BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 11 2012, pages 1471–1479
doi:10.1093/bioinformatics/bts142

Gene expression

Advance Access publication April 6, 2012

Detection of differentially expressed segments in tiling array data

Christian Otto [1,2], Kristin Reiche [1,3,4] and Jörg Hackermüller [1,3,4,*]

[1]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04107 Leipzig, [2]LIFE Leipzig Research Center for Civilization Diseases, Universität Leipzig, 04103 Leipzig, [3]Young Investigators Group Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research—UFZ, 04318 Leipzig and [4]RNomics Group, Department of Diagnostics and New Technologies, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Tiling arrays have been a mainstay of unbiased genome-wide transcriptomics over the last decade. Currently available approaches to identify expressed or differentially expressed segments in tiling array data are limited in the recovery of the underlying gene structures and require several parameters that are intensity-related or partly dataset-specific.

Results: We have developed `TileShuffle`, a statistical approach that identifies transcribed and differentially expressed segments as significant differences from the background distribution while considering sequence-specific affinity biases and cross-hybridization. It avoids dataset-specific parameters in order to provide better comparability of different tiling array datasets, based on different technologies or array designs. `TileShuffle` detects highly and differentially expressed segments in biological data with significantly lower false discovery rates under equal sensitivities than commonly used methods. Also, it is clearly superior in the recovery of exon–intron structures. It further provides window *z*-scores as a normalized and robust measure for visual inspection.

Availability: The R package including documentation and examples is freely available at http://www.bioinf.uni-leipzig.de/Software/TileShuffle/

Contact: joerg.hackermueller@ufz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 8, 2011; revised on March 4, 2012; accepted on March 22, 2012

1 INTRODUCTION

During the last decade, tiling arrays have been a mainstay of unbiased transcriptomics (e.g., Bertone *et al.*, 2004; Kapranov *et al.*, 2002; Rinn *et al.*, 2003) and continue to contribute to novel findings. Tiling arrays have recently been applied, e.g., in the discovery of novel long non-coding RNAs (Guttman *et al.*, 2009), to the identification of spatio-temporal patterns of gene expression (Spencer *et al.*, 2011), to the characterization of the transcriptome in 30 distinct developmental stages as well as in 25 cell lines of *Drosophila melanogaster* (Cherbas *et al.*, 2011; Graveley *et al.*, 2011), and to the identification of a 'large complement of novel loci' with stage-specific expression in *Caenorhabditis elegans* (Wang *et al.*, 2011). High-throughput sequencing methods have recently shown distinct advantages over array-based approaches (Agarwal *et al.*, 2010; Bradford *et al.*, 2010). However, due to the availability of large tiling array reference datasets, e.g., from ENCODE and a clear statistical understanding on how to model differential expression in microarray data, tiling arrays are an important experimental approach in transcriptomics, and tiling array data analysis is a relevant topic in computational biology.

One of the most widely used methods in tiling array expression analysis was introduced by (Kampa *et al.*, 2004) and is implemented in the Tiling Array Software (`TAS`). In brief, the local expression levels of probes are estimated by calculating the pseudo-median or Hodge–Lehmann estimator over intensities of probes within genomic distance of *bandwidth*. Transcribed segments are collections of expressed probes, i.e., probes with a smoothed intensity above a given threshold, with maximal genomic distance of *maxgap* and minimal length of *minrun*. `TAS` extends the method of Kampa *et al.* by estimating the significance of differential expression using a Wilcoxon signed-rank test. It tests for significant changes of probe intensities among states applied to local windows of given width centered around each probe. Hence, *p*-values for differential expression are assigned to each probe.

More recently, Johnson *et al.* introduced an approach that models the expected probe behavior. It is available in the tool `MAT` (Johnson *et al.*, 2006). Originally, it was designed to detect regions enriched by ChIP-chip but has also been applied to detect transcriptional activity (Kadener *et al.*, 2009; Lee *et al.*, 2009). In contrast to `TAS`, `MAT` uses a mixture model to normalize probe intensities by estimating the expected binding affinity on the basis of the composition and copy number of their nucleotide sequence on the corresponding genome. To identify (differentially) expressed probes, the score over all normalized intensities of probes within a local window, given by a *bandwidth* parameter, is compared with a null distribution. This distribution is composed of all non-overlapping window scores that can be calculated on the same array or the array in a different state during expression or differential expression analysis, respectively. Hence, it uses a two-step approach with different background distributions to normalize the probe intensity and assess its significance within a probe-centered window. In the detection of (differentially) expressed segments, positive probes are joined if their genomic distance is below a given *maxgap* parameter and segments enclosing more than *minprobe* probes are then reported. `TileProbe` is a variant of `MAT`, which models residual probe effects that cannot be explained by the `MAT` model by incorporating publicly available datasets (Judy and Ji, 2009).

*To whom correspondence should be addressed.

TileProbe has been successfully applied to detect enriched motifs in ChIP-chip tiling array data, but in contrast to MAT, no application to detect differential expression has been reported. HAT uses a hypergeometric distribution to assess the probability to observe a specific number of probes within a window. It is less sensitive but more specific than MAT and cannot directly be used to detect differential expression (Taskesen *et al.*, 2010). Lastly, HMMTiling models probe-specific effects by a normal distribution defined for each probe individually compared with a control group (Li *et al.*, 2005), but requires many samples in order to estimate the variance for a probe correctly which may not be available for arbitrary types of tiling arrays.

gSAM, is a powerful framework for analyzing differential response of time series tiling array data (Ghosh *et al.*, 2007). It generalizes SAM (Tusher *et al.*, 2001) from a gene-centric view to genomic intervals in an underlying piece-wise model. Under this model, the time series is subdivided into logical segments and differential changes are analyzed on each of these segments separately. gSAM requires replicates which are often not available for whole genome tiling data. Another method suitable to detect differential expression on tiling array data is TileMap which assesses the significance of each probe by averaging over moderated *t*-statistics within a pre-defined window size (Ji and Wong, 2005). (Kechris *et al.*, 2010) propose the averaging of *p*-values instead of test statistics providing a more flexible framework to evaluate more complicated experimental designs and to overcome the problem that the length of a sliding window may not be large enough to assume normal distribution. However, both methods again require replicates because probe-wise expression changes are assessed by hypothesis tests. An HMM-based approach was introduced by Munch *et al.* (2006) that adaptively models tiling array data on given annotation and subsequently predicts expression on the genomic sequence. It does not require *ad hoc* parameters but is limited to expression analysis and hence cannot predict differential expression.

