Published in final edited form as:

Stat Appl Genet Mol Biol. ; 11(1): Article—4. doi:10.2202/1544-6115.1726.

Fast Wavelet Based Functional Models for Transcriptome Analysis with Tiling Arrays

Lieven Clement, Kristof De Beuf, Olivier Thas, Marnik Vuylsteke, Rafael A. Irizarry, and Ciprian M. Crainiceanu

Lieven Clement, I-BioStat, Katholieke Universiteit Leuven and Universiteit Hasselt. Kristof De Beuf, Department of Mathematical Modelling, Statistics and Bioinformatics, Universiteit Gent. Olivier Thas, Department of Mathematical Modelling, Statistics and Bioinformatics, Universiteit Gent. Marnik Vuylsteke, VIB Department of Plant Systems Biology, Universiteit Gent. Rafael A. Irizarry, Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics. Ciprian M. Crainiceanu, Bloomberg School of Public Health, Department of Biostatistics, Johns Hopkins University

Abstract

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 are used for this purpose. They allow for the discovery of novel transcripts and the assessment of differential expression between two or more experimental conditions such as genotype, treatment, tissue, etc. In tiling array literature, many efforts are devoted to transcript discovery, whereas more recent developments also focus on differential expression. To our knowledge, however, no methods for tiling arrays have been described that can simultaneously assess transcript discovery and identify differentially expressed transcripts. In this paper, we adopt wavelet based functional models to the context of tiling arrays. The high dimensionality of the data triggered us to avoid inference based on Bayesian MCMC methods. Instead, we introduce a fast empirical Bayes method that provides adaptive regularization of the functional effects. A simulation study and a case study illustrate that our approach is well suited for the simultaneous assessment of transcript discovery and differential expression in tiling array studies, and that it outperforms methods that accomplish only one of these tasks.

Keywords

tiling microarray; wavelets; adaptive regularization; transcript discovery; differential expression; genomics; *Arabidopsis thaliana*

1 Introduction

In the last decade the genomes of many organisms have been entirely sequenced. A detailed description of all genomic regions that are actively transcribed is needed for enhancing the knowledge of the organism's 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, van Batenburg, Perusse, Hua, Lu, White, and Bussemaker, 2006). Thus, expression profiling has to be assessed experimentally. Genomic tiling arrays can provide an unbiased quantification of transcriptional activity (e.g. Bertone, Stolc, Royce, Rozowsky, Urban, Zhu, Rinn, Tongprasit, Samanta, Weissman, Gerstein, and Snyder, 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 because they contain probes that only map to annotated regions.

1.1 Motivation

Methods for transcriptome analysis with tiling arrays initially focused on transcript discovery (TD). They are often based on sliding windows and a thresholding criterion for identifying transcriptionally active regions (TARs) (e.g. Bertone et al., 2004, Kampa, Cheng, Kapranov, Yamanaka, Brubaker, Cawley, Drenkow, Piccolboni, Bekiranov, Helt, Tammana, and Gingeras, 2004, Royce, Rozowsky, Bertone, Samanta, Stolc, Weissman, Snyder, and Gerstein, 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, Vandepoele, Lammens, Casneuf, Zeller, van Hummelen, Weigel, Raetsch, Inzé, Kuiper, De Veylder, and Vuylsteke, 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 (DE) with tiling arrays is to group probes in probesets by mapping them to a gene model. The constructed probesets enable the researcher to use standard techniques developed for differential expression analysis in classical microarray experiments. This approach, however, does not provide for detecting differentially expressed transcripts in unannotated regions. Therefore, other methods are needed that can use expression data of tiling array experiments to their 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. Our main aim is to propose a method that can assess TD and DE unbiasedly from the existing annotation. Another issue that we bear in mind is related to the computational complexity of analyzing very large datasets. As high-throughput methods become much cheaper, scientists tend to design larger studies. Therefore, we also pursue the development of fast and efficient algorithms for estimation and statistical inference on high-throughput tiling array data.

1.2 *Arabidopsis thaliana* tiling array study

The methods proposed 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 Inzé, 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_{OE} plants. Expression profiling was performed with Affymetrix GeneChip *Arabidopsis* Tiling 1.0R arrays. A single array contains 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_{OE} strains are used. The wild type refers to the typical form of *Arabidopsis thaliana* as it occurs in nature. The replicates 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 WT and E2F-DPa_{OE} plants.

Figure 1 shows $\log_2$ transformed intensities from the E2F-DPa_{OE} and WT plant hybridizations as a function of the genomic coordinate. In theory, the measured intensity at a certain probe is proportional to the amount of the target that hybridized to this particular probe (specific binding). In practice, however, the data look very noisy. Microarray signals are known to be obscured by optical noise, background noise, non-specific binding (with non-complementary molecules), differences in probe affinity, small changes in experimental conditions of the different arrays, etc. Due to the complexity of the signal composition, some degree of preprocessing is required. For tiling array data, this typically involves (1) background correction that accounts for optical noise and non-specific binding and (2) normalization for making data from different arrays comparable. The data described in this manuscript were preprocessed with the background correction step from RMA, followed by quantile normalization across arrays (Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf, and Speed, 2003). Besides these probe to probe fluctuations our main interest is the detection of sudden jumps in the measured intensities associated with differences in transcriptional activity among exonic regions, and, between exonic, intronic and intergenic regions. They reflect the underlying transcriptional activity within each experimental unit.

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 effect functions without losing the ability to model local features. The huge number of features on tiling arrays triggered us to avoid fully Bayesian MCMC methods. We propose a very fast empirical Bayes procedure that provides adaptive regularization of the wavelet coefficients in functional fixed effects models.

The paper is organized as follows. In Section 2, we introduce the wavelet based functional model for transcriptome analysis and the parameter estimation procedure. We continue in Section 3 with an empirical Bayes false discovery rate (FDR) procedure for identifying both expressed and differentially expressed regions. In Section 4, our method is compared to existing methods in a simulation study and it is applied to the *Arabidopsis E2F-DPa_{OE}* case study. Finally, we present conclusions and some possible directions for further research in Section 5.

2 Fast wavelet based functional models for transcriptome analysis

We first present a functional model in the genomic space that can assess transcript discovery (TD) and differential expression (DE), simultaneously. The model is then transformed to the wavelet space in which an efficient regularization of the functional effects is accomplished.

2.1 Functional model

Let $N(\mu, \sigma^2)$ denote the density function of a normal distribution with mean μ and variance σ^2 and let $MVN(\mu, \Sigma)$ denote the density function of a multivariate normal distribution with mean μ and variance covariance matrix Σ .

Suppose that N_1 and N_2 tiling arrays are collected for two distinct experimental conditions C_1 and C_2 , respectively. Let $N = N_1 + N_2$. The expression functions $Y_i(t)$ are evaluated on

an equally spaced grid, say $t = (1, \dots, T)$, corresponding to the genomic locations of the probes within the same chromosome. We consider the functional model

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

with $i = 1, \dots, N$, $Y_i(t)$ 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 for which it is assumed that $E_i = [E_i(1), \dots, E_i(T)]^T$ are i.i.d. $MVN(\mathbf{0}, \Sigma_e)$, with Σ_e a $T \times T$ covariance matrix defined on grid $t \times t$. Hence, the intensities are assumed to be correlated within the same sample and are assumed to be independent across samples. 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 $\widehat{\beta}_1(t)$, 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, $FC(t)$, between the two distinct experimental conditions.

Model (1) can be written in matrix form as

$$Y = XB + E. \quad (2)$$

Here Y is an $N \times T$ matrix whose rows contain the $\log_2$ transformed intensities of one array observed on t . X is an $N \times q$ design matrix of the covariates. For Model (1) the number of effect functions q is defined as $q = 2$. However, all derivations below also hold for more complex designs with any arbitrary number of predictors q . Each row of the $q \times T$ matrix B contains one of the effect functions evaluated in t . The rows of E, E_i with $i = 1, \dots, N$, consist of the error processes evaluated on t , corresponding to each of the N observed tiling arrays and they are i.i.d. $MVN(0, \Sigma_e)$.

