Supplementary materials for this article are available at 10.1007/s13253-013-0159-5.

# Comparing Segmentation Methods for Genome Annotation Based on RNA-Seq Data

Alice CLEYNEN, Sandrine DUDOIT, and Stéphane ROBIN

Transcriptome sequencing (RNA-Seq) yields massive data sets, containing a wealth of information on the expression of a genome. While numerous methods have been developed for the analysis of differential gene expression, little has been attempted for the localization of transcribed regions, that is, segments of DNA that are transcribed and processed to result in a mature messenger RNA. Our understanding of genomes, mostly annotated from biological experiments or computational gene prediction methods, could benefit greatly from re-annotation using the high precision of RNA-Seq.

We consider five classes of genome segmentation methods to delineate transcribed regions, including intron/exon boundaries, based on RNA-Seq data. The methods provide different functionality and include both exact and heuristic approaches, using diverse models, such as hidden Markov or Bayesian models, and diverse algorithms, such as dynamic programming or the forward-backward algorithm. We evaluate the methods in a simulation study where RNA-Seq read counts are generated from parametric models as well as by resampling of actual yeast RNA-Seq data. The methods are compared in terms of criteria that include global and local fit to a reference segmentation, Receiver Operator Characteristic (ROC) curves, and coverage of credibility intervals based on posterior change-point distributions. All compared algorithms are implemented in packages available on the Comprehensive R Archive Network (CRAN, *http://cran.r-project.org*). The data set used in the simulation study is publicly available from the Sequence Read Archive (SRA, *http://www.ncbi.nlm.nih.gov/sra*).

While the different methods each have pros and cons, our results suggest that the EBS Bayesian approach of Rigaill, Lebarbier, and Robin (2012) performs well in a re-annotation context, as illustrated in the simulation study and in the application to actual yeast RNA-Seq data.

This article has supplementary material online.

**Key Words:** Change-point detection; Confidence intervals; Count data; Genome annotation; Negative binomial distribution; RNA-Seq; Segmentation.

---

Alice Cleynen (✉) is PhD Student (E-mail: *alice.cleynen@agroparistech.fr*) and Stéphane Robin is Senior Researcher (E-mail: *robin@agroparistech.fr*), AgroParisTech, UMR 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France. Alice Cleynen is PhD Student and Stéphane Robin is Senior Researcher, INRA, UMR 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France. Alice Cleynen is PhD Student and Sandrine Dudoit is Professor (E-mail: *sandrine@stat.berkeley.edu*), Division of Biostatistics and Department of Statistics, University of California, Berkeley, 185 Li Ka Shing Center, #3370, Berkeley, CA 94720-3370, USA.


*Journal of Agricultural, Biological, and Environmental Statistics, Volume 19, Number 1, Pages 101–118*
*DOI: 10.1007/s13253-013-0159-5*

## 1. INTRODUCTION

Many genomes have been annotated, using approaches ranging from *in vitro* biological experiments to *in silico* gene prediction. Today, with the low cost and high precision of high-throughput sequencing, the question of re-annotation arises. In this context, an interesting problem is the following: given base-level read counts from transcriptome sequencing (RNA-Seq) and approximate knowledge of a gene's location from prior heuristic annotation, is it possible to precisely localize a transcribed region, that is, the set of nucleotides leading to a mature messenger RNA (mRNA). This involves identifying the set of nucleotides defining the $5'$ and $3'$ untranslated regions (UTRs), i.e., the start and end of transcription, as well as the boundaries between exons and introns. In this paper, we use as lax and inclusive definition for a gene, the set of all genomic regions that are transcribed to eventually form a mature transcript (including all exons and introns and the $5'$ and $3'$ UTRs) and that can be represented as a discrete interval. Additional motivation for genome re-annotation based on RNA-Seq data is the ability to localize UTRs: while available annotation typically only provides the location of *translated* regions (corresponding to a protein), we consider the annotation of *transcribed* regions, which are usually larger than and include translated regions.

