BIOINFORMATICS ORIGINAL PAPER
Vol. 22 no. 16 2006, pages 1963–1970
doi:10.1093/bioinformatics/btl289

Gene expression

Transcript mapping with high-density oligonucleotide tiling arrays

Wolfgang Huber[1,*], Joern Toedling[1] and Lars M. Steinmetz[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK and
[2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received on April 3, 2006; accepted on May 24, 2006
Advance Access publication June 20, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: High-density DNA tiling microarrays are a powerful tool for the characterization of complete transcriptomes. The two major analytical challenges are the segmentation of the hybridization signal along genomic coordinates to accurately determine transcript boundaries and the adjustment of the sequence-dependent response of the oligonucleotide probes to achieve quantitative comparability of the signal between different probes.

Results: We describe a dynamic programming algorithm for finding a globally optimal fit of a piecewise constant expression profile along genomic coordinates. We developed a probe-specific background correction and scaling method that employs empirical probe response parameters determined from reference hybridizations with no need for paired mismatch probes. This combined analysis approach allows the accurate determination of dynamical changes in transcription architectures from hybridization data and will help to study the biological significance of complex transcriptional phenomena in eukaryotic genomes.

Availability: R package tilingArray at http://www.bioconductor.org.

Contact: huber@ebi.ac.uk

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

1 INTRODUCTION

High-density genomic tiling microarrays cover a complete genome or a large fraction of it with densely tiled oligonucleotide probes. The major applications of these arrays are for transcriptome analysis, DNA–protein-binding and chromatin modification assays (ChIP-chip) and DNA variation detection (Bertone *et al*., 2004; Carroll *et al*., 2005; David *et al*., 2006; Gendrel *et al*., 2005; Gresham *et al*., 2006; Kampa *et al*., 2004; Kapranov *et al*., 2002; Mockler *et al*., 2005; Royce *et al*., 2005; Samanta *et al*., 2006; Schadt *et al*., 2004; Selinger *et al*., 2000; Shoemaker *et al*., 2001; Stolc *et al*., 2004; Sun *et al*., 2003; Yamada *et al*., 2003).

The current highest-density tiling microarrays contain 6.5 million distinct features on a single chip and are produced by the company Affymetrix. Each feature measures 5 μm × 5 μm in size and typically contains olignucleotide probes 25 bases in length. In this paper, we focus on the specific analytical challenges posed by the application of short oligonucleotide tiling arrays to transcriptome analysis (Royce *et al*., 2005).

The first task is the detection of transcript boundaries, i.e. transcript start and stop sites. The challenge is to obtain optimal estimates of the genomic coordinates of transcript boundaries from the tiling array data. The hybridization signal corresponds to the sum of target molecules at each probe position. The maximal precision is determined by the offset between the tiling features and can be as fine as a few bases, however in practice, it is often limited by noise and the transcriptional activity of the genomic region surrounding the transcripts. In addition, we want to quantify the relative level of transcript abundance and have a statistical measure of uncertainty for each estimated transcript boundary.

Second, we would like to detect the presence of architectural features within genes such as alternative transcription start and stop sites, alternative splicing and alternative partial degradation. For this, we need to compare the signal from probes that target different parts of the same gene. To achieve this, we must address the problem of differential probe response: different oligonucleotide probes may report consistently different intensities even if the abundance of their target molecules is the same. Such sequence dependent variation can extend over several orders of magnitude. If not accounted for, this variation leaves a great deal of apparent noise in the data and will obstruct the reliable detection of transcript boundaries, levels and architectures.

Here we describe a segmentation method that addresses the first of the above challenges and a DNA reference normalization method that addresses the second. These methods were successfully used in David *et al*. (2006) and promise to advance the extraction of biological meaning from tiling array data.

2 METHODS

2.1 Example data

We employed an Affymetrix oligonucleotide array that contains 6 553 600 probes and interrogates both strands of the complete genomic sequence of *Saccharomyces cerevisiae* with 25mer probes tiled at intervals of 8 nt on each strand (17 nt overlap) and a 4 nt offset of the tile between strands. This design enables a 4 bp resolution for hybridization of double stranded targets and an eight base resolution for strand-specific targets.

RNA was isolated from yeast cells during the exponential growth phase in rich medium (YPD) and was doubly enriched for polyadenylated molecules. First-strand cDNA was synthesized using random primers and labeled.

*To whom correspondence should be addressed.