Huber and colleagues presented a powerful segmentation approach for tiling array data, which controls for probe-specific effects by normalizing probe-wise intensities to a reference experiment with genomic DNA (Huber *et al.*, 2006). Recently, Karpikov *et al.* (2011) introduced a wavelet transformation to tiling array ChIP-chip data in order to discriminate regions of activity from noisy data.

Our aim is to use tiling array data for identifying novel ncRNAs, which are differentially expressed in response to critical signaling pathways or cellular processes. For this purpose, a data analysis method is required to (i) analyze differential expression in tiling array data for genome-wide approaches; (ii) allow the latter without using replicate tiling array experiments due to limitations in the availability of sample material; (iii) identify boundaries of differentially expressed segments sufficiently precise to allow transcript annotation; and (iv) avoid the use of dataset-specific parameters which may hamper analyzing differential expression between arrays of different experiments. In our opinion, none of the state-of-the-art methods sufficiently fulfills all these requirements.

Here, we present TileShuffle, a novel tiling array analysis approach that identifies transcribed and differentially expressed segments in terms of significant differences from the background distribution by using a permutation test statistic. Significance is assessed on minimal expected transcriptional units rather than on a single-probe level. TileShuffle does not require any dataset-specific parameters, e.g., intensity-related thresholds or parameters concerning collection of expressed probes. This is particularly favorable since in common tiling array experiments neither spike-ins to control the false discovery rate [FDR; as in (Kampa *et al.*, 2004)] nor sufficiently large positive and negative sets to optimally adjust these *ad hoc* parameters might be available.

We compare TileShuffle to TAS and MAT in analyzing differential expression in one human whole genome tiling array dataset and one spike-in dataset (Sasaki *et al.*, 2007). TAS is the most widely used tool in tiling array expression analysis and although MAT was originally designed for ChIP-chip data, it was successfully applied to detect transcriptional activity. All, TileShuffle, TAS, and MAT, do not require replicates to detect differentially expressed transcripts which is in particular favorable for studies with limited material and costs. At the same FDR, TileShuffle achieves significantly higher sensitivities than the other methods. Also, it detects boundaries of differentially expressed exons with higher precision than TAS and MAT.

2 METHODS

2.1 Expression detection

To determine transcribed segments in tiling array data, we apply a statistical approach that differentiates expression signals from the background distribution under consideration of common tiling array biases. Given the array design of nearly uniformly distributed probe sequences over the non-repetitive genome, hybridization affinity and hence signal intensity is highly dependent on the probe sequence itself, i.e., nucleotide composition and nucleotide positioning (Johnson *et al.*, 2006; Royce *et al.*, 2007). Analogously, in absence of specific transcripts, a detected probe signal may solely originate from non-specific hybridization, e.g., background noise and cross-hybridization, causing single spikes in the tiling array data. Here, cross-hybridization refers to the hybridization of DNA/RNA fragments to probe sequences that are similar or even equal to their actual target, but originate from different genomic loci.

Handling common tiling array biases: Even though transcripts are expected to be detected by several neighboring probes in similar scale, non-specific hybridization and sequence-specific effects like nucleotide composition and positioning can largely increase the detected signal intensity of single probes while having no effect on the neighboring probes and hence roughen the signal across the tiling array. For example, probes with high GC content tend to exhibit increased signal intensities compared to probes with low GC content. In addition to the GC content, Royce *et al.* highlighted the influence of position-specific effects of each nucleotide on the probe intensities, e.g., higher average intensities of probes with Gs toward the probe start or Cs toward the probe end (Royce *et al.*, 2007). These sequence-specific biases introduce a disparity in the binding affinity among different probe sequences, subsequently denoted as sequence-specific affinity.

We therefore assess the significance of expression on windows of length l with respect to the background distribution rather than on single probes. A score $S_e(w)$ is assigned to each sliding window w by applying a scoring function (arithmetic mean trimmed by maximal and minimal value or median) over the signal intensities of all probes within the window. Due to the robustness of the two scoring functions, window scores are less susceptible to

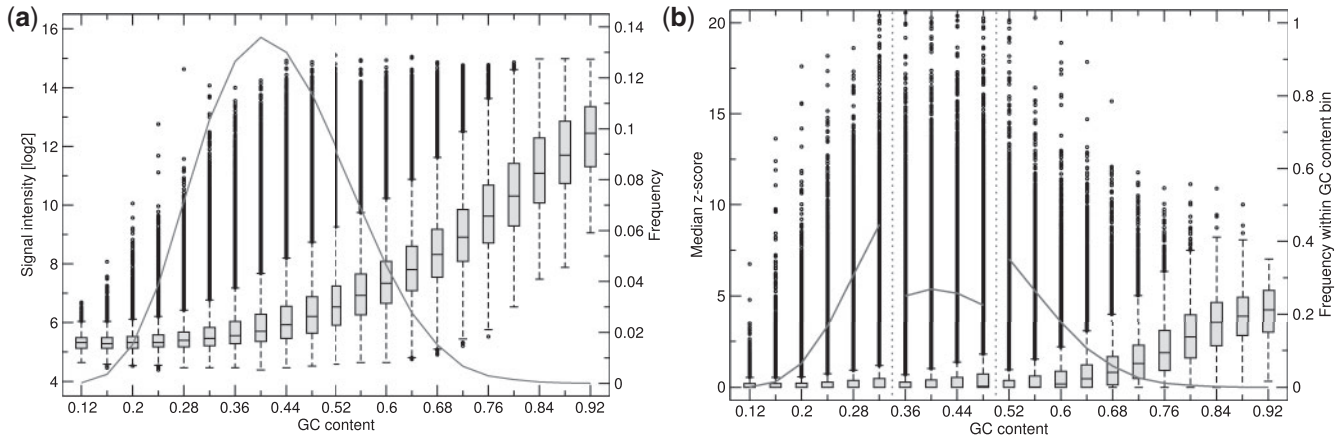

Fig. 1. (**a**) Boxplot of probe intensities on a tiling array for different GC content in the probe sequence. The relative frequency of probes with each GC content bin on the tiling array is shown in the overlay graph (red solid line). (**b**) Boxplot of probe median z-scores for different GC content in the probe sequences. The probe median z-score is defined as the median over the z-scores of all windows enclosing the probe where z-scores were estimated by `TileShuffle` using three GC bins during permutation. Vertical dotted red lines display the boundaries of different bins whereas solid red lines indicate the relative frequency of probes with the respective GC content in their bin.

signal intensity variation within a given window originating solely from outliers. In addition, probes are subdivided into affinity bins with similar expected sequence-specific affinity and bins are processed independently from each other. Accordingly, intensities of probes that belong to different affinity bins must not be interchanged. Otherwise, the expression analysis might favor windows simply due to the sequence compositions of their probes, e.g., high GC content.