A segmentation of a discrete interval $\{1, \ldots, n\}$ of size $n$ (e.g., set of $n$ consecutive nucleotides) is a partition of this interval into disjoint intervals, or segments, whose union is the original interval. The segmentation is usually summarized by a set of change-points, i.e., boundaries between segments. In a statistical inference context, a segmentation is based on random variables indexed by the elements of the interval to be segmented (e.g., RNA-Seq base-level read counts). The random variables have segment-specific distributions and the change-points correspond to changes in distribution (e.g., in mean parameter). Segmentation methods are particularly adapted to transcript localization using RNA-Seq data: the exons expressed in a given transcript are separated by intronic and non-expressed exonic regions expected to have low read counts (reflecting low transcriptional activity), thus allowing the variation in read counts to be exploited to define the transcript. Because of the discrete nature of RNA-Seq data (number of sequenced reads beginning at each position of the genome), segmentation is based on discrete distributions, such as the Poisson or negative binomial distributions. Figure 1 of the Supplementary Materials displays RNA-Seq base-level read counts for a few representative genes in *Saccharomyces cerevisiae*.

This paper is dedicated to the comparison of segmentation methods for the annotation of genomes based on RNA-Seq data. Each segmentation method involves a combination of three choices: (i) a model for the segment-specific read count distributions (e.g., Poisson, negative binomial); (ii) criteria for inferring parameters of the segment-specific distributions (e.g., log-likelihood) and for selecting the number of segments (e.g., penalized log-likelihood); (iii) optimization methods for the criteria in (ii). For ease of implementation, we have limited our comparison to segmentation methods available in R or Bioconductor packages.

We distinguish between two main classes of segmentation methods: those that return a segmentation into a fixed number of segments and those that return a probability for

the existence of a change-point at each location. Note that, in some cases, the number of segments might be known (e.g., in the context of re-annotation) and in others it might be part of the statistical inference problem (e.g., in the context of *de novo* annotation or transcript discovery).

The first class includes algorithms that are usually fast enough to deal with long sequences ($10^5$ to $10^9$ base-pairs) and that can be applied to recover an entire set of expressed transcripts or to localize novel transcripts. The *Dynamic Programming Algorithm* (DPA) is an exact algorithm that returns the optimal segmentation into $K$ segments, according to the log-likelihood criterion, for each $K$ ranging from 1 up to a user-supplied $K_{\max}$ (Guthery 1974). Its fast (but still exact) version, the *Pruned Dynamic Programming Algorithm* (PDPA), is implemented in the R package `Segmentor3IsBack` for the negative binomial and Poisson distributions (Cleynen, Koskas, and Rigaill under review). Segmentation by binary classification with *CART* (Scott and Knott 1974; Breiman et al. 1984) is an efficient and extremely fast heuristic algorithm that returns a non-optimal segmentation into $K$ segments, for each $K$ ranging from 1 up to a user-supplied $K_{\max}$, by drastically reducing the number of segmentations explored but still yielding good results when the signal is not too noisy. When the number of segments is unknown, these algorithms have to be combined with a model selection strategy. Finally, *Pruned Exact Linear Time* (PELT) is an exact algorithm that returns the optimal segmentation according to a penalized log-likelihood criterion and where the number of segments is estimated within the algorithm. These last two algorithms are implemented for the Poisson distribution in the R package `changepoint` (Killick and Eckley 2011).

The second class of segmentation approaches includes algorithms with a longer runtime, but that provide credibility intervals (a.k.a. Bayesian confidence intervals) for the location of change-points. They usually deal with shorter sequences ($10^3$ to $10^4$ base-pairs), but can be applied for precise re-annotation of the genome with high confidence. The constrained hidden Markov model (HMM) approach implemented in the package `postCP` (Luong, Rozenholc, and Nuel 2013) uses the PDPA for its parameter initialization. The exact Bayesian approach proposed by Rigaill, Lebarbier, and Robin (2012) is implemented in the R package `EBS` (which is available on the CRAN). Both methods are applicable to the Poisson and negative binomial distributions.

Note that all segmentation methods mentioned thus far are also available for the Gaussian distribution, which is widely-used, for instance for the identification of copy-number variation based on Comparative Genomic Hybridization (CGH) microarray data.

