De Beuf *et al. BMC Bioinformatics* 2012, **13**:234
http://www.biomedcentral.com/1471-2105/13/234

BMC Bioinformatics

SOFTWARE Open Access

# Analysis of tiling array expression studies with flexible designs in Bioconductor (waveTiling)

Kristof De Beuf[1*], Peter Pipelers[1], Megan Andriankaja[2,3], Olivier Thas[1], Dirk Inzé[2,3], Ciprian Crainiceanu[4] and Lieven Clement[1,5*]

## Abstract

**Background:** Existing statistical methods for tiling array transcriptome data either focus on transcript discovery in one biological or experimental condition or on the detection of differential expression between two conditions. Increasingly often, however, biologists are interested in time-course studies, studies with more than two conditions or even multiple-factor studies. As these studies are currently analyzed with the traditional microarray analysis techniques, they do not exploit the genome-wide nature of tiling array data to its full potential.

**Results:** We present an R Bioconductor package, waveTiling, which implements a wavelet-based model for analyzing transcriptome data and extends it towards more complex experimental designs. With waveTiling the user is able to discover (1) group-wise expressed regions, (2) differentially expressed regions between any two groups in single-factor studies and in (3) multifactorial designs. Moreover, for time-course experiments it is also possible to detect (4) linear time effects and (5) a circadian rhythm of transcripts. By considering the expression values of the individual tiling probes as a function of genomic position, effect regions can be detected regardless of existing annotation. Three case studies with different experimental set-ups illustrate the use and the flexibility of the model-based transcriptome analysis.

**Conclusions:** The waveTiling package provides the user with a convenient tool for the analysis of tiling array trancriptome data for a multitude of experimental set-ups. Regardless of the study design, the probe-wise analysis allows for the detection of transcriptional effects in both exonic, intronic and intergenic regions, without prior consultation of existing annotation.

## Background

In the last few years tiling microarrays have become a well-established tool for whole-genome transcriptome analysis. They have shown to be very useful for exploring and unraveling the complex genome-wide trancriptional landscape of higher organisms, in which not only protein coding genes, but also non-coding RNAs play an important role [1-4]. The methods that have been developed for transcriptome analysis with tiling arrays either focus on segmentation and transcript discovery within a single biological condition [5-8], or on the detection of differential expression between two distinct conditions [9,10]. Recently, the focus in tiling array studies has shifted towards more complex experimental designs, such as studies with more than two conditions [11] and studies with several experimental factors [12]. Furthermore, it is recognized that expression is a dynamic rather than a static phenomenon. Hence, more and more time-course experiments are designed to provide insights into the whole-genome transcript regulation of species during different developmental stages or external periodic changes in the environment [13,14].

Currently, most tiling array transcriptome analysis pipelines start with summarization of the probe-level data. This can be done by constructing probesets from the groups of probes that map to known annotated genes, (e.g. [11,15]). Hereby unannotated regions are disregarded. In [12,13,16] a sliding window-based approach is

*Correspondence: kristof.debeuf@ugent.be; lieven.clement@gmail.com
[1] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B9000 Ghent, Belgium
[5] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven and Universiteit Hasselt, Kapucijnenvoer 35, Blok D, bus 7001, B3000 Leuven, Belgium
Full list of author information is available at the end of the article

adopted, combined with a thresholding rule for selecting transcriptional units, whereas in [14] segments with piece-wise constant intensity levels are constructed first [17]. After the summarization a statistical test or a more heuristic analysis technique is conducted on the summarized expression values of the transcriptional units. In current time-course and single-factor studies this is merely done by directly applying traditional microarray analysis methods, such as a pairwise moderated t-test (Limma) [18] conducted in [11] or a permutated t-test (SAM) [19] conducted in [16]. Other studies adopt ad-hoc approaches to filter the genes or transcriptional units of interest. Transcriptional units in a time-course experiment, for example, can be filtered based on thresholding the amplitude of the signal [20]. In an alternative approach the correlations between temporal expression patterns are explored and a clustering is performed of genomic regions based on expression profiles in different gene classes showing expression at different time-points [21]. The tests reported in [13] and [14] on the other hand are less ad hoc, but very specific for the periodic time-course design apparent in these studies [22-24]. The aforementioned methods either lack flexibility by only focusing on one specific experimental design, or they first summarize probes to probesets based on existing annotation, hence not exploiting the genome-wide nature of the data to the full extent.

Here, we present waveTiling, a R Bioconductor package for transcriptome analysis of tiling arrays with flexible designs. The package is based on and provides an extension to a recently introduced wavelet-based functional model for transcriptome analysis [25]. While the methodology in [25] was initially developed to conduct the simultaneous tasks of transcript discovery and detection of differential expression, their framework can be easily extended by adapting the model design matrix. After modeling the specific effect function of interest, probe-wise inference can be conducted for detecting affected regions. The probe-wise analysis allows for the detection of transcriptional units in both exonic, intronic and intergenic regions, without prior consultation of existing annotation. Currently, waveTiling provides a standard analysis flow for transcriptome analysis on single-factor experiments with two or more biological conditions, the detection of linear and quadratic effects and circadian rhythms in time-course experiments, and the analysis of two-factor experiments, while more experienced users can also specify customized designs. Furthermore, it generates along-genome plots and contains functions to easily extract the detected genes and unannotated regions. The *Implementation* section gives an overview of the main functionalities of the waveTiling package and describes the model for the different designs, as well as the associated inference procedures. In *Results and Discussion* we illustrate the use of the package and the model on three different case studies with very distinct experimental designs.

## Implementation

The waveTiling package is an add-on package to the Bioconductor project [26] written in the programming language and statistical environment R [27]. It provides all the tools necessary to conduct a full analysis of tiling microarray experiments for flexible designs based on the recently introduced wavelet-based functional model for trancriptome analysis [25]. The package uses the standard Bioconductor S4-class data structures making it fully compatible with existing packages. The data is imported with the aid of the oligo-package [28] and the resulting object inherits from *TilingFeatureSet*, which is specifically designed for representing tiling array data and in turn extends *ExpressionSet*. Existing instance methods from oligo and other Bioconductor packages supporting this structure are therefore applicable as well. Before starting the analysis the probes can be remapped to the existing annotation. Moreover, probes that contain duplicated sequences for perfect match and mismatch probes or for probes on different strands can be filtered because they are deemed unreliable due to cross-hybridization effects. The main transcriptome analysis consists of two consecutive steps: (1) fitting the wavelet-based functional model to the data, and (2) model-based inference to identify transcriptionally affected regions. The fitted model is stored in a *WfmFit*-class object. Depending on the design of the study a *WfmFitFactor* (factorial design), *WfmFitTime* (time-course design), *WfmFitCircadian* (circadian rhythm design) or *WfmFitCustom* (custom design) subclass is used. Part of the code for fitting the model is implemented in C to speed up computation. In the second step, different inference procedures can be conducted depending on the research question. The inference procedure that can be conducted depends on the *WfmFit*-subclass. The results are stored as a *WfmInf*-class object. There are 3 main subclasses: *WfmInfCompare* which contains the results of a pairwise comparison between two groups or time points; *WfmInfMeans* with the results of transcript discovery for each individual group or time point; and *WfmInfEffects* which contains results with linear or quadratic time effects for time-course designs and circadian rhythm effects for circadian designs. All transcriptionally affected regions can be extracted from the *WfmInf*-class objects and are stored as *IRanges*-class objects [29]. The model fitting and inference steps are described in more detail in the *Statistical Methods* part.

The results can be visually explored by means of a general plot function. The implementation is based on the GenomeGraphs package [30]. For any genomic region the fitted expression values and transcriptionally affected

regions can be plotted along the genomic coordinate. Furthermore, two functions are available for further post-processing of the results. Provided a suitable annotation file is given, the transcriptionally affected regions are mapped against the existing annotation. The first function outputs the genes that are transcriptionally affected, while the second function provides a list of the detected unannotated regions. The output of both functions is a list of *GRanges*-class objects [31].

## Statistical Methods

We start by presenting an overview of the basic model introduced by [25]. Subsequently, we show how we accomodate for several sampling schemes in time-course experiments or other experiments with more flexible designs.

### Basic wavelet-based model for transcriptome analysis

We consider the functional model designed for the detection of (differentially) expressed regions in experiments with two biological conditions. It is given by [25]

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

with $i = 1, ..., N$, $Y_i(t)$ the measured $\log_2$-transformed expression values for the probe with position $t$ ($t = 1, ..., T$) on array $i$ ($i = 1, ..., N$). $T$ is the number of probes that are more or less equally spaced along the genomic position of the chromosome, and $N = N_1 + N_2$ is the number of tiling arrays in the experiment, with $N_1$ the number for biological condition 1, say $C_1$, and $N_2$ the number for biological condition 2, say $C_2$. Further, $X_{1,i}$ is a dummy variable which is 1 for $C_1$ and $-1$ for $C_2$, and $E_i(t)$ is a zero mean error term. It is assumed that $E_i(1), ..., E_i(T)$ are jointly $MVN(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$. Here, $MVN(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}$.

