

and $\rho(j) = c2^{-j/2}$. Given the values for the hyperparameters and the data, the posterior distributions of the effect functions in the wavelet space are given in Equation (2.7) with

$$\hat{\beta}^*_{0,0}(j,k) = \left\{\boldsymbol{X}_0^T\boldsymbol{X}_0 + 1/\rho(j)\right\}^{-1}\boldsymbol{X}_0^T\boldsymbol{D}(j,k), \tag{A.6}$$

$$\begin{bmatrix}\hat{\beta}_{0,1}(j,k)\\ \hat{\beta}_{1,1}(j,k)\end{bmatrix} = \left\{\boldsymbol{X}_{01}^T\boldsymbol{X}_{01} + 1/\rho(j)\boldsymbol{I}\right\}^{-1}\boldsymbol{X}_{01}^T\boldsymbol{D}(j,k), \tag{A.7}$$

$$\hat{\beta}^*_{0,2}(j,k) = \hat{\beta}^*_{0,1}(j,k), \tag{A.8}$$

$$\hat{\beta}^*_{1,2}(j,k) = \hat{\beta}^*_{2,1}(j,k), \tag{A.9}$$

$$\hat{\beta}^*_{2,2}(j,k) = \left\{\boldsymbol{X}_2^T\boldsymbol{X}_2 + 1/\rho(j)\right\}^{-1}\boldsymbol{X}_2^T\boldsymbol{D}(j,k), \tag{A.10}$$

$$\omega_0(j,k) = 1 - \omega_1(j,k) - \omega_2(j,k), \tag{A.11}$$

$$\omega_1(j,k) = \frac{\pi_1(j)g_1\left(\boldsymbol{D}(j,k)\right)}{f\left(\boldsymbol{D}(j,k)\right)}, \tag{A.12}$$

$$\omega_2(j,k) = \frac{\pi_2(j)g_2\left(\boldsymbol{D}(j,k)\right)}{f\left(\boldsymbol{D}(j,k)\right)}, \tag{A.13}$$

and the non-zero elements of the covariance matices $\boldsymbol{\Sigma}^*_m(j)$:

$$\sigma^2_{0,0}(j) = \sigma^2\left\{\boldsymbol{X}_0^T\boldsymbol{X}_0 + 1/\rho(j)\right\}^{-1} \tag{A.14}$$

$$\begin{aligned}\begin{bmatrix}\sigma^2_{0,1}(j) & \sigma_{01,1}(j)\\ \sigma_{01,1}(j) & \sigma^2_{1,1}(j)\end{bmatrix} &= \begin{bmatrix}\sigma^2_{0,2}(j) & \sigma_{01,2}(j)\\ \sigma_{01,2}(j) & \sigma^2_{1,2}(j)\end{bmatrix}\\ &= \sigma^2\left\{\boldsymbol{X}_{01}^T\boldsymbol{X}_{01} + 1/\rho(j)\boldsymbol{I}\right\}^{-1}\end{aligned} \tag{A.15}$$

$$\sigma^2_{2,2}(j) = \sigma^2\left\{\boldsymbol{X}_2^T\boldsymbol{X}_2 + 1/\rho(j)\right\}^{-1}. \tag{A.16}$$

APPENDIX B: CUMULANTS OF MIXTURE DISTRIBUTION

For a Gaussian mixture distribution with density

$$f(x) = (1-\pi)\delta(0) + \pi N(\mu, \sigma^2), \tag{B.1}$$

the first four cumulants are given by

$$\kappa_1 = \pi\mu \tag{B.2}$$

$$\kappa_2 = \pi\sigma^2 + \pi\mu^2 - \pi^2\mu^2 \tag{B.3}$$

$$\kappa_3 = 3\pi\sigma^2\mu - 3\pi^2\sigma^2\mu + \pi\mu^3 - 3\pi^2\mu^3 + 2\pi^3\mu^3 \tag{B.4}$$

$$\kappa_4 = 3\pi\sigma^4 + 6\pi\sigma^2\mu^2 - 18\pi^2\sigma^2\mu^2 + 12\pi^3\sigma^2\mu^2 - 3\pi^2\sigma^4 + \pi\mu^4 - 7\pi^2\mu^4 + 12\pi^3\mu^4 - 6\pi^4\mu^4. \tag{B.5}$$