Numerous other segmentation approaches exist, such as, to only mention a few, least squares regression (Bai and Perron 2003), Bayesian inference based on product partition models and Markov sampling (Barry and Hartigan 1993), adaptive weights smoothing (Hupé et al. 2004), or wavelets (Hsu et al. 2005). Since they are not adapted to count data, we do not consider them in our comparison study. Though `FREEC` (Boeva et al. 2011) was developed for discrete sequencing data, it applies a Gaussian segmentation method to transformed read counts and is thus not considered here.

The paper is organized as follows. In the next section, we describe our segmentation framework, the methods to be evaluated, and our simulation study design. Then, we present

results of the comparison of segmentation methods for different types of biological questions and examine the effect of a classical log-transformation of the data. Finally, we discuss the results and consider extensions to other problems such as copy-number variation.

## 2. METHODS

### 2.1. SEGMENTATION FRAMEWORK

In the context of re-annotation, the segmentation framework can be formulated as follows. Suppose we have RNA-Seq base-level read counts for a region of the genome represented by nucleotide positions $t \in \{1, \ldots, n\}$ and which contains, for simplicity, only one transcript (i.e., we do not consider alternative splicing). For a transcript with $K_e$ exons, the segmentation for the sequence has $K = 2 \times K_e + 1$ segments, where each even (odd) segment corresponds to an exon (intron). Let $\tau_k$, $k = 0, \ldots, K$, denote the $k$th change-point, with the convention that $\tau_0 = 1$ and $\tau_K = n + 1$. Then, the $k$th segment is defined as the interval $[\![\tau_{k-1}, \tau_k[\![$ and the corresponding segmentation can be summarized by $\tau = \{\tau_k : k = 0, \ldots, K\}$. Finally, the set of all possible segmentations into $K$ segments is denoted by $\mathcal{M}_K$.

Let $Y_t$ and $y_t$ denote, respectively, the random variable and its realization for the number of aligned reads with first base at position $t$ and let $Y = \{Y_t : t = 1, \ldots, n\}$ and $y = \{y_t : t = 1, \ldots, n\}$ denote the signal over the entire region to be segmented. Note that strand-specific reads are mapped and counted separately for each strand and that distinct segmentations are performed on each strand. When comparing segmentation results for actual RNA-Seq data to existing annotation, read length is taken into account by extending the change-point locations $\tau_k$ accordingly (this is unnecessary for simulated data sets). We assume that the $Y_t$ are independent random variables with distributions affected by $K - 1$ abrupt changes in their parameters at each of the change-points $\tau_k$. Specifically, the model can be written as

$$Y_t \sim \mathcal{G}(\theta_k, \phi), \quad \forall t \in [\![\tau_{k-1}; \tau_k[\![, \ k = 1, \ldots, K,$$

where $\mathcal{G}$ is a parametric distribution (e.g., Poisson or negative binomial), $\theta_k$ are segment-specific parameters (such as, but not limited to the mean $\mu_k$), and $\phi$ is a global parameter (e.g., dispersion).

Three statistical inference questions are therefore pertinent in the context of segmentation: (i) the estimation of the number of segments $K$; (ii) the estimation of the parameters $\theta = \{\theta_k : k = 1, \ldots, K\}$ and $\phi$ of the distribution $\mathcal{G}$; (iii) the estimation of the location $\tau = \{\tau_k : k = 0, \ldots, K\}$ of the change-points. Our main concern is the localization of exon/intron boundaries and hence the estimation of $\tau$. While it can be hard in general to estimate $K$, this parameter is often known in the context of re-annotation. Additionally, although the parameters $\{\theta_k\}$ are typically not of interest, they can often be estimated trivially by maximum likelihood given estimates of $K$ and $\tau$.