Assessing significance of expression: In order to estimate the significance of a window score $S_e(w)$, we repeatedly permute all probe intensities across the array while interchanging only those that belong to the same sequence-specific affinity bin, recompute the window scores, and compare them with the original ones. We use random permutations of probe intensities to remain independent of any annotation or underlying gene structure.

By counting the number of permuted windows with higher score, we estimate empirical p-values of windows. Following a Benjamini–Hochberg multiple testing correction (Benjamini and Hochberg, 1995), all windows of high significance, i.e., the ones with corrected p-values (q-values) below a given threshold, are deemed 'expressed'. Since permutations necessitate sufficiently large groups, the binning is only based on the GC content of each probe sequence as the most dominant bias on hybridization affinity (Fig. 1a). In accordance with the findings of Johnson *et al.* (2006), the copy number of probes, i.e., number of perfect matches of the probe sequence to the genomic sequence and hence the extent of potential signal overlay, showed only a minor impact on signal intensity (Supplementary Fig. S1a). We therefore refrain from controlling for copy number in favor of larger bins during permutation. The described algorithm to detect expressed segments in tiling array data is illustrated in Supplementary Figure S4.

2.2 Differential expression detection

In many cases, tiling array data is available from different cellular states or other biological conditions and one might be interested in structural changes in the expression between different conditions. To avoid that signal intensity variation at the detection limit is classified as differential expression, we require that differentially expressed intervals must also be significantly expressed relative to the background distribution in at least one of the investigated conditions, and call these intervals *highdiff*. This is analogous to the frequently performed unspecific filtering in conventional microarray data analysis.

Assessing significance of differential expression: On the contrary to one-state expression analyses, signal intensities are normalized using a quantile-normalization across each tiling array in both considered conditions (Bolstad *et al.*, 2003). Expression shifts are then measured in terms of log-fold changes (i.e., differences of log signals) between probe intensities in both cellular conditions. In consequence, sequence-specific effects cancel out and affinity classification as it is done for expression detection is rendered unnecessary (Supplementary Fig. S1b). Fold changes assume constant variance among probes, which might not be valid in any case. However, if replicate data is not available, fold changes are the only applicable measure for differential signal changes. Otherwise, it is possible to use moderated t-statistics in `TileShuffle`, an empirical Bayes method to shrink the probe-wise variance toward a common value. Hence, it is preferable over ordinary t-statistics (Witten and Tibshirani, 2007).

Due to the two-tailed distribution of fold changes, the estimation of p-values needs to be adapted. We implemented and compared two different variants to detect significant changes. In variant A, window scores $S_d(w)$ of differential expression are calculated following the same outline as described for the expression analysis with the exception that two-tailed p-values are estimated in order to regard both regulation directions, up and down. The multiple testing correction is then adjusted to account for these additional comparisons. In variant B, it is assumed that entire windows represent the smallest unit of expression and are either constant, or up- or downregulated between two conditions. Converse behavior of neighboring probes is considered a consequence of non-specific hybridization. In order to correct for this bias, the presumed direction of regulation is initially assigned to

each window w on the basis of the sign of its expression score $S_e(w)$. Subsequently, all converse probes, i.e., probes with negative log-fold change within positive windows or vice versa, are ignored and neither permuted nor incorporated into the score calculation for differential expression. Consequently, positive and negative windows are compared with different background distributions. To assess the significance of a window score $S_d(w)$ of differential expression, a one-tailed empirical p-value is estimated (according to the corresponding background distribution) and corrected for multiple testing, similar to the one-state analysis. The assignment of the significance to a window in case of the differential expression analysis with both variants is illustrated in Supplementary Figures S5 and S6. Overall, both variants merely differ in the window score calculation (independently from the used scoring function) and multiple testing correction in differential expression analysis. Due to their difference in treating converse probe behavior possibly leading to more robustness of variant B, we implemented and included both of them in our comparative analyses.

2.3 Estimating z-scores

In addition to the statistical significance, a normalized score can be reported for each processed window on the tiling array. Since the score distribution of the permuted windows is a sample from the background distribution, a z-score of a window w is calculated by

$$z(w) = \frac{x - \mu}{\sigma}, \quad (1)$$

where x is either the score $S_e(w)$ or $S_d(w)$, while μ and σ are the mean and standard deviation (SD) of the permuted window scores, respectively. To obtain a probe-wise measure, the probe median z-score $z(p)$ is defined as the median over the z-scores of all windows enclosing the probe p. In consequence, probe median z-scores may be used as a normalized measure of probe expression in order to visually inspect regions of interest.

2.4 Validation

A custom microarray based on a different manufacturer, labeling procedure, and probe length has been designed to validate the tiling array results as an alternative experimental approach. We used the *Agilent eArray* procedure (https://earray.chem.agilent.com/earray/) to ensure that probe-specific biases are minimized and designed probes of 60 mer length for both reading directions of all *highdiff* regions that have been identified by `TileShuffle` and `TAS`. We, furthermore, verified that the custom microarray also covers an unbiased sample of the regions identified by `MAT` (Supplementary Table S6). In addition, the custom microarray also includes probes for genomic regions, determined independently of the tiling array experiment: probes for all human mRNAs, for genomic regions predicted to contain a conserved secondary structure identified by RNAz (Washietl *et al.*, 2005) or Evofold (Pedersen *et al.*, 2006), and known ncRNAs from public databases.

The custom microarray was run in triplicates and differentially expressed probes were identified using the statistical software package R and Bioconductor (Gentleman *et al.*, 2004). Expression intensities were quantile normalized (Bolstad *et al.*, 2003) and a linear model was fitted using the `Limma R` package (Smyth, 2005). Reliable variance estimations were obtained by empirical Bayes moderated t-statistics and the FDR was controlled by Benjamini–Hochberg adjustment (Benjamini and Hochberg, 1995). A probe on the custom microarray is called significant in case the adjusted p-value is found to be < 0.05.

In addition to the custom microarray, we tested the performance of `TileShuffle`, `TAS` and `MAT` on the outcome of a spike-in dataset comprising 162 full-length cDNA clones at two concentrations, 0.0055 µg and 0.055 µg, in the gene-dense regions of chromosome 22 (Sasaki *et al.*, 2007).

