

RESEARCH ARTICLE

Open Access

# Wavelet-based detection of transcriptional activity on a novel *Staphylococcus aureus* tiling microarray

Víctor Segura<sup>1\*</sup>, Alejandro Toledo-Arana<sup>2</sup>, Maite Uzqueda<sup>1</sup>, Iñigo Lasa<sup>2</sup> and Arrate Muñoz-Barrutia<sup>3</sup>

## Abstract

**Background:** High-density oligonucleotide microarray is an appropriate technology for genomic analysis, and is particularly useful in the generation of transcriptional maps, ChIP-on-chip studies and re-sequencing of the genome. Transcriptome analysis of tiling microarray data facilitates the discovery of novel transcripts and the assessment of differential expression in diverse experimental conditions. Although new technologies such as next-generation sequencing have appeared, microarrays might still be useful for the study of small genomes or for the analysis of genomic regions with custom microarrays due to their lower price and good accuracy in expression quantification.

**Results:** Here, we propose a novel wavelet-based method, named ZCL (zero-crossing lines), for the combined denoising and segmentation of tiling signals. The denoising is performed with the classical SUREshrink method and the detection of transcriptionally active regions is based on the computation of the Continuous Wavelet Transform (CWT). In particular, the detection of the transitions is implemented as the thresholding of the zero-crossing lines. The algorithm described has been applied to the public *Saccharomyces cerevisiae* dataset and it has been compared with two well-known algorithms: pseudo-median sliding window (PMSW) and the structural change model (SCM). As a proof-of-principle, we applied the ZCL algorithm to the analysis of the custom tiling microarray hybridization results of a *S. aureus* mutant deficient in the sigma B transcription factor. The challenge was to identify those transcripts whose expression decreases in the absence of sigma B.

**Conclusions:** The proposed method archives the best performance in terms of positive predictive value (PPV) while its sensitivity is similar to the other algorithms used for the comparison. The computation time needed to process the transcriptional signals is low as compared with model-based methods and in the same range to those based on the use of filters. Automatic parameter selection has been incorporated and moreover, it can be easily adapted to a parallel implementation. We can conclude that the proposed method is well suited for the analysis of tiling signals, in which transcriptional activity is often hidden in the noise. Finally, the quantification and differential expression analysis of *S. aureus* dataset have demonstrated the valuable utility of this novel device to the biological analysis of the *S. aureus* transcriptome.

\*Correspondence: vsegura@unav.es

<sup>1</sup> Genomics, Proteomics and Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona, Spain

Full list of author information is available at the end of the article

## Background

The complete deciphering of the information contained in the genome would be helpful to improve our understanding of the biological processes occurring in living organisms. High-density oligonucleotide-based whole-genome microarray is an extensively used technology to detect the expression of all RNA species including protein coding RNAs and non-coding RNAs. It is particularly suitable for the analysis of whole small-sized genomes such as those corresponding to bacteria. For these organisms high resolution can be achieved with the microarrays currently provided by the manufactures.

Applications of tiling array technology include the generation of transcriptional maps and annotations of genomes, the identification of transcription factor binding sites, the analysis of alternative splicing events, the analysis of methylation states, the discovery of genotyping and polymorphism, and the re-sequentionation of genomes [1].

The emerging high-throughput next generation DNA sequencing (NGS) technologies [2] have revolutionized transcriptomics by allowing RNA analysis through cDNA sequencing on a massive scale (RNA-seq). Several limitations inherent to microarray technologies are overcome by NGS technologies, in particular, it is not necessary to design appropriate probes and the experimental reproducibility is guaranteed. However, the microarray design presented in [3] allowed a comprehensive examination of gene expression and genome-wide identification of alternative splicing as well as detection of coding and non-coding transcripts. This microarray (Affymetrix GG-H array) was compared with RNA-seq in [3]. The reproducibility in the estimation of gene and exon abundance was high and even more sensitive than RNA-seq at the exon level. This microarray design contains as targets near 50000 highly transcribed fragments of unknown functions from Affymetrix tiling microarray data [4]. The NGS experiments highlighted that 49% of these fragments had uniquely mapped reads, revealing a high degree of concordance between both technologies.

The analysis of a tiling microarray experiment starts with a two-step process that generates a discrete signal. First, the DNA or RNA samples are hybridized in the custom designed tiling array. Second, for each probe, the raw intensities are converted to a score [5]. The result is a discrete intensity signal with a value per probe.

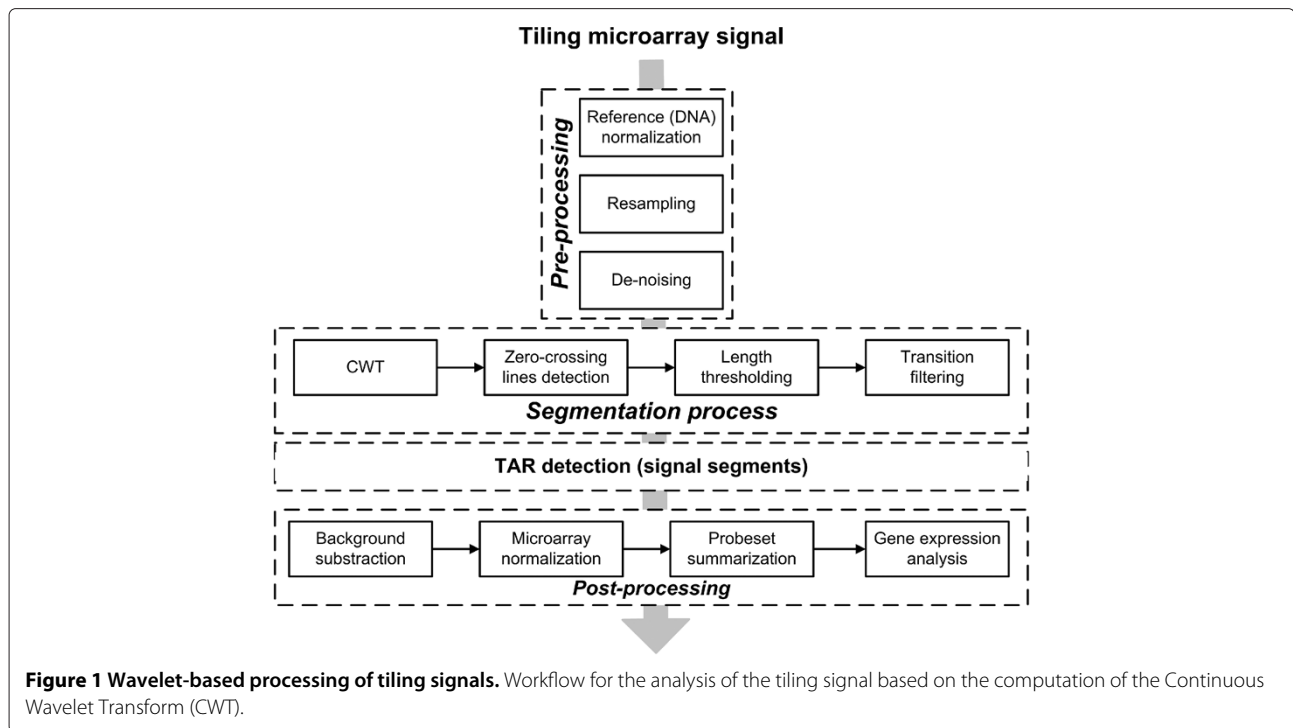
The workflow shown in Figure 1 summarizes the methodology. It consists of four basic blocks: (1) signal pre-processing (DNA normalization, non-uniform to uniform resampling and de-noising); (2) segmentation to detect abrupt intensity changes; (3) definition of transcriptionally active regions (TARs) and (4) biological knowledge extraction (for example, differential expression analysis of genes).

Transcriptome analysis refers to the detection of segments where the noisy tiling signal is constant. The start and end points of these segments correspond to transcript start and end sites. Several approaches have been deployed in the segmentation of tiling signals: pseudo-median or Hodges-Lehmann estimator [6,7], local non parametric smoothing [8,9], hidden Markov models [10-13], circular binary segmentation [14] and structural change model [15,16].