Because of the discrete nature of RNA-Seq data, we consider methods that model read counts using a Poisson ($\mathcal{P}$) or negative binomial ($\mathcal{NB}$) distribution, that is, assume that

$$\mathcal{P} : \ \mathcal{G}(\theta_k, \phi) = \mathcal{P}(\theta_k)$$

$$\mathcal{NB}: \mathcal{G}(\theta_k, \phi) = \mathcal{NB}(\theta_k, \phi).$$

Note that, for the Poisson distribution, $\theta_k$ coincides with the mean parameter $\mu_k$. For the negative binomial distribution, $\theta_k$ denotes the probability parameter ($0 \leq \theta_k \leq 1$) and $\phi > 0$ the dispersion parameter, so that the mean signal on the $k$th segment is $\mu_k = \phi(1-\theta_k)/\theta_k$ and the variance $\mu_k(1 + \mu_k/\phi) \geq \mu_k$. Because RNA-Seq read counts typically exhibit overdispersion, the negative binomial model is most appropriate (Robinson, McCarthy, and Smyth 2010; Risso et al. 2011). When $\phi \to +\infty$, with $\theta_k$ such that the ratio $\mu_k = \phi(1-\theta_k)/\theta_k$ remains constant, one recovers the Poisson distribution with parameter $\mu_k$. Our model requires that the dispersion parameter be constant over all segments.

Since the numbers of reads $Y_t$ are assumed to be independent at each position, the log-likelihood can be decomposed into the sum of the log-likelihoods for each segment, i.e.,

$$\log p(y|K, \tau, \theta, \phi) = \sum_{k=0}^{K-1} \sum_{t=\tau_k}^{\tau_{k+1}-1} \log\big(g(y_t; \theta_k, \phi)\big),$$

where $g(\cdot; \theta_k, \phi)$ is the probability density function (PDF) of distribution $\mathcal{G}$. In order to work in a Bayesian framework, one further needs to specify prior distributions $p(K)$, $p(\tau|K)$, and $p(\theta|K, \tau)$. Their choice is discussed in Rigaill, Lebarbier, and Robin (2012).

### 2.2. SEGMENTATION METHODS

Most segmentation methods comprise two steps. The first combines inference questions (ii) and (iii) by estimating, for a given number of segments $K$, the location of the change-points $\{\tau_k\}$ and the parameters $\{\theta_k\}$ and $\phi$ using, for example, maximum likelihood. The second step is then to estimate the number of segments $K$, resolving inference question (i). Note that some methods such as PELT combine the two steps into one, estimating the parameters of $\mathcal{G}$, the change-point locations, and the number of segments directly, using, for example, a penalized version of the likelihood.

Estimating $K$ can be viewed as a model selection problem, for which natural approaches include cross-validation and penalized likelihood criteria. Although cross-validation methods have been proposed in the context of segmentation (Arlot and Celisse 2010), the interpretation of cross-validation is problematic due to the spatial structure and hence dependence of the data. Furthermore, the approach is time-consuming and no software implementation is currently available. We therefore focus on likelihood-based goodness-of-fit criteria, where the estimator $\hat{K}$ of the number of segments $K$ maximizes some function $crit(K; y)$ of the data $y$ with respect to $K$ (for simplicity, we adopt the shorter notation $c(K)$). We consider specifically the following three criteria, corresponding to three different penalties for the likelihood.

- The natural approach in a Bayesian framework is to maximize the posterior probability of $K$ given the data, i.e., select the $\hat{K}$ maximizing $c(K) = \log p(K|y)$, which in the case of EBS can be computed exactly. A crude approximation leads to a penalized version of the likelihood, $c(K) = \log p(y|K, \tau, \theta, \phi) - K \log(n)$. In the sequel, we refer to both criteria as the Bayesian Information Criterion (BIC), $BIC(K)$.

- The penalized likelihood criterion of Cleynen and Lebarbier (under review), proposed in a non-asymptotic framework that takes into account the complexity of the visited segmentation, is defined as $PL(K) = \log p(y|K, \tau, \theta, \phi) - \beta K(1 + 4\sqrt{1.1 + \log(\frac{n}{K})})^2$, where $\beta$ is a constant tuned according to the data.
- The Integrated Completed Likelihood (ICL) criterion is defined in a Bayesian framework as

$$ICL(K) = \log p(K|y) + \mathcal{H}(K),$$

where $\mathcal{H}(K) = -\sum_{m \in \mathcal{M}_K} p(\tau|y, K) \log p(\tau|y, K)$ is the posterior entropy. Indeed, the segmentation $\tau$ can be viewed as an unobserved variable, in the sense that the segment labels of each data point $y_t$ are unknown. Rigaill, Lebarbier, and Robin (2012) introduced this criterion in the context of change-point detection and showed that it performs better than other criteria such as the BIC or DIC (Deviance Information Criterion, i.e., the expected deviance of the model). In a frequentist framework, the ICL criterion can be approximated by $ICL(K) = BIC(K) - \sum_{m \in \mathcal{M}_K} p(\tau|y, K, \theta, \phi) \log p(\tau|y, K, \theta, \phi)$.