3 RESULTS AND DISCUSSION

3.1 Control of tiling array specific biases

We evaluate the capability of `TileShuffle` to cope with the most dominant sequence-specific affinity effects in tiling array data such as GC content and nucleotide positioning of a probe. Assuming that most probes show only non-specific hybridization, the correlation between GC content of probe sequences and their detected signal intensities ($R^2 = 0.383$, Fig. 1a) indicates a measurable bias that needs to be taken into account. Otherwise, intensity-based analyses may favor windows simply due to their GC-richness. A signal smoothing as realized by windowing and calculating the probe median z-score $z(p)$, does not correct for the bias sufficiently ($R^2 = 0.266$, Supplementary Fig. S2a).

In theory, the use of affinity-based binning with respect to the GC content of probe sequences should reduce the general effect whereas the intensity of outliers and hence potentially expressed probes remains relatively stable. Supplementary Figure S2b illustrates a strong reduction of the sequence-specific affinity bias with merely two GC content bins ($R^2 = 0.037$). Higher numbers of bins further attenuate the correlation between GC content of probe sequences and their probe median z-score, e.g., $R^2 = 0.019$ with three bins (Fig. 1b). In each case, the distribution of the outliers (black dots) differs from the original data only to a minor extent. According to these findings, three bins may already suffice to efficiently attenuate this bias while retaining sufficiently large permutation bins.

To illustrate the influence of position-specific effects of each nucleotide on the probe intensities, we use the R package Starr (Zacher *et al.*, 2010) on probe intensities (Fig. 2a) and on probe median z-scores $z(p)$ after applying `TileShuffle` with three GC content bins (Fig. 2b). Using Starr, we can assess the position-specific bias of every nucleotide in each of the 25 positions within the probe sequence for given probe scores (e.g., probe intensities or probe median z-scores). More precisely, for any position and nucleotide, it calculates the difference between the mean score of probes, where the nucleotide is at this particular position within the probe sequence, and the overall mean probe score. To obtain comparable scales, the changes of probe intensities and probe median z-scores are normalized by dividing them by the SD of their distributions. Overall, even though position-specific biases are not explicitly considered in our framework, the combination of affinity-based permutations and overlapping windows is capable of greatly reducing position-specific biases in the tiling array data (Fig. 2b). Correction of this bias is not only a consequence of windowing, but also depends on affinity-based permutations: performing the analysis on probe median z-scores after applying `TileShuffle` with only one GC bin and hence

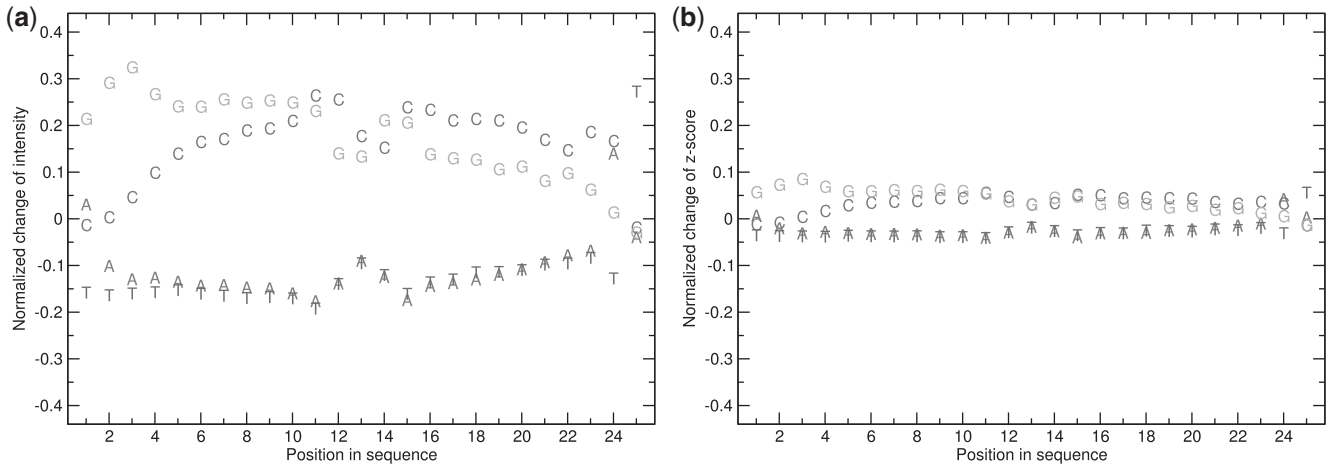

Fig. 2. Position-specific bias of every nucleotide in each of the 25 positions within the probe calculated on probe signal intensities (**a**) and on probe median *z*-scores (**b**) by use of the Starr R package (Zacher *et al.*, 2010). The distances of probe intensities and probe median *z*-scores are further normalized by dividing them by the SD of the intensity and median *z*-score distribution, respectively. The probe median *z*-score is calculated as the median over the *z*-scores of all windows enclosing the probe where *z*-scores were estimated by `TileShuffle` using three GC content bins.