The model can also be written as

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}. \tag{2}$$

In this model, $\boldsymbol{Y}$ is an $N \times T$ matrix of measured $\log_2$-transformed expression values, containing the elements $Y_i(t)$ for the probe with genomic position $t$ ($t = 1, .., T$) on array $i$ ($i = 1, ..., N$). Further, $\boldsymbol{E}$ is an $N \times T$ error matrix containing the errors terms $E_i(t)$ for probe position $t$ on array $i$. The $N \times 2$ design matrix $\boldsymbol{X}$ is constructed as

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} \end{bmatrix},$$

where the upper row represents the dummy coding for the $N_1$ arrays in $C_1$ and the lower row the dummy coding for the $N_2$ arrays in $C_2$. The $2 \times T$ effect function matrix $\boldsymbol{B}$ contains the probe-wise effect functions $\beta_1(t)$ and $\beta_2(t)$ on the respective rows. Column 1 of $\boldsymbol{X}$ will be used to find regions with a mean expression level above some threshold, whereas the coding in column 2 allows for assessing differential expression between the two conditions. Note that the coding in $\boldsymbol{X}$ implies that two effect functions are estimated orthogonally for a balanced study design. This can be seen from

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} N/2 & 0 \\ 0 & N/2 \end{bmatrix},$$

with $N_1 = N_2 = \frac{N}{2}$.

Before estimating the effect functions, the expression data are projected onto the wavelet space by using the discrete wavelet transform (DWT). This linear projection can be written as the matrix multiplication $\boldsymbol{D} = \boldsymbol{YW}^T$, where $\boldsymbol{W}$ is an orthogonal DWT matrix. This allows us to rewrite model (2) in the wavelet space as

$$\boldsymbol{D} = \boldsymbol{XB}^* + \boldsymbol{E}^*, \tag{3}$$

where the rows of the $N \times T$ matrix $\boldsymbol{D}$ contain the wavelet coefficients for each array, double-indexed by location $k = 1, ..., K_j$ and scale $j = 0, ..., J$. The $2 \times T$ and $N \times T$ matrices $\boldsymbol{B}^*$ and $\boldsymbol{E}^*$ contain the wavelet coefficients for the effect functions and the error terms, respectively. By putting a normal prior on the effect functions in the wavelet space, this model can also be written as

$$\begin{cases} \boldsymbol{D}(j,k)|\boldsymbol{\beta}^*(j,k) & \sim & MVN\left\{\boldsymbol{X\beta}^*(j,k), \boldsymbol{I}\sigma_\epsilon^2(j,k)\right\} \\ \beta_m^*(j,k)|\tau_m(j,k) & \sim & N\left\{0, \tau_m(j,k)\sigma_\epsilon^2(j,k)\right\} \end{cases}, \tag{4}$$

where $\beta_m^*(j,k)$ is the element of $\boldsymbol{B}^*$ corresponding to scale $j$ and location $k$ and $m = 1, 2$. In (4) $N(\mu, \sigma^2)$ denotes the density function of a normal distribution with mean $\mu$ and variance $\sigma^2$. The smoothing parameters $\tau_m(j,k)$ and the error variances $\sigma_\epsilon^2(j,k)$ are estimated by marginal maximum likelihood using a Gauss-Seidel algorithm. The estimated $\hat{\tau}_m(j,k)$ induce a regularization of the wavelet coefficients of the effect functions. When backtransforming the modified coefficients to the original data space, this leads to a denoised expression signal whereby the main features are retained. The method has proven to be very fast which is essential when analyzing large datasets. For more details, see [25].

### Wavelet-based models for transcriptome analysis in more flexible designs

To extend the modeling framework reviewed in the previous section and to make it suitable for the analysis of tiling array data with more flexible designs, the design matrix

$\boldsymbol{X}$ needs to be adapted in an appropriate way. Firstly, the adaptation must enable the model to answer the specific research questions provoked by the experimental design. Secondly, it must allow us to use the same fast algorithms introduced in [25]. This second argument comes down to the preservation of the orthogonality of $\boldsymbol{X}$. In the first part of this section we focus on general time-course designs and single-factor designs for more than 2 groups. The second part aims at specific time-course designs for assessing circadian rhythms in the transcriptome. The section concludes with looking specifically at non-orthogonal designs, typically encountered in multi-factor studies.

#### *General time-course designs*

In tiling array time-course experiments one is often interested in the detection of differentially expressed regions between any two different time points. An additional concern might be to detect significant effects of transcriptional activity in time, e.g. linearly increasing or decreasing transcriptional expression of certain regions. These two possible research aims can be dealt with by considering a functional relationship of the designed time points described by orthogonal polynomials. This approach has also been used in quantitative trait associated expression studies based on traditional microarrays [32]. In that paper the functional relationship with phenotype is considered instead of with time.

Consider a time-course experiment with whole-genome expression levels measured at $q$ time points. Let $N$ be the total number of arrays used in the experiment. The number of arrays used for each time point is represented by $N_1, \ldots, N_q$, with $N_1 + \ldots + N_q = N$. In this exposition we only consider balanced designs, i.e. $N_1 = \ldots = N_q$, with equidistant time points. However, it is rather straightforward to obtain orthogonal polynomials when dealing with non-balanced and non-equidistance designs. A simple procedure is discussed in [33]. The design matrix $\boldsymbol{X}$ in model (2) now has dimensions $N \times q$ and can be written as

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \psi_1(\boldsymbol{X}_1) & \psi_2(\boldsymbol{X}_1) & \cdots & \psi_{q-1}(\boldsymbol{X}_1) \\ \mathbf{1} & \psi_1(\boldsymbol{X}_2) & \psi_2(\boldsymbol{X}_2) & \cdots & \psi_{q-1}(\boldsymbol{X}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \psi_1(\boldsymbol{X}_q) & \psi_2(\boldsymbol{X}_q) & \cdots & \psi_{q-1}(\boldsymbol{X}_q) \end{bmatrix}, \quad (5)$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ are the $N_1, \ldots, N_q$-valued vectors that correspond with the $q$ respective designed time points in the experiment. In (5) each function $\psi_j(\boldsymbol{x})$ is a polynomial of degree $j$, with $j = 0, \ldots, q-1$, and is orthogonal to $\psi_k(\boldsymbol{x})$ $(k = 0, \ldots, q-1)$ if $j \neq k$. Note that each $\mathbf{1}$ in the first column of $\boldsymbol{X}$ can also be seen as $\psi_0(\boldsymbol{X}_i)$ $(i = 1, \ldots, q)$. The orthogonality of $\boldsymbol{X}$ is clear from

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} N & 0 & 0 & 0 & \ldots & 0 \\ 0 & \sum_{i=1}^N \psi_1^2(\boldsymbol{X}_i) & 0 & 0 & \ldots & 0 \\ 0 & 0 & \sum_{i=1}^N \psi_2^2(\boldsymbol{X}_i) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \sum_{i=1}^N \psi_{q-2}^2(\boldsymbol{X}_i) & 0 \\ 0 & 0 & 0 & \ldots & 0 & \sum_{i=1}^N \psi_{q-1}^2(\boldsymbol{X}_i) \end{bmatrix}. \quad (6)$$

With this design matrix a $q \times N$ matrix $\boldsymbol{B}$ with the $q$ effect functions is associated. The first row of $\boldsymbol{B}$ corresponds with an overall mean expression level over all time points, while row 2 until $q$ are associated with a linear, quadratic, cubic, ..., $(q-1)$-th order polynomial effect respectively between the different time points. The fitted expression levels for each time point are obtained by a linear combination of all effect functions in accordance with model (2). This allows for a straightforward comparison between any two time points. When combining several effect functions, it may be desirable to induce the same amount of smoothing for each of them. This implies the estimation of one general smoothing parameter $\tau(j,k)$, instead of a separate $\tau_m(j,k)$ for each effect function $(m = 1, ..., q)$. To retain the fast algorithms of [25], however, the diagonal elements of $\boldsymbol{X}^T\boldsymbol{X}$ need to be identical in this case. This can be obtained by normalizing each column vector of $\boldsymbol{X}$ to give the normalized design matrix $\boldsymbol{X'}$. This leads to the property

$$\boldsymbol{X'}^T\boldsymbol{X'} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix} = \boldsymbol{I}_q, \quad (7)$$

where $\boldsymbol{I}_q$ is an $q \times q$ identity matrix. For this orthonormal design matrix $\boldsymbol{X'}$ it can be shown that the common smoothing parameter can be estimated by

$$\hat{\tau}(j,k) = \left[ \frac{\boldsymbol{D}^T(j,k)\boldsymbol{X'}\boldsymbol{X'}^T\boldsymbol{D}(j,k)}{q\sigma_\epsilon^2(j,k)} \right]_+. \quad (8)$$