If one is not concerned with obtaining an estimate for the number of segments $K$ or if one does not trust the estimation of $K$, the following two Bayesian approaches are available. The first applies the BIC directly to the segmentation, so that $BIC(\tau) = \log p(\tau|y)$, and chooses the $\hat{\tau}$ that maximizes this criterion. In our study, results using $BIC(\tau)$ and $ICL(K)$ were very similar and only the later will be discussed. The second approach is to integrate posterior probabilities of interest (as those mentioned next) over the possible values of $K$ rather than choose an optimal one (e.g., method EBS-a discussed below). Such model averaging presupposes the ability to compute posterior distributions for $K$ and $\tau$.

Finally, to allow precise and confident re-annotation, it is useful to obtain credibility intervals. The posterior distribution of the $j$th change-point of a segmentation into $k$ segments is

$$p_{\tau_j,y,k}(t) = \mathbf{P}\{\tau_j = t|Y = y, K = k\}, \quad \forall t \in [\![1, n + 1[\![, \tag{2.1}$$

from which we can derive, by model averaging, the probability of, say, the first change-point occurring at position $t$,

$$\mathbf{P}\{\tau_1 = t|Y = y\} = \sum_k p_{\tau_1,y,k}(t)\mathbf{P}\{K = k|Y = y\}, \tag{2.2}$$

where $\mathbf{P}\{A\}$ is the probability of event $A$. One can then define 95 % credibility intervals by, for instance, selecting values of highest posterior probability until 95 % coverage.

In our comparison of segmentation algorithms, we are therefore interested in the following functionalities: (i) the ability to model RNA-Seq read counts using a discrete distribution, such as the Poisson or negative binomial; (ii) the ability to estimate the number of segments $K$ according to criteria such as those mentioned above; (iii) the possibility to obtain credibility intervals. The left part of Table 1 summarizes the available functionality for the algorithms introduced in Section 1.

Table 1. Properties of segmentation algorithms. Left: × indicate available functionality in terms of distribution and model selection; ⊗ indicate methods retained for the simulation study (see RESULTS AND DISCUSSION section). Right: × indicate comparison criteria that can be computed for each method.

| | Functionality | | | | | | | Comparison criteria | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Distribution | | Model selection | | | | Model | Global | Local | ROC | Credibility | Run- |
| Algorithm | $\mathcal{P}$ | $\mathcal{NB}$ | $BIC(K)$ | $BIC(\tau)$ | $PL(K)$ | ICL | averaging | fit | fit | curves | intervals | time |
| CART | ⊗ | | ⊗ | | | | | × | × | | | × |
| PELT | ⊗ | | ⊗ | | | | | × | × | | | × |
| PDPA | × | ⊗ | × | | ⊗ | | | × | × | | | × |
| postCP | × | ⊗ | × | | | ⊗ | | × | × | × | × | × |
| EBS | × | ⊗ | × | × | | ⊗ | ⊗ | × | × | × | × | × |

### 2.3. SIMULATION STUDY DESIGN

*Data Sets* The simulation study was conceived to mimic typical RNA-Seq data. We used as benchmark strand-specific, poly(A)-selected *S. cerevisiae* RNA-Seq data from the Sherlock Laboratory at Stanford University (Risso et al. 2011) and publicly available from the NCBI's Sequence Read Archive (SRA, *http://www.ncbi.nlm.nih.gov/sra*, accession number SRA048710). Reads were mapped to the reference genome using Bowtie (Langmead et al. 2008) and strand-specificity and read-length information was taken into account in our analysis. We selected a set of five genes (YAL038W, YAL035W, YAL030W, YAL019W, and YAR008W) that were previously annotated by Nagalakshmi et al. (2008) and that span representative scenarios for yeast RNA-Seq data, in terms of gene length, number of exons, and read counts. Figure 1 of the Supplementary Materials shows the original unnormalized base-level read counts for these five genes.