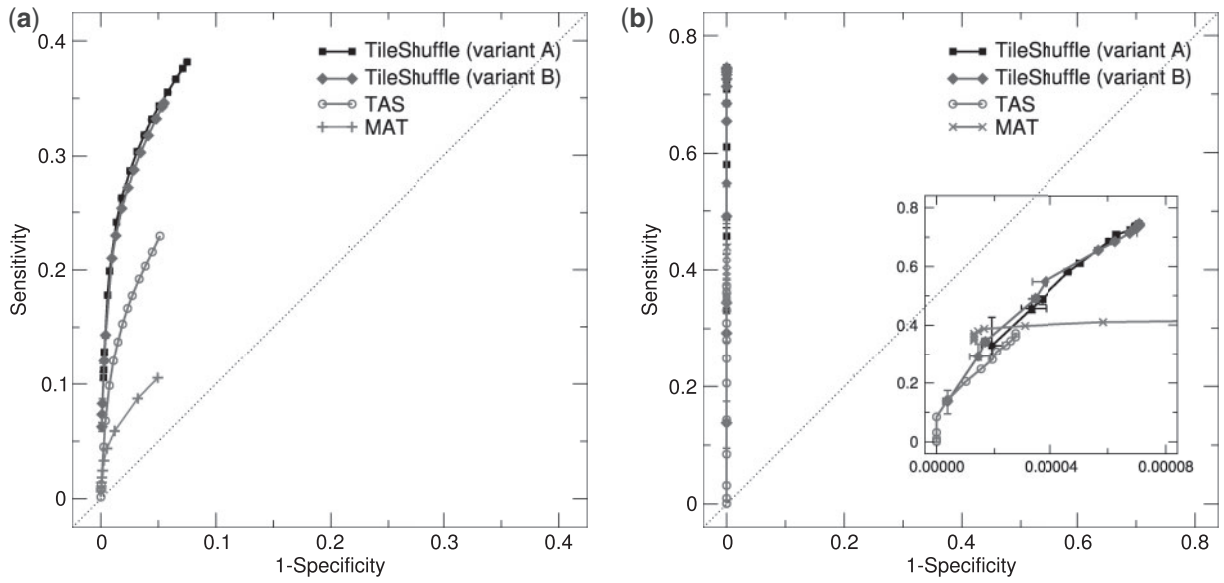

Fig. 3. ROC curve after evaluating the outcome of `TAS`, `MAT`, and `TileShuffle` applied to the G0/G1 transition of the cell cycle tiling array dataset (**a**) and to the spike-in tiling array dataset comparing hybridizations of 0.0055 µg and 0.055 µg cDNA (**b**) over a range of different *p*/*q*-value cutoffs in the differential analysis. In the cell cycle dataset, the positive set is obtained by conducting and evaluating verification experiments using a custom-designed microarray in triplicate. In the spike-in dataset, the positive set is comprised of regions covered by the 162 full-length cDNA clones which were spiked in. Note that the whiskers express the variation in the outcome of `TileShuffle` after five repetitions, i.e., smallest and highest value on the *x*-axis (or *y*-axis) for each differential significance threshold, with the median result shown on the solid line. The inlay in the right panel magnifies the area with an *x*-coordinate close to zero (same units on axes). Due to the intersection of high and differential intervals in *highdiff* at fixed parameters for high, some intervals are never identified and thus the curves do not reach (1,1). Sensitivity versus FDR curves are given in Supplementary Figure S7.

without affinity-based permutations does not sufficiently remove the bias (Supplementary Fig. S3).

3.2 Comparison with MAT and TAS

We evaluate the potential of `TileShuffle` to detect differentially expressed regions in comparison to `MAT` and `TAS`, which are the two most widely used algorithms for analyzing expression tiling array data and are both applicable to non-replicated data which is frequently the case for expensive whole genome tiling experiments. The three algorithms are evaluated in two different scenarios. In the first, we apply the different algorithms to a tiling array dataset of human foreskin fibroblasts which are synchronized by serum starvation in G0 or in G1 phase of the cell cycle. This transcriptome-wide variation study is based on the Affymetrix

Human Tiling 1.0 array set[1]. It consists of 14 arrays where probes are tiled at approx. 35 base pair (bp) intervals across the whole human genome with gaps of approx. 10 bp. The tiling array data is compared with a custom array experiment with considerably lower FDR as a reference. This allows to assess the performance of the algorithms applied to real biological data and to perform statistics on a large number of differentially expressed elements. In the second scenario, we apply all three algorithms to a spike-in dataset of 162 full-length cDNA clones, which are hybridized at two, 10-fold different concentrations to an Affymetrix chr21/22 array (Sasaki *et al.*, 2007). In this scenario, positives and negatives are more clearly defined than above, but the number of differentially expressed intervals is comparably low and the extent of differential expression and complexity of the sample is more artificial.

For `MAT` and `TAS`, the expression and differential expression analysis is carried out independently from each other: *Highdiff* regions are obtained by intersecting intervals identified as differentially expressed with those intervals deemed as 'expressed' in at least one of the compared biological states. `TileShuffle`, in contrast, takes regions found to be significantly expressed in at least one of the compared states (one-state analysis) as input for the two-state analysis, assesses differential expression solely on the expressed segments and directly reports *highdiff* regions.

For one-state analyses, i.e., determination of expressed regions, parameters for `TAS` have been set following Kampa *et al.* (2004). Parameters for `MAT` as given in Johnson *et al.* (2006) are geared toward ChIP-chip analysis and not suitable for expression analysis. Upon inspection of positive control transcripts, we identified optimal parameters for `MAT` as the same or analogous values as used for `TAS`. In summary, we set bandwidth = 35, i.e., on average the probe intensities are smoothed by calculating the Hodges–Lehman estimator over three probes, and the maximal gap between positive probes to be included in a positive interval maxgap = 40. The minimal length or minimal probe count of segments to be reported were set to minrun = 90 and minprobe = 3, for `TAS` and `MAT`, respectively. Perfect match (PM) and mismatch (MM) probe intensities were utilized in `TAS` using an intensity threshold of 150.

For expression analyses with `MAT`, which uses only PM intensities, a *p*-value threshold is set to 0.05 which yielded the best results in terms of sensitivity and FDR in the analysis of the cell cycle tiling array dataset. *P*-value cutoffs were tested in the range of 10^{-10} to 0.05. `TileShuffle` was applied using only PM probes, the arithmetic mean trimmed by maximal and minimal value as scoring function, 10 000 permutations, and a *q*-value threshold of 0.05. `TileShuffle` was applied using window sizes 20, 200, and 400 and different numbers of GC bins ranging from 1 to 9, to assess the effect of these two parameters. The intermediate window size of 200 was chosen in order to include an adequate number of probes in the calculation of the window scores S_e and S_d, and to ensure that the majority of known exons is spanned by one single window. The median exon length of known protein-coding genes is 118 bp, while 90% of the exons are shorter than 228 bp according to GENCODE version 3c (Harrow *et al.*, 2006).

Analysis of differential expression was performed with the same parameters, except bandwidth = 150 for `TAS` and `MAT` and 100 000 permutations for `TileShuffle`, both aiming at accommodating the more rugged nature of the expression difference signal (log-fold change).