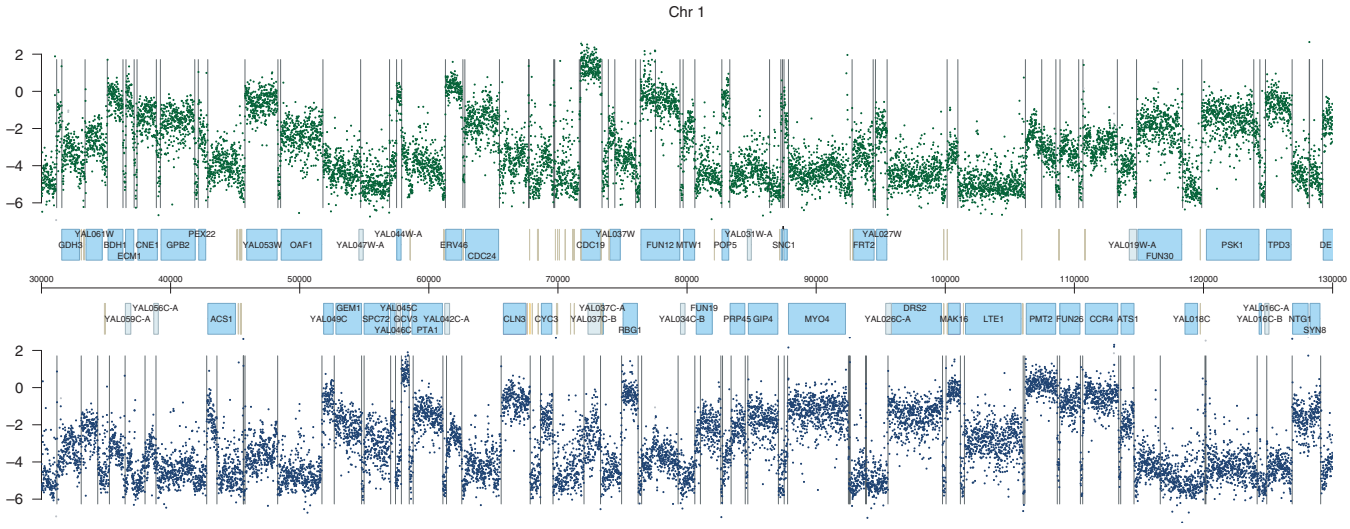

Fig. 1. Visualization of yeast tiling array intensities along 100 kb of chromosome 1, corresponding to ∼1% of the genome. The plot shows the normalized $\log_2$ hybridization intensities (*y*-axis) along genomic coordinates (*x*-axis in bp). Each dot corresponds to a unique probe, Watson (+) strand in green and Crick (−) strand in blue. Annotated open reading frames (ORFs) are shown as blue boxes, dubious ORFs as light blue boxes, transcription factor binding sites as grey bars. Vertical lines are segment boundaries. The use of the data to map the boundaries and levels of all transcripts, including the untranslated regions (UTRs) of protein-coding genes, antisense transcripts, and currently uncharacterized non-coding RNAs was described by David *et al*. (2006). A browsable on-line database of such plots for the whole yeast genome is available at http://www.ebi.ac.uk/huber-srv/queryGene. A colour version of this figure is available as part of the supplementary data.

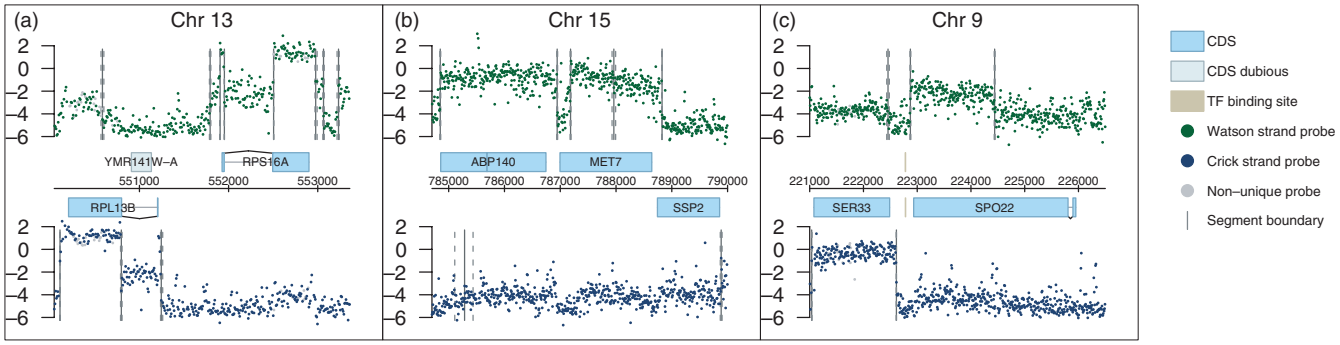

Fig. 2. Detailed views on segmentation results. The 95%-confidence intervals are shown by vertical dashed lines. Note that the confidence intervals are calculated in terms of data sampling points, not genomic coordinates. Hence, in cases where the data are unequivocal, the interval boundaries coincide with the change-point estimate itself, e.g. see the 5′ end of *SER33* in panel (c). (**a**) Spliced transcripts *RPS16A* and *RPL13B*. (**b**) Complex transcript architecture of *MET7*. (**c**) Transcript antisense to *SPO22*. CDS refers to coding sequence; TF, transcription factor. A colour version of this figure is available as part of the supplementary data.

Genomic DNA was isolated from the same yeast strain. RNA and DNA samples were hybridized in three replicates each. The data are available at ArrayExpress (accession number E-TABM-14) and in the Bioconductor data package davidTiling. The biological findings from this study are described in David *et al*. (2006).

2.2 Structural change model segmentation

Transcript boundaries can be identified from sudden changes of the hybridization signal plotted along a linear genomic coordinate axis (Fig. 1). The signal is affected by noise, which suggests that smoothing or probabilistic modeling of the signal is beneficial. The signal could be either the hybridization intensities from a single RNA sample or it could be a per-probe summary statistic for the comparison of multiple conditions or time points.

Perhaps the most obvious approach is to move a sliding window along the coordinate axis and to measure the evidence for the presence of transcripts by computing a scan statistic at each window step. As we discuss in Section 3.1, such approaches are not well-suited to precisely determine the boundaries of transcripts. An improvement is provided by hidden Markov models. Discrete state hidden Markov models are powerful and popular tools in biological sequence analysis, and there are efficient and elegant dynamic programming algorithms for fitting them to data (Durbin *et al*., 2002; Rabiner, 1989).

Tjaden *et al*. (2002) applied a two-state hidden Markov model to the detection of untranslated region (UTR) boundaries, where the two states corresponded to presence or absence of transcription. However, transcript abundance is a continuous-valued quantity, and there are biological effects such as alternative transcription start and end sites and partial transcript degradation that result in complex signal patterns (Fig. 2). These patterns are richer than can be detected by a simple ‘on/off’ model. We propose a continuous-state model whose hidden state can be any real number.