Wavelet analysis using the Discrete Wavelet Transform (DWT) [17] has demonstrated excellent performance in the analysis of ChIP-chip experiments using tiling array technology [18,19]. In this paper, we propose a wavelet transform based method for the identification of TARs in tiling signals (ZCL). We have chosen the SUREShrink algorithm for denoising and a method based on the computation of the Continuous Wavelet Transform (CWT) for detection of transcription start and end sites. In particular, the sharp transitions of the tiling signal are identified as the zero-crossing lines of a multiresolution decomposition using as the mother wavelet the second derivative of a Gaussian [20]. We applied the proposed analysis to the public *Saccharomyces Cerevisiae* dataset to validate our analytical approach. The novel identification algorithm was compared with two well-known methods: pseudo-median sliding window (PMSW) and structural change model (SCM). The absence of a biologically validated ground truth to evaluate the resulting segmentations prevent the use of specificity and sensitivity as performance metrics. Consequently, the evaluation has been made in terms of positive predictive value (PPV), sensitivity and computation time. We also evaluated the segmentation quality resulting from the combination of the TARs detected by several of the methods under study.

We also used this algorithm for the identification of the subset of transcripts whose expression decreases in a *S. aureus* strain deficient in the sigma B transcription factor. SigB has been shown to be involved in the stress response to different stimuli, the regulation of *sarA*, *sarHI*, and *agr* that control a wide array of virulence factors, biofilm formation, the ability of *S. aureus* to bind to various host-cell matrix proteins such as fibrinogen and fibronectin, and in the development of resistance to the antibiotics methicillin and teicoplanin [21-26].

We applied the segmentation methods to this high quality dataset and we have demonstrated its usefulness for the analysis of the tiling array derived transcriptome map. The results demonstrate that ZCL not only allows a rapid identification of the transcript based on the segmentation procedure but also a more accurate estimation of the expression level of each transcript.



## Results and discussion

All the steps needed to obtain a transcriptional map from the raw data (read the CEL files, normalize, denoise and segment the tiling signal) have been implemented using the statistical language *R/Bioconductor* [27]. The CRAN packages *Rwave*, *wavethresh* and *wmtsa* have been used for wavelet analysis. All the R functions described are available as Supplementary Material (Additional file 1, Additional file 2 and Additional file 3). The R code to perform the example analyses and the generation of the figures included in the paper can also be found as Supplementary Material. The results show that wavelets compare well with the rest of methods in terms of segmentation accuracy and time consumed in the analysis.

## Experimental datasets

### *Saccharomyces cerevisiae* dataset

The dataset is described in [16]. An oligonucleotide array for *S. cerevisiae* was developed. It contains 6.5 million probes and interrogates both strands of the full genomic sequence with 25-mer probes tiled at an average of 8 nucleotide intervals on each strand and 4 nucleotide tile offset between strands. The first-strand cDNA was synthesized using random primers from poly(A) and total RNA. A set of genomic DNA was also hybridized for normalization purposes [15]. Their analysis of the transcription map identified the transcript boundaries, its structure and the intensity level of coding and non-coding transcripts [16]. All data (CEL files, bmap files for both

strands and annotation file) was deposited in ArrayExpress database with accession number E-TABM-14.

### *Staphylococcus aureus* dataset

The *Staphylococcus aureus* custom tiling microarray (NA-Staph-b520729F) was designed in collaboration with Affymetrix (Santa Clara, CA, USA). Specifically, the microarray (format 49-7875 with 11  $\mu\text{m}$  features) contains a total of 522,406 probes, divided into two parts. The first part corresponds to the tiling array containing a total of 384,932 probes (25-mer), which are further divided into eight sets. The set used in our analysis covers both strands of the *S. aureus* NCTC 8325 genome (2,821,347 bp covered by 363,127 probes). Each 25-mer probe was tiled each 14-nt across the whole genome, resulting in 11-nt overlaps and a 7-nt tile offset between strands. The microarray design has been deposited in the ArrayExpress Archive at EMBL-EBI (<http://www.ebi.ac.uk/microarray-as/ae/>), ArrayExpress accession no. A-AFFY-165.

Before cDNA synthesis, RNA integrity from each sample was confirmed on Agilent RNA Nano LabChips (Agilent Technologies). 10  $\mu\text{g}$  of RNAs extracted from bacterial strains grown until exponential phase (OD<sub>600nm</sub> = 0.8) were reverse transcribed using SuperScript II reverse transcriptase (Invitrogen Life Technologies). They were processed following the protocol of the Affymetrix GeneChip Expression Analysis Technical Manual (P/N 702232 Rev. 2) in the presence of 6 ng/ml Actinomycin D to avoid spurious second-strand cDNA synthesis during the reverse transcription

reaction [28]. Sense RNA corresponding to *B. subtilis* poly-A *lys*, *phe*, *thr*, *trp*, *dap* genes were spiked into sample RNA as a control for the labeling and hybridization steps. cDNA was digested by DNase I (PIERCE) in 10X DNase I buffer (USB-Affymetrix) and the size of digestion products was analyzed in the Agilent Bioanalyser 2100 using RNA Nano LabChips to ensure that the fragmentation resulted in a majority of products in the range of 50 to 200 base-pairs. The fragmented cDNA were then biotinylated using terminal deoxynucleotidyl transferase (Promega) and the GeneChip DNA labeling reagent (Affymetrix) following the manufacturer's recommendations. Biotinylated cDNA (5  $\mu$ g per array) were hybridized for 16 hours according to the Affymetrix protocol in a total volume of 200  $\mu$ l per hybridization chamber. Following incubation, the arrays were washed and stained in the Fluidics station 450 (Affymetrix) using the protocol FS450.0005. The arrays were then scanned using the GeneChip scanner 3000 (Affymetrix). The intensity signals of each probe cell were computed by the GeneChip operating software (GCOS) and stored in cell intensity files (.CEL extension) before preprocessing and analysis. All microarray data described in this study have been deposited in the ArrayExpress Archive at EMBL-EBI (<http://www.ebi.ac.uk/microarray-as/ae/>), ArrayExpress accession no. E-MEXP-2778.

### Probe annotation and normalization

The annotation of the PM probe sequences was obtained with the alignment to the genome sequence of *S. cerevisiae* strain S288c (SGD of August 7, 2005) as provided in the package *davidTiling* of Bioconductor. Available data correspond to 3 replicates of poly(A), 2 replicates of total RNA and 3 replicates of genomic DNA. The CEL files were read and the normalized signals (poly(A) and total RNA) were obtained using Equation 4. The analysis steps (denoising, segmentation and detection of TARs) were performed on the poly(A) signal as it showed an improved hybridization quality [16]. Once the signal is constructed from CEL and annotation files we used *tilingArray* package functions to obtain equally-spaced samples. Other resampling methods can be applied without loss of generality.

The annotation files for *S. aureus* microarray are provided in the ArrayExpress database (A-AFFY-165). The microarrays of the experiment correspond to three replicates of genomic DNA, three replicates of RNA of the 15981 wild-type strain, and three replicates of the *sigB* deletion. All the preprocessing steps were performed as previously described for *S. cerevisiae* dataset.

### Denoising

The denoising was evaluated using the signal to noise ratio (SNR), a quantitative measure of its performance. In order

to compare the results obtained with those from Huber et al. [15] based on a variance stabilization and normalization transformation, the same definition of SNR was used. We looked at a set of control regions, two positive control regions (pos) within the ORFs of RPN2 and SER33 at coordinates 217860 – 220697 and 221078 – 222487 and two negative control regions (neg) in the background at coordinates 216800 – 217700 and 222800 – 227000 of *S. cerevisiae* (see Figure 2). We assumed (as in [15]) that the differences between positive and negative controls give an estimation of the signal level, whereas variations from the mean intensity within each region are due to noise.

The SNR was computed as

$$\text{SNR} = \frac{\Delta\mu}{\sigma} = \frac{1}{\sigma} \left( \sum_{r \in \text{pos}} \frac{\mu_r}{|\text{pos}|} - \sum_{r \in \text{neg}} \frac{\mu_r}{|\text{neg}|} \right), \quad (1)$$

with the noise standard deviation  $\sigma$  calculated as the average of the differences between 0.975 and 0.025 quantiles of the data within each of the control regions. Namely,