For the whole genome scenario, *highdiff* intervals were generated with all three tools over a range of *q*- and *p*-value cutoffs, respectively. The custom microarray was run in triplicates for each of the biological conditions of the tiling array experiment and was used as a reference to estimate sensitivity, specificity, and FDR, defined as follows:

$$\text{sensitivity} = \frac{\text{TP}}{P} \tag{2}$$

$$\text{specificity} = 1 - \frac{\text{FP}}{N} \tag{3}$$

$$\text{FDR} = \frac{\text{FP}}{\text{FP}+\text{TP}} \tag{4}$$

The number of true positives (TP) corresponds to the number of nucleotides which are *highdiff* in the tiling array analysis and overlap with a probe that was found significantly differentially expressed in the corresponding custom microarray experiment. The number of false positives (FP) is defined as the number of those nucleotides in *highdiff* intervals that overlap a probe that is not significantly differentially expressed in the custom microarray experiment. The number of positive nucleotides (P) is defined as the sum of all nucleotides of probes that are significantly differentially expressed in the custom microarray experiment (FDR < 0.05), whereas the number of negative nucleotides (N) corresponds to the sum of all nucleotides of probes that are not significantly differentially expressed in the custom microarray experiment (FDR ≥ 0.05).

The results for each algorithm are illustrated as receiver operating characteristic (ROC) curve and as a function of sensitivity versus FDR (Fig. 3a and Supplementary Fig. S7a). Overall, `TileShuffle` in both tested variants A and B clearly outperforms the two other algorithms. For example, at a maximal FDR of 20%, both variants of `TileShuffle` yield a sensitivity of approx. 23%, which is approx. 4- and 11-fold increase compared with `TAS` and `MAT`, respectively. Both `TileShuffle` variants differ only to a minor extent from each other but variant B is generally more restrictive and hence recommended as the default choice. Evaluating the three algorithms based on counts of intervals rather than on nucleotides yields concordant results with the latter (Supplementary Fig. S8).

In this test scenario, we also investigated the influence of the number of GC bins and different window sizes on the ROC curve. The worst performance is observed for one GC bin. This shows that probes with low GC content tend to exhibit lower signal intensities than probes with high GC content and hence are less likely to be found in the right tail of the signal intensity distribution (Supplementary Fig. S16). A number of three GC bins results in higher sensitivity at similar FDR, while increasing the number of GC bins further yields only minor improvements at high FDR values. Following Occam's razor, we hence select the simpler model, and recommend to use three GC bins as the default for the one-state analysis. A window size of 400 bp leads to the best ROC curve, but exhibits to fail in exon boundary detection described in Section 3.3 (Supplementary Figs S17–S20). A window size of 20 bp delivers only very few *highdiff* regions, resulting in very low sensitivities. Thus, a window size of 200 bp seems to be the

[1]Array data and experimental details can be accessed at GEO (Supplementary Table S1).

optimal trade-off between good sensitivity and good recovery of exon–intron structures at low FDR values.

In similar manner, we estimated the sensitivity, specificity, and FDR in case of the spike-in dataset. Therein, the positive set comprises all genomic regions covered by the 162 full-length cDNA clones, i.e., 877 exonic regions, which were spiked in at two different concentrations. The set of negative regions comprises all unique protein coding exons annotated in GENCODE version 3c that do not overlap with any positive region. The GENCODE annotation was converted from human genome version hg18 to hg17 using the UCSC liftover tool. The number of TP corresponds to the number of nucleotides in positive regions which are *highdiff* in the tiling array analysis. The number of FP is defined as the number of nucleotides in negative regions which are in *highdiff* in the tiling array analysis. Accordingly, the number of positive nucleotides (P) is defined as the sum of all nucleotides in positive regions, while the number of negative nucleotides (N) corresponds to the sum of all nucleotides in negative regions. The resulting ROC curves are depicted in Figure 3b and Supplementary Figure S7b. In summary, all three methods recover the differentially expressed exons as all reach high sensitivity values at high specificity or low FDR values. However, `TileShuffle` reaches maximal sensitivity at comparable FDR values. Even though a spike-in experiment allows to precisely define TP and FP rates, it is artificial and different from real expression perturbation studies as much less noise is observed (Supplementary Fig. S22 for an exemplary region).

Due to the resampling step in `TileShuffle`, results may vary between runs with different random number generator seeds. We therefore plot the median of five different runs where the number of permutations was set to 10 000 for the one-state and 100 000 for the two-state analysis, and illustrate minimal and maximal values as whiskers in x and y direction. Only negligible variation in sensitivity and FDR is found for the most restrictive significance thresholds. This is an expected consequence of increasing variability in sampling when the tails of the background distribution are estimated. Hence, the numbers of required permutations of 10 000 and 100 000 for the one-state and two-state analysis, respectively, mark a sufficient trade-off between running time and variation in sensitivity and FDR. Due to the high degree of variation observed for fold changes, the tails of the background distributions for two-state analysis must be well estimated with an increased number of permutations. We adapted the code for the two-state analysis to ensure that a sufficiently large number of permutations can be computed within a feasible time scale. On a single 2.66 GHz 64-Bit Intel Xeon CPU, a one-state analysis of a single array under the given parameters took approx. 12 h whereas a single two-state analysis took around 9 h and 14 h with variant A and B, respectively. Since an array comprises sufficient information to sample from the background distribution and hence eliminate array-wide effects, the arrays can be analyzed independently from each other.

3.3 Detection of transcript structures

One of the advantages of tiling arrays over conventional expression arrays is information on the intron–exon-structure of transcripts, as probes are tiled in an unbiased way across the genome. We manually inspected a small set of genes that are known to be cell cycle regulated (Bar-Joseph *et al.*, 2008). In several cases, we observed that `TileShuffle` is capable of detecting a higher fraction of exons of a transcript as *highdiff* and identifies the intron-exon boundaries more accurately than `TAS` or `MAT`. Supplementary Figure S21 displays examples of known cell cycle regulated genes where the three algorithms perform remarkably different.

To substantiate this finding and to exclude that the above mentioned observation is merely a consequence of the increased sensitivity of `TileShuffle`, we studied the accuracy in detecting intron-exon boundaries on a global scale.

All unique exons of all protein-coding transcripts annotated in GENCODE version 3c (Harrow *et al.*, 2006) were extracted, resulting in 293 000 annotations. *Highdiff* intervals of the G0/G1 transition of the cell cycle dataset were computed with all three methods. To increase comparability, significance thresholds were adjusted to yield comparable FDR values, i.e., 18% FDR in case of `TAS` ($q=0.05$), 17% in case of `MAT` ($p=$1e-6), and 19% and 18% in case of `TileShuffle` variant A ($q=0.05$) and variant B ($q=0.1$), respectively. For each method, the overall reported nucleotides identified as *highdiff* in the G0/G1 transition of the cell cycle dataset including the absolute and relative base pair coverage with GENCODE version 3c annotations is given in Supplementary Table S5. The absolute number of reported nucleotides and their length greatly differs among the methods (see Supplementary Table S3). An analogous analysis for high intervals is shown in Supplementary Tables S4 and S2.