2.2.1 The model The SCM model is well known in econometrics (Bai and Perron, 2003; Zeileis *et al*., 2002) for the modeling of sudden jumps in financial time series and has been applied to the segmentation of array-CGH data (Picard *et al*., 2005). It models the data as a piecewise constant function of chromosomal coordinates,

$$z_{ki} = \mu_s + \varepsilon_{ki} \quad \text{for } t_s \le k < t_{s+1}, \tag{1}$$

where $k = 1,2,\ldots,n$ indexes the probes in ascending order along the chromosome, i indexes replicate experiments, z_{ki} is the signal from the k-th probe in the i-th replicate, $t_2,\ldots,t_S$ parameterize the segment boundaries, $t_1 = 1$ and $t_{S+1} = n+1$, S is the total number of segments, μ_s is the mean signal level of the s-th segment, and ε_{ki} are the residuals. $\mu_1,\ldots,\mu_S$ can be any set of real numbers. $t_2,\ldots,t_S$ are also called the change-points. Model (1) is applied separately to each chromosome and, if the signal is strand-specific, to each of its two strands.

2.2.2 Parameter estimation Fitting the model (1) can be accomplished by minimizing the sum of squared residuals

$$G(t_1,\ldots,t_S) = \sum_{s=1}^{S}\sum_{i=1}^{I}\sum_{k=t_s}^{t_{s+1}-1}(z_{ki}-\hat{\mu}_s)^2, \tag{2}$$

where S is the number of segments, I is the number of replicate arrays and $\hat{\mu}_s$ is the arithmetic mean of the z_{ki} in segment s.

There is a dynamic programming algorithm that allows the globally optimal set of parameters $\hat{t}_1,\ldots,\hat{t}_S$, for all values of S between 1 and $S_{\max}$, to be obtained in quadratic time $O(n^2)$. Recent presentations of the algorithm include Bai and Perron (2003), Picard *et al*. (2005). If one bounds the maximal length of individual segments to a fixed size (e.g. $l = 20$ kb), then the complexity of the algorithm can be reduced to $O(nl)$. With this approach, which we have taken, sequences of several hundred thousand probes can be processed in a single run. We provide a C implementation in the function `segment` in the *tilingArray* package of the Bioconductor project (Gentleman et al., 2004).

2.2.3 Confidence intervals Bai and Perron (1998) present an asymptotic theory for inference on the SCM model (1), and in a companion paper (Bai and Perron, 2003) they provide a comprehensive and detailed discussion of the associated computational aspects. The calculation of the confidence intervals on estimated change-points $\hat{t}_s$ involves the distribution of the argmax functional of a process composed of two independent Brownian motions. The drift and scale parameters of these processes depend on the difference between the segment means and on the standard deviation and serial correlation of the residuals. Owing to the limitations of floating-point arithmetic, the correct numerical evaluation of this distribution function is not trivial. Zeileis and Kleiber (2005) point out some of the caveats. The package *tilingArray* makes use of their implementation of the confidence interval estimation in the R package *strucchange* (Zeileis *et al*., 2002, 2003).

2.2.4 Model selection The only user-defined parameter of the SCM model (1) is the number of segments S ($1 \le S \le S_{\max}$). In principle, it can be chosen by a penalized likelihood approach (Picard *et al*., 2005), which we now discuss. Assuming that the residuals ε_{ki} in Equation (1) are independent and identically normal, the log-likelihood is

$$\log L = -\frac{N}{2}\left(1+\log\frac{2\pi G}{N}\right), \tag{3}$$

where G is the sum of squared residuals from Equation (2), and $N = nI$ is the number of data points. Since the class of models with parameter $S-1$ is contained in that with parameter S, $\log L$ is a monotonically increasing function of S.

To penalize model complexity, we can consider the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). They are defined, e.g. in Hastie *et al*. (2001), as AIC $= -2\log L + 2p$ and BIC $= -2\log L + p\log N$, where p is the number of parameters of the model. In our case, $p = 2S$, since for a segmentation with S segments, we estimate $S-1$ change-points, S mean values and the standard deviation of the ε_{ki}. Hence the penalized likelihood functions are

$$\log\tilde{L}_{\mathrm{AIC}} = \log L - 2S, \tag{4}$$

$$\log\tilde{L}_{\mathrm{BIC}} = \log L - S\log N. \tag{5}$$

Since the probes on the array can overlap and some sources of noise are correlated for successive probes (for example, cross-hybridization or random fluctuations in the abundance of specific target fragments), the data will usually be serially correlated. Serial correlation is not a substantial problem for the point estimates $\hat{t}_s$ and $\hat{\mu}_s$, and it is explicitly taken into account in Bai's and Perron's confidence intervals. However, it needs to be considered when making inference based on the log-likelihoods (3–5).

2.3 DNA reference normalization

2.3.1 The model The fluorescence intensity values obtained from an oligonucleotide microarray hybridization do not directly correspond to interpretable physical units. The same abundance of a target transcript can result in systematically different values when measured with different oligonucleotide probes. This is due to a variety of reasons, among them the different thermodynamic properties of different polynucleotide sequences and biases in labeling efficiency.

The fluorescence intensity response of a probe k to the abundance x_k of its target molecule in the sample can be modeled as $y_k = \beta_k + \alpha_k x_k$, where y_k is the intensity of probe k, β_k is a term that represents unspecific (background) intensity, and α_k is a proportionality factor for the specific part of the signal (Rocke and Durbin, 2001). The specific part of the signal is (to reasonable approximation) proportional to the abundance x_k of the target molecule, while the unspecific part corresponds to the background fluorescence that is observed even in the absence of the intended target. Here we are not concerned with stochastic measurement error ('noise'), for which we refer to Huber *et al*. (2004). Furthermore, we assume that non-linear saturation effects are negligible. α_k and β_k are not usually known, and they can be different for each probe. This explains why even for $x_k = x_j$, in general $y_k \neq y_j$. The goal of DNA reference normalization is to estimate parameters a_k and b_k such that

$$y'_k = \frac{y_k - b_k}{a_k} \tag{6}$$

quantifies the target abundance in a way that is to sufficient approximation independent of k, the probe identity.