In order to choose realistic values for the parameters of the distributions used to simulate read counts, we considered the annotation of the five genes in Nagalakshmi et al. (2008) as the reference or "true" segmentation. For each gene, we first fit a negative binomial distribution $\mathcal{NB}(\theta_k, \phi)$ to each segment $k$ and estimated $\theta_k$ using the method of moments and $\phi$ using a modified version of the Jonhson and Kotz's estimator (Johnson, Kemp, and Kotz 2005). Specifically, for each sliding window of size $h$ equal to twice the size of the longest zero band, we computed the method of moments estimator of $\phi$, using the formula $\phi = \mathbf{E}^2(X)/(\mathbf{V}(X) - \mathbf{E}(X))$, and retained the median over all windows. We also computed the maximal value of the read counts over the entire region. The results are given in Table 2.

Following Lai et al. (2005), we created an artificial four-exon gene, with $K = 9$ segments defined by $\tau = (1, 101, 121, 221, 271, 371, 471, 571, 1071, 1171)$. An odd segment corresponds to an intronic region (average size 100 bases), while an even segment corresponds to an exon (length varying from 20 to 500 bases).

We considered three simulation scenarios, corresponding to two parametric distributions and one resampling-based distribution.

- For the *Negative Binomial* (NB) scenario, with $\mathcal{G}(\theta_k, \phi) = \mathcal{NB}(\theta_k, \phi)$, we used the artificial four-exon gene segmentation and set the dispersion parameter $\phi$ to 0.27 for all segments. For odd segments (i.e., introns), we chose $\theta_{2k+1} = 0.9$, and for even

Table 2. Estimates of model parameters for each of the five yeast genes. For each segment, parameters correspond to a negative binomial distribution with mean $\mu_k = \phi(1-\theta_k)/\theta_k$ and dispersion $\phi$.

| Gene | Length | Dispersion, $\hat{\phi}$ | Empirical mean, $\hat{\mu}_k$ | max $y_t$ |
|---|---|---|---|---|
| YAL038W | 2003 | 0.3121 | (0.2928, 325, 0.2738) | 6216 |
| YAL035W | 3509 | 0.2523 | (0.2174, 7.81, 0.1232) | 690 |
| YAL030W | 967 | 0.2966 | (0.0044, 1.55, 0.08, 3.62, 0.103) | 167 |
| YAL019W | 3896 | 0.2721 | (0, 1.196, 0.0266) | 25 |
| YAR008W | 1328 | 0.2758 | (0, 1.325, 0.0466) | 34 |

segments (i.e., exons), we allowed $\theta_{2k}$ to vary smoothly between 0.2 and 0.001. For each value of $\theta_{2k}$, we simulated 100 data sets.

- For the *Mixture of discrete Uniforms* (MU) scenario, with $\mathcal{G}(\theta_k, \phi) = \frac{1}{2}\mathcal{U}([\![0, \theta_k/2]\!]) + \frac{1}{2}\mathcal{U}([\![0, \theta_k]\!])$, we again used the artificial four-exon gene segmentation and set $\theta_{2k+1} = 4$ and allowed $\theta_{2k}$ to vary smoothly between 24 and 6250. For each value of $\theta_{2k}$, we simulated 100 data sets.

- For the *Resampling* (RS) framework, we considered the true segmentation of each of the five yeast genes (Nagalakshmi et al. 2008) and resampled the counts of each segment at random, with replacement, i.e., $\mathcal{G}(\theta_k, \phi) = \text{sample}([\![y_{\tau_k}; y_{\tau_{k+1}}[\![)$. For each gene, we repeated this procedure 100 times.