2.2 Wavelet based functional model

Functional models are commonly estimated by using basis functions. We propose to use a wavelet basis. We can construct a wavelet basis by choosing a suitable mother (ψ) and father wavelet function (ϕ) and considering all dilations and translations

$$\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k), \quad (3)$$

$$\phi_{jk}(t) = 2^{j/2} \phi(2^j t - k), \quad (4)$$

for integers j and k (e.g. Ramsay and Silverman, 2005). The mother wavelet is chosen to ensure that the basis is orthogonal, i.e. the integral of the product of any two distinct basis functions is zero. Typically, the mother wavelet and all derived basis functions have a compact support and the wavelet decomposition provides a location and scale decomposition of the underlying function. Therefore, wavelets are well suited to deal with irregular functional data that are characterized by a high number of local features. Only the coefficients of those basis functions whose support includes the region of the local feature are affected. Hence, wavelets often can provide a very economic representation of a function with relatively few non-zero coefficients. The Haar wavelet is the most simple wavelet. Its mother wavelet can be described as

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 0.5 \\ -1 & \text{if } 0.5 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

and its father wavelet or scaling function as

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

Note, that a decomposition with Haar wavelets will represent the underlying function by a piecewise constant (e.g. Bruce and Gao, 1996). This is very convenient in our setting because the process of gene transcription can be considered as a piecewise constant function within the genomic domain: in theory probes that cover the same exonic region should measure the same expression signal. The methods derived in this paper, however, are not restricted to the Haar wavelet basis. Hence, the reader can adopt our algorithms in conjunction with the wavelet basis of his or her choice.

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 $\mathbf{D} = \mathbf{Y}\mathbf{W}^T$, where $\mathbf{W}$ is a $T \times T$ orthogonal DWT matrix. The rows of the matrix $\mathbf{D}$ contain the wavelet coefficients for each of the observed curves and they are double indexed by the location $k = 1, \dots, K_j$ within the wavelet scale $j = 0, \dots, J$. In practice, the projection from the data space onto the wavelet space is performed by a more efficient pyramid-based algorithm (Mallat, 1989).

The wavelet transform allows us to rewrite the model within the wavelet space by post-multiplying both sides of Model (2) with the DWT matrix $\mathbf{W}^T$, resulting in

$$\mathbf{D} = \mathbf{X}\mathbf{B}^* + \mathbf{E}^*. \quad (7)$$

Hence, $\mathbf{B}^*$ and $\mathbf{E}^*$ 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, the rows of $\mathbf{E}^*$ are i.i.d. multivariate normal with mean zero and covariance matrix $\mathbf{S}^* = \mathbf{W}\Sigma_{\epsilon}\mathbf{W}^T$. Similar to many contributions in wavelet literature, we will assume that the wavelet coefficients within a given curve are independent across locations k and wavelet scales j , i.e. $\mathbf{S}^*$ is a diagonal matrix. This assumption, however, does not imply independence in the data space unless identical variance components are assumed across all wavelet scales j and locations k . Johnstone and Silverman (1997), for instance, argue that the variance of the wavelet coefficients for stationary correlated noise only depends on the scale j of the wavelet decomposition, but remains constant across locations k within the same scale. Obviously, the assumptions on the correlation structure are further relaxed when the variance components are allowed to vary both within and across wavelet scales.

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 to zero, which have an absolute value below the threshold, and it leaves the remaining coefficients unchanged. This results in a discontinuity at the threshold. Soft thresholding also replaces coefficients (ω) with an absolute value below the threshold δ by zero, but, it shrinks the remaining coefficients towards zero by subtracting the threshold from the absolute value of the wavelet coefficients, i.e. $\omega_{\text{soft}} = \text{sign}(\omega)[\omega - \delta]_+$, where $[\cdot]_+$ stands for the “positive part”. Hence, the

remaining coefficients after soft thresholding form a continuous distribution that is centered around 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 in the original data space.

2.3 Regularization and parameter estimation

Let $D(j, k)$ denote the wavelet coefficients at scale j and location k , which are assumed to be independent (cfr. Section 2.2), and, let $\beta_m^*(j, k)$, $m = 1, \dots, q$ be the wavelet coefficient of the m^{th} functional effect. Within the wavelet space, Model (1) can be written as

$$D(j, k) | \beta^*(j, k) \sim MVN \{ X \beta^*(j, k), I \sigma_\epsilon^2(j, k) \} \quad (8)$$

where $j = 0, \dots, J$ and $k = 1, \dots, K_j$.

Regularization can be imposed by putting a Gaussian prior on the wavelet coefficients $\beta^*(j, k)$:

$$\beta_m^*(j, k) | \tau_m(j, k) \sim N \{ 0, \tau_m(j, k) \sigma_\epsilon^2(j, k) \}. \quad (9)$$

Model (8)–(9) is a linear mixed model within the wavelet space, which also has a Bayesian model interpretation. A fully Bayesian modeling approach would involve the specification of a prior distribution on the variance components $\sigma_\epsilon^2(j, k)$. However, the hierarchical Model (8)–(9) also accommodates empirical Bayesian methods, i.e. the fully Bayesian analysis chain can be broken by replacing the unknown smoothing parameters $\tau_m(j, k)$ and variances $\sigma_\epsilon^2(j, k)$ by estimates and then performing a Bayesian analysis with the previously unknown parameters regarded as fixed (e.g. Section 16.3 of Ruppert, Wand, and Carroll, 2003). The joint likelihood for D and β^* within this empirical Bayes setting becomes

$$p \{ D, \beta^* | \tau, \sigma_\epsilon^2 \} \propto \prod_{j=0}^J \prod_{k=1}^{K_j} \left([\sigma_\epsilon^2(j, k)]^{-N/2} \exp \left\{ -\frac{[D(j, k) - X \beta^*(j, k)]^T [D(j, k) - X \beta^*(j, k)]}{2 \sigma_\epsilon^2(j, k)} \right\} \right) \prod_{m=1}^q \left[\tau_m(j, k) \sigma_\epsilon^2(j, k) \right]^{-1/2} \exp \left[-\frac{\beta_m^{*2}(j, k)}{2 \tau_m(j, k)} \right]$$

with $\tau = [\tau_1(1, 1), \dots, \tau_q(J, K_J)]^T$ and $\sigma_\epsilon^2 = [\sigma_\epsilon^2(j, k), \dots, \sigma_\epsilon^2(J, K_J)]^T$.

Marginal maximum likelihood (MML) is commonly used for deriving empirical Bayes estimators in wavelet based scatterplot smoothing, e.g. Figueiredo and Nowak (2001). Here, we generalize their approach towards a functional data analysis context. The marginal likelihood corresponding to Model (8)–(9) is defined as

$$p \{ D | \tau, \sigma_\epsilon^2 \} \propto \prod_{j=0}^J \prod_{k=1}^{K_j} |V(j, k) \sigma_\epsilon^2(j, k)|^{-1/2} \exp \left[-\frac{D^T(j, k) V^{-1}(j, k) D(j, k)}{2 \sigma_\epsilon^2(j, k)} \right], \quad (11)$$

with

$$\mathbf{V}(j, k) = \mathbf{I} + \sum_{m=1}^q \tau_m(j, k) \mathbf{X}_m \mathbf{X}_m^T. \quad (12)$$

In the case of an orthogonal design matrix, $\mathbf{X}$, the determinant of $\mathbf{V}(j, k)$ simplifies to

$$|\mathbf{V}(j, k)| = \prod_{m=1}^q [\mathbf{X}_m^T \mathbf{X}_m \tau_m(j, k) + 1], \quad (13)$$

and its inverse equals

$$\mathbf{V}^{-1}(j, k) = \mathbf{I} - \sum_{m=1}^q \frac{\mathbf{X}_m \mathbf{X}_m^T}{\mathbf{X}_m^T \mathbf{X}_m + 1/\tau_m(j, k)}. \quad (14)$$

After replacing $|\mathbf{V}(j, k)|$ and $\mathbf{V}^{-1}(j, k)$, $-2 \times \log$ -likelihood becomes

$$-2 \times L \sum_{j=0}^J \sum_{k=1}^{K_j} \left(\sum_{m=1}^q \left\{ \log [\mathbf{X}_m^T \mathbf{X}_m \tau_m(j, k) + 1] \right\} + N \log [\sigma_\epsilon^2(j, k)] + \frac{1}{\sigma_\epsilon^2(j, k)} \mathbf{D}^T(j, k) \left[\mathbf{I} - \sum_{m=1}^q \frac{\mathbf{X}_m \mathbf{X}_m^T}{\mathbf{X}_m^T \mathbf{X}_m + 1/\tau_m(j, k)} \right] \mathbf{D}(j, k) \right) \quad (15)$$

When the variance of the noise $\sigma_\epsilon^2(j, k)$ is known, the MML-estimator of $\tau_m(j, k)$ can be obtained in closed form:

$$\widehat{\tau}_m(j, k) = \left[\frac{\mathbf{D}(j, k) \mathbf{X}_m \mathbf{X}_m^T \mathbf{D}(j, k)}{(\mathbf{X}_m^T \mathbf{X}_m)^2 \sigma_\epsilon^2(j, k)} - \frac{1}{\mathbf{X}_m^T \mathbf{X}_m} \right]_+. \quad (16)$$

A detailed derivation of this result is given in Appendix A.