Although design matrix (5) can also be used for non-ordered single factor studies, one may choose to use a design matrix specifically constructed for unordered factors, e.g. a Helmert contrast design matrix. Helmert contrasts are basically designed to compare the mean expression at a specific time point with the overall mean over all preceding time points. The main reason why we use them here, however, is that they also lead to estimation orthogonalities for the effect functions. This is seen from

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} N & 0 & 0 & 0 & \dots & 0 \\ 0 & \sum_{i=1}^{2} N_i & 0 & 0 & \dots & 0 \\ 0 & 0 & 2\sum_{i=1}^{3} N_i & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & (q-2)\sum_{i=1}^{q-1} N_i & 0 \\ 0 & 0 & 0 & \dots & 0 & (q-1)\sum_{i=1}^{q} N_i \end{bmatrix}. \tag{9}$$

Just like for the polynomials, the design matrix $\boldsymbol{X}$ based on Helmert contrasts still needs to be normalized if the same smoothing for all factor effects is desired.

#### *Designs for circadian rhythms*

Suppose now that we are interested in the detection of a certain circadian rhythm in the transcriptome of an organism, based on an equally spaced time-course experiment. A natural way to model the circular effect is to construct $\boldsymbol{X}$ by means of Fourier basis functions, instead of polynomial basis functions. The design matrix is then given by

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \sin(\mathbf{0}) & \cos(\mathbf{0}) \\ \mathbf{1} & \sin(\frac{\mathbf{2\pi}}{\boldsymbol{q}}) & \cos(\frac{\mathbf{2\pi}}{\boldsymbol{q}}) \\ \mathbf{1} & \sin(\frac{\mathbf{4\pi}}{\boldsymbol{q}}) & \cos(\frac{\mathbf{4\pi}}{\boldsymbol{q}}) \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \sin(\mathbf{2\pi} - \frac{\boldsymbol{\pi}}{\boldsymbol{q}}) & \cos(\mathbf{2\pi} - \frac{\boldsymbol{\pi}}{\boldsymbol{q}}) \end{bmatrix} \tag{10}$$

Again the separate effect functions can be estimated orthogonally, which is seen from

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} N & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & q \end{bmatrix}. \tag{11}$$

To estimate a common smoothing parameter for inducing the same amount of smoothing for all effect functions, $\boldsymbol{X}$ can again be normalized as described previously.

#### *Non-orthogonal designs*

Design matrices for two- or multiple-factor designs are typically non-orthogonal. Using these in the wavelet-based model would imply that the fast algorithms presented in [25] would have to be adapted. This would lead to undesirably increased computation time during parameter estimation. A solution to this problem is to apply the Gram-Schmidt process to orthogonalize $\boldsymbol{X}$ and subsequently estimate the model parameters based on the orthogonalized design matrix. The Gram-Schmidt orthogonalization comes down to a QR-decomposition [34] of $\boldsymbol{X}$ into an upper-triangular matrix $\boldsymbol{X}_{tri}$ and an orthogonal matrix $\boldsymbol{X}_{orth}$, which is now used to fit the model. Afterwards, the estimated parameters have to be transformed back to obtain the parameter values for the original $\boldsymbol{X}$. This is possible by premultiplying them with $(\boldsymbol{X}_{orth}^T\boldsymbol{X})^{-1}$. Similar to single-factor and time-course designs, the coding of the initial design matrix $\boldsymbol{X}$ still determines how the parameters can be interpreted, and may thus be constructed according to the specific research interest.

### Statistical inference: detection of transcriptional effect regions

Depending on the study design and the aim of the analysis, either the parameters themselves or a function of the parameters are used to detect transcriptional effect regions. In both instances, the effect of interest can be represented by $F\{\boldsymbol{\beta}(t)\}$. For general time-course designs one can be interested in detecting genomic regions that show a linear or a quadratic trend in time. In this situation $F\{\boldsymbol{\beta}(t)\}$ is just the effect function $\boldsymbol{\beta}(t)$ that corresponds with either the linear polynomial term $\psi_1(\boldsymbol{X})$ or the quadratic polynomial term $\psi_2(\boldsymbol{X})$ in (5). On the other hand, if interest lies in the detection of differentially expressed regions between different time points, inference is performed on each row of a $\frac{q(q-1)}{2} \times T$ matrix $\boldsymbol{ZXB}$, where $\boldsymbol{Z}$ is a $\frac{q(q-1)}{2} \times N$ contrast matrix indicating the specific time points to be contrasted. Hence, each row of $\boldsymbol{ZXB}$ corresponds with one of the $\frac{q(q-1)}{2}$ possible pairwise comparisons between two time points and gives rise to an effect function $F\{\boldsymbol{\beta}(t)\}$ for each desired comparison. In circadian rhythm designs the sine and the cosine effect functions are combined to give the amplitude $\boldsymbol{A}(t)$ of the circadian rhythm per probe position, i.e.

$$F\{\boldsymbol{\beta}(t)\} = \boldsymbol{A}(t) = \sqrt{\boldsymbol{\beta}_2^2(t) + \boldsymbol{\beta}_3^2(t)}. \tag{12}$$

Based on the size of $\boldsymbol{A}(t)$ circadian effect regions can be detected. In the case of non-orthogonal designs in multiple-factor studies, there are several possibilities for the choice of $F\{\boldsymbol{\beta}(t)\}$, depending on the aim of the analysis. The idea remains the same, however.

For each genomic location $t$, $F\{\boldsymbol{\beta}(t)\}$ is compared to a certain threshold value $\delta$ which can be chosen freely by the biological researcher. A Bayesian FDR procedure [35]

is adopted to evaluate statistical significance. This may be written as

$$FDR_F(t) = P\left[F\left\{\boldsymbol{\beta}(t)\right\} < \delta | \boldsymbol{Y}\right]. \quad (13)$$

It basically involves the calculation of a probability mass from a univariate normally distributed random variable if $F\{\boldsymbol{\beta}(t)\}$ contains only one $\boldsymbol{\beta}(t)$, or from a multivariate normally distributed random variable if $F\{\boldsymbol{\beta}(t)\}$ contains a linear combination of $\boldsymbol{\beta}(t)$'s [25]. The variance-covariance matrix is readily available if $\boldsymbol{X}$ is orthogonal. For non-orthogonal designs it can be calculated by

$$\left\{(\boldsymbol{X}_{orth}^T\boldsymbol{X})^{-1}\right\}^T \mathrm{Var}\left[F\left\{\boldsymbol{\beta}(t)\right\}\right]\left\{(\boldsymbol{X}_{orth}^T\boldsymbol{X})^{-1}\right\}.$$

For the circadian rhythms design however, this approach is not possible because of the non-linear dependence of $\boldsymbol{A}(t)$ on the $\boldsymbol{\beta}(t)$'s. In this case $FDR_F(t)$ can be approximated by simulation. In each simulation step we sample from the normal sine and cosine effect functions and calculate $\boldsymbol{A}_{\mathrm{sim}}(t)$. $FDR_F(t)$ is now given by the proportion of simulations for which $\boldsymbol{A}_{\mathrm{sim}}(t) < \delta$. Specifically for differential expression, (13) is used to detect overexpression at probe $t$, while for detecting underexpression at probe $t$ we use

$$FDR_F(t) = P\left[F\left\{\boldsymbol{\beta}(t)\right\} > -\delta | \boldsymbol{Y}\right]. \quad (14)$$

## Results and discussion

The use and flexibility of the waveTiling package is illustrated in three case studies for transcriptome analysis with different experimental set-ups.

### Case study 1: Time-course experiment

The first data set consists of a tiling array expression study for identifying the molecular events associated with early leaf development of the plant species *Arabisopsis thaliana* [11]. Unraveling the underlying mechanisms of on one hand the transition from cell division to cell expansion and on the other hand the transition from non-photosynthetic to photosynthetic leaves, was the focus of this study. Trancriptome analysis for six developmental time points (day 8 to day 13) was conducted with AGRONOMICS1 tiling arrays [36], with three biological replicates per time point. Primarily, the researchers were focusing on the detection of differentially expressed regions between any two pairs of developmental time points. This specific study design, however, also allows for the detection of expression regions that change linearly over time. The functions and code used for this case study are described in more detail in the package vignette (see Additional file 1).