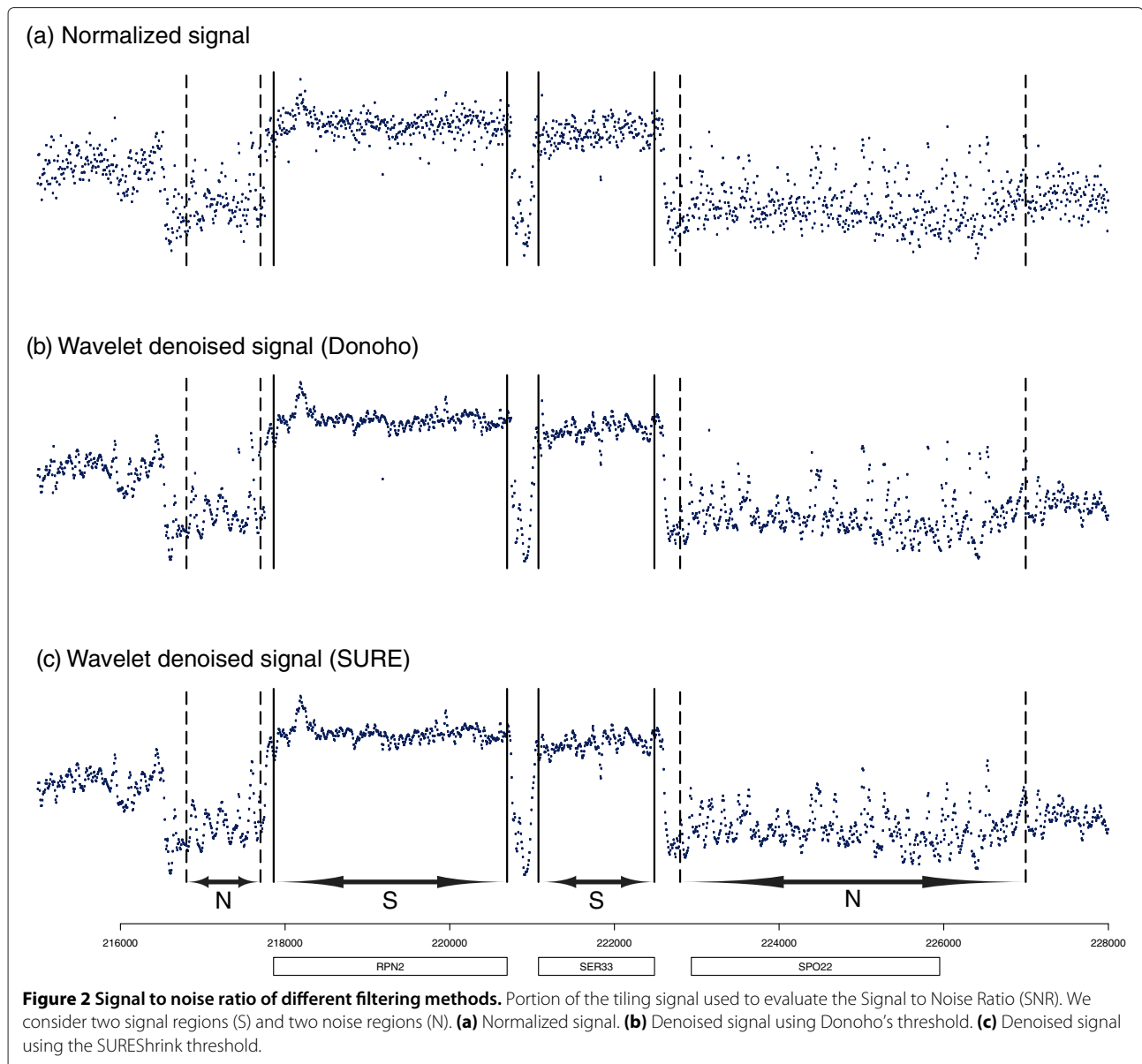
$$\sigma = \frac{\sum_{r \in \text{pos, neg}} (Q_r^{0.975} - Q_r^{0.025})}{(Q_N^{0.975} - Q_N^{0.025})(|\text{pos}| + |\text{neg}|)} \quad (2)$$

where the symbol  $r$  counts over the different regions and  $Q_N$  refers to the standard normal distribution  $\mathcal{N}(0, 1)$ . Table 1 shows the SNR of the normalized signal and the wavelet-based denoised signal using Donoho's method [29] and the SUREShrink approach [30] in relation to the best SNR obtained in [15]. We observe that the use of wavelets for denoising results in a large increase in the SNR (18.94% with Donoho's method and 30.63% with SUREShrink approach), especially when the SUREShrink denoising is applied. This could be due to the elimination of most part of the non-Gaussian noise component and the consequent reduction in the estimated variance. In the rest of the paper, the SUREShrink is the method applied for denoising.

### Segmentation

A descriptive example of the denoising and segmentation for *S. cerevisiae* is shown in Figure 3. The analysis corresponds to a 140 Kb segment of chromosome 1 from position 20000 to position 160000. The results are given for the three algorithms compared (SCM, PMSW, ZCL). The CWT computation of ZCL used as mother wavelet the second derivative of a Gaussian with 100 scales. Zero-crossing lines were calculated and only those with a length greater than a pre-defined threshold were considered to correspond to signal transitions.

The TAR start and end positions were defined as the transition locations for which the difference between the mean intensity of neighboring segments is greater than 10% of the dynamic range of the tiling signal. Moreover, the inspection of the intensity histogram of chromosome



1 forward strand was used to set the minimum normalized transcription level value to  $-2$ . The same parameters were adopted to process the other strands of the organism. The R function *segmentZCL* (provided as Supplementary Material) implements the whole segmentation procedure.

Another representative example of segmentation results is given in Figure 4. In this case, the *S. aureus* signal of 15981 and *sigB* mutant were segmented using ZCL and fixing only the number of scales to 100. No denoising was applied prior to the segmentation. The figure represent a fragment of the signals from position 2.1 Mb to 2.3 Mb. In Additional file 4: Figure S1 and Additional file 5: Figure S2, the results for the PMSW algorithm are shown. In Additional file 6: Figure S3 and Additional file 7: Figure S4, the results for the SCM algorithm are presented.

#### Segmentations comparison using *S. cerevisiae* dataset

The results from the ZCL segmentation were compared to those obtained with PMSW [6] and SCM [15]. The robust PMSW method is based on the calculation of a pseudo-median within a sliding window. The local expression level is computed with the Hodges-Lehmann estimator [31] on the RNA normalized signal. To be able to do this, the *Telescope* pipeline [7] was implemented. Once the candidate transcript regions were determined, the TARs were assembled by the combination of a normalization intensity threshold and a max-gap and min-run criteria. The former is defined as the maximum distance below which two adjacent transcribed probes are included in the same TAR. The later as the minimum length of a feature to be classified as a transcribed region.

**Table 1 Estimated SNR values of the tiling signal shown in Figure 3**

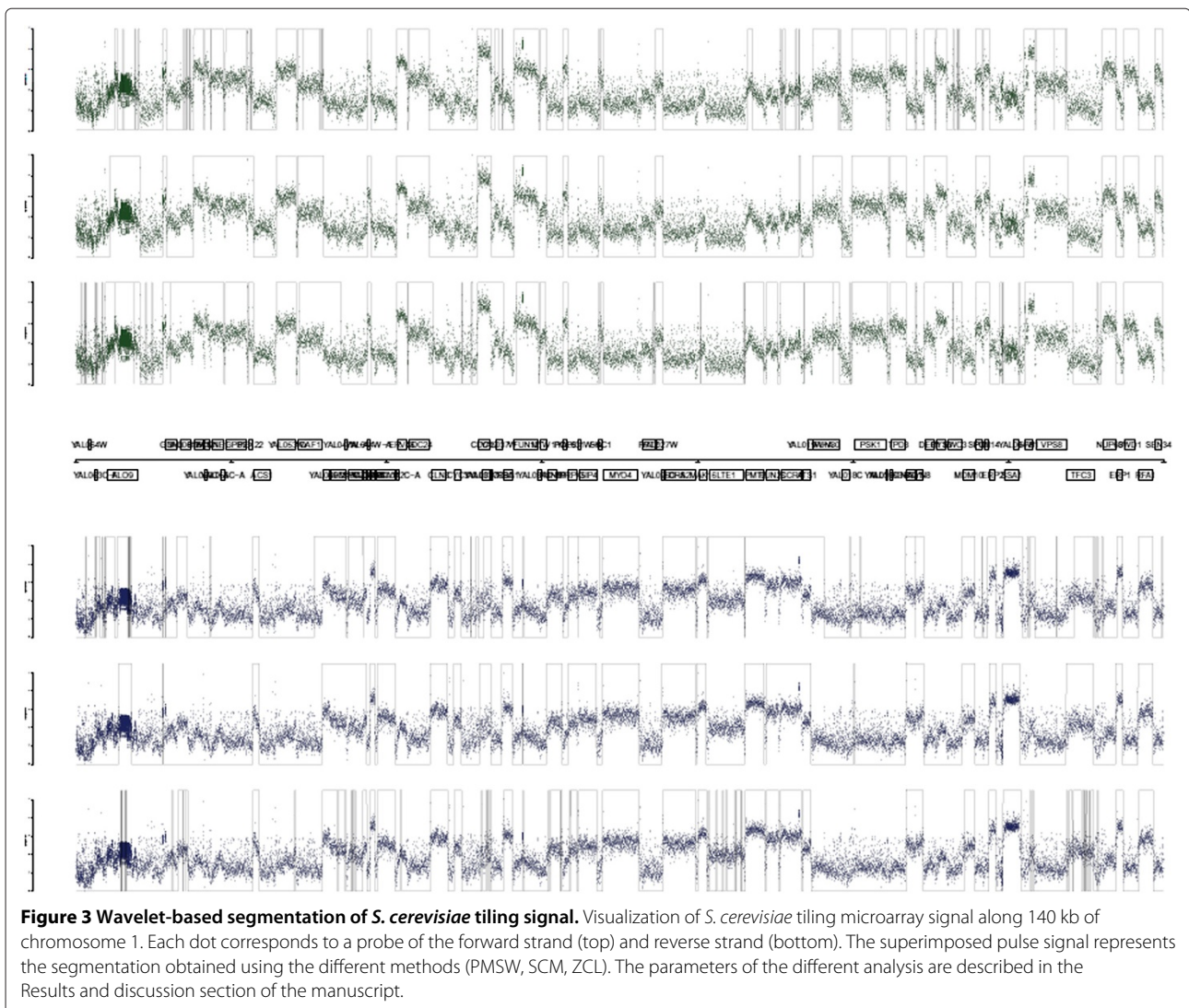
SNR results	
Signal	SNR
Best SNR in [15]	4.58
Normalized signal	4.28
Wavelet denoising (Donoho's)	5.28
Wavelet denoising (SUREShrink)	<b>6.17</b>