In contrast to Figueiredo and Nowak (2001), we did not put a Jeffreys prior on the smoothing parameters $\tau_m(j, k)$. For our model, the uninformative Jeffreys prior is defined as

$$p(\tau) \propto |I(\tau)|^{1/2}, \quad (17)$$

where $|I(\tau)|$ is the determinant of the expected Fisher information matrix of the marginal model (e.g. Ibrahim and Laud, 1991). With similar derivations as presented in Appendix A it can be easily verified that the Jeffreys rule prior for τ becomes

$$p(\tau) \propto \prod_{j=0}^J \prod_{k=1}^{K_j} \prod_{m=1}^q [\mathbf{X}_m^T \mathbf{X}_m \tau_m(j, k) + 1]^{-1}, \quad (18)$$

and that the use of this additional prior on τ , simply alters estimator (16) by dividing its first term by a factor 3:

$$\widehat{\tau}_m(j, k) = \left[\frac{\mathbf{D}(j, k) \mathbf{X}_m \mathbf{X}_m^T \mathbf{D}(j, k)}{3(\mathbf{X}_m^T \mathbf{X}_m)^2 \sigma_\epsilon^2(j, k)} - \frac{1}{\mathbf{X}_m^T \mathbf{X}_m} \right]_+, \quad (19)$$

which imposes additional shrinkage. Note, that in the special case of $q = 1$, $N = 1$ and $X = 1$,

estimator (19) reduces to $\left[\frac{D^2(j, k)}{3\sigma_\epsilon^2(j, k)} - 1 \right]_+$, which is equivalent to the result of Figueiredo and Nowak (2001) for wavelet denoising in a single curve setting.

Above, the variance $\sigma_\epsilon^2(j, k)$ was assumed to be known, however, in real applications it has to be estimated. In the wavelet literature, $\sigma_\epsilon(j, k)$ is often estimated by the mean absolute deviation (MAD) scheme of Donoho and Johnstone (1995), assuming equal variances across locations and wavelet scales. However, within a functional data context, this would lead to the restrictive assumption of independence in the original data space. We suggest using MML for the estimation of the variances of the noise component at the different wavelet scales j and/or locations k . Given the smoothing parameters $\tau(j, k)$, the MML-estimator of $\sigma_\epsilon^2(j, k)$ becomes

$$\widehat{\sigma}_\epsilon^2(j, k) = \frac{\mathbf{D}^T(j, k) \mathbf{V}^{-1}(j, k) \mathbf{D}(j, k)}{N}, \quad (20)$$

if the variances are allowed to vary across and within each wavelet scale, or

$$\widehat{\sigma}_\epsilon^2(j, \cdot) = \frac{\sum_{k=1}^{K_j} \mathbf{D}^T(j, k) \mathbf{V}^{-1}(j, k) \mathbf{D}(j, k)}{NK_j}, \quad (21)$$

if fixed variances are assumed within each wavelet scale. Detailed derivations can be found in Appendix A. We suggest a Gauss-Seidel type algorithm that iterates between expressions (16) and (20), or (16) and (21) until convergence (e.g. Thisted, 1988).

Once we have estimated $\widehat{\tau}_m(j, k)$ and $\widehat{\sigma}_\epsilon^2(j, k)$, we can plug them into the posterior densities of the effect functions. For an orthogonal design matrix $\mathbf{X}$ they are given by

$$p\{\beta_m^*(j, k) | D(j, k), \tau_m(j, k), \sigma_\epsilon^2(j, k)\} \propto N\left\{ \frac{\mathbf{X}_m^T \mathbf{D}(j, k)}{\mathbf{X}_m^T \mathbf{X}_m + 1/\tau_m(j, k)}, \frac{\sigma_\epsilon^2(j, k)}{\mathbf{X}_m^T \mathbf{X}_m + 1/\tau_m(j, k)} \right\}. \quad (22)$$

A derivation of the posterior distribution is given in appendix B. This result illustrates that Model (8)–(9) performs adaptive regularization of the wavelet coefficients $\beta_m^*(j, k)$ by specifying location and scale dependent regularization parameters $\tau_m(j, k)$.

The empirical Bayes method, however, ignores the extra variability in the posterior distribution caused by estimating the variance components. The posterior variance of the effect functions should be calculated from the joint posterior distribution of $\{\tau_m(j, k), \beta_m^*(j, k)\}$. We use the standard identity (e.g. Ruppert et al., 2003)

$$\text{var}\{\beta_m^*(j, k)\} = E[\text{var}\{\beta_m^*(j, k) | \theta\}] + \text{var}[E\{\beta_m^*(j, k) | \theta\}], \quad (23)$$

with $\theta = (\tau, \sigma^2)$ the parameter vector of all parameters that need to be estimated in the empirical Bayes procedure. Note, that the first term in (23) is well approximated by the posterior variance of $\beta_m^*(j, k)$ where $\tau_m(j, k)$ and $\sigma_\epsilon^2(j, k)$ are treated as known and fixed at their posterior mode (Kass and Steffey, 1989). The second term thus corrects for the extra variability in posterior distribution of $\beta_m^*(j, k)$ that is not accounted for by the approximate posterior variance. We estimate $\text{var}[E\{\beta_m^*(j, k) | \theta\}]$ by the following two steps.

1. Use a parametric bootstrap to estimate the covariance matrix of θ : $\widehat{\Sigma}_\theta$

2. Plug the results from step 1 into the delta-method formula:

$$\text{var} \left[E \left[\widehat{\beta}_m^* (j, k) | \theta \right] \right] = \left\{ \frac{\partial \widehat{\beta}_m^* (j, k)}{\partial \theta} \right\}^T \bigg|_{\widehat{\Sigma}_\theta} \left\{ \frac{\partial \widehat{\beta}_m^* (j, k)}{\partial \theta} \right\} \bigg|_{\widehat{\theta}}.$$

The partial derivatives of $\widehat{\beta}_m^* (j, k)$ are available analytically. Since the correction is a relatively small portion of the corrected posterior variance, it needs not be estimated by the bootstrap with as great a precision as when a variance is estimated entirely by the bootstrap (Ruppert et al., 2003). In Section 4.1 we illustrate that the correction has only a minor impact on the results.

Note, that our algorithms are based on the orthogonality of the predictors. However, non-orthogonal designs can be orthogonalized by using the Gram-Schmidt process. The Gram-Schmidt process performs a linear transformation of the original predictor variables. Thus, the least squares solution after fitting a linear model based on the original predictors and the orthogonalized design matrix are equivalent. Moreover, the Gram-Schmidt process also provides a straightforward backtransformation of the results towards the original predictor space. For regularization purposes, however, we introduced a Gaussian prior on the effect functions. Shrinkage estimators can be influenced by transformations of the design matrix. The high-dimensionality of the data, however, justifies this sacrifice as the available closed form solutions provide a tremendous decrease in the computational complexity of the algorithms for parameter estimation and inference.

2.4 Wavelet thresholding

The MML-thresholding rule associated with (16) is displayed in Figure 2 along with soft and hard thresholding rules (using the same threshold). The MML-estimators of the effect functions are plotted against the least squares solution. Note, that our MML thresholding rule is positioned in between the classical soft and hard thresholding functions. It imposes strong shrinkage on small values and it approaches the hard thresholding rule for large values. Hence, MML avoids the bias associated with soft thresholding. More shrinkage can be imposed by putting another prior on the effect functions, e.g. the use of a mixture prior between a point mass on zero and a normal distribution (Clyde, Parmigiani, and Vidakovic, 1998). This approach, however, is computationally less attractive. Alternatively, one could use another variance estimator such as the MAD estimator, or putting hyperpriors on the variances of the effect functions. When using carefully chosen priors, closed form solutions can also be obtained. We showed that the use of an improper Jeffreys prior, for instance, imposes additional shrinkage and generalizes the maximum a posteriori wavelet thresholding rule of Figueiredo and Nowak (2001) towards a functional data analysis context.