#### *Pairwise comparison*

Figure 1 gives an example of a genomic region on chromosome 1 of *Arabidopsis thaliana* found to be differentially expressed between different time points. The threshold value used here was $|\log_2(1.2)|$. For the most significant time point pairs the detected regions clearly resemble the exons of gene *AT1G04350*, encoding a putative 2-oxoglutarate-dependent dioxygenase (Figure 1).

We evaluate the regions detected by the wavelet-based analysis against the genes produced by the well-established and often used RMA method [37]. This is done by comparing the results of a gene set enrichment analysis based on both methods. By mapping the genomic regions found by the wavelet-based method to the *Arabidopsis thaliana* TAIR9 annotation [38], a list of genes is created for this method. Only genes that showed an overlap of at least 15% with the detected regions were retained. The enrichment analysis as performed with Plaza [39] revealed a strong overlap in the processes detected by both methods. A total of 483 enrichments were identified using both genesets of which 360 common enrichments were shared. The RMA gene list had 75 specific enrichments, while the wavelet-based gene list had 48.

The enrichment analyses revealed a high similarity of genes in common by the two methods for identifying differentially expressed regions of the genome that have previously been annotated. However, we could also discover non-annotated regions that were differentially expressed. We identified a total of 109 unannotated and differentially expressed regions in the genome with a length of at least 200 bp. Selected regions were validated with qRT-PCR to confirm these findings. These regions were chosen based on the following criteria:

1. Region was not in or near an exon or promoter from an annotated gene.
2. Longer regions containing more differentially expressed probes were preferentially selected.
3. Regions showing homogeneous probe directionality (all probes going in the same direction) across the entire region of differential expression were preferentially selected.

Using these criteria 12 regions were selected and qRT-PCR analysis was performed (see Additional file 2: Table S1). Of the 12 regions, 11 could be confirmed to contain differentially expressed transcripts during the time-course analysis. Only 1 region had no detectable transcriptional products. Log fold changes were calculated for confirming the expression and differential expression, as well as the directionality of the differential expression. From this analysis 9 of the 11 regions showed the same log fold change directionality as previously identified from the tiling arrays, and 2 regions showed opposite log fold change directionality. However, these 2 regions had the lowest log fold changes in the wavelet-based analysis. More details about the methods

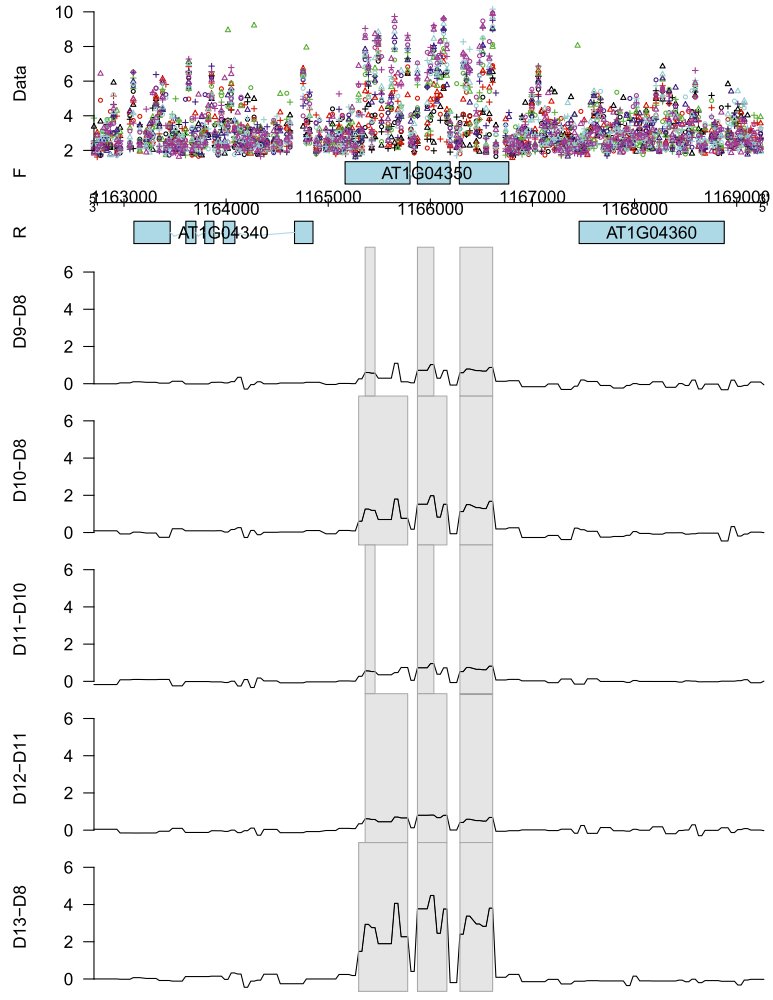


**Figure 1 Pairwise day-to-day differentially expressed genomic region.** Fitted differential expression effect for the genomic region of gene *AT1G04350* on the forward strand of chromosome 1 between selected pairs of developmental time points varying from day 8 (D8) to day 13 (D13). The grey rectangles indicate the detected regions showing a significant differential expression effect. The different replicates are indicated by ∘, + and △, while the different days are represented by different colors: black (D8), red (D9), green (D10), blue (D11), cyan (D12) and magenta (D13).

of enrichment and qRT-PCR analysis can be found in Additional file 2.

### *Linear and quadratic time effects*

In addition to a pairwise comparison analysis, the wavelet-based functional model using the orthogonal polynomial design matrix is also useful for detecting genes with linear and quadratic expression patterns over time. In fact, the estimated parameters now give direct interpretations in terms of the different order time effects. Figure 2 gives some example plots of genes from the forward strand of chromosome 1 with a clear linear or quadratic time effect. From the plots, it is clear that the fitted probe-wise $\log_2$ intensities at the different time points (orange lines) are squeezed to some extent towards the mean fitted $\log_2$ intensities over all probes in the whole detected region at these time points (purple line). The main reason for this is that in the wavelet domain strength is borrowed from the neighboring probes in the genomic region to provide a more reliable estimate for each probe-wise effect.

For two of the genes shown in Figure 2 a more detailed visualization is given of the fitted linear or quadratic time effect along the genomic coordinate of chromosome 1. Figure 3 shows the regions with significant decreasing linear time effects overlapping with gene *AT1G62500*, encoding a putative lipid transfer protein, while Figure 4 shows those regions with a significant quadratic time effect overlapping with gene *AT1G16410*, encoding a cytochrome P450. It is also possible to look at the fitted $\log_2$ intensities at the different time points. This means that we are still able to perform transcript discovery at each time point separately. Figure 5 gives the corresponding plots for the linearly decreasing gene *AT1G62500*. The