ACKNOWLEDGEMENTS

Part of this research was supported by IAP research network grantnr. P6/03 of the Belgian government (Belgian Science Policy) and by BOF grantnr. 01517607 of the Flemish government. The authors also would like to thank Thomas Lumley and Giovanni Parmigiani for the fruitful discussions on their work and the Research Foundation - Flanders for providing a traveller grant.

REFERENCES

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet Thresholding via a Bayesian Approach. *Journal of the Royal Statistical Society: Series B* **60** 725–749.

Antoniadis, A. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys* **1** 16-55.

Barber, S., Nason, G. P. and Silverman, B. W. (2002). Posterior Probability Intervals for Wavelet Thresholding. *Journal of the Royal Statistical Society: Series B* **64** 189–205.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M. (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science* **306** 2242-2246.

Blais, A. and Dynlacht, B. D. (2007). E2F-associated chromatin modifiers and cell cycle control. *Current Opinion in Cell Biology* **19** 658 - 662. Cell differentiation / Cell division, growth and death.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. and Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Science* **103** 5320-5325.

De Veylder, L., Beeckman, T. and Inze, D. (2007). The ins and outs of the plant cell cycle. *Nature Reviews Molecular Cell Biology* **8** 655-665.

Halasz, G., van Batenburg, M., Perusse, J., Hua, S., Lu, X.-J., White, K. and Bussemaker, H. (2006). Detecting transcriptionally active regions using genomic tiling arrays. *Genome Biology* **7** R59.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

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

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes Selection of Wavelet Thresholds. *Annals of Statistics* **33** 1700–1752.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H. and Gingeras, T. R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research* **14** 331–342.

Kim, , Sik, J., Yun, H., amd Hyung Seok Choi, H. U. K., Kim, T. Y. and an Sang Yup Lee, H. M. W. (2006). Resources for Systems Biology Research. *Journal of Microbiology and Biotechnology* **16** 832848.

Laubinger, S., Zeller, G., Henz, S., Sachsenberg, T., Widmer, C., Naouar, N., Vuylsteke, M., Scholkopf, B., Ratsch, G. and Weigel, D. (2008). At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biology* **9** R112.

Morris, J. and Carroll, R. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B* **68** 179 - 199.

Morris, J., Brown, P., Herrick, R., Baggerly, K. and Coombes, K. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479 - 489.

Naouar, N., Vandepoele, K., Lammens, T., Casneuf, T., Zeller, G., van Hummelen, P., Weigel, D., Raetsch, G., Inze, D., Kuiper, M., De Veylder, L. and Vuylsteke, M. (2009). Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. *Plant Journal* **57** 184-194.

Royce, T. E., Carriero, N. J. and Gerstein, M. B. (2007). An efficient pseudomedian filter for tiling microarrays. *BMC Bioinformatics* **8** 186-193.

Royce, T. E., Rozowsky, J. S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M. and Gerstein, M. (2005). Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics* **21** 466–475.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. and Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **99** 909-917.

Biostat, Dep. of Appl. Math., Biometrics and Process Control
Ghent University, Coupure Links 653
9000 Ghent
Belgium
E-mail: lieven.clement@ugent.be
kristof.debeuf@ugent.be
olivier.thas@ugent.be

Dep. of Biostatistics
Johns Hopkins University
615 N. Wolfe E3620
Baltimore, Maryland 21205
USA
E-mail: ccrainic@jhsph.edu
ririzarr@jhsph.edu

VIB Dep. of Plant Systems Biology
Ghent University, Technologiepark 927
9052 Ghent
Belgium
E-mail: marnik.vuylsteke@psb.ugent.be