3 Empirical Bayes inference for tiling array data

Here, we describe an inference procedure for tiling array experiments using an 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 Newton, Noueiry, Sarkar, and Ahlquist (2004). However, we avoid the use of computationally intensive Bayesian MCMC methods.

In tiling microarray experiments differentially expressed regions across treatments can be identified by statistical hypothesis testing. However, they are often found to be only weakly related to the magnitude of the fold change (e.g. McCarthy and Smyth, 2009). Therefore, many experimenters prefer to identify the set of differentially expressed regions that exhibit a fold change above a threshold, say δ_{FC} . Two-fold is typically considered a worthwhile cutoff (e.g. DeRisi, Penland, Brown, Bittner, Meltzer, Ray, Chen, Su, and Trent, 1996, Schena, Shalon, Heller, Chai, Brown, and Davis, 1996). The use of thresholds is more intuitive for biologists, but until recently, it was lacking statistical rigor. McCarthy and Smyth (2009) developed an empirical Bayes moderated t-statistic for inferring if the fold change is above the threshold and Morris, Brown, Herrick, Baggerly, and Coombes (2008) provided a procedure for functional 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). Here, we will use a similar approach. At each probe position t , three different DE statuses exist and their corresponding posterior probabilities are given below,

- Overexpression at a particular probe t ,

$$p_{DE,1}(t) = Pr\{2 \times \beta_2(t) > \log_2(\delta_{FC}) | Y\} \quad (24)$$

- No biologically relevant DE at a particular probe t :

$$p_{DE,0}(t) = Pr\{-\log_2(\delta_{FC}) \leq 2 \times \beta_2(t) \leq \log_2(\delta_{FC}) | Y\} \quad (25)$$

- Underexpression at a particular probe t :

$$p_{DE,2}(t) = Pr\{2 \times \beta_2(t) < -\log_2(\delta_{FC}) | Y\} \quad (26)$$

A probe at a certain position can be classified according to the highest posterior probability and the local Bayesian false discovery rate corresponding to over-or under-expression is given by

$$BFDR_{DE,r}(t) = 1 - p_{DE,r}(t), \quad (27)$$

with $r = 1, 2$ (e.g. Newton et al., 2004, Efron, 2003). Setting $\log_2(\delta_{FC}) = 0$ will return all statistical significant results for which the null hypothesis of equality can be rejected. Hence, the method can be used for obtaining all probes for which the log-fold change is statistically significantly different from zero, as well as for returning results that are both of statistical and practical significance. Model (1) can also be used for inferring on transcript discovery by using the mean function $\beta_1(t)$ and a threshold δ_{TD} for the background intensity. The local Bayesian FDR becomes

$$BFDR_{TD}(t) = 1 - Pr\{\beta_1(t) > \delta_{TD} | Y\}. \quad (28)$$

Quantities (27) and (28) can be used for identifying probes, which correspond to significantly (differentially) expressed genomic targets, i.e. probes for which $BFDR(t) < \alpha$). The significant probes can be combined in significantly (differentially) expressed regions, which are defined as all probe sets ϕ_m that can be constructed by joining neighboring locations t_j for which

$$\phi_m = \{t_l: [t_l - (t_{l-1})] = 1 \vee BFDR(t_l) < \alpha\}. \quad (29)$$

They correspond to consecutive genomic regions with a BFDR below the significance level α . Morris et al. (2008) would evaluate both $\text{BFDR}_{DE,i}(t)$ and $\text{BFDR}_{TD}(t)$ using an MCMC approach. We adopt the empirical Bayes methods from the previous section.

4 Results and discussion

We first assess our procedure in a simulation study and compare our method with the methods of Kampa et al. (2004) and Huber et al. (2006) for transcript discovery (TD), and an RMA based method for the analysis of differential expression (DE). Next, we apply our model to the *Arabidopsis E2F-DPaOE* dataset.

4.1 Simulation study

Tiling array data are simulated by adapting the model of Purdom, Simpson, Robinson, Conboy, Lapuk, and Speed (2008):

$$y_i(t) = \log_2 \left[B(t) + I_i(t) \times 2^{c_i(t) + p(t)} \right] + \epsilon_i(t), \quad (30)$$

with $y_i(t)$ the $\log_2$ transformed intensity of probe t on array i and where the additive background, $B(t)$, is modeled as a log-normal variable, the chip effect $c_i(t)$ as a normal variable, the probe affinity $p(t)$ as mean-zero normal variable and the errors $\epsilon_i(t)$ are serially correlated. The indicator variable $I_i(t)$ is 1 when probe t is expressed on chip i , and 0 otherwise. This model features additive background, multiplicative noise, probe specific affinities and serial correlation in the data domain. We chose values for the simulation parameters by obtaining rough estimates of typical values from real data. With the model we simulate *Arabidopsis thaliana* data for the entire first chromosome. In the simulation study 3 WT and 3 E2F-DPaOE were considered. The average hybridization signal of annotated transcripts is chosen at random from the set (0, 3, 4, 5, 6, 7, 8, 9, 10, 11). All unannotated regions of the genome were not expressed and only consist of background noise. Six different FC levels are used (0.95, 1.5, 2, 3, 4 of 5) and each of these are randomly attributed to 100 genes. For the study we assume that the researchers are only interested in detecting genes that are differentially expressed above a fold change $\delta_{FC} = 2$ (or $\log_2 \delta_{FC} = 1$). The threshold for transcript discovery was set at the 90 percentile of log transformed, background corrected and quantile normalized simulated intensities of the non-annotated regions in the genome. In a post-processing step, we only maintain regions consisting of at least two consecutive probes. More details on the simulation study can be found in Appendix C. Figure 6 of Appendix C shows that the marginal distribution of the simulated intensities in annotated regions closely resembles the distribution of the raw intensities of annotated regions in the E2F-DPaOE experiment. Moreover, the simulated data also possess a similar within-sample dependence structure as the original data. An example of a region with simulated data is given in Figure 7 in Appendix C. The top panel shows data from the simulation run and in the bottom panel real data from the E2F-DPaOE experiment is displayed for comparison. The plot shows that the characteristics of the observed data are realistically preserved in the simulated data.

In the simulation study, five different versions of the wavelet model are used, (1) EB: method using variance estimator (21), (2) EB_{MAD} : based on the MAD estimator for the standard deviation, (3) EB_{jk} : using variance estimator (20), (4) EB_b : EB with bootstrap correction and (5) EB_{imp} : method using an improper prior on the random effect variances $\tau_m(j, k)$ and variance estimator (21). In this study, the local BFDR is controlled at the 5% level. The performance for transcript discovery is compared with: (6) the method of Kampa et al. (2004), which is based on the calculation of the pseudomedian or Hodges-Lehmann estimator (Hollander and Wolfe, 1999) 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; (7) the method of Huber et al. (2006) that uses a structural change model for performing the segmentation of the genomic expression profile. A segment is 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 in non-annotated regions.

For differential expression (8) an RMA based method is used as a benchmark. The RMA (Irizarry et al., 2003) algorithm is designed for the analysis of classical microarray studies. The probes on these arrays are organised in probesets for each gene. When applying RMA we first have to construct probesets from the individual probes. We use the existing annotation for this purpose. Therefore, it is hard to view the RMA method as a fair competitor to our method, which does not rely on the annotation. Moreover, RMA is unable to detect novel transcripts in a real tiling array experiment: the probes of intronic and intergenic regions cannot be organized into probesets. RMA involves three steps: (a) background correction, (b) quantile normalization and (c) summarization using the median polish algorithm. DE is assessed using the empirical Bayes moderated t-statistic of McCarthy and Smyth (2009) for inferring if the contrast between E2F-DPa_{OE} and WT is higher than the $\log_2$ FC-threshold ($\log_2 \delta_{FC} = 1$). The FDR is controlled at the 5% level by using the method of Benjamini and Hochberg (1995).

In table 1, the methods are compared in terms of positive predictive value (ppv), sensitivity (sens), specificity (spec) and of computation time measured on a 2×6 Core Intel® Xeon® X7460, 2.66 GHz Processors GNU/Linux server system with 128 GB RAM. The ppv, sensitivity and specificity are calculated on probe level, except for RMA which acts on gene level. We also tabulate the sensitivity for the wavelet based functional models on a gene level: a gene is called DE if it contains probes that are flagged as DE by our method.