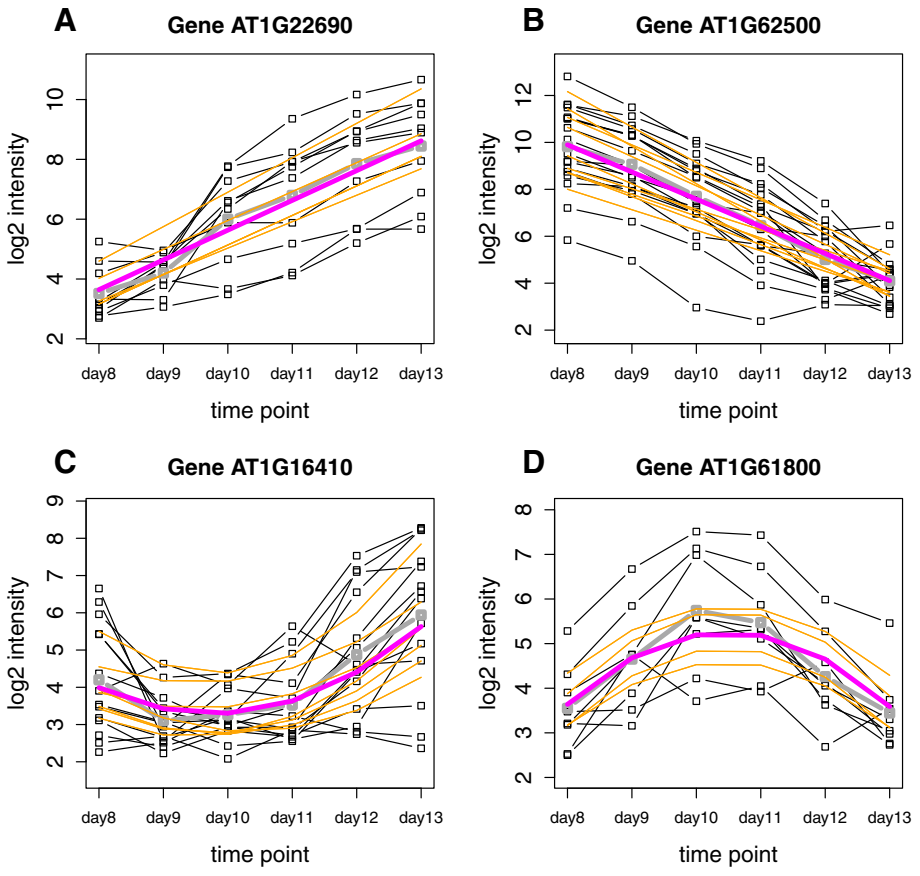


**Figure 2 Gene-wise linear and quadratic effects of transcription levels.** Example plot for two genes showing a linearly increasing (**A**) and decreasing (**B**) mean $\log_2$ intensity level as a function of the 6 days in the time-course. These genes map to two of the top detected regions with a linear time effect for the forward strand of chromosome 1. The mean of the linear time effect parameter estimates corresponding with the probes in these regions are 1.08 and -1.16 respectively. Plots **C** and **D** give two examples of genes with a strong quadratic effect on the forward strand of chromosome 1. The dotted black lines represent the mean observed $\log_2$ expression for the probes over the three biological replicates at the different time points. The dotted grey line is the mean observed $\log_2$ expression over all the probes in the region. The orange lines are the probe-wise fitted $\log_2$ expression values when only considering the intercept and the linear time effect in the model for the two upper-part genes, and considering the intercept, the linear and the quadratic time effect in the model for the two lower-part genes. The purple line gives the corresponding mean fitted $\log_2$ expression values at the different time points over all the probes in the region.

trend apparent in the example plots of Figure 2 is also clear from this figure. The grey rectangles in Figure 5 indicate the discovered regions with mean $\log_2$ intensities significantly above a certain threshold chosen according to the procedure described in [25]. This illustrates that for the discussed models, it is possible to simultaneously detect differentially affected regions between groups as well as transcriptionally active regions for each group - in this case for each day - separately.

### Case study 2: Circadian rhythms

The second case study concerns an expression analysis to examine circadian rhythms in *Arabisopsis thaliana* [13]. It is known that photosynthetic organisms anticipate changes in the daily environment with an internal oscillator, called the circadian clock. The aim of the study was to explore the genome-wide extent of the rhythmic expression patterns governed by this oscillator. In this experiment, 12 samples were collected from *Arabidopsis thaliana* seedlings that were placed under a 12 h light / 12 h dark cycles regime. Every 4 hours 2 samples were taken and hybridized to the Affymetrix AtTile 1.0F and 1.0R tiling arrays. More information about the experiment can be found in [13].

Figure 6 shows an example of the model fit for gene *AT2G46830* with a clear strong circadian effect. This gene has been previously described and is known under the name *CIRCADIAN CLOCK ASSOCIATED1 (CCA1)*. Besides the circadian effects, no other time-dependent effects are considered in the model. Therefore, the fitted $\log_2$ intensities for time points at identical moments in the 24h day/night cycle always coincide. This strong circadian effect is confirmed by Figure 7, which shows the

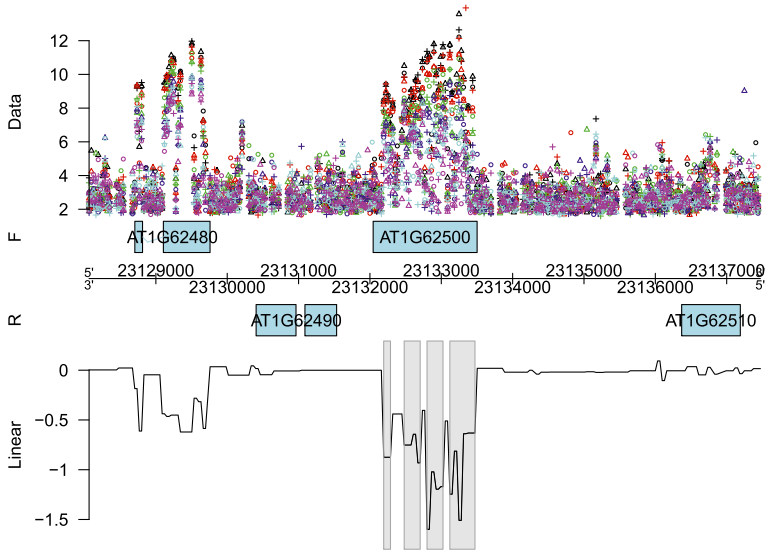


**Figure 3 Genomic region with a linear time effect.** Fitted linear time effect for the genomic region of gene *AT1G62500* on the forward strand of chromosome 1. The different replicates are indicated by ∘, + and △, while the different days are represented by different colors: black (D8), red (D9), green (D10), blue (D11), cyan (D12) and magenta (D13). The grey rectangles indicate the detected regions showing a significant linear time effect, while the black line corresponds with the coefficient function of the linear effect. The negative sign of the coefficients implies a decreasing effect over time. More specifically, the effect at probe $t$ is $\hat{\beta}_1(t) \times time$.

fitted effect for the genomic region of *CCA1*. This effect corresponds with the amplitude of the circadian rhythm, $A(t) = \sqrt{\beta_2^2(t) + \beta_3^2(t)}$, as estimated by the model.

The performance of the wavelet-based method for circadian rhythms is further tested by examining some specific circadian clock associated genes on the forward strand of the *Arabidopsis thaliana* genome [40]. The genes that we consider here were also reported in [13]. The results are shown in Table 1. All genes show a considerable overlap with the genomic regions for which a circadian effect was detected significantly above the threshold value $\log_2(1.1)$, except *TIME FOR COFFEE* (*AT3G22380*). They also have a quite high maximum estimated effect or amplitude size, except *TIME FOR COFFEE* and *ZEITLUPE* (*AT5G57360*). These latter two genes are the only genes from the list that do not fall within the top 20 genes

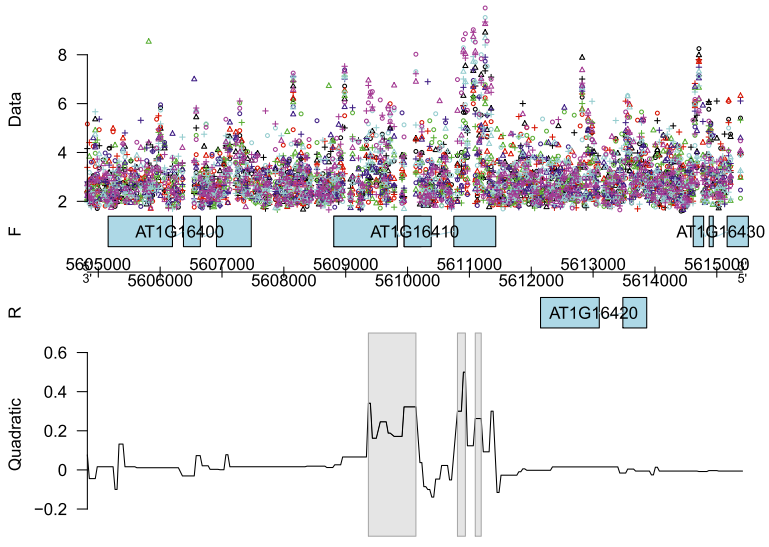


**Figure 4 Genomic region with a quadratic time effect.** Fitted quadratic time effect for the genomic region of gene *AT1G16410* on the forward strand of chromosome 1. The different replicates are indicated by ∘, + and △, while the different days are represented by different colors: black (D8), red (D9), green (D10), blue (D11), cyan (D12) and magenta (D13). The grey rectangles indicate the detected regions showing a significant quadratic time effect.