We calculated the overlap of all tiling array intervals (either highly expressed intervals or *highdiff* intervals) with all annotated exons no matter of the annotated reading strand direction for exons, since strand information cannot be inferred from the Affymetrix Human Tiling 1.0 array set. For each overlapping pair of tiling array interval and annotated exon, the genomic distances between the inferred and annotated 5′- and 3′-ends, respectively, are summarized in an empirical cumulative distribution function (ecdf). We do not only include the pair with minimal distance but consider all overlaps of several tiling array intervals with one exon, as well as all overlaps of several exons with one tiling array interval in the ecdf. This penalizes the distance distribution in cases where one exon is represented by many small tiling array intervals. It also penalizes intronic tiling array intervals that partly overlap with an exon. Due to the higher sensitivity of `TileShuffle`, the number of regions included in this analysis is significantly higher compared with the other methods. We therefore normalize the ecdf to the total number of overlaps of the respective method.

`TileShuffle` clearly outperforms the other two methods in detecting exon–intron boundaries in *highdiff* data (Fig. 4). The results are more balanced for expression analysis, where `TAS` finds a higher proportion of exons boundaries with an offset below the window size of `TileShuffle`, while overall, `TileShuffle` identifies a higher proportion of boundaries (Supplementary Fig. S9). Of all window sizes tested for `TileShuffle`, 200 bp performs best. A window of 400 bp further extends exons and a window size of 20 bp, i.e., comprising just one probe, shortens exons remarkably. Different GC bins for the one-state analysis do not have a considerable impact on exon boundary detection (Supplementary Figs. S17–S20).

Supplementary Figures S10 and S11 further illustrate the orientation in the offset to annotated exons. Both, for expression and differential expression analysis, `TileShuffle` has a tendency to extend the reported region beyond the exon boundaries with the largest extension observed for long window sizes as, i.e., 400 bp.

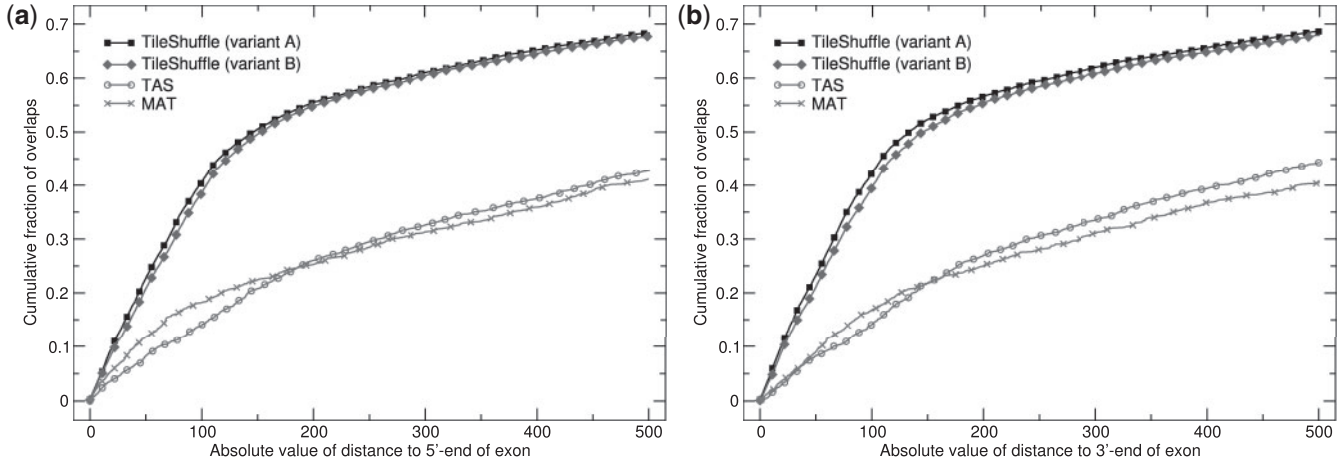

Fig. 4. Empirical cumulative distribution function of the absolute distances between 5′- (**a**) and 3′-end (**b**), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol. For each method, the set of *highdiff* intervals in the G0/G1 transition of the cell cycle tiling array dataset is used as input dataset. The significance thresholds of the three methods for differential analysis were adjusted to obtain similar FDRs as estimated before using the custom microarray, i.e., 18% FDR in case of `TAS` ($q=0.05$), 17% in case of `MAT` ($p=$1e-6), and 19% and 18% in case of `TileShuffle` with variant A ($q=0.05$) and variant B ($q=0.1$), respectively. The absolute number of overlaps is 15 835 and 13 479 with `TileShuffle` and variant A and B, respectively, 4337 with `TAS`, and 2381 with `MAT`.

Again, if the window includes just one probe (window size set to 20 bp), `TileShuffle` tends to shorten exons (Supplementary Figs. S17–S20). `TAS` and `MAT` tend to find exons shorter than annotated, caused by a comparable offset at 5′- and 3′-boundaries in the case of expression analysis (Supplementary Fig. S10). Boundaries of differentially expressed exons are hardly detected correctly by `TAS` and `MAT`, but again with a tendency to shortening (Supplementary Fig. S11). This bias in the offset to the correct exon boundary is not unexpected: considering windows of length 200, `TileShuffle` will always extend expressed exons smaller than the window size, which constitutes a significant proportion of exons in the human genome. `TAS` and `MAT` extend regions probe-wise and thus can detect exons more precisely if the signal across the exon is smooth. On the other hand, exon signals, strongly affected by sequence-specific affinities or cross-hybridization across the exon, may prevent correct extension and lead to fragmentation into several intervals or shortening. This may explain, why overall, `TileShuffle` identifies a greater proportion of boundaries. Probe-wise extension largely fails in detecting *highdiff* exons. The expression difference signal is rugged and can reverse signs within one exon. `TileShuffle`, which combines a robust windowing approach and scoring function with 'window-wise' extension, is clearly advantageous over the other methods that rely on probe-wise extension only.

We finally investigated whether the observed differences in detecting boundaries of *highdiff* exons are biased by the selected significance thresholds. Over a range of q-value thresholds, `TileShuffle` displays only minor variation in the ecdf of distances to exon boundaries and nearly constant results for distances below the window size (Supplementary Figs S12 and S13). In contrast, the ecdf of `TAS` and `MAT` vary strongly between the different thresholds (Supplementary Figs S14 and S15) and for significance threshold < 0.5, `TAS` and `MAT` obtain significantly lower accuracies than `TileShuffle` at any q-value threshold.