The EB and the EB_k wavelet based methods outperform the Kampa method and Huber method for TD. They have very high specificity and positive predictive values while maintaining a high sensitivity. Among the wavelet based methods, the EB_{MAD} and EB_{imp} slightly suffer from a loss in sensitivity. For DE, all wavelet methods seem to be more sensitive than RMA and they strongly control for false positives. The wavelet methods without bootstrap correction are also very competitive in terms of computation time.

Receiver Operating Characteristic (ROC)-curves for TD and DE are given in Figure 3. The sensitivity or true positive rate (TPR) is plotted against the false positive rate (FPR, 1-specificity). An optimal test would detect all true positives without any false positives. All wavelet methods seem to have a similar performance and clearly outperform the Huber and Kampa methods. Among the wavelet based methods, the EB_{MAD} and the EB_{imp} seem to be superior for DE. In practice, however, they are less sensitive than the EB and EB_b method when using a low FDR, e.g. the $FDR = 5\%$ used in Table 1. Based on Table 1 we prefer the EB-method for the case study. However, code for all methods is available upon request from the first author, so the reader may use the method of his choice.

Finally, example plots of the resulting segmentation for TD and DE are given in Figure 9 and Figure 10 of Appendix C, respectively. The top panel displays the simulated background corrected and quantile normalized data along with the regions that were truly (differentially) expressed. The bottom panels show the model tracks. Since the wavelet based methods resulted in approximately the same segmentation, we only show results for method EB in Figure 9 and for EB, EB_b and EB_{imp} in Figure 10. The regions that are discovered by our wavelet based model correspond very well with the underlying exonic structure. The Kampa method also mimics the exonic structure, while the method of Huber cannot distinguish as

good between intronic and exonic regions. Note, that the effect of the bootstrap correction on the variance can be seen in the FC-model tracks of Figure 10. The credible intervals in non-differentially expressed regions are much wider. The correction has a minor effect in regions with signal. Hence, the variance is mainly underestimated in regions that are not of interest. That explains why the empirical Bayes methods without bootstrap correction also attain a high ppv and specificity. The small impact of the bootstrap correction in regions of interest combined with the drastic increase in the computation time indicate that the bootstrap correction is not worth the effort in our application. The introduction of the improper prior on the smoothing parameters τ clearly imposes additional regularization. Hence, researchers can adopt the EB_{imp} method when they prefer more smooth estimates for the FC.

4.2 Case study: the *Arabidopsis* E2F-DPa_{OE} tiling experiment

The tiling data are obtained by hybridizing $N_1 = 3$ arrays for the E2F-DPa_{OE} plants and $N_2 = 3$ arrays for the WT plants. We remapped the PM probes to the *Arabidopsis thaliana* genome annotation TAIR 9 (TAIR resources can be found on <http://www.arabidopsis.org/>). The design matrix X in (2) equals

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T.$$

Tiling arrays contain probes at exonic, intronic and intergenic regions. Hence, sudden jumps in the measured intensity can be expected due to differences in transcriptional activity among exonic regions, and, between exonic, intronic and intergenic regions. In theory this should result into a piecewise constant signal along the genomic coordinate. The tiling array signal, however, is largely affected by noise, which suggests that smoothing or probabilistic modeling of the signal is beneficial. The structural change model is a popular choice for this purpose since it models the data as a piecewise constant function of chromosomal coordinates (Huber et al., 2006). In our contribution, we adopt wavelets for denoising the data. We use the Haar basis for the wavelet transformation because the underlying transcriptional activity and Haar are both piecewise constant (Bruce and Gao, 1996). For large datasets, the DWT is typically stopped at a certain level J . Similar to Morris et al. (2008), we perform the DWT down to $J = 10$. Hence, the father wavelet spans around 60 kb, which is much larger than a typical gene of *Arabidopsis*. (The average gene size in *Arabidopsis* is about 1.9 kb and large genes are defined as genes with a length > 3 kb, e.g. Meinke, Meinke, Showalter, Schissel, Mueller, and Tzafrir, 2003).

Early microarray publications inferred on differential expression by only considering the fold change, with 2-fold typically considered a worthwhile cutoff (e.g. DeRisi et al., 1996, Schena et al., 1996). But, fold change cutoffs do not account for variability nor guarantee reproducibility (McCarthy and Smyth, 2009). Most statistical methods for assessing differential expression, however, allow for genes with small fold changes to be considered statistically significant. Hence, they report significant genes that are not biologically relevant. In this contribution, both FDR procedures for transcript discovery and for the detection of differential expression can 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. Similar to the early microarray studies, we consider a fold change between the E2F-DPa_{OE} and WT arrays to be relevant as soon as it exceeds $\delta_{FC} = 2$ ($\log_2(\delta_{FC}) = 1$). Researchers who want to avoid the use of an arbitrary threshold for the fold change, can use $\log_2(\delta_{FC}) = 0$. With this threshold all regions are recovered that exhibit a $\log_2$ fold change that is significantly different from 0. Similar to most existing methods in

tiling array literature for transcript discovery (e.g. David, Huber, Granovskaia, Toedling, Palm, Bofkin, Jones, Davis, and Steinmetz, 2006, Kampa et al., 2004), we cannot avoid the use of a threshold for transcript discovery. If we adopt a similar approach as David et al. (2006), the threshold for the mean function $\beta_1(t)$ becomes $\delta_{TD} = 3.7$. More details on this procedure can be found in Appendix D. The biologists involved in this study, however, proposed to set the threshold for the mean function $\beta_1(t)$ at $\delta_{TD} = 4.5$. This approximately corresponds to the median of the $\log_2$ -transformed, background corrected and quantile normalized (LBCQQ) data, as well as to the location of the minimum between the two modes of the distribution of the LBCQQ intensities in annotated regions (Figure 11 in Appendix D). Note, that with the latter TD-threshold slightly more conservative results are obtained. The local BFDR is controlled at 5%. In a postprocessing step, we only maintain regions consisting of at least two consecutive probes.

With our method we find 77663 transcribed regions and 3885 differentially expressed regions. Of these discovered TARs and differentially expressed TARs, 15149 and 765 do not overlap with existing annotation, respectively. They can be considered as potential discoveries that have to be biologically validated. A more detailed overview of the results for each chromosome is given in Table 2.

Figure 4 shows a genomic region of chromosome 1. The top panel consists of the $\log_2$ transformed background corrected and normalized E2F-DPaOE and WT intensities. In the middle panel, the genomic coordinate and the annotation are displayed. The bottom panel shows the posterior means (black lines) of $\beta_1(t)$ and $\log_2$ transformed fold change $FC(t) = 2 \times \beta_2(t)$ along with 95% credible intervals (light blue lines). Discovered transcripts and differentially expressed regions are indicated by black boxes on the M and FC track, respectively. An annotated gene is transcribed to the same level in both strains and a novel transcript is discovered, which is upregulated in E2F-DPaOE plants. The discovered region spans approximately 1500 bp and seems to have an exonic structure. It is an interesting region for further biological validation.

5 Conclusions and further research

We have adopted a wavelet based functional model for transcriptome analysis with tiling arrays. In contrast to existing methods for the analysis of tiling arrays, it can assess transcript discovery and identify differentially expressed transcripts, simultaneously. It is as powerful in detecting existing as well as novel differentially expressed transcripts. The wavelet based functional model thus exploits tiling array data to their full potential. It seems a promising tool to assist biologists in further unraveling transcriptional networks. In a simulation study we showed that our method is superior to existing methods and performs very good in terms of numerical speed. The case study showed its use for finding potential targets without being biased by the existing annotation.

The extension of our method towards other members of the exponential family is tempting and would open new possibilities for analyzing expression data generated by next generation sequencers such as Solexa, 454 and SOLiD®.

Acknowledgments

Part of this research was supported by IAP research network grant no. P6/03 of the Belgian government (Belgian Science Policy) and by BOF grant no. 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 traveller grants.

Appendix A: Estimation of the variance components by empirical Bayes

The empirical Bayes procedure is based on marginal maximum likelihood.