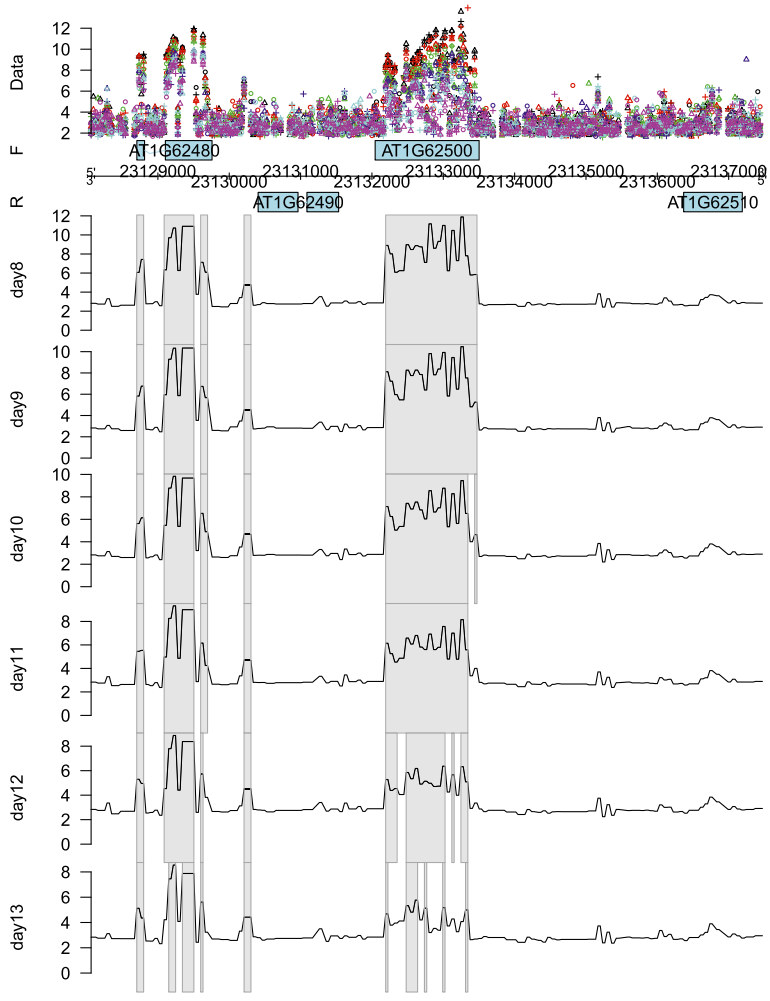


**Figure 5 Day-by-day expression levels for a genomic region show the linear effect.** Fitted $\log_2$ intensities per time point of the genomic region of gene *AT1G62500* on the forward strand of chromosome 1. The different replicates are indicated by ∘, + and △, while the different days are represented by different colors: black (D8), red (D9), green (D10), blue (D11), cyan (D12) and magenta (D13). The grey rectangles indicate the detected regions showing a significant mean expression. The decreasing trend of the fitted $\log_2$ intensities over the different time points exemplified in Figures 2 and 3 is clearly apparent in this figure.

with the strongest estimated circadian effect for their corresponding chromosome. The gene *TIME FOR COFFEE* is known as a clock gene that does not cycle at the transcriptional level [41]. Hence, it is as expected that both the overlap between detected region and gene annotation, and the effect size are very small. The gene *ZEITLUPE* is reported as having weak rhythms at the transcriptional level [40]. This is confirmed by the low maximum effect size, while still showing a considerable overlap of the significant region with the existing annotation. The results of Table 1 are thus completely in line with what was expected from literature.

### Case study 3: Non-orthogonal two-factor design

The third data set is used to illustrate the analysis of a two-factor design tiling array experiment. The data are taken from a study of the genome-wide analysis of endogenous abscisic acid (ABA)-mediated transcription in dry and imbibed seeds of *Arabidopsis thaliana* [12]. ABA is a phytohormone that is important for the induction and maintenance of seed dormancy. To understand how endogenous ABA regulates the transcriptome in seeds, whole-genome expression analyses were conducted in two ABA metabolism mutants, an ABA-deficient mutant (*aba2*) and an ABA over-accumulation mutant (*cyp707a1a2a3* triple mutant), and compared to a wild type. This is the first factor in the design. Since endogenous levels of ABA often change drastically during seed imbibition [12], these experiments were done both for dry and for 24-h imbibed seeds. This is the second factor in the design. For each design point, three biological replicates were hybridized using the Affymetrix AtTile

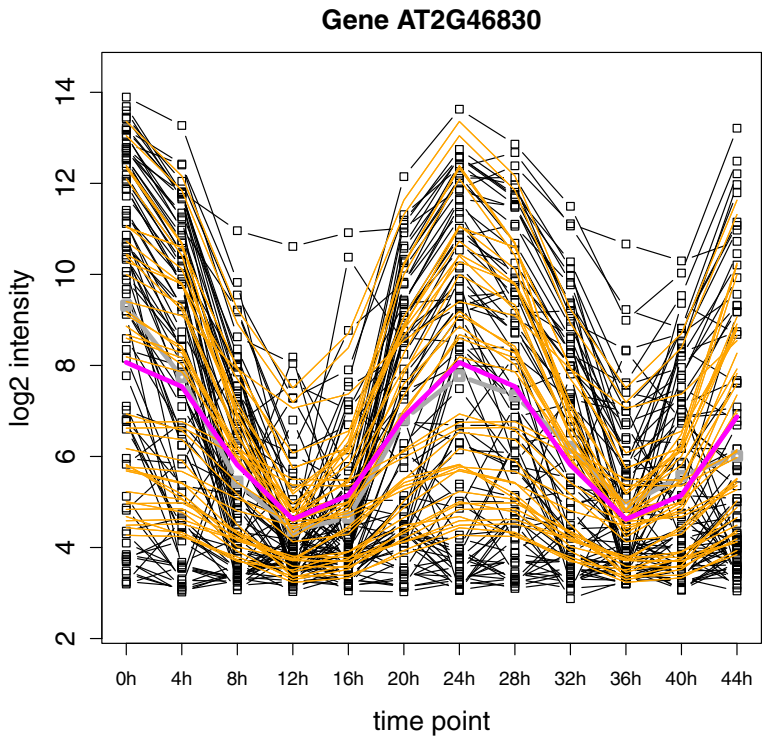


**Figure 6 Gene-wise circadian effect of transcription levels.** Example plot for gene *AT2G46830*, better known as *CIRCADIAN CLOCK ASSOCIATED1*, showing a clear circadian rhythm effect of the mean $\log_2$ intensity level over the 48h time course. The dotted black lines represent the observed $\log_2$ expression for the probes at the different time points. The dotted grey line is the mean observed $\log_2$ expression over all the probes in the region. The orange lines are the probe-wise fitted $\log_2$ expression values, while the purple line gives the corresponding mean fitted $\log_2$ expression values at the different time points over all the probes in the region.

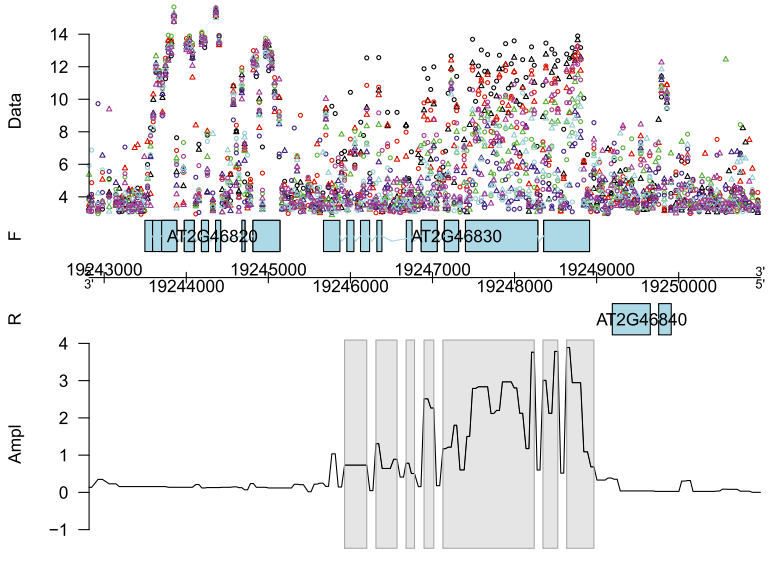


**Figure 7 Genomic region with a circadian effect.** Fitted circadian effect for the genomic region of gene *AT2G46830* on the forward strand of chromosome 2. On the *Y*-axis the amplitude of the circadian rhythm $\boldsymbol{A}(t) = \sqrt{\boldsymbol{\beta}_2^2(t) + \boldsymbol{\beta}_3^2(t)}$ is given. The grey rectangles indicate the detected regions showing a significant circadian effect. The different replicates are indicated by ∘ and △, while the different samples in the 12 h light / 12 h dark cycles regime are represented by different colors.