2.3.2 Parameter estimation We estimate a_k by the geometric mean of the intensities from three replicate array hybridizations of genomic DNA. This procedure is motivated by the fact that the abundance of the target is the same for all probes that have a unique match to the genome. Note that we are excluding probes with multiple matches to the genome. We are also not considering probes without any perfect matches in the genome and in particular, we are ignoring the so-called mismatch (MM) probes.

We have no direct way to obtain a detailed estimate of b_k for every probe, but we can assume that some of its probe to probe variability can be explained through a functional dependence on a_k. We use as an estimate $\hat{b}_k = f(\hat{a}_k)$ with a smooth function f, which we obtain as follows. Probes are grouped into 10 strata corresponding to the 10, 20, ..., 100% quantiles of $\hat{a}_k$. Within each stratum we calculate the midpoint of the shorth of the intensities of those probes whose target sequence is not annotated to be within a transcribed region on either strand. The shorth of a univariate distribution is defined as the shortest interval that contains at least half of the data, and its midpoint is a robust estimator of the location of the distribution. An estimate of the function f can be obtained from these values by linear interpolation or smoothing.

2.3.3 Between array normalization In order to deal with data from multiple arrays, we need to adjust for systematic variations in the intensities

between different arrays, which can be caused, for example, by varying amounts of sample material. If we now denote y'_k from Equation (6) by y'_{ki}, with the index i counting the different arrays, then we are looking for an affine transformation $y''_{ki} = (y'_{ki} - d_i)/c_i$, where d_i is an array-specific offset and c_i a scaling factor. Microarray intensities are usually transformed to a logarithmic scale in order to make the distributions of the stochastic noise components in the data more symmetric and more homogeneous (Dudoit *et al*., 2002; Durbin *et al*., 2002; Huber *et al*., 2002; Irizarry *et al*., 2003). Because for probes with weakly or unexpressed targets y''_{ki} can be close to zero or even non-positive, we apply the so-called generalized logarithmic transformation $z_{ki} = \mathrm{glog}_\Delta(y''_{ki}) = \log_2(y''_{ki} + \sqrt{y''^2_{ki} + \Delta^2})/2$. This transformation depends on a parameter Δ which is related to the size of the background noise.

The parameters c_i, d_i, Δ can be estimated from the data, and for this we use the robustified maximum-likelihood method provided by the Bioconductor package *vsn* (Huber *et al*., 2002).

2.3.4 Exclusion of non-responding probes We observed that a certain fraction of probes respond poorly to their target and are not informative. We allow for the exclusion of such data. In particular, we discard the probes whose estimated a_k is smaller than a user-defined quantile of the a_k—distribution.

The DNA reference normalization method described in this section is implemented in the function `normalizeByReference` of the *tilingArray* package.

3 RESULTS AND DISCUSSION

3.1 SCM segmentation

The main results of this paper are visualized in Figures 1 and 2, which show the application of SCM segmentation and DNA normalization to the yeast tiling array data. The segmentation clearly picks up the major change-points in the data, many of which correspond to the beginning or end of annotated genes. In addition to the change-point estimates, Figure 2 also includes 95%-confidence intervals as described in Section 2.2.3.

3.1.1 Comparison to sliding windows Previous high-density tiling array studies used sliding window methods in combination with a thresholding criterion for the identification of transcripts (Bertone *et al*., 2004; Kampa *et al*., 2004; Royce *et al*., 2005; Schadt *et al*., 2004). In contrast to SCM, which optimizes a clearly defined objective function, the sum of squares (Bai and Perron, 2003), sliding window methods are defined algorithmically. One of the main problems with sliding window approaches is shown in Figure 3. Such methods tend to produce biased estimates of the start and end points of transcribed regions, depending on the level of signal above background (Hastie *et al*., 2001).

3.1.2 Model selection SCM segmentation has one parameter, the number of segments S, which controls the model complexity. The data can be fit better by increasing S, and this will decrease the number of missed, real change-points (false negatives) for the cost of increasing the number biologically irrelevant change-points (false positives). In Section 2.2.4 we have described a standard penalized likelihood approach. A potential method of choosing S would be to use the value that maximizes a suitably penalized likelihood function. Figure 4 shows a plot of the log-likelihood as a function of the parameter S, together with two possible choices of penalized log-likelihoods according to the AIC and the BIC. While in particular $\tilde{L}_{\mathrm{BIC}}$ works well on the simulated data, both $\tilde{L}_{\mathrm{AIC}}$ and $\tilde{L}_{\mathrm{BIC}}$ would choose substantially higher values for S than