Estimated SNR values of the tiling signal. The normalized and the wavelet denoised signal using Donoho's and SUREShrink on which the calculation was performed are shown in Figure 3.

Huber's method is based on the structural change model (SCM). The SCM model [15,16] is used in econometrics for the modeling of sharp transitions in financial time series. It has been applied to the segmentation of comparative genomics hybridization (CGH) data [32]. The signal

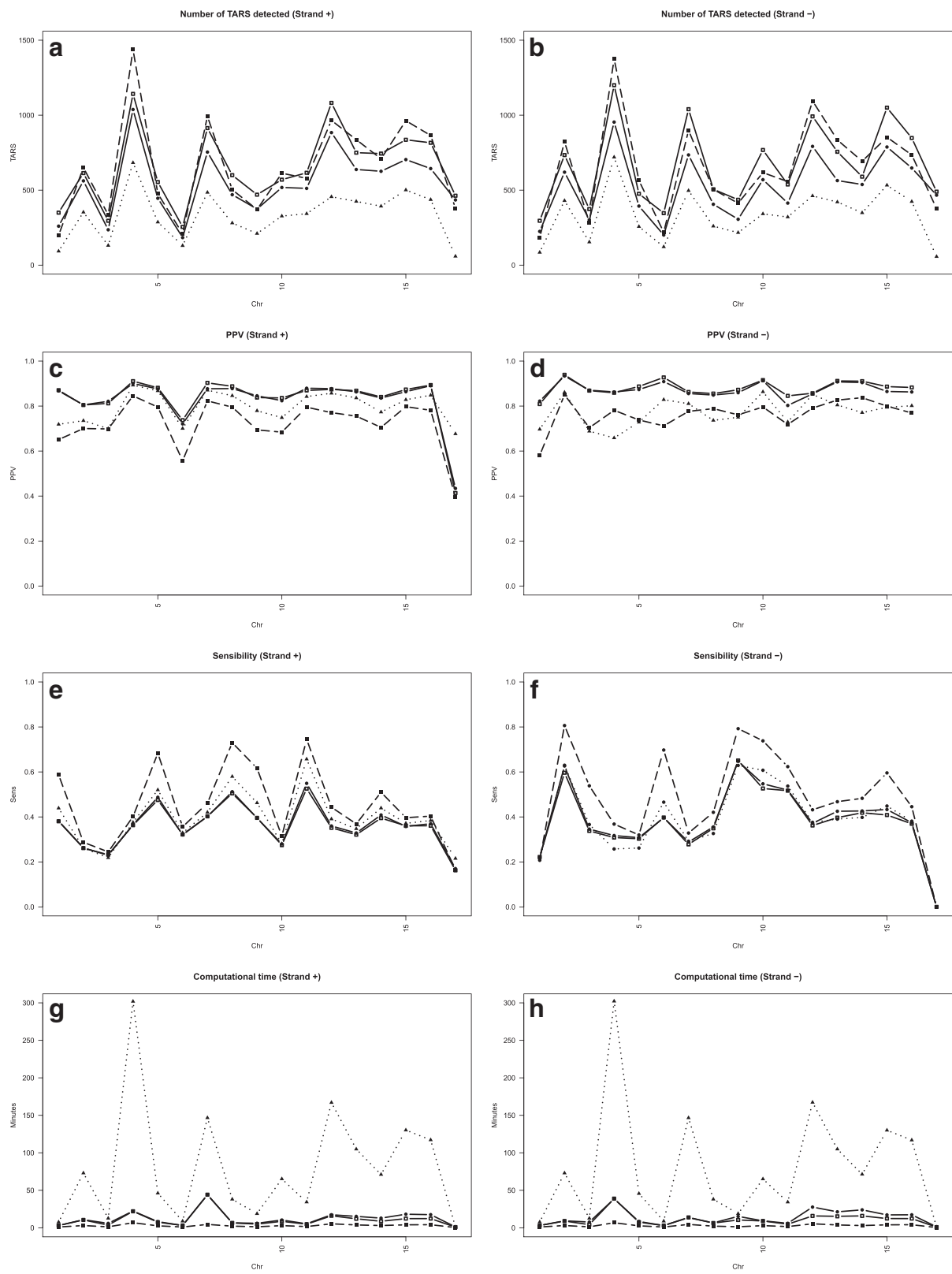
is modeled as a piecewise constant function of chromosomal coordinates described using the segment boundaries, the maximum number of segments and the mean signal value for each segment. The method is applied independently to each chromosome and, if the signal is strand-specific, to each of its two strands. A dynamic programming algorithm part of the *tilingArray* package of Bioconductor computes a globally optimal set of parameters for segmentations of increasing number of segments.

Due to the lack of a biologically validated ground truth to evaluate the outputs, we compared the methods in terms of two metrics, sensitivity and positive predictive value (PPV) at probe-level. We define sensitivity as the number of probes in the detected TARs that overlap with annotated regions (true positives, *TP*) divided by the total number of probes in the annotated regions (sum of true positives and false negatives, *TP + FN*): Sensitivity =  $TP / (TP + FN)$ . The PPV is defined as the number of



**Figure 3 Wavelet-based segmentation of *S. cerevisiae* tiling signal.** Visualization of *S. cerevisiae* tiling microarray signal along 140 kb of chromosome 1. Each dot corresponds to a probe of the forward strand (top) and reverse strand (bottom). The superimposed pulse signal represents the segmentation obtained using the different methods (PMSW, SCM, ZCL). The parameters of the different analysis are described in the Results and discussion section of the manuscript.





**Figure 5 Results for the identification of TARs.** Number of detected TARs, probe-level PPV and sensitivity, and computation time for the proposed (solid line), PMSW (dashed line) and SCM (dotted line) methods. The analysis was performed for the forward (left) and the reverse (right) strands of all chromosomes of *S. cerevisiae* tiling microarray data.



**Table 2 Evaluation metrics for *S. cerevisiae* dataset**

Evaluation metrics ( <i>S. cerevisiae</i> )				
Method	TARs	PPV	Sensitivity	Time (min)
PMSW	22114	0.7416	<b>0.4700</b>	2.88
SCM	11246	0.7847	0.3904	79.09
ZCL	18209	0.8486	0.3760	13.02
ZCLSure	22513	<b>0.8547</b>	0.3686	10.70

Mean number of detected TARs, probe-level PPV, probe-level sensitivity and computational time for PMSW, SCM and ZCL methods (all chromosome and strands of *S. cerevisiae*).

different methods would improve the performance of the segmentation.

#### Combination of TAR probes candidates

We evaluated the improvement in performance obtained by the combination of the different segmentations. We chose different strategies to define the sets (intersection of two or three methods and majority voting system). After a decision is taken on the candidates, TARs are constructed to create the transcriptional map. In Table 3, we give evaluation metrics (PPV and sensitivity) for both strands of *S. cerevisiae* chromosome 1. As individual methods, ZCL gives the best PPV and the best sensitivity for the reverse strand. The best performing combination considering a compromise between PPV and sensitivity is given by the majority voting system.