**Table 1 Circadian effect for 9 genes put forward in the Hazen study**

| Gene ID | Name | Overlap | Max. Eff. | Top 20 |
|---|---|---|---|---|
| *AT1G22770* | *GIGANTEA* | 0.529 | 2.28 | yes |
| *AT1G68050* | *FLAVIN-BINDING KELCH DFB PROTEIN1* | 0.867 | 2.90 | yes |
| *AT2G25930* | *EARLY FLOWERING3* | 0.562 | 1.46 | yes |
| *AT2G46790* | *PSEUDO RESPONSE REGULATOR9* | 0.473 | 1.38 | yes |
| *AT2G46830* | *CIRCADIAN CLOCK ASSOCIATED1* | 0.867 | 3.89 | yes |
| *AT3G22380* | *TIME FOR COFFEE* | 0.040 | 0.06 | no |
| *AT3G46640* | *LUX ARRHYTHMO* | 0.717 | 1.69 | yes |
| *AT5G57360* | *ZEITLUPE* | 0.350 | 0.41 | no |
| *AT5G61380* | *TIMING OF CAB2 EXPRESSION1* | 0.797 | 1.74 | yes |

Analysis results for 8 circadian clock associated genes and for *TIME FOR COFFEE*, a clock gene that does not cycle at the transcriptional level. These are the genes on the forward strand that were also tested in [13]. *Overlap* indicates the proportion of overlap between the regions detected by the wavelet-based method and the gene annotation; *Max. Eff.* gives the maximum estimated effect or amplitude size for this gene; *Top 20* indicates whether the gene is within the top 20 genes with the strongest circadian effect for the associated chromosome, as produced by the wavelet-based method.

1.0F and 1.0R tiling arrays, resulting in 18 samples. For this case, model (2) can be written as

$$\begin{aligned} Y_i(t) = {} & \beta_0(t) + \beta_1(t)\ imbibed + \beta_2(t)\ mutant1 \\ & + \beta_3(t)\ mutant2 + \beta_4(t)\ imbibed * mutant1 \\ & + \beta_5(t)\ imbibed * mutant2 + E_i(t), \end{aligned}$$

where $imbibed = 1$ if the seed was imbibed and $imbibed = 0$ if the seed was dry, $mutant1 = 1$ for the *aba2*-mutant and $mutant1 = 0$ otherwise, and $mutant2 = 1$ for the *cyp707a1a2a3* triple mutant and $mutant2 = 0$ otherwise. This model specification implies that the design matrix $\boldsymbol{X}$ used for this model is

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{bmatrix}.$$

Column 1 of $\boldsymbol{X}$ corresponds with an overall mean expression level over all samples. The main imbibition effect is coded in column 2. Note that this corresponds with the imbibition effect for wild types, which is the reference species. Columns 3 and 4 are associated with the main ABA mutation effects, whereas columns 5 and

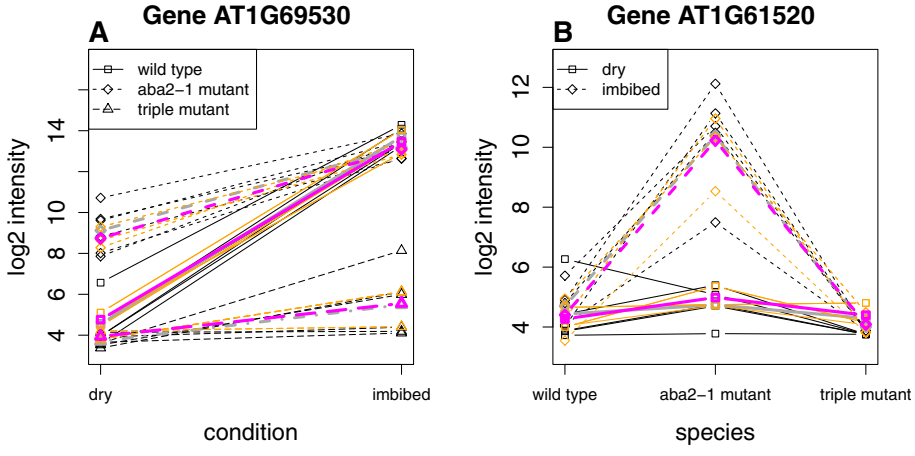


**Figure 8 Gene-wise main and interaction effects of transcription levels for a two-factor model.** Interaction plots for genes *AT1G69530* (**A**) and *AT1G61520* (**B**). The black lines represent the observed $\log_2$ expression for the probes at the different combinations of the two factor levels. The dotted grey line is corresponding the mean observed $\log_2$ expression over all the probes in the region. The orange lines are the probe-wise fitted $\log_2$ expression values, while the purple line gives the corresponding mean fitted $\log_2$ expression values for all the probes in the region.

**Table 2 Two-factor model gene-wise effects**

| | $\hat{\beta}_{0,gene}$ | $\hat{\beta}_{1,gene}$ | $\hat{\beta}_{2,gene}$ | $\hat{\beta}_{3,gene}$ | $\hat{\beta}_{4,gene}$ | $\hat{\beta}_{5,gene}$ |
|---|---|---|---|---|---|---|
| *AT1G69530* | 4.76 | 8.70 | 3.98 | −0.82 | −4.34 | −7.09 |
| *AT1G61520* | 4.27 | 0.13 | 0.72 | 0.13 | 5.11 | −0.44 |

Gene-wise mean parameter estimates for genes *AT1G69530* and *AT1G61520*. The estimates indicate a clear interaction effect between *condition* and *species* for these genes, which is further visualized in Figure 8.

6 allow to examine an interaction effect between imbibition and ABA mutation statuses. Figure 8 shows two examples plots for representing the model fit for the genes *AT1G69530*, encoding an expansin, and *AT1G61520*, encoding a chlorophyll a/b binding protein, on the forward strand of chromosome 1. Table 2 gives the associated gene-wise mean parameter estimates for these genes. The left panel plot of Figure 8 suggests a larger mean expression level of gene *AT1G69530* for imbibed seeds compared to dry seeds. The increase in mean expression level, however, is larger for wild types than for ABA-related mutants. The increase in mean expression level between imbibed seeds compared to dry seeds is given by $\hat{\beta}_{1,gene} = 8.70$ for wild types, while for *aba2* mutants this increase is given by $\hat{\beta}_{1,gene} + \hat{\beta}_{4,gene} = 4.36$ and for *cyp707a1a2a3* triple mutants by $\hat{\beta}_{1,gene} + \hat{\beta}_{5,gene} = 1.61$. On the right panel of Figure 8 we see an increased mean expression level of gene *AT1G61520* for *aba2* mutants as compared to wild types and *cyp707a1a2a3* triple mutants. In addition, this increase is much stronger for imbibed seeds.

## Conclusions

In this paper, we have described the R package waveTiling for model-based analysis of tiling array expression studies with flexible designs. It implements the recently proposed wavelet-based model for transcriptome analysis [25] and extends its applicability towards more complex experimental set-ups. Unlike most currently applied methods, transcriptional activity is modeled at probe-level instead of gene- or exon-level. This probe-wise analysis allows for the detection of transcriptional units in both exonic, intronic and intergenic regions, without prior consultation of existing annotation. By appropriate adaptations of the basic model design matrix it becomes possible to easily analyze the transcriptome for single-factor experiments with more than two biological conditions, to detect linear and quadratic time effects or a circadian rhythm effect in time-course experiments, and to even conduct two- or multiple-factor studies. The package's use and flexibility are illustrated with three case studies on the reference plant *Arabidopsis thaliana*. These cases show the potential of the package and method to cope with a multitude of study designs and associated specific research questions and still provide reliable results. The waveTiling package will be freely available as part of the Bioconductor project.

### Availability and requirements

- **Project name:** waveTiling
- **Project home page:** http://r-forge.r-project.org/projects/wavetiling/
- **Operating system(s):** Platform independent
- **Programming language:** R
- **Other requirements:** R >= 2.14
- **License:** GNU GPL
- **Any restrictions to use by non-academics:** None

## Additional files

**Additional file 1: waveTiling package vignette.** Package vignette containing detailed information on how to perform a transcriptome analysis using a wavelet-based functional model with the waveTiling package. The data set of case study 1 (leaf development data) is used in the vignette.

**Additional file 2: Methods for biological validation.** Detailed information about the gene set enrichment and qRT-PCR analysis for case study 1 (leaf development data).

#### Competing interests

The authors declare that they have no competing interests.

#### Author's contributions

KDB and LC conceived the study and developed the model. KDB implemented the model, conducted the case studies and statistical analyses and wrote the manuscript. PP helped in the design and implementation of the package. MA conducted the biological validation experiments and analyses and helped write the manuscript. OT and CC took part in several discussions related to the model. DI took part in several discussions regarding the biological data. All authors read and approved the final manuscript.

#### Acknowledgements

Part of this research was supported by IAP research network grant no. P6/03 of the Belgian government (Belgian Science Policy) and Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks").

#### Author details

[1]Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B9000 Ghent, Belgium. [2]Department of Plant Systems Biology, Flanders Institute for Biotechnology, Ghent, Belgium. [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD, USA. [5]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven and Universiteit Hasselt, Kapucijnenvoer 35, Blok D, bus 7001, B3000 Leuven, Belgium.