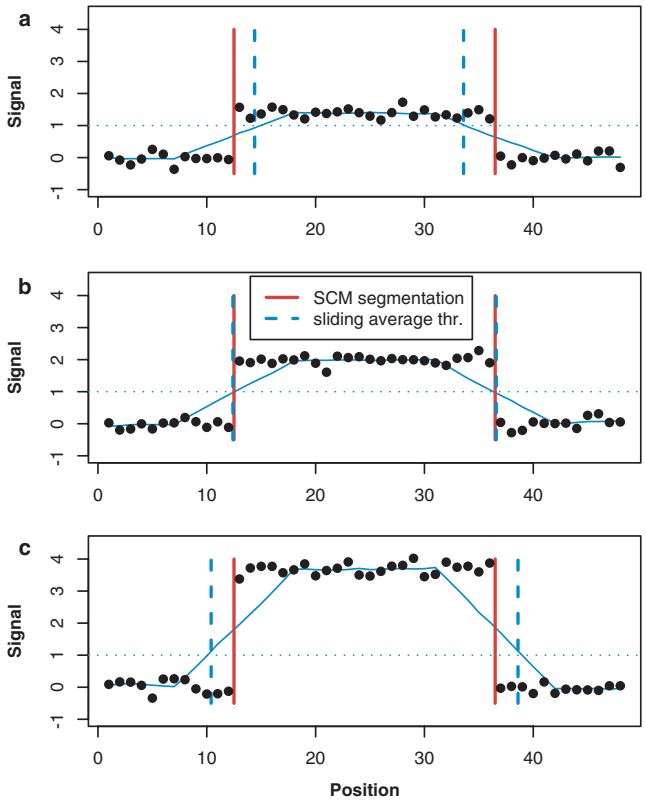

Fig. 3. Comparison between SCM segmentation and sliding average thresholding (SAT). (**a**) The dots correspond to simulated data for a weakly expressed transcript starting at position 13 and ending at 36. The vertical solid red lines show the change-points found by SCM segmentation. The blue line shows a sliding average (window width 11). The vertical dashed blue lines show the change-points found by thresholding the sliding average at a threshold of $y = 1$ (horizontal dotted line). The size of the transcript is underestimated. (**b**) As in panel a, for a moderately expressed transcript. The change-points from SCM segmentation and SAT coincide. (**c**) As in panel a, for a strongly expressed transcript. The size of the transcript is overestimated by SAT. SCM segmentation produces unbiased estimates in all cases. A colour version of this figure is available as part of the supplementary data.

that which we decided upon for the analysis presented in David *et al*. (2006) based on comparison of the segmentation with biological expectations.

We hypothesize that this discrepancy may be the consequence of the model of Equation (1) being too simple, in two ways. First, there are biological phenomena that lead to more complex hybridization profiles than the piecewise constant shape assumed by the model. Second, the residuals ε_{ki} are in practice not independent, as discussed in Section 2.2.4. While the model is evidently useful to estimate meaningful change-points and confidence intervals, when S is given, it might not be powerful enough to also let us infer S.

We recommend the following strategy. Since the algorithm produces not just the optimal segmentation for a given number $S_{\max}$ of segments, but also all optimal segmentations with $S = 2,3,\ldots,S_{\max} - 1$, a practical approach is to do the computation with a choice of $S_{\max}$ that is comfortably too large. The results can be visualized

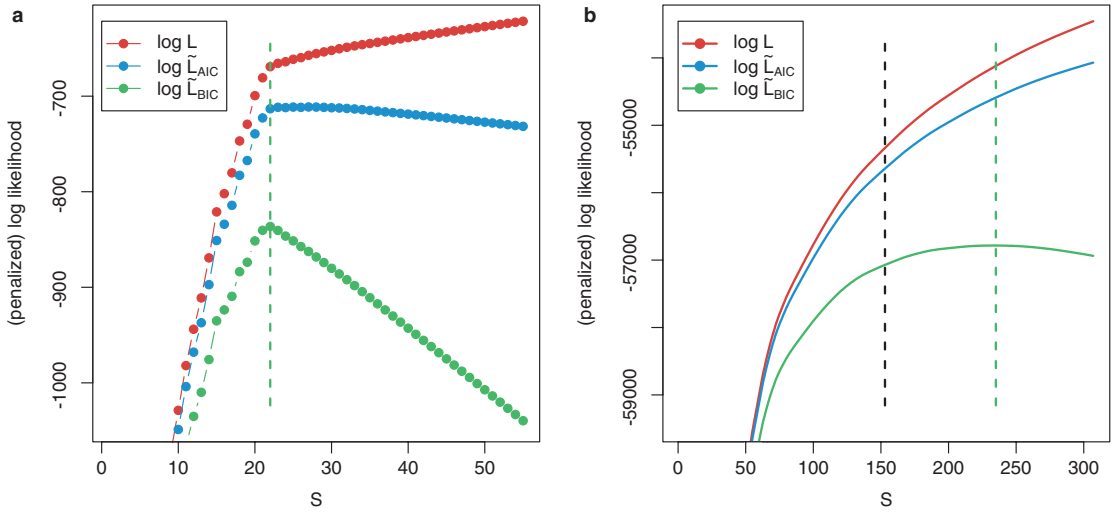

Fig. 4. Model selection. The plots show the log-likelihood (red) and penalized log-likelihoods according to the AIC (blue) and the Bayes information criterion (BIC; green) as functions of the parameter S. (**a**) Simulated data according to model (1) with independent Normal ε_{ki} and $S = 22$. The vertical dashed green line is at the maximum of $\log \tilde{L}_{\mathrm{BIC}}$ and correctly identifies the true value of S. (**b**) Tiling array data from the Watson (+) strand of chromosome 1. The vertical dashed grey line at $S = 153$ corresponds to the parameter value that was used for Figures 1 and 2. The vertical dashed green line at $S = 235$ is at the maximum of $\log \tilde{L}_{\mathrm{BIC}}$. A colour version of this figure is available as part of the supplementary data.

for different values of S using the visualization tool provided in the *tilingArray* package. This tool was also used to create Figures 1 and 2 and the online supplement of David *et al*. (2006) (http://www.ebi.ac.uk/huber-srv/queryGene). By examining the results in control regions where one has clear expectations about the transcript structures, it is possible to identify an S that has a desirable trade-off between sensitivity and specificity, and to gain confidence in the algorithm's results in lesser known regions of the transcriptome. Often it is reasonable to expect that the segment length distribution should be approximately the same on different chromosomes, hence the choice of S is equivalent to the choice of the average segment length L_S, with S being the integer closest to L_C/L_S and L_C the length of the region to be segmented, typically a chromosome. In David *et al*. (2006), this procedure let us choose $L_S = 1500$ bases uniformly for all chromosomes.