In the remainder of the paper, we let $\mu$ represent the mean signal intensity over even segments (i.e., exons), so that $\mu_{2k}$ is equal to $\phi(1-\theta_{2k})/\theta_{2k}$ in the NB simulations and $3\theta_{2k}/8$ in the MU simulations and refers to the qualitative level of expression of the genes in the RS simulations. With parameters chosen as above in the NB and MU simulations, the different $\theta_{2k}$ yield comparable signal intensities $\mu_{2k}$. Note that we have associated $\mu$ with the level of expression of a gene, but that it can also relate to the sequencing coverage of an experiment. While we will only refer to the former in the manuscript, low-expressed genes from experiments with higher coverage might present the same characteristics as highly-expressed genes from experiments with lower coverage.

*Comparison Criteria* In the simulation study, the segmentation methods are compared according to the following criteria.

- The *global fit* index *gf* assesses the global quality of a proposed segmentation, in the sense that it reflects the agreement between the true segmentation $\tau$ and the estimated segmentation $\hat{\tau}$ over all pairs of bases in the region. Specifically, let $C_t$ be the true index of the segment to which base $t$ belongs and let $\hat{C}_t$ be the index estimated by the method, then

$$gf = \frac{2}{(n-1)(n-2)} \sum_{s=1}^{n} \sum_{t=s+1}^{n} [\mathbf{1}_{C_t=C_s}\mathbf{1}_{\hat{C}_t=\hat{C}_s} + \mathbf{1}_{C_t\neq C_s}\mathbf{1}_{\hat{C}_t\neq\hat{C}_s}].$$

- The *local fit* index $lf$ assesses the ability to recover a particular change-point $c$ and is defined by

$$lf(c) = \delta_c / P_{k(\hat{\tau})},$$

where $\delta_c$ is equal to 1 if the method finds a change-point at most three bases away from $c$ and 0 otherwise, $k(\hat{\tau})$ is the number of segments of the segmentation $\hat{\tau}$, and $P_k$ is the probability that a segmentation into $k$ segments has a change-point at $c$, i.e., $P_k = \frac{k-1}{n-1}$. Note that while the choice of a three-base tolerance threshold is somewhat subjective and allows change-points to be detected more easily, the ranking of the methods is robust to the value of the threshold (results not shown).

- *Receiver Operator Characteristic* (ROC) curves for methods yielding change-point probabilities as in Equations (2.1) and (2.2).

- For the Resampling Simulation scenario and methods yielding change-point probabilities, the percentage of true change-points covered by 95 % credibility intervals defined by starting from the mode of the distributions in Equation (2.1) or (2.2) and adding the next most probable location until 95 % (or slightly more because of the discrete nature of the distribution) of the mass has been reached. This leads to intervals that may not be contiguous, but have the smallest possible length.

- The average run-time.

The right part of Table 1 indicates which criteria are applicable for each of the algorithms to be evaluated.

## 3. RESULTS AND DISCUSSION

### 3.1. PRELIMINARY REMARKS

We first compared all algorithms with every available distribution. Our results show (see Figure 2 of Supplementary Materials) that when an algorithm was implemented for both the Poisson and negative binomial distributions (PDPA, postCP, and EBS), the latter always performed better. This is to be expected, as read counts typically exhibit overdispersion. For this reason, as well as to simplify the reporting of results and figures, we only retained PDPA, postCP, and EBS with the negative binomial distribution and CART and PELT with the Poisson distribution, as indicated by $\otimes$ symbols in the left part of Table 1. Although this may appear to bias the results in favor of PDPA, postCP, and EBS, the comparison is still fair, as the restriction of CART and PELT to the Poisson distribution and their inability to accommodate overdispersion is a clear limitation of these methods. Furthermore, the Poisson distribution is included as special case of the negative binomial implementation of PDPA, postCP, and EBS.

Method postCP failed to return a segmentation for a number of simulations (221 times for the mixture of uniforms scenario, 111 times for the negative binomial scenario and, for the RS scenario, 21 times for gene YAL030W, 20 times for gene YAL035W, and 5 times for gene YAR008W). The results presented in this section exclude these cases for postCP.