#### Computational performance

*S. cerevisiae* analyses were executed in an Intel(R) Xeon(R) processor server (64 bits, 4 cores, 2 GHz) with 32 Gb installed memory running Red Hat Enterprise Linux AS release 4 and R 2.13.0. Computing times needed to process each chromosome strand with the described methods are shown in Figure 5. For the same signal length, longer computation time is required for SCM, while comparable times are needed for PMSW and ZCL. The mean time to

segment the whole transcriptome is 2.88 mins for PMSW, 13.02 mins for ZCL and 79.09 mins for SCM.

#### Differential expression analysis of *S. aureus sigmaB* mutant

Comparative segmentation analysis using ZCL and PMSW and SCM algorithms was applied for the hybridization data obtained with a custom designed Affymetrix tiling array of *S. aureus*. Segmentation results for *S. aureus* are summarized in Table 4. In this case, the performance measures are almost identical for all methods. These results suggest that the performance of the methods depends on the quality of the signals, decreasing for PMSW and SCM algorithms as the SNR of the signal get worse. In spite of this, other advantages such as computation time, automatic selection of parameters and the possibility of parallel computation makes ZCL our preferred option to segment tiling signals.

The most frequent transcriptional analysis is the detection of genes that have changed their expression in the conditions under study (differential expression analysis). As sigma B affects the expression of more than one hundred genes, we decided to test whether it is possible to use the intensity of all the probes included in each detected TAR with the ZCL segmentation procedure to calculate the expression level of the transcript in a particular environmental condition. In order to carry out this analysis

**Table 3 Evaluation of segmentation combinations for both strand of chromosome 1**

Integrative transcriptional analysis				
Method	PPV Forward	Sensitivity Forward	PPV Reverse	Sensitivity Reverse
PMSW	0.6511	0.5873	0.5811	0.2073
SCM	0.7188	0.4390	0.6968	0.2146
ZCL	<b>0.8675</b>	0.3821	<b>0.8220</b>	<b>0.2208</b>
PMSW $\cap$ SCM $\cap$ ZCL	0.6312	<b>0.5984</b>	0.5441	0.2030
PMSW $\cap$ ZCL	0.6448	0.5921	0.5744	0.2076
PMSW $\cap$ SCM	0.6370	0.5957	0.5504	0.2043
SCM $\cap$ ZCL	0.7053	0.4409	0.6626	0.2116
Majority voting	0.7247	0.4396	0.6993	0.2164

PPV and sensitivity for both strand of chromosome 1 using individual TAR detection algorithms and the combination of their results.

**Table 4 Evaluation metrics for *S. aureus* dataset**

Tiling Signal	Metric	Evaluation metrics ( <i>S. aureus</i> )			
		PMSW	SCM	ZCL	ZCLSure
WT Forward	PPV	0.6298	<b>0.6498</b>	0.6248	0.6407
WT Forward	Sens	0.8657	<b>0.8766</b>	0.8715	0.8719
WT Reverse	PPV	0.6867	0.6993	<b>0.7050</b>	0.6989
WT Reverse	Sens	0.8506	<b>0.8560</b>	0.8388	0.8535
$\Delta$ sigB Forward	PPV	0.6238	<b>0.6388</b>	0.6227	0.6308
$\Delta$ sigB Forward	Sens	<b>0.9054</b>	0.9035	0.9027	0.9036
$\Delta$ sigB Reverse	PPV	0.6664	<b>0.6815</b>	0.6765	0.6748
$\Delta$ sigB Reverse	Sens	<b>0.8697</b>	0.8684	0.8667	0.8515

Mean number of detected TARs, probe-level PPV, probe-level sensitivity and computational time for PMSW, SCM and ZCL methods.

using tiling microarrays we need to compress the intensity of all the probes included in each detected TAR into one value. Standard methods for microarray normalization can be applied, for example RMA (Robust Multichip Average) algorithm in the case of Affymetrix microarrays [34]. This processing can be performed using the packages *affxparser*, *affy* and *limma* of *Bioconductor* for CDF (chip definition file) generation, normalization and differential expression analysis.

We introduced a simple analytical tool to be used independently of the microarray platform to measure the gene expression level based on the median value of the TAR probe intensities. We calculated this value for each wild-type and sigmaB mutant sample. We applied a statistical analysis (t-test) to obtain the p-value associated with the expression change taking into account the biological variability of the samples. Considering well-defined TARs in the *S. aureus* annotation, we found previously described alterations in several genes [21]. In Figure 6, we show the boxplots that represent these expression level changes. We confirmed the down-regulation of sigB and other  $\sigma^B$ -regulated genes, as the alkaline shock protein 23 (*asp23*) [22,23] and lysine-specific permease (*lysP*) [21], although the latter is not statistically significant ( $p > 0.05$ ). We also found genes up-regulated in the sigmaB mutant, as the staphylococcal nuclease (*nuc*) [23,24], the zinc metalloprotease aureolysin (*aur*) [24,25] and the  $\alpha$ -hemolysin (*hla*) [24,26], the latter without a statistically significant p-value.

## Conclusions

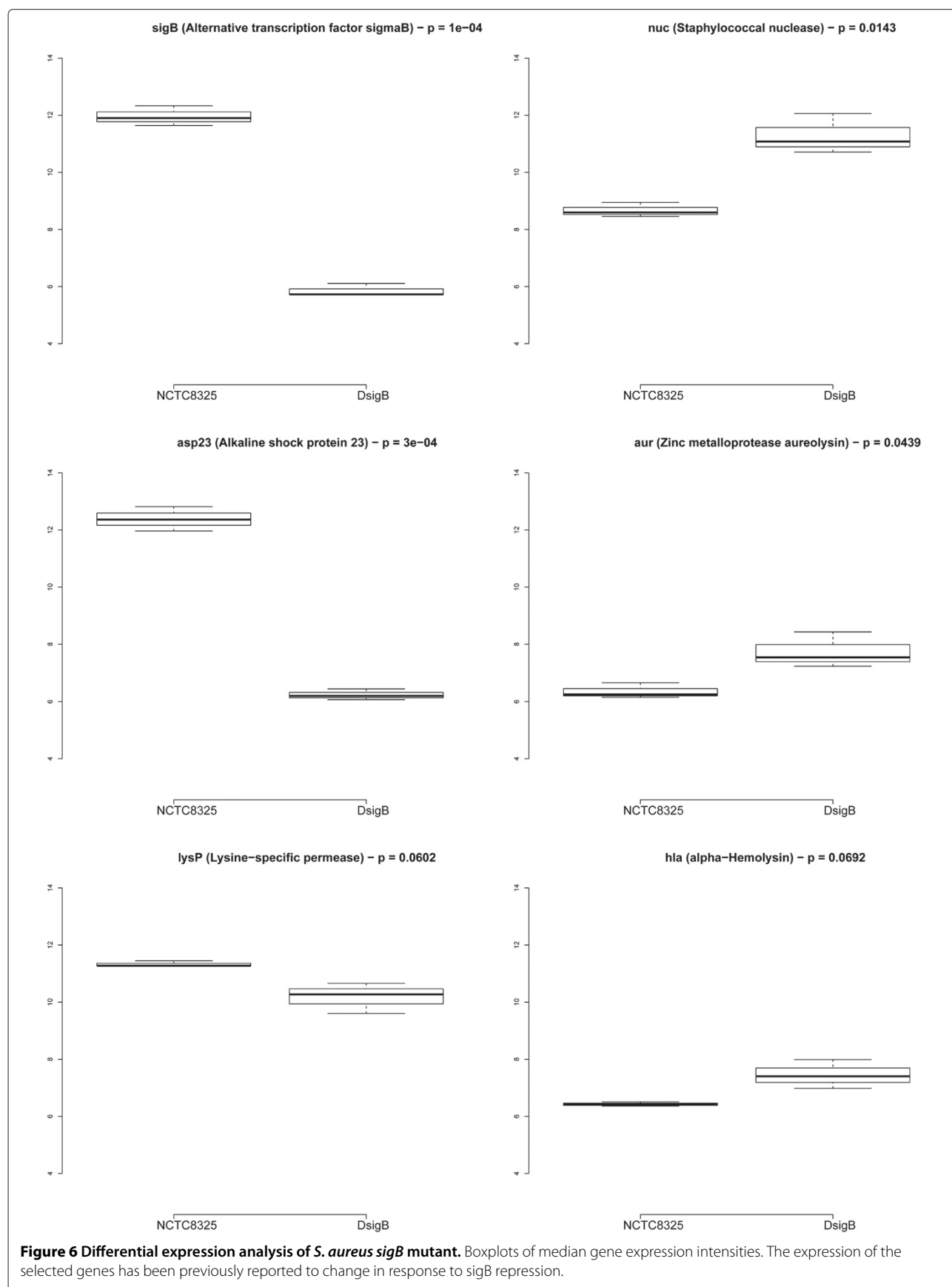
Transcriptomics is a powerful technology for the study of gene structures and RNA-based regulation in any organism. Genome-wide transcriptome analysis of prokaryotes can be carried out with any of these two techniques: RNA-seq and genomic tiling arrays [35]. High-resolution tiling arrays have been used, among others, to

study the transcriptomes of *Caulobacter crescentus* [36], *Escherichia coli* and [37], *Listeria monocytogenes* [38].