4 CONCLUSION

Most published tiling array studies have focused on discovery of novel expressed transcripts rather than unbiased detection of differential expression and the choice of software for the latter task is limited. Variants of the *maxgap/minrun* algorithm (Kampa *et al.*, 2004; Royce *et al.*, 2005) like `TAS` require dataset-specific cutoff parameters and `MAT` has been developed for ChIP-chip data analysis and requires adapted parameters to be applicable to expression tiling array data. Both hampers the applicability of these methods in different scenarios without manually inspecting a small set of expected positive regions.

We have presented `TileShuffle`, a method specifically designed for expression and differential expression analysis of tiling array data. It implements a statistical approach to detect expression or differential expression in terms of differences from the background distribution that avoids any intensity-related parameters. `TileShuffle` reduces the most dominant tiling array biases using an affinity-dependent permutation in conjunction with a windowing approach. A related resampling approach has been used by Guttman *et al.* (2009), which does, however, not consider probe affinities and is not applied to detection of differential expression.

We compared `TileShuffle`, `TAS`, and `MAT` in two different test scenarios. In the cell cycle dataset, where a custom array was used for validiation, `TileShuffle` achieved significantly lower false discovery rates under equal sensitivities. This test scenario has the advantages of building on a biologically meaningful experiment with the associated noise in expression signals and transcriptome complexity and of calculating sensitivity and specificity on a large number of intervals. However, the custom array data has an FDR

itself, which is better controlled and significantly lower than for the tiling array, but still providing a surrogate for a true reference.

In the second scenario, the algorithms are compared using a spike-in dataset (Sasaki *et al.*, 2007). The differences between the three algorithms are smaller than in the previous scenario. `TileShuffle`, however, is the only one obtaining sensitivities >50%. The spike-in experiment has the advantage of a clear definition of positve and negative intervals for calculating sensitivity and specificty. However – though large for a spike-in experiment – 162 differentially expressed elements is a small number compared with the cell cycle experiment, the noise is low, the basal expression level is already high and a 10-fold differential expression is a strong effect in biological experiments. The scenario is thus rather artificial.

Apart from the ROCs, `TileShuffle` clearly outmatches `TAS` and `MAT` in the recovery of transcript structures by identifying the intron–exon structure more accurately. However, `TileShuffle` fails to detect very short exons because of the windowing approach.

Additionally, `TileShuffle` can incorporate replicate experiments and supports input data as custom-formatted files and hence is not dependent on any technology or tiling array design and can also be applied to ChIP-chip data by selecting a larger window size. The required computation time of `TileShuffle` is considerably higher than for `TAS` and `MAT`. However, it is negligible compared with efforts for the genome-wide tiling array experiment and thus does not constitute a bottleneck in the analysis work flow.

Funding: This publication was supported in part by the Initiative and Networking Fund of the Helmholtz Association (VH-NG-738), by LIFE Leipzig Research Center for Civilization Diseases, Universität Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Agarwal,A. *et al*. (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, **11**, 383.

Bar-Joseph,Z. *et al*. (2008) Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci. USA*, **105**, 955–960.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, **57**, 289–300.

Bertone,P. *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bradford,J.R. *et al*. (2010) A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics*, **11**, 282.

Cherbas,L. *et al*. (2011) The transcriptional diversity of 25 Drosophila cell lines. *Genome Res.*, **21**, 301–314.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ghosh,S. *et al*. (2007) Differential analysis for high density tiling microarray data. *BMC Bioinformatics*, **8**, 359.

Graveley,B.R. *et al*. (2011) The developmental transcriptome of Drosophila melanogaster. *Nature*, **471**, 473–479.

Guttman,M. *et al*. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Harrow,J. *et al*. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7** (Suppl. 1), S4.1–S4.9.

Huber,W. *et al*. (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**, 1963–1970.

Ji,H. and Wong,W.H. (2005) Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.

Johnson,W.E. *et al*. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.

Judy,J.T. and Ji,H. (2009) Tileprobe: modeling tiling array probe effects using publicly available data. *Bioinformatics*, **25**, 2369–2375.

Kadener,S. *et al*. (2009) Genome-wide identification of targets of the drosha-pasha/DGCR8 complex. *RNA*, **15**, 537–545.

Kampa,D. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, **14**, 331–342.

Kapranov,P. *et al*. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Karpikov,A. *et al*. (2011) Tiling array data analysis: a multiscale approach using wavelets. *BMC Bioinformatics*, **12**, 57.

Kechris,K.J. *et al*. (2010) Generalizing moving averages for tiling arrays using combined p-value statistics. *Stat. Appl. Genet. Mol. Biol.*, **9**, Article 29.

Lee,S.J. *et al*. (2009) Cellular stress created by intermediary metabolite imbalances. *Proc. Natl Acad. Sci. USA*, **106**, 19515–19520.

Li,W. *et al*. (2005) A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21** (Suppl. 1), i274–i282.

Munch,K. *et al*. (2006) A hidden markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7**, 239.

Pedersen,J.S. *et al*. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

Rinn,J.L. *et al*. (2003) The transcriptional activity of human chromosome 22. *Genes Dev.*, **17**, 529–540.

Royce,T.E. *et al*. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.*, **21**, 466–475.

Royce,T.E. *et al*. (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, **23**, 988–997.

Sasaki,D. *et al*. (2007) Characteristics of oligonucleotide tiling arrays measured by hybridizing full-length cDNA clones: causes of signal variation and FP signals. *Genomics*, **89**, 541–551.

Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R. *et al*. (eds) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Spencer,W.C. *et al*. (2011) A spatial and temporal map of C. elegans gene expression. *Genome Res.*, **21**, 325–341.

Taskesen,E. *et al*. (2010) Hat: hypergeometric analysis of tiling-arrays with application to promoter-genechip data. *BMC Bioinform.*, **11**, 275.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,Y. *et al*. (2011) The Caenorhabditis elegans intermediate-size transcriptome shows high degree of stage-specific expression. *Nucleic Acids Res.*, **39**, 5203–5214.

Washietl,S. *et al*. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Witten,D. and Tibshirani,R. (2007) A comparison of fold-change and the t-statistic for microarray data analysis. *Technical Report*. Department of Statistics, Stanford University. http://www-stat.stanford.edu/ tibs/ftp/FCTComparison.pdf

Zacher, B., Kuan, P. F. and Tresch, A. (2010) Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. *BMC Bioinformatics,* **11**, 194.