3.2 Normalization

3.2.1 Visual assessment Figure 5 shows scatterplots of different types of signal along genomic coordinates. Each dot corresponds to a microarray feature. The intensities from a hybridization of genomic DNA are shown in Figure 5a. The y–axis is on the logarithmic scale to base 2. Ideally, all features should show the same intensity, since the copy number of genomic DNA is the same throughout. Some of the variation in Figure 5a can be explained by stochastic noise, but the larger part of it is systematic and is due to sequence-specific properties of either the probes or the target DNA (Naef and Magnasco, 2003; Wu *et al*., 2004; Zhang *et al*., 2003). The y-coordinate of each dot is also encoded using a pseudo-color scheme. Red corresponds to features that have a weak response, blue to those with the strongest response. The same coloring for each feature is also used in panels 5b–f.

Figure 5b shows the intensites resulting from hybridization of RNA, again on a logarithmic scale to base 2. One can clearly distinguish between transcribed regions, corresponding to the annotated genes RPN2 and SER33, and background intergenic regions. However, the signal appears rather noisy, with many individual features that map into the transcribed region showing weak signal, and a large spread of values even in the background region. Notably, this variation is not random, as can be seen from the coloring of the dots: to a large part, it can be explained by the probe response as encoded by the color. This motivates the use of the DNA intensities for adjusting the probe sequence related signal variation.

Figure 5c shows the result of dividing the RNA-signal by the DNA-signal, then taking the logarithm to base 2. Since the overall scaling is arbitrary, we have shifted the data in panels 5c–f such that the 5% quantile is at 0. While the distribution of the data within the transcribed regions is now much tighter, there is still considerable variability in the background region. Remarkably, this background variability is not random, one can see a pattern that correlates with the coloring of the dots. This motivates a probe specific background correction that again employs the DNA intensities from panel 5a.

Figure 5d shows the z_{ki} values resulting from the DNA reference normalization. While the spread of the data in the background region is not substantially different compared to panel 5c, we note two important aspects: the distribution of the noise in the background is now more symmetric, and, more importantly, the difference between the mean signal in the background regions and the transcribed regions is increased. Background correction does not reduce the variance, but it increases the dynamic range and hence the sensitivity to detect weak signals (Irizarry *et al*., 2003).

Figure 5e shows the same data as in panel 5d, but with the 5% of features that had the weakest signal in the DNA hybridization removed, as described in Section 2.3.4. This removes many of the outliers at little cost of good data.

For comparison, Figure 5f shows the result of a normalization that is similar to Figure 5e, with the only difference that for the