In this paper, we propose a combined WT-based method for the denoising and segmentation of tiling signals. For illustrative and evaluative purposes, we applied the proposed analysis to the public *S. cerevisiae*. Our denoising results show an increase in the SNR of the filtered signal with respect to Huber's method [15]. We believe it is advisable to properly denoise the tiling signal before segmentation as the number of false positives induced by signal variability is thus reduced. Even when constructing a manual segmentation, it seems an advisable choice to mark the transitions on the denoised signal, as its improved quality could help the expert to better discriminate between low expression transcripts and noise.

Our segmentation algorithm (ZCL) calculates all the possible break points based on the zero-crossing lines of the second derivative of the Gaussian wavelet. The results show that our method achieves the best compromise between accuracy (evaluated in terms of PPV and sensitivity) and computation time. The R code provided can be used to apply our algorithm as well as to combine the resulting segmentation with other methods as PMSW and SCM.

We also designed a new tiling microarray for the analysis of *S. aureus* genome, publicly available in the Array-Express database (accession number A-AFFY-165). This platform has been used for the comparison of the gene expression pattern of the *S. aureus* 15981 wild type and its isogenic sigB mutant. We selected this mutant because it is one of the most studied staphylococcal regulatory factors and consequently it was a useful gold standard to compare the accuracy of our algorithms. The relevance of the segmentation results comes from the fact that a correct analysis of the tiling signals could improve the matching between the probes and the corresponding transcriptional units. In particular, it could help to more precisely localize



the start and end transcription sites or even, include units that are not annotated in the current genome definition.

Once the TARs are properly detected, differentially expressed transcripts can be identified by well-known methods (such as Linear Models for Microarray Data (LIMMA) [39]) with a previous probe summarization algorithm to generate the transcript annotation (using, for example, Robust Microarray Analysis (RMA) [34]). In practice, this means that differential gene expression analysis could benefit from an enhanced analysis of tiling signals such as the one proposed here. To confirm the accuracy of the proposed method, we introduced a simple measure based on the median of TAR probe intensity. Using this approach, known up-regulated (*nuc*, *aur*, *hla*) and down-regulated (*asp23*, *lysP*) genes in sigB mutant were verified.

In conclusion, we present a novel method for denoising and segmentation of tiling microarray signals based on wavelet multiresolution analysis that outperforms previous methods in terms of SNR, positive predictive value and computation time. The R code that implements the method is given as supplementary material and can be easily adapted to a parallel computing schema. Also, we have introduced the possibility of combining the results of ZCL with those obtained with other two well-known approaches (PMSW and SCM) for the segmentation of tiling signals.

## Methods

### WT-based analysis

The CWT of a continuous signal  $s(x)$  is defined as [40]

$$CWT(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(x) \psi^* \left( \frac{b-x}{a} \right) dx \quad (3)$$

where  $a \in \mathbb{R}^+ - \{0\}$  is the scale,  $b \in \mathbb{R}$  is the translation,  $\psi(x)$  is the mother wavelet,  $\psi^*((b-x)/a)$  is the complex conjugated, scaled and translated wavelet and  $CWT$  is the 2D matrix of wavelet coefficients. The continuous input signal  $s(x)$  interpolates the discrete input samples  $s[k], k = 1, \dots, n$  where  $n$  is the length of the signal.

The CWT can be interpreted as the correlation of the input signal with a position reversed version of  $\psi$  rescaled by a factor  $a$ . For an 1D input signal, the result is a 2D description of the signal with respect to the position  $b$  and scale  $a$  and shifted by  $b$ . The scale  $a$  is inversely proportional to the central frequency of the dilated wavelet  $\psi_a = \psi(x/a)$ , which is typically a bandpass function;  $b$  represents the position location at which we analyze the signal. The larger the scale  $a$ , the wider the analyzing function  $\psi_a$ , and hence the smaller the corresponding analyzed frequency. The output value is maximized when the frequency of the signal matches that of the corresponding dilated wavelet. The CWT computation for arbitrary

scales can be easily adapted to a parallel implementation with a linear computational complexity [41].

Mallat's fast wavelet algorithm [42] uses the multiresolution properties of the wavelet to compute the CWT at dyadic scales  $a = 2^i$  and time shifts  $b = 2^i k, k \in \mathbb{Z}$ , resulting in what is known as DWT. For additional information about the wavelet transform and its properties the reader is referred to [17].

### Normalization of tiling microarray data

The analysis starts with background correction and quantile normalization as describe by the RMA algorithm [34]. Next, we calculate the geometric mean of the RNA intensities and the geometric mean of the DNA replicates to get a signal score  $s[k]$  at position  $k$  proportional to the transcription level in the reference genome [16]

$$s[k] = \frac{\sum_{j=1}^n \log \text{RNA}_j[k]}{\sum_{j=1}^m \log \text{DNA}_j[k]}, \quad (4)$$

where  $n$  is the number of RNA samples and  $m$  is the number of DNA samples.

### WT-based denoising

One of the most established methods of wavelet-based denoising was proposed by Donoho and Johnstone [29] and it is based on the thresholding of the DWT coefficients at scale  $a = 2$ . This method is composed of three steps: (i) calculate the DWT of the tiling signal  $s[k]$  at scale  $a = 2$ ; (ii) threshold the wavelet coefficients; (iii) compute the inverse wavelet transform of the thresholded coefficients. A universal threshold,  $T$ , is proposed [29] to remove white noise which it is given by

$$T = \sigma \sqrt{2 \log(n)} \quad \text{with} \quad \sigma = \text{MAD}/0.6745, \quad (5)$$

where  $n$  is the length of  $s$ ,  $\sigma$  is the noise level and MAD is the estimated median absolute deviation in the first scale. An important issue is the selection of a suitable wavelet function. As the signal can be roughly approximated to a zero-order polynomial, a boxcar-like function such as the Haar wavelet gives a reasonable level of correlation (i.e., a good pattern matching) with the target signal.

Another well established method of wavelet shrinkage is SUREShrink [30]. This is based on Stein's Unbiased Estimator for Risk (SURE). A subband adaptive threshold is applied. If the wavelet coefficients in the  $j$ th subband are  $\{x_i : i = 1, \dots, n\}$ , we consider a soft thresholding procedure and apply Stein's result. The quantity

$$\text{SURE}(T; x) = n - 2 \cdot \#\{i : |x_i| \leq T\} + \sum_{i=1}^n (x_i \wedge T)^2 \quad (6)$$

is an unbiased estimate of risk, where  $T$  is the threshold and  $x_i \wedge t = \min(x_i, T)$ . This estimator can be used to select a threshold:

$$T_{\text{SURE}} = \operatorname{argmin}_{0 \leq T \leq \sqrt{2 \log n}} \text{SURE}(T; x) \quad (7)$$