Estimation $\tau_m(j, k)$ given $\sigma_\epsilon^2(j, k)$

In case of an orthogonal design matrix $\mathbf{X}$, the marginal likelihood in Equation (11) simplifies to Equation (15). When the variance $\sigma_\epsilon^2(j, k)$ is known, an estimator for the smoothing parameters $\tau_m(j, k)$ can be derived as follows (note, that we will suppress the index (j, k) for notational convenience):

$$\begin{aligned} \frac{\partial L}{\partial \tau_m} &= 0 \\ \Downarrow \\ \frac{\mathbf{X}_m^T \mathbf{X}_m}{\mathbf{X}_m^T \mathbf{X}_m \tau_m + 1} - \frac{\mathbf{D}^T \mathbf{X}_m \mathbf{X}_m^T \mathbf{D}}{[\mathbf{X}_m^T \mathbf{X}_m + 1/\tau_m]^2 \tau_m^2 \sigma_\epsilon^2} &= 0 \\ \Downarrow \\ [\mathbf{X}_m^T \mathbf{X}_m]^2 \tau_m \sigma_\epsilon^2 + \mathbf{X}_m^T \mathbf{X}_m \sigma_\epsilon^2 - \mathbf{D}^T \mathbf{X}_m \mathbf{X}_m^T \mathbf{D} &= 0 \\ \Downarrow \\ \tau_m &= \frac{\mathbf{D}^T \mathbf{X}_m \mathbf{X}_m^T \mathbf{D}}{[\mathbf{X}_m^T \mathbf{X}_m]^2 \sigma_\epsilon^2} - \frac{1}{\mathbf{X}_m^T \mathbf{X}_m} \end{aligned}$$

By definition $\tau_m \in [0, \infty]$, thus, the MML estimate after reintroducing the index (j, k) becomes

$$\hat{\tau}_m(j, k) = \left[\frac{\mathbf{D}(j, k) \mathbf{X}_m \mathbf{X}_m^T \mathbf{D}(j, k)}{(\mathbf{X}_m^T \mathbf{X}_m)^2 \sigma_\epsilon^2(j, k)} - \frac{1}{\mathbf{X}_m^T \mathbf{X}_m} \right]_+.$$

Estimation of $\sigma_\epsilon^2(j, k)$ given $\tau(j, k)$

We start the derivation of an estimator for the variance components $\sigma_\epsilon^2(j, k)$ given the smoothness parameters $\tau(j, k)$ from the formulation of the likelihood in Equation (11).

$$\begin{aligned} \frac{\partial L}{\partial \sigma_\epsilon^2(j, k)} &= 0 \\ \Downarrow \\ \frac{N}{\sigma_\epsilon^2(j, k)} - \frac{\mathbf{D}^T(j, k) \mathbf{V}^{-1}(j, k) \mathbf{D}(j, k)}{\sigma_\epsilon^4(j, k)} &= 0 \\ \Downarrow \\ \hat{\sigma}_\epsilon^2(j, k) &= \frac{\mathbf{D}^T(j, k) \mathbf{V}^{-1}(j, k) \mathbf{D}(j, k)}{N}. \end{aligned}$$

Appendix B: Derivation of the posterior distribution of the functional effects in the wavelet space

The wavelet coefficients $D(j, k)$ are assumed to be independent. When $\sigma_\epsilon^2(j, k)$ and $\tau_m(j, k)$ are known, the posterior of $\beta^*(j, k)$ only involves $\mathbf{D}(j, k)$, $\sigma_\epsilon^2(j, k)$, $\tau_m(j, k)$. For notational convenience, we suppress index (j, k) .

$$\begin{aligned}
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &= \frac{p\{\beta^*, D|\tau, \sigma_\epsilon^2\}}{p\{D|\tau, \sigma_\epsilon^2\}} \\
&\Downarrow \text{Equations (10) – (11)} \\
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &\propto \left\{(\sigma_\epsilon^2)^{-N/2} \exp\left[-\frac{(D - X\beta^*)^T(D - X\beta^*)}{2\sigma_\epsilon^2}\right] \times \prod_{m=1}^a (\tau_m \sigma_\epsilon^2)^{-1/2} \exp\left(-\frac{\beta_m^2}{2\tau_m \sigma_\epsilon^2}\right)\right\} \times |V\sigma_\epsilon^2|^{1/2} \left(\frac{D^T V^{-1} D}{2\sigma_\epsilon^2}\right) \\
&\Downarrow \text{Using the orthogonal property of design matrix } X \\
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &\propto \exp\left(-\frac{D^T D}{2\sigma_\epsilon^2}\right) \prod_{m=1}^q \left\{(\tau_m \sigma_\epsilon^2)^{-1/2} \times \exp\left[-\frac{(X_m^T X_m + 1/\tau_m)\beta_m^{*2} - 2X_m^T D\beta_m^*}{2\sigma_\epsilon^2}\right]\right\} \times \exp\left(\frac{D^T D}{2\sigma_\epsilon^2}\right) \prod_{m=1}^q \left[(X_m^T X_m \tau_m + 1)^{1/2} \times \exp\left(-\frac{D^T X_m}{X_m^T X_m + 1/\tau_m}\right)\right] \\
&\Downarrow \\
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &\propto \prod_{m=1}^q \left(\frac{\sigma_\epsilon^2}{X_m^T X_m + 1/\tau_m}\right)^{-1/2} \times \exp\left[-\frac{\beta_m^{*2} - 2\frac{X_m^T D}{X_m^T X_m + 1/\tau_m}\beta_m^* + \frac{D^T X_m X_m^T D}{(X_m^T X_m + 1/\tau_m)^2}}{2\sigma_\epsilon^2(X_m^T X_m + 1/\tau_m)^{-1}}\right] \\
&\Downarrow \\
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &\propto \prod_{m=1}^q \left(\frac{\sigma_\epsilon^2}{X_m^T X_m + 1/\tau_m}\right)^{-1/2} \times \exp\left[-\frac{\left(\beta_m^* - \frac{X_m^T D}{X_m^T X_m + 1/\tau_m}\right)^2}{2\sigma_\epsilon^2(X_m^T X_m + 1/\tau_m)^{-1}}\right] \\
&\Downarrow \\
p\{\beta^*|D, \tau, \sigma_\epsilon^2\} &\sim \prod_{m=1}^q N\left(\frac{X_m^T D}{X_m^T X_m + 1/\tau_m}, \frac{\sigma_\epsilon^2}{X_m^T X_m + 1/\tau_m}\right)
\end{aligned}$$

Appendix C: Simulation study

Data are simulated by adapting the model of Purdom et al. (2008):

$$y_i(t) = \log_2 \left[B(t) + I_i(t) \times 2^{c_i(t) + p(t)} \right] + \epsilon_i(t), \quad (31)$$

with $y_i(t)$ the $\log_2$ transformed intensity of probe t on array i and where

$$\begin{cases} \log_2 \{B(t)\} & \mathcal{N}(\mu_B, \sigma_B^2) \\ c_i(t) & \mathcal{N}(\mu_{c,i}, \sigma_c^2) \\ p(t) & \mathcal{N}(0, \sigma_p^2) \\ \epsilon_i(t) & = \text{ARMA}(m, n) + w_i(t) \\ w_i(t) & \mathcal{N}(0, \sigma_w^2) \end{cases}, \quad (32)$$

$\log_2 B(t)$ is the $\log_2$ transformed background hybridization signal at probe t , $c_i(t)$ denotes the mean effect of an expressed exon on chip i , hence $c_i(t)$ is constant for all probes within a particular exon on the same array, $p(t)$ accounts for the differences in probe affinity and $\epsilon_i(t)$ is the noise component of probe t on chip i , which follows an autoregressive moving average process ARMA(m, n) with Gaussian white noise $w_i(t)$. This model features additive background, multiplicative noise, probe specific affinities and serial correlation in the data domain. We follow the approach of Purdom et al. (2008) for tuning the simulation parameters by rough estimates of “typical” values from real data. We fit an ANOVA model to the E2F-DPaOE data for this purpose. The model consists of a fixed probe effect and a fixed group effect (WT, E2F-DPaOE) nested within probe. The errors of the ANOVA model are serially correlated and could be modeled by an ARMA(1,1) process. They were used for estimating the AR, MA parameters and the variance parameter of the white noise. We set the mean of the background μ_B at 6.5 and σ_B at 0.2, which is slightly larger than the empirical estimate $\widehat{\sigma}_{0.15}$ derived from the raw intensities of the E2F-DPaOE experiment. An empirical method for characterizing the distribution of background noise can be found in Appendix D. Figure 5 illustrates the method using the raw intensities of the E2F-DPaOE

($\widehat{\mu}_B=6.35$ and $\widehat{\sigma}_B=0.15$). The indicator variable $I_\lambda(t)$ is used to add hybridization signal, i.e. $I_\lambda(t) = 1$ in all exonic regions that are assumed to be expressed and $I_\lambda(t) = 0$ for all probes covering intronic, intergenic regions and exons of genes that are non-expressed in the simulation study. The mean chip effect in exonic regions, $\mu_{c,\lambda}(t) = \mu_{c,\lambda}(t) + FC_\lambda(t)$ is arbitrary. It is the mean expression level of the simulated gene in the treatment group corresponding to chip i , with $\mu_{c,\lambda}(t)$ the average expression level for a certain gene over all the arrays and $FC_\lambda(t)$ the average $\log_2$ fold change in the group of chip i . σ_c is set at 0.2, which approximately equals the standard deviation of mean expression levels between exonic regions within each group of the E2F-DPaOE experiment. The standard deviation of the probe effect $p_\lambda(t)$ is set at $\sigma_p = 1.5$, which corresponds to the 75% quantile of the empirical distribution of the standard deviations of probes that target the same exonic region.