Received: 22 May 2012 Accepted: 5 September 2012
Published: 14 September 2012

#### References

1. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MMH, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, et al.: **Empirical analysis of transcriptional activity in the arabidopsis genome.** *Science* 2003, **302**(5646):842–846.
2. 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 Res* 2004, **14:**331–342.

3. Schadt E, Edwards S, GuhaThakurta D, Holder D, Ying L, Svetnik V, Leonardson A, Hart K, Russell A, Li G, Cavet G, Castle J, McDonagh P, Kan Z, Chen R, Kasarskis A, Margarint M, Caceres R, Johnson J, Armour C, Garrett-Engele P, Tsinoremas N, Shoemaker D: **A comprehensive transcript index of the human genome generated using microarrays and computational approaches.** *Genome Biol* 2004, **5**(10):R73.
4. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S, Ulrich EL, Zhao Q, Wrobel RL, Newman CS, Fox BG, Phillips GN, Markley JL, Sussman MR: **Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays.** *Proc Nat Acad Sci U S A* 2005, **102**(12):4453–4458.
5. Toyoda T, Shinozaki K: **Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models.** *Plant J* 2005, **43**(4):611–621.
6. Zeller G, Henz SR, Laubinger S, Weigel D, Rätsch G: **Transcript normalization and segmentation of tiling array data.** *Pac Symp Biocomput* 2008, **13:**527–538.
7. Nicolas P, Leduc A, Robin S, Rasmussen S, Jarmer H, Bessires P: **Transcriptional landscape estimation from tiling array data using a model of signal shift and drift.** *Bioinformatics* 2009, **25**(18):2341–2347.
8. Munch K, Gardner P, Arctander P, Krogh A: **A hidden Markov model approach for determining expression from genomic tiling micro arrays.** *BMC Bioinformatics* 2006, **7:**239.
9. Piccolboni A: **Multivariate segmentation in the analysis of transcription tiling array data.** *J comput biol: a j comput mol cell biol* 2008, **15**(7):845–856.
10. Otto C, Reiche K, Hackermüller J: **Detection of differentially expressed segments in tiling array data.** *Bioinformatics* 2012, **28:**1471–1479.
11. Andriankaja M, Dhondt S, De Bodt, S, Vanhaeren H, Coppens F, De Milde, L, Mühlenbock P, Skirycz A, Gonzalez N, Beemster GT, Inzé D: **Exit from proliferation during leaf development in *Arabidopsis thaliana*: a not-so-gradual process.** *Dev Cell* 2012, **22:**64–78.
12. Okamoto M, Tatematsu K, Matsui A, Morosawa T, Ishida J, Tanaka M, Endo TA, Mochizuki Y, Toyoda T, Kamiya Y, Shinozaki K, Nambara E, Seki M: **Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays.** *Plant J* 2010, **62:**39–51.
13. Hazen S, Naef F, Quisel T, Gendron J, Chen H, Ecker J, Borevitz J, Kay S: **Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays.** *Genome Biol* 2009, **10**(2):R17.
14. Granovskaia M, Jensen L, Ritchie M, Toedling J, Ning Y, Bork P, Huber W, Steinmetz L: **High-resolution transcription atlas of the mitotic cell cycle in budding yeast.** *Genome Biol* 2010, **11**(3):R24.
15. Naouar N, Vandepoele K, Lammens T, Casneuf T, Zeller G, Van Hummelen, P, Weigel D, Rätsch G, Inzé D: **Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes.** *Plant J* 2009, **57:**184–194.
16. Matsui A, Ishida J, Morosawa T, Mochizuki Y, Kaminuma E, Endo TA, Okamoto M, Nambara E, Nakajima M, Kawashima M, Satou M, Kim JM, Kobayashi N, Toyoda T, Shinozaki K, Seki M: **Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array.** *Plant and Cell Physiol* 2008, **49**(8):1135–1149.
17. Huber W, Toedling J, Steinmetz LM: **Transcript mapping with high-density oligonucleotide tiling arrays.** *Bioinformatics* 2006, **22**(16):1963–1970.
18. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet and Mol Biol* 2004, **3:**Iss. 1, Article 3.
19. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Nat Acad Sci* 2001, **98**(9):5116–5121.
20. Samanta MP, Tongprasit W, Sethi H, Chin CS, Stolc V: **Global identification of noncoding RNAs in Saccharomyces cerevisiae by modulating an essential RNA processing pathway.** *Proc Nat Acad Sci U S A* 2006, **103**(11):4192–4197.
21. Assarsson E, Greenbaum JA, Sundström M, Schaffer L, Hammond JA, Pasquetto V, Oseroff C, Hendrickson RC, Lefkowitz EJ, Tscharke DC, Sidney J, Grey HM, Head SR, Peters B, Sette A: **Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes.** *Proc Nat Acad Sci* 2008, **105**(6):2140–2145.
22. Wijnen H, Naef F, Young MW: **Molecular and statistical tools for circadian transcript profiling.** *Methods Enzymol* 2005, **393:**341–365.
23. Ahdesmaki M, Lahdesmaki H, Pearson R, Huttunen H, Yli-Harja O: **Robust detection of periodic time series measured from biological systems.** *BMC Bioinf* 2005, **6:**117.
24. de Lichtenberg U, Jensen L, Fausboll A, Jensen T, Bork P, Brunak S: **Comparison of computational methods for the identification of cell cycle-regulated genes.** *Bioinformatics* 2005, **21:**1164–1171.
25. Clement L, De Beuf K, Thas O, Vuylsteke M, Irizarry RA, Crainiceanu C: **Fast wavelet based functional models for transcriptome analysis with tiling arrays.** *Stat Appl Genet and Mol Biol* 2012, **11:**Iss. 1, Article 4.
26. Gentleman RC, Carey VJ, Bates DM, et al.: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
27. R Development Core Team: *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Austria: Vienna; 2011. http://www.R-project.org.
28. Carvalho BS, Irizarry RA: **A framework for oligonucleotide microarray preprocessing.** *Bioinformatics* 2010, **26:**2363–2367.
29. Pages H, Aboyoun P, Lawrence M: *IRanges: Infrastructure for manipulating intervals on sequences (R package version 1.12.5)*. [http://www.bioconductor.org/packages/2.10/bioc/html/IRanges.html]
30. Durinck S, Bullard J, Spellman P, Dudoit S: **GenomeGraphs: integrated genomic data visualization with R.** *BMC Bioinf* 2009, **10:**2.
31. Aboyoun P, Pages H, Lawrence M: *GenomicRanges: Representation and manipulation of genomic intervals (R package version 1.6.6)*. [http://www.bioconductor.org/packages/2.10/bioc/html/GenomicRanges.html]
32. Qu Y, Xu S: **Quantitative trait associated microarray gene expression data analysis.** *Mol Biol and Evol* 2006, **23**(8):1558–1573.
33. Narula SC: **Orthogonal polynomial regression.** *Int Stat Rev* 1979, **47:**31–36.
34. Golub GH, Loan CFV: *Matrix Computations, third ed*. Baltimore: The Johns Hopkins University Press; 1996.
35. Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR: **Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models.** *Biometrics* 2008, **64**(2):479–489.
36. Rehrauer H, Aquino C, Gruissem W, Henz SR, Hilson P, Laubinger S, Naouar N, Patrignani A, Rombauts S, Shu H, Van de Peer Y, Vuylsteke M, Weigel D, Zeller G, Hennig L: **AGRONOMICS1: a new resource for arabidopsis transcriptome profiling.** *Plant Physiol* 2010, **152**(2):487–499.
37. 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**(2):249–264.
38. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic Acids Res* 2008, **36**(suppl 1):D1009–D1014.
39. Proost S, Van Bel M, Stercka L, Billiaua K, Van Parysa T, Van de Peer Y, Vandepoele K: **PLAZA: a comparative genomics resource to study gene and genome evolution in plants.** *The Plant Cell* 2009, **21**(12):3718–3731.
40. Gardner MJ, Hubbard KE, Hotta CT, Dodd AN, Webb AAR: **How plants tell the time.** *Biochem J* 2006, **397:**15–24.
41. Ding Z, Millar AJ, Davis AM, Davis SJ: **TIME FOR COFFEE encodes a nuclear regulator in the *arabidopsis thaliana* circadian clock.** *Plant Cell* 2007, **19**(5):1522–1536.

doi:10.1186/1471-2105-13-234
**Cite this article as:** De Beuf *et al.*: Analysis of tiling array expression studies with flexible designs in Bioconductor (waveTiling). *BMC Bioinformatics* 2012 **13**:234.