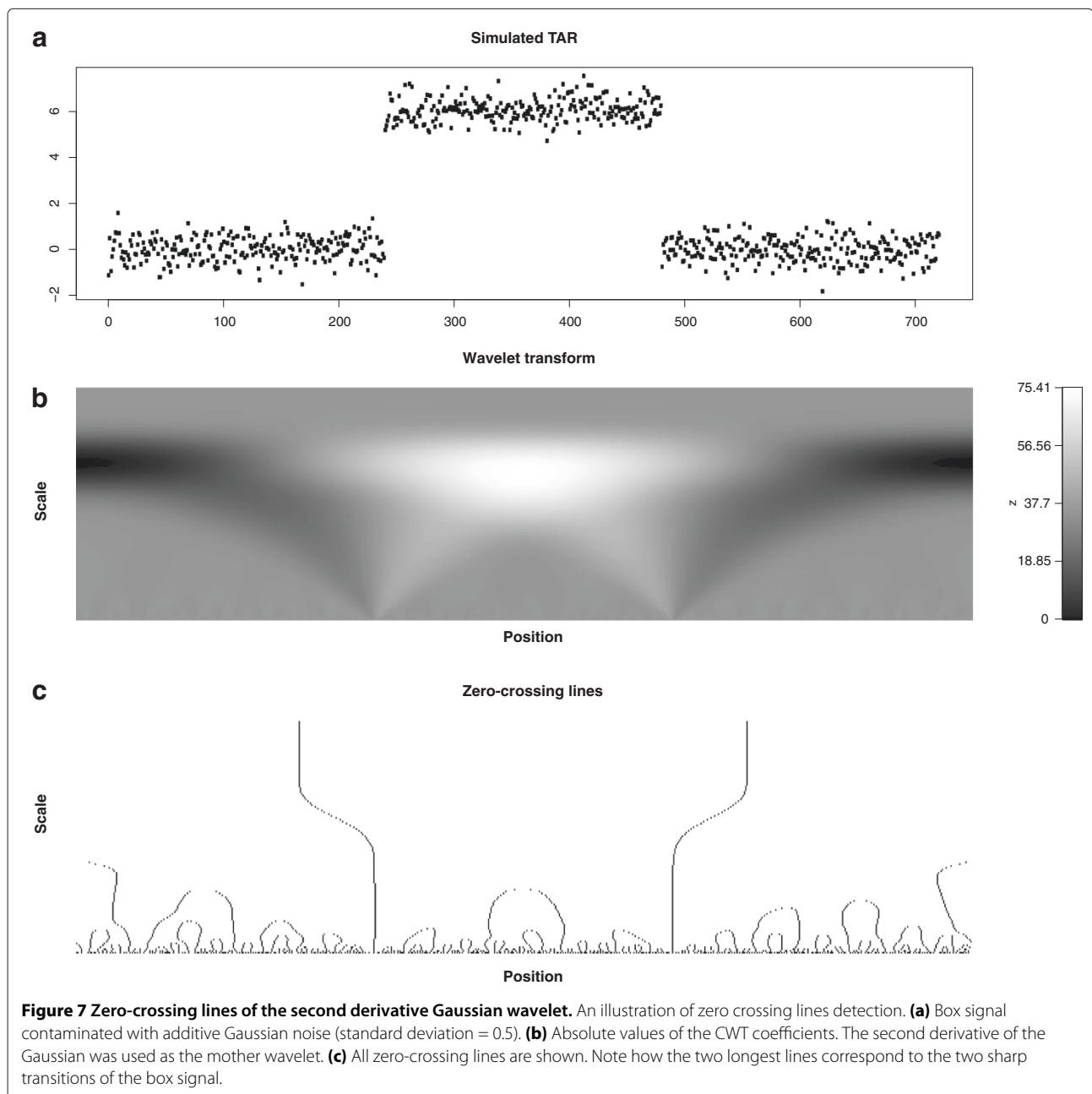
For a large dimension  $n$  the law of large numbers will ensure that  $T_{\text{SURE}}$  will be almost the optimal threshold [30].

### WT-based segmentation

An important issue in signal processing is to define an appropriate representation able to compress most of the

signal information into few representative features. Sharp variations in amplitude (i.e., transitions and peaks) are among the most meaningful features of a signal. For that reason, many segmentation algorithms rely on their detection. Previous studies have detected the peaks in mass spectrometry data using either the ridge lines [43] or the zero-crossing lines [44] in a multi-scale decomposition of the signal. Zero-crossing lines seems a more consistent description as they belong to connected curves, are more robust to noise and easier to detect than ridge lines [44].

It has been previously shown that the position of multiscale sharp transitions can be obtained from the



zero-crossings of the signal convolved with the Laplacian of a Gaussian [45]. We define a wavelet at scale  $a$  as

$$\psi(x) = \frac{d^2\theta_a(x)}{dx^2} \quad (8)$$

where  $\theta_a$  is a Gaussian function dilated by a factor  $a$ . Since the wavelet transform can be represented as

$$CWT(a, b) = (s * \psi_a)(x)|_{x=b} \quad (9)$$

we derive that

$$\begin{aligned} CWT(a, b) &= \left[ s * \left( a^2 \frac{d^2\theta_a}{dx^2} \right) \right] (x)|_{x=b} \\ &= a^2 \frac{d^2}{dx^2} (s * \theta_a)(x)|_{x=b} \end{aligned} \quad (10)$$

Hence, the wavelet transform of  $s(x)$  is proportional to the second derivative of  $s(x)$  smoothed by  $\theta_a(x)$ . The zero-crossings of  $CWT(a, b)$  correspond to the inflection points of  $s * \theta_a$ . The identification of transcript start and end sites is achieved by computation of the redundant CWT over a wide scale range followed by zero-crossing line detection and length thresholding. The chosen mother wavelet is the second derivative of a Gaussian. The redundancy of the CWT yields enhanced information on the position-scale localization of the features of interest (in this case, the transitions) [46].

An illustrative example is given in Figure 7. We generated a simulated transcriptional unit with a rectangular pulse signal of 721 samples with additive Gaussian noise (mean 0 and standard deviation 0.5) (see Figure 7(a)). The absolute values of the wavelet transform coefficients and the zero crossing lines are shown in Figure 7(b) and (c), respectively. Observe how the position of these lines corresponds to abrupt intensity transitions in the noisy signal and the longest connected curves identify the start and end points of the rectangular pulse. The R functions provided as supplementary material detect the zero-crossing lines and identify them as transcription start sites (TSS) and transcription end sites (TES) depending on the slope sign.

#### Identification of transcriptional active regions

The candidates start and end sites detected as described in the previous section, are filtered to remove incorrect assignments. The purpose of this procedure is to filter those transitions that do not correspond to variations in signal intensity. For the generation of TARs we considered the signal transitions in which variation in intensity is at least 10% of the dynamic range of the analyzed signal. We also eliminate from the list of detected TAR all the start and end points that are not correctly paired off. We use the sign of the zero-crossing lines to separate start and end points and we match each start site with its corresponding end site. Finally, we define the minimum

normalized intensity threshold required for the segments to be considered as transcriptional active regions. This value is calculated as the median of the signal intensity distribution, but this threshold can also be user-defined. In order to improve the definition of TARs, we cluster together consecutive segments for which the mean normalized intensity value is over the threshold.

#### Additional files

##### Additional file 1: R code: Segmentation and visualization functions.

Implemented functions in R language to perform PMSW and SCM segmentation and the proposed wavelet-based method for denoising and segmentation. In addition, functions are provided for proper visualization of data, integration of analysis results and evaluation of the obtained transcriptional maps.

##### Additional file 2: R code: Segmentation analysis of *S. cerevisiae*.

R script for segmentation of the *S. cerevisiae* dataset and the generation of the figures included in the manuscript.

##### Additional file 3: R code: Segmentation analysis of *S. aureus*.

R script for the segmentation of the *S. aureus* dataset and the generation of the figures included in the manuscript.

##### Additional file 4: Figure S1.

Visualization of *S. aureus* tiling microarray signal along 200 kb of NCTC8325 genome. Each dot corresponds to a probe of the forward strand (top) and reverse strand (bottom) for NCTC8325 wild-type normalized signal. Superimposed pulse signal represents the segmentation obtained using PMSW method. The parameters of the analysis are described in the Results and discussion section of the manuscript.

##### Additional file 5: Figure S2.

Visualization of *S. aureus* tiling microarray signal along 200 kb of NCTC8325 genome. Each dot corresponds to a probe of the forward strand (top) and reverse strand (bottom) for sigmaB mutant normalized signal. Superimposed pulse signal represents the segmentation obtained using PMSW method. The parameters of the analysis are described in the Results and discussion section of the manuscript.

##### Additional file 6: Figure S3.

Visualization of *S. aureus* tiling microarray signal along 200 kb of NCTC8325 genome. Each dot corresponds to a probe of the forward strand (top) and reverse strand (bottom) for NCTC8325 wild-type normalized signal. Superimposed pulse signal represents the segmentation obtained using SCM method. The parameters of the analysis are described in the Results and discussion section of the manuscript.

##### Additional file 7: Figure S4.

Visualization of *S. aureus* tiling microarray signal along 200 kb of NCTC8325 genome. Each dot corresponds to a probe of the forward strand (top) and reverse strand (bottom) for sigmaB mutant normalized signal. Superimposed pulse signal represents the segmentation obtained using SCM method. The parameters of the analysis are described in the Results and discussion section of the manuscript.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

VS and AMB conceived the idea, developed the methods and implemented the software. ATA and IL design the NA-Staph-b520729F microarray and MU and ATA carried out the processing and hybridization of samples. ATA, IL and VS made the biological interpretation of the results. All authors participated in writing and revising the manuscript.

#### Acknowledgements

We thank Prof. Fernando J. Corrales and Lourdes Ortiz (Genomics Core Facility) for technical support and all the useful comments about the manuscript. This work was supported by the Spanish Torres-Quevedo fellowship [PTQ-08-03-07769] to VS. ATA and AMB were supported by Spanish Ministry of Science and Innovation 'Ramón y Cajal' contracts. This work was supported by

the Spanish Ministry of Science and Innovation Grants BIO2008-05284-C02-01, BFU2011-23222, ERA-NET Pathogenomics PIM2010EPA-00606 and the agreement between 'Fundación para la Investigación médica aplicada' (FIMA) and the 'UTE project CIMA'.