With this model we simulate *Arabidopsis thaliana* data for the entire first chromosome. In the simulation study 3 WT and 3 E2F-DPaOE were considered. The average hybridization signal of annotated transcripts is chosen at random from the set (0, 3, 4, 5, 6, 7, 8, 9, 10, 11) with probabilities (1/15, 1/15, 1/15, 1/15, 1/15, 2/15, 2/15, 2/15, 2/15, 2/15). The lower values were set at a lower weights so that the distribution of the simulated data corresponds better with the empirical distribution of the raw data in annotated regions. Six different FC levels are used (0.95, 1.5, 2, 3, 4 of 5) and each of these are attributed to 100 genes. The marginal distribution in the annotated regions of the simulated and the real data are displayed in the left panel of Figure 6. The distribution of the simulated data corresponds well to the empirical distribution of the raw data. The simulated data also possess a similar serial correlation structure as the E2F-DPaOE data (see middle and right panels of Figure 6). An example of a region with simulated data is given in Figure 7. The top panel shows data from the simulation run and in the bottom panel real data from the E2F-DPaOE experiment is displayed for comparison. The plot shows that the characteristics of the observed data are realistically preserved in the simulated data.

Figure 8 provides more insight in the simulation model. In the left panel empirical density curves are displayed for simulated data of one particular exon on chromosome 1:

1. distribution when the exon is not expressed (Bg + noise): $\log_2[B(t)] + \epsilon_\lambda(t)$, where $I_\lambda(t) = 0$,
2. hypothetical distribution of an expressed exon when the probe effect is negligible (Bg + signal + noise): $\log_2[B(t) + 2^{c_\lambda(t)}] + \epsilon_\lambda(t)$, with $\sigma_p = 0$ and $\mu_{c,\lambda}(t) = 7$,
3. distribution of a simulated expressed exon with probe effect (Bg + signal + probe + noise): $\log_2[B(t) + 2^{c_\lambda(t) + p_\lambda(t)}] + \epsilon_\lambda(t)$, using $I_\lambda(t) = 1$, $\mu_{c,\lambda}(t) = 7$ and $\sigma_p = 1.5$.

For comparison reasons, the same values are used for the background $B(t)$, mean exonic signal $c_\lambda(t)$ and noise $\epsilon_\lambda(t)$. The signal of the expressed exon $c_\lambda(t)$ shifts the background distribution towards higher values and sharpens the distribution. The additional probe effect clearly introduces a huge variability among the probe intensities, which target the same exon. In the right panel of Figure 8, the distributions of the simulated probe intensities for the same exon on different arrays are given ($I_\lambda(t) = 1$, $\mu_{c,\lambda}(t) = 7$ for WT, $\mu_{c,\lambda}(t) = 9$ for E2F, $\sigma_p = 1.5$). The difference in mean exon effect $c_\lambda(t)$ on each array affects both the shape and the location of the distribution.

Figure 9 and 10 display the segmentation results for TD and DE in a simulated genomic region using different methods.

Appendix D: Huber method for determining the threshold for transcript discovery (δ_{TD})

Following David et al. (2006), we consider the distribution of $\log_2$ transformed background corrected intensities for probes that do not overlap with annotated features (Figure 11, black dashed curve). The distribution has a sharp peak at the left, corresponding to background probes, and a flat shoulder on the right, corresponding to probes targeting unannotated transcripts. We consider this distribution a mixture of a normal distribution (the “peak”) and some other distribution (the “shoulder”). The mean, μ_{BG} , of the normal component is estimated by the mode of the mixture distribution and its variance by the MAD of the distribution obtained by mirroring the part of the mixture with values μ_{BG} about $x = \mu_{BG}$. If the normal component (solid red line in Figure 11) is assumed to be the null distribution of probes that measure background intensities, then we can derive the BFDR for background probes. We select the background threshold δ_{TD} that corresponds to a BFDR of 0.1%.

References

- Antoniadis A. “Wavelet methods in statistics: Some recent developments and their applications,”. *Statistics Surveys*. 2007; 1:16–55. URL [doi:10.1214/07-SS014](https://doi.org/10.1214/07-SS014).
- Benjamini Y, Hochberg Y. “Controlling the false discovery rate: A practical and powerful approach to multiple testing,”. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995; 57:289–300. URL <http://dx.doi.org/10.2307/2346101>.
- Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. “Global Identification of Human Transcribed Sequences with Genome Tiling Arrays,”. *Science*. 2004; 306:2242–2246. [PubMed: 15539566]
- Blais A, Dynlacht BD. “E2f-associated chromatin modifiers and cell cycle control,”. *Current Opinion in Cell Biology*. 2007; 19:658–662. cell differentiation / Cell division, growth and death. [PubMed: 18023996]
- Bruce AG, Gao H-Y. “Understanding waveshrink: Variance and bias estimation,”. *Biometrika*. 1996; 83:727–745.
- Clyde M, Parmigiani G, Vidakovic B. “Multiple shrinkage and subset selection in wavelets,”. *Biometrika*. 1998; 85:391–401. URL <http://dx.doi.org/10.1093/biomet/85.2.391>.
- David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. “A high-resolution map of transcription in the yeast genome,”. *Proceedings of the National Academy of Science*. 2006; 103:5320–5325.
- De Veylder L, Beeckman T, Inzé D. “The ins and outs of the plant cell cycle,”. *Nature Reviews Molecular Cell Biology*. 2007; 8:655–665.
- DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. “Use of a cDNA microarray to analyse gene expression patterns in human cancer,”. *Nature genetics*. 1996; 14:457–460. URL <http://dx.doi.org/10.1038/ng1296-457>. [PubMed: 8944026]
- Donoho D, Johnstone IM. “Adapting to unknown smoothness via wavelet shrinkage,”. *Journal of the American Statistical Association*. 1995; 90:1200–1224.
- Efron B. “Robbins, empirical bayes and microarrays.pdf,”. *Annals of Statistics*. 2003; 31:366–378.
- Figueiredo M, Nowak R. “Wavelet-based image estimation: An empirical bayes approach using jeffreys noninformative prior,”. *IEEE Transactions on Image Processing*. 2001; 10:1322–1331. [PubMed: 18255547]
- Halasz G, van Batenburg M, Perusse J, Hua S, Lu X-J, White K, Bussemaker H. “Detecting transcriptionally active regions using genomic tiling arrays,”. *Genome Biology*. 2006; 7:R59. [PubMed: 16859498]
- Hollander, M.; Wolfe, DA. *Nonparametric Statistical Methods*, 2nd Edition. Wiley-Interscience; New York: 1999.
- Huber W, Toedling J, Steinmetz LM. “Transcript mapping with high-density oligonucleotide tiling arrays,”. *Bioinformatics*. 2006; 22:1963–1970. [PubMed: 16787969]