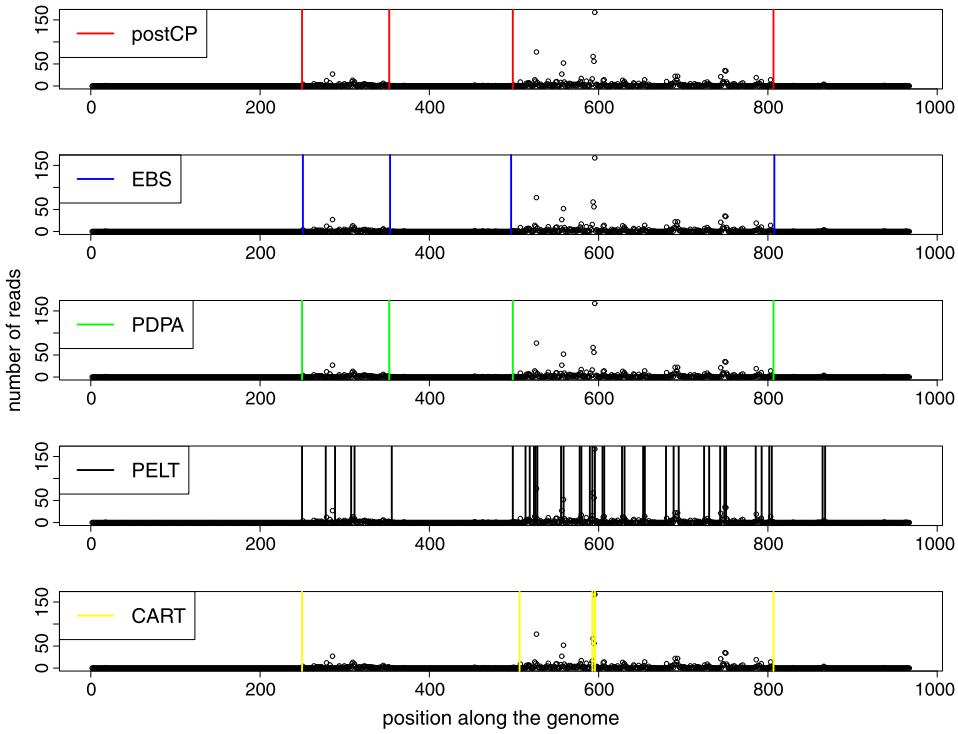


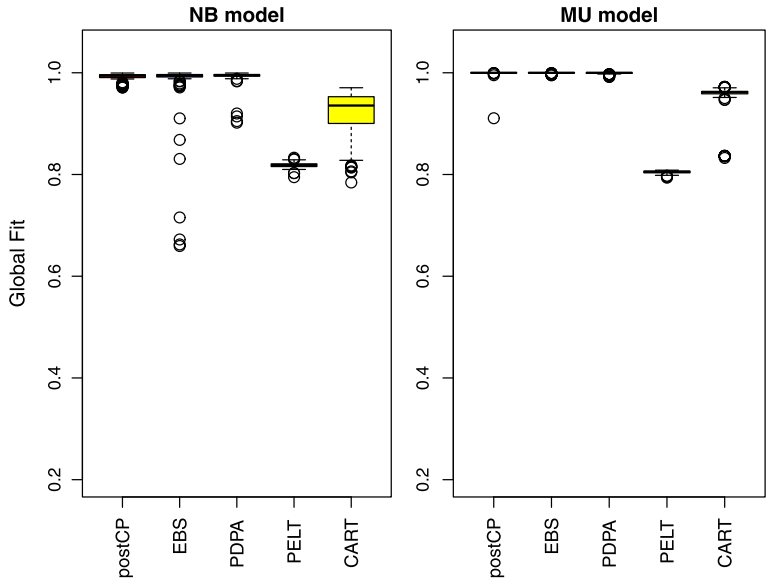

Figure 1. Segmentation of gene YAL030W. For each of five methods, segmentation based on actual RNA-Seq read counts. Vertical lines indicate a change-point found by the method. Note that no true segmentation is available, but the change-points proposed by each method can be related to the reference annotation in Nagalakshmi's paper, which has four change-points. PELT and CART propose additional change-points, probably due to their