#### Author details

<sup>1</sup>Genomics, Proteomics and Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona, Spain. <sup>2</sup>Laboratory of Microbial Biofilms, Instituto de Agrobiotecnología, Universidad Pública de Navarra-Consejo Superior de Investigaciones Científicas-Gobierno de Navarra, Pamplona 31006, Spain. <sup>3</sup>Cancer Imaging Laboratory, Center for Applied Medical Research, University of Navarra, Pamplona, Spain.

Received: 27 June 2012 Accepted: 16 August 2012

Published: 5 September 2012

#### References

1. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR: **Applications of DNA tiling arrays for whole-genome analysis.** *Genomics* 2005, **85**:1–15.
2. Oszolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities.** *Nat Rev Genet* 2011, **12**(2):87–98.
3. Xu W, Seok J, Mindrinos MN, Schweitzer AC, Jiang H, Wilhelm J, Clark TA, Kapur K, Xing Y, Faham M, Storey JD, Moldawer LL, Maier RV, Tompkins RG, Wong WH, Davis RW, Xiao W, the Inflammation and Host Response to Injury Large-Scale Collaborative Research Program: **Human transcriptome array for high-throughput clinical studies.** *Proc Natl Acad Sci U S A* 2011, **108**(9):3707–3712.
4. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bel II, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammanna H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**(5830):1484–1488.
5. Emanuelsson O, Nagalakshmi U, Zheng D, Rozowsky JS, Urban AE, Du J, Lian Z, Stolc V, Weissman S, Snyder M, Gerstein MB: **Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome.** *Genome Res* 2007, **17**(6):886–897.
6. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammanna H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14**(3):331–342.
7. Zhang ZD, Rozowsky J, Lam HY, Du J, Snyder M, Gerstein M: **Tilescope: online analysis pipeline for high-density tiling microarray data.** *Genome Biol* 2007, **8**(5).
8. Royce TE, Carriero NJ, Gerstein MB: **An efficient pseudomedian filter for tiling microarrays.** *BMC Bioinformatics* 2007, **8**:186–186.
9. Wang LY, Abyzov A, Korbel JO, Snyder M, Gerstein M: **MSB: a mean-shift-based approach for the analysis of structural variation in the genome.** *Genome Res* 2009, **19**:106–117.
10. Ji H, Wong WH: **TileMap: create chromosomal map of tiling array hybridizations.** *Bioinformatics* 2005, **21**(18):3629–3636.
11. Du J, Rozowsky JS, Korbel JO, Zhang ZD, Royce TE, Schultz MH, Snyder M, Gerstein M: **A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chip-chip experiments: systematically incorporating validated biological knowledge.** *Bioinformatics* 2006, **22**(24):3016–3024.
12. Toyoda T, Shinozaki K: **Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models.** *Plant J* 2005, **43**(4):611–621.
13. Nicolas P, Leduc A, Robin S, Rasmussen S, Jarmer H, Bessières P: **Transcriptional landscape estimation from tiling array data using a model of signal shift and drift.** *Bioinformatics* 2009, **25**(18):2341–2347.
14. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5**(4):557–572.
15. Huber W, Toedling J, Steinmetz LM: **Transcript mapping with high-density oligonucleotide tiling arrays.** *Bioinformatics* 2006, **22**(16):1963–1970.
16. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome.** *Proc Natl Acad Sci U S A* 2006, **103**(14):5320–5325.
17. Mallat S: *A Wavelet Tour of Signal Processing.* San Diego: Academic Press; 1998.
18. Karpikov A, Rozowsky J, Gerstein M: **Tiling array data analysis: a multiscale approach using wavelets.** *BMC Bioinformatics* 2011, **12**:57–57.
19. Buck MJ, Lieb JD: **ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.** *Genomics* 2004, **83**(3):349–360.
20. Mallat S: **Zero-crossings of a wavelet transform.** *IEEE Trans Info Th* 1991, **37**(4):1019–1033.
21. Bischoff M, Dunman P, Kormanec J, Macapagal D, Murphy E, Mounts W, Berger-Bächli B, Projan S: **Microarray-based analysis of the Staphylococcus aureus sigmaB regulon.** *J Bacteriol* 2004, **186**(13):4085–4099.
22. Gertz S, Engelmann S, Schmid R, Ohlsen K, Hacker J, Hecker M: **Regulation of sigmaB-dependent transcription of sigB and asp23 in two different Staphylococcus aureus strains.** *Mol Gen Genet* 1999, **261**(3):558–566.
23. Kullik I, Giachino P, Fuchs T: **Deletion of the alternative sigma factor sigmaB in Staphylococcus aureus reveals its function as a global regulator of virulence genes.** *J Bacteriol* 1998, **180**(18):4814–4820.
24. Ziebandt AK, Weber H, Rudolph J, Schmid R, Höper D, Engelmann S, Hecker M: **Extracellular proteins of Staphylococcus aureus and the role of SarA and sigma B.** *Proteomics* 2001, **1**(4):480–493.
25. Karlsson A, Arvidson S: **Variation in extracellular protease production among clinical isolates of Staphylococcus aureus due to different levels of expression of the protease repressor sarA.** *Infect Immun* 2002, **70**(8):4239–4246.
26. Giachino P, Engelmann S, Bischoff M: **Sigma(B) activity depends on RsbU in Staphylococcus aureus.** *J Bacteriol* 2001, **183**(6):1843–1852.
27. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health).* Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2005.
28. Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D.** *Nucleic Acids Res* 2007, **35**(19):e128.
29. Donoho D, Johnstone I, Johnstone IM: **Ideal Spatial Adaptation by Wavelet Shrinkage.** *Biometrika* 1993, **81**:425–455.
30. Donoho D, Johnstone IM: **Adapting to Unknown Smoothness via Wavelet Shrinkage.** *J Am Stat Assoc* 1995, **90**(432):1200–1224.
31. Hollander M, Wolfe DA: *Nonparametric Statistical Methods, 2nd Edition.* Wiley-Interscience; 1999.
32. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for array CGH data analysis.** *BMC Bioinformatics* 2005, **6**:27–27.
33. Kim Y, Bekiranov S, Lee JK, Park T: **Double error shrinkage method for identifying protein binding sites observed by tiling arrays with limited replication.** *Bioinformatics* 2009, **25**(19):2486–2491.
34. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**(4):e15.
35. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.** *Nat Rev Genet* 2010, **11**:9–16.
36. Landt SG, Abeliuk E, McGrath PT, Lesley JA, McAdams HH, Shapiro: **Small non-coding RNAs in Caulobacter crescentus.** *Mol Microbiol* 2008, **68**(3):600–614.
37. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ: **The transcription unit architecture of the Escherichia coli genome.** *Nat Biotechnol* 2009, **27**(11):1043–1049.
38. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, Barthelemy M, Vergassola M, Nahori MA, Soubigou G, Régnauld B, Coppée JY, Lecuit M, Johansson J, Cossart P: **The Listeria transcriptional landscape from saprophytism to virulence.** *Nature* 2009, **459**(7249):950–956.

39. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
40. Daubechies I: *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics)*. 1 edition. SIAM: Society for Industrial and Applied Mathematics; 1992.
41. Munoz Barrutia, A, Ertle R, Unser M: **Continuous Wavelet Transform with Arbitrary Scales and  $\mathcal{O}(N)$  Complexity.** *Signal Processing* 2002, **82**(5):749–757.
42. Mallat SG: **A theory for multiresolution signal decomposition: the wavelet representation.** *Pattern Anal and Machine Intelligence, IEEE Trans on* 1989, **11**(7):674–693.
43. Du P, Kibbe WA, Lin SM: **Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.** *Bioinformatics* 2006, **22**(17):2059–2065.
44. Nguyen N, Huang H, Oraintara S, Vo A: **Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet.** *Bioinformatics* 2010, **26**(18):659–665.
45. Marr D, Hildreth E: **Theory of Edge Detection.** *Proc R Soc London Ser B* 1980, **207**:187–217.
46. Rioul O, Vetterli M: **Wavelets and signal processing.** *IEEE Signal Process Mag* 1991, **8**(4):14–38.

doi:10.1186/1471-2105-13-222

**Cite this article as:** Segura et al.: Wavelet-based detection of transcriptional activity on a novel *Staphylococcus aureus* tiling microarray. *BMC Bioinformatics* 2012 **13**:222.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