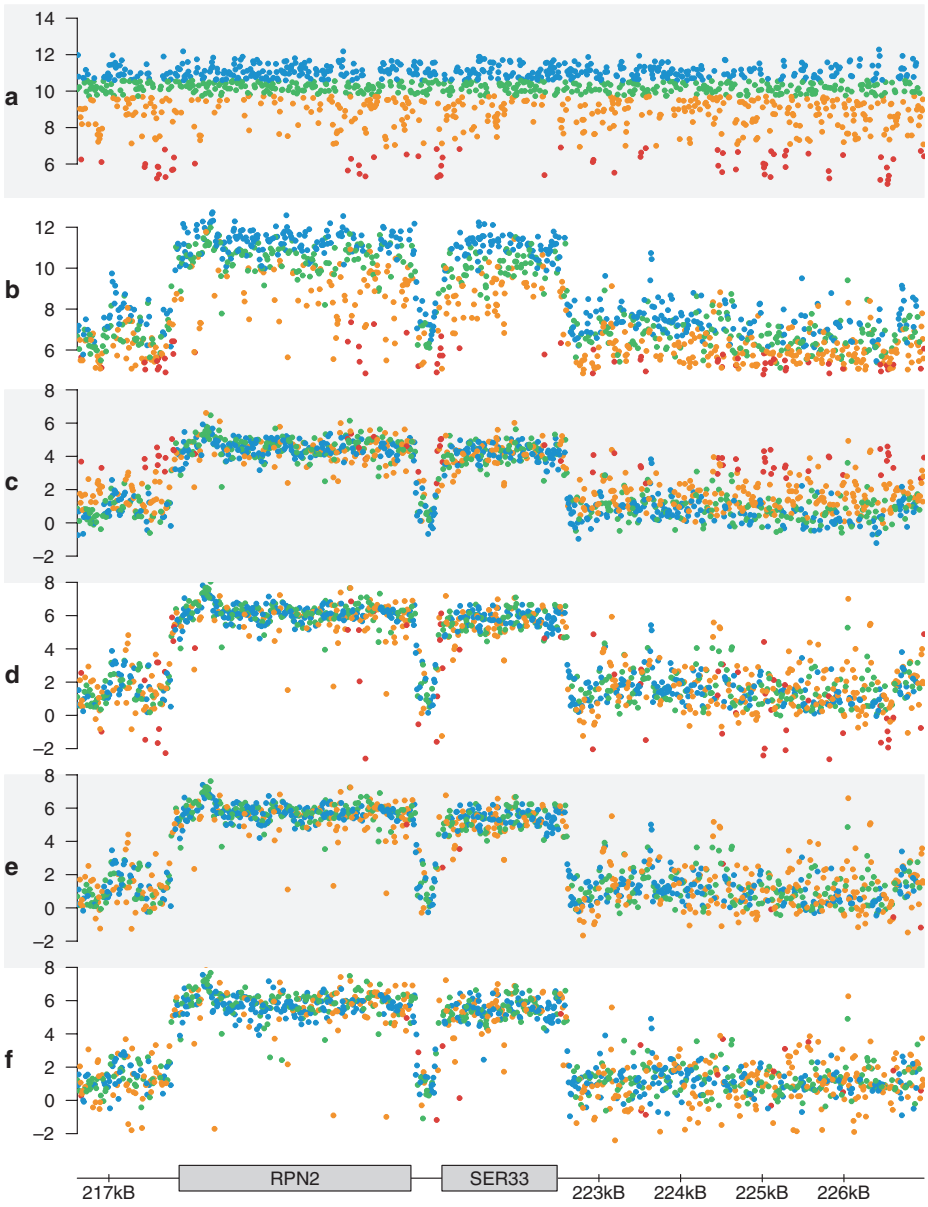

Fig. 5. Along-chromosome plot of array intensities from hybridization of DNA (**a**), RNA (**b**), RNA values divided by DNA values (**c**), with background subtraction with (**d**) and without non-responding probes (**e**), alternative background correction by paired MM probes (**f**). A colour version of this figure is available as part of the supplementary data.

estimation of the background parameters b_k the intensity of the MM probe that is paired with each perfect match (PM) probe is used instead of the interpolated values from background-level PM probes. Using paired MM probes for background correction can increase the dynamic range of the signal, but one of the main limitations of this approach is that it also greatly increases the signal's variance, which can lead to a net increase of mean squared error.

3.2.2 Quantitative assessment In order to assess quantitatively the results of the different procedures 5b–f, we consider a signal/noise ratio. We look at a set of control regions, two positive control regions within the ORFs of RPN2 and SER33 at coordinates 217860—220697 and 221078—222487 and two negative control regions in the background at coordinates 216800—217700 and 222800—227000. The assumption is that the signal within a region should be constant and deviations from that are noise, while the difference between positive and negative controls should be large and is counted as signal. Noise σ is calculated as the average of the differences between 97.5 and 2.5% quantiles of the data within each of the control regions. The range between the 97.5 and 2.5% quantiles contains 95% of the data.

$$\sigma = \frac{1}{Q_N^{0.975} - Q_N^{0.025}} \cdot \frac{\sum_{r \in \text{pos, neg}} Q_r^{0.975} - Q_r^{0.025}}{|\text{pos}| + |\text{neg}|}. \qquad (7)$$

Table 1. Signal to noise ratio $\Delta\mu/\sigma$ of different data processing methods b–f as described in Section 3.2 and Figure 5

Method	b	c	d	e	f
$\Delta\mu/\sigma$	3.22	3.47	4.04	4.58	4.36

Here, the symbol r counts over the different regions. The constant in the denominator is the differences between 97.5 and 2.5% quantiles for the standard normal distribution, hence σ is equal to 1 if the data come from the standard Normal distribution.

Signal Δ_μ is calculated as the difference between the averages of positive and negative control regions, $\Delta\mu = \sum_{r\in\text{pos}} \mu_r/|\text{pos}| - \sum_{r\in\text{neg}} \mu_r/|\text{neg}|$.

The result is shown in Table 1. We have explored many variations of this calculation, using different definitions of σ, $\Delta\mu$, and of the control regions. The ranking of the methods was always the same as shown in Table 1. The data and the code for the calculations that produce Figure 5 and Table 1 are available in the online documentation of the package *tilingArray*.

4 CONCLUSION

The adjustment of probe sequence related signal variation is a fundamental problem in the analysis of oligo-nucleotide microarray data (Hubell, 2005; Irizarry *et al*., 2003; Li and Wong, 2001; Naef and Magnasco, 2003; Wu *et al*., 2004; Zhang *et al*., 2003). Current methods rely on pre-specified groupings of probes each matching the same transcript, so-called probe sets, and their focus has been on the accurate detection of differential expression between different conditions rather than the detection of internal structure within the probe set. The advent of high-density tiling arrays enables us to observe transcript architecture, including transcription start and stop sites, splicing and degradation and possibly their differential regulation. A prerequisite is the quantitative comparability, at least approximately, of microarray signals from different regions of one transcript. The DNA reference normalization of Section 2.3 responds to this requirement.

Our method does not employ the data from the paired MM probes. The manufacturer's intention for these features is to serve as controls for unspecific background hybridization. However, as we have shown in Figure 5 and Table 1, a much smaller set of control probes is sufficient and even produces slightly better results. In particular, we show that the background component of a probe's signal can be estimated from a control population of similar probes that may perfectly match to the genome, but whose target does not appear to be transcribed. Since we use a robust estimation technique, it does not matter if this distribution contains a minority of probes with specific, non-background signal. We can save half of the real estate and about half of the cost of a chip at practically no loss for the analysis.

For the estimation of the probe-specific scaling and background parameters, we have opted to characterize each probe by a single value empirically obtained from the DNA reference hybridization. In addition, or indeed alternatively, one could try to use the probe sequence information to build a regression model on sequence-related variables for the background and scaling factor of each probe (Johnson *et al*., 2006; Naef and Magnasco, 2003; Wu *et al*., 2004; Zhang *et al*., 2003). Such a model can collect strength by smoothing across probes that are similar in sequence space and could also employ information on unspecific hybridization that is provided by so-called 'antigenomic' probes on recent Affymetrix GeneChip designs. Our attempts at such a model with the present data have not produced results that were better than the DNA reference normalization described in Section 2.3, but clearly there is room for more research.