- Ibrahim JG, Laud PW. "On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior,". *Journal of the American Statistical Association*. 1991; 86:981–986.
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data,". *Biostatistics*. 2003; 4:249–264. URL <http://dx.doi.org/10.1093/biostatistics/4.2.249>. [PubMed: 12925520]
- Johnstone IM, Silverman BW. "Wavelet threshold estimators for data with correlated noise,". *Journal of the Royal Statistical Society: Series B*. 1997; 59:319–351.
- Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR. "Novel rnas identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22,". *Genome Research*. 2004; 14:331–342. [PubMed: 14993201]
- Kass R, Steffey D. "Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayesian models)," . *Journal of the American Statistical Association*. 1989; 84
- Mallat S. "A theory for multiresolution signal decomposition: The wavelet representation,". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1989; 11:674–693.
- McCarthy DJ, Smyth GK. "Testing significance relative to a fold-change threshold is a TREAT,". *Bioinformatics*. 2009; 25:765–771. [PubMed: 19176553]
- Meinke D, Meinke L, Showalter T, Schissel A, Mueller L, Tzafrir I. "A sequence-based map of arabidopsis genes with mutant pheno-types,". *Plant Physiology*. 2003; 132:409–418. [PubMed: 12586866]
- Morris J, Brown P, Herrick R, Baggerly K, Coombes K. "Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models,". *Biometrics*. 2008; 64:479–489. [PubMed: 17888041]
- Naouar N, Vandepoele K, Lammens T, Casneuf T, Zeller G, van Hummelen P, Weigel D, Raetsch G, Inzé D, Kuiper M, De Veylder L, Vuylsteke M. "Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes,". *Plant Journal*. 2009; 57:184–194. [PubMed: 18764924]
- Newton M, Noueiry A, Sarkar D, Ahlquist P. "Detecting differential gene expression with a semiparametric hierarchical mixture method,". *Biostatistics*. 2004; 5
- Purdum E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. "Firma: a method for detection of alternative splicing from exon array data,". *Bioinformatics*. 2008; 24:1707–1714. URL <http://dx.doi.org/10.1093/bioinformatics/btn284>. [PubMed: 18573797]
- Ramsay, J.; Silverman, B. *Functional data analysis*. Springer; 2005. Springer series in statistics
- Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M. "Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping,". *Trends in Genetics*. 2005; 21:466–475. [PubMed: 15979196]
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric regression*. Cambridge University Press; Cambridge, New York: 2003.
- Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes,". *Proceedings of the National Academy of Sciences*. 1996; 93:10614–10619. URL <http://www.pnas.org/content/93/20/10614.abstract>.
- Thisted, R. *Elements of Statistical Computing*. Chapman and Hall; London: 1988.

Figure 1.
 Along-chromosome plot of array intensities of E2F-DPa_{OE} plants (red) and wild type plants (black). The different replicates for WT and E2F-DPa_{OE} are indicated by °, + and Δ.

Figure 2. MML thresholding (solid line) rule versus soft (dash dotted line) and hard thresholding (dashed line).

Figure 3.

ROC curves for transcript discovery (left panel) and differential expression (right panel). The symbol indicates the result corresponding to Table (1). (EB: method using variance estimator (21), EB_{MAD}: method using the MAD estimator, EB_{JK}: method with variance estimator (20), EB_b: EB with bootstrap correction, EB_{imp}: method using an improper prior on $\tau_m(j, k)$ and variance estimator (21), Huber: Huber method for transcript discovery and Kampa: Kampa method for transcript discovery).

Figure 4.

Along-chromosome plot of an unannotated region, which seems to be upregulated in E2F-DPa_{OE} plants. Array intensities of E2F-DPa_{OE} plants (red) and wild type plants (black) are given in the top panel and different replicates for WT and E2F-DPa_{OE} are indicated by °, + and Δ. The bottom panels depict the mean model track $\beta_1(t)$ (M-Track) and a track for the log₂ fold change $FC(t) = 2 \times \beta_2(t)$ (FC-Track). 95% credible intervals are indicated with light blue lines, discovered transcripts and differentially expressed regions are indicated by black boxes on the M and FC track, respectively.

Figure 5.

Estimation of the parameters of the background signal using E2F-DPa_{OE} data. The solid line represents the empirical distribution of the E2F-DPa_{OE} data in non-annotated regions. The dashed red line represents the estimated background distribution ($\hat{\mu}_B=6.35$ and $\hat{\sigma}_B=0.15$).

Figure 6.

Comparison of simulated data and E2F-DPa_{OE} data. The left panel displays the empirical distribution of $\log_2$ -transformed raw probe intensities in annotated regions on chromosome 1 (Solid black line: simulated data, red dashed line: E2F-DPa_{OE} data). The middle and right panel show an auto-correlation plot of data on the first chromosome from a real and a simulated array, respectively.

Figure 7.

Along-chromosome plot of a simulated region (top panel) and E2F-DPa_{OE} experiment (bottom panel). The different replicates for WT (black) and E2F-DPa_{OE} (red) are indicated by °, + and ▲. In the simulation experiment the third gene was differentially expressed.

Figure 8.

Empirical density curves for simulated data of one particular exon on chromosome 1. The left panel displays the empirical distribution of $\log_2$ -transformed raw probe intensities when (1) the exon is not expressed: $\log_2[B(t)] + \epsilon_f(t)$, ($I_f(t) = 0$, dashed line), (2) the probe effect is negligible: $\log_2[B(t) + 2^{c_f(t)}] + \epsilon_f(t)$, ($I_f(t) = 1$, $\mu_{c_f(t)} = 7$, $\sigma_p = 0$, dashed-dotted line), (3) expressed exon with probe effect: $\log_2[B(t) + 2^{c_f(t)+p(t)}] + \epsilon_f(t)$, ($I_f(t) = 1$, $\mu_{c_f(t)} = 7$, $\sigma_p = 1.5$, solid line). For comparison reasons, the same values are used for the background $B(t)$, mean exonic signal $c_f(t)$ and noise $\epsilon_f(t)$. The right panel shows the distribution of the simulated probe intensities for the same exon on 6 different arrays ($I_f(t) = 1$, $\mu_{c_f(t)} = 7$ for WT, $\mu_{c_f(t)} = 9$ for E2F, $\sigma_p = 1.5$; WT plants are represented with a solid line and E2F plants with a dashed line).

Figure 9.

Along-chromosome plot of a simulated region with model tracks for transcript discovery. Top panel: background corrected and quantile normalized simulated array intensities of E2F-DPaOE plants (black) and WT plants (red), boxes indicate the annotation of the expressed exons. The different replicates for WT and E2F-DPaOE are indicated by °, + and Δ. Bottom panels: Model tracks with black boxes indicating the discovered expressed transcripts. (EB: method using variance estimator (21), Huber: Huber method for transcript discovery, Kampa: Kampa method for transcript discovery, the other wavelet based functional models resulted in the same segmentation as EB).

Figure 10.

Along-chromosome plot of a simulated region with model tracks for differential expression. Top-panel: background corrected and quantile normalized simulated array intensities of E2F-DPaOE plants (black) and WT plants (red), boxes indicate the annotation of the differentially expressed exons. The different replicates for WT and E2F-DPaOE are indicated by °, + and Δ. Bottom panels: Different model tracks with boxes indicating the discovered differentially expressed regions. (EB: method using variance estimator (21), EB_b: EB with bootstrap correction and EB_{imp}: method using an improper prior on the random effect variances $\tau_m(j, k)$ and variance estimator (21)).

Figure 11.

Distribution of the $\log_2$ transformed background corrected intensities for unannotated (back dashed) and annotated probes (blue dashed). The estimated distribution for background probes (solid red) is derived from the distribution of the unannotated probes. The red and blue vertical lines represent the threshold for transcript discovery (δ_{TD}) by the Huber method and the one proposed by the biologists.

Table 1

Comparison of the performance of wavelet based methods with the Kampa, Huber and RMA method in terms of positive predictive value (ppv), sensitivity (sens), specificity (spec) and computational time per chromosome (time). ppv, sens and spec are given at probe level except for RMA for which they are calculated on gene level. Note, that RMA method heavily relies on the existing annotation while the other methods or unbiased towards existing annotation.

Model	Biased by annot.	Transcript Discovery			Differential Expression			Time (s/chr)	
		ppv	sens	spec	ppv	sens	spec		
						probe	gene		
EB	no	98.9	82.3	99.4	98.8	82.4	98.1	100	34.0
EB _{MAD}	no	99.6	80.0	99.8	99.3	75.0	97.1	100	13.5
EB _{jk}	no	99.0	81.8	99.5	99.0	81.0	97.8	100	35.4
EB _b	no	99.0	81.9	99.4	99.5	78.1	97.6	100	1023.1
EB _{imp}	no	99.2	81.5	99.6	99.9	75.7	96.9	100	34.0
Kampa	no	85.0	82.4	90.5	-	-	-	-	27.3
Huber	no	85.3	76.7	91.5	-	-	-	-	6321.1
RMA	yes	-	-	-	100	-	83.0	100	32.1

Table 2

Transcript discovery and differential expression for all chromosomes of *Arabidopsis thaliana* in the E2F-DPaOE experiment.

Chromosome	Transcript discovery		Differential expression	
	Detected	Non-annotated	Detected	Non-annotated
1	20420	3664	975	204
2	12050	2801	576	86
3	15105	2958	727	178
4	12017	2382	626	109
5	18071	3344	981	188
1–5	77663	15149	3885	765