We have described a simple structural change model (SCM) for RNA hybridization profiles along the genome, namely a piecewise-constant function. This model lends itself to an efficient dynamic programming algorithm for optimally estimating the change-point positions, which together with the segment levels are of primary biological interest. In addition to the change-point positions, the theory of SCMs also provides estimators for their confidence intervals. Figure 2 shows how the calculated confidence intervals adequately reflect the uncertainty in the change-point position depending on the steepness and height of the change and the noise level. The confidence intervals are useful for the ranking and interpretation of the fitted change-points.

SCMs also allow the modeling of general linear relationships between a possibly vector-valued regressor along a linear coordinate and a possibly vector-valued dependent variable (Bai and Perron, 2003; Zeileis *et al*., 2002, 2003). Among the uses for such a generalized approach could be, for example, the modeling of decaying flanks (Bourgon and Speed, 2006). Remember that a linear decay on the logarithmic data scale corresponds to an exponential decay on the fluorescence scale, which is often a good approximation for many naturally occurring length distributions.

The methods that we have presented here were successfully used in David *et al*. (2006) to identify the boundary, structure and level of coding and non-coding transcripts of yeast. Apart from expected transcripts, this study found operon-like transcripts, transcripts from neighboring genes that are not separated by intergenic regions and genes with complex transcriptional architecture. It mapped the positions of 3′-and 5′-UTRs of coding genes and identified hundreds of RNA transcripts both antisense to, and distinct from, annotated genes. The methods presented here, DNA reference normalization and SCM segmentation, were instrumental for the analysis, by providing a clean normalized signal and accurate transcript boundary identifications. We expect that the methods will also be useful in the study of transcriptional complexity under dynamic conditions and in other organisms. With suitable adaption they should also be valuable for the application of tiling arrays in the detection of genomic regions purified in chromatin-immunoprecipitation experiments. Owing to their ability to interrogate the entire genome we expect tiling microarrays soon to become a widely used tool.

ACKNOWLEDGEMENTS

The authors thank Lior David for fruitful discussions regarding the DNA reference normalization, Achim Zeileis for insightful comments about the estimation of confidence intervals and his software package *strucchange* and Richard Bourgon for helpful comments on the manuscript. The authors also thank the contributors to the Bioconductor (www.bioconductor.org) and R (www.R-project.org) projects for their software. This work has been supported by the

Deutsche Forschungsgemeinschaft and the National Institutes of Health (L.M.S.) and by the European Community's Sixth Framework Programme contract (HeartRepair) LSHM-CT-2005-018630 (J.T.). Funding to pay the Open Access publication charges for this article was provided by European Community's Sixth Framework Programme contract (HeartRepair) LSHM-CT-2005-018630.

Conflict of Interest: none declared.

REFERENCES

Bai,J. and Perron,P. (1998) Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

Bai,J. and Perron,P. (2003) Computation and analysis of multiple structural change models. *J. Appl. Econom.*, **18**, 1–22.

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

Bourgon,R. and Speed,T.P. (2006) A model for chromatin immuno-precipitation/high density tiling array experiments: implications for data analysis. In Takahashi, M. and Winegarden, N. (eds), *Profiling Transcriptional Activity with Promoter and CpG Microarrays*. DNA Press, Eagleville, PA.

Carroll,J.S. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

David,L. *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA*, **103**, 5320–5325.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, **12**, 111–139.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Durbin,B.P. *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

Gendrel,A.-V. *et al.* (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat. Methods*, **2**, 213–218.

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

Gresham,D. *et al.* (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science*, **311**, 1932–1936.

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

Hubbell,E. (2005) *PLIER White Paper*. Affymetrix, Santa Clara, California.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Huber,W., von Heydebreck,A. and Vingron,M. (2004) Error models for microarray intensities. In Quackenbush, J. (ed), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Wiley, New York.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Johnson,J.M. *et al.* (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.

Johnson,W.E., Li,W., Meyer,C.A., Gottardo,R., Carroll,J.S., Brown,M. and Liu,X.S. (2006) Model-based Analysis of Tiling-array for ChlP-chip. *Proc. Natl Acad. Sci. USA*, In Press.

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.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Mockler,T.C. *et al.* Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.

Naef,F. and Magnasco,M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E*, **68**, 011906.

Picard,F. *et al.* (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Rocke,D.M. and Durbin,B.P. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

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

Samanta,M.P. *et al.* (2006) Global identification of noncoding RNAs in *Saccharomyces cerevisiae* by modulating an essential RNA processing pathway. *Proc. Natl Acad. Sci. USA*, **103**, 4192–4197.

Schadt,E.E. *et al.* (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.*, **5**, R73.

Selinger,D.W. *et al.* (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.*, **18**, 1262–1268.

Shoemaker,D.D. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.

Stolc,V. *et al.* (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, **306**, 655–660.

Sun,L.V. *et al.* (2003) Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **100**, 9428–9433.

Tjaden,B. *et al.* (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, **18** (Suppl. 1), 337–344.

Wu,Z. *et al.* (2004) A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.

Yamada,K. *et al.* (2003) Empirical analysis of transcriptional activity in the *Arabidopsis genome*. *Science*, **302**, 842–846.

Zeileis,A. *et al.* (2002) strucchange: an R package for testing for structural change in linear regression models. *J. Stat. Software*, **7**, 1–38.

Zeileis,A. *et al.* (2003) Testing and dating of structural changes in practice. *Comput. Stat. Data Anal.*, **44**, 109–123.

Zeileis,A. and Kleiber,C. (2005) Validating multiple structural change models—a case study. *J. Appl. Econom.*, **20**, 685–690.

Zhang,L. *et al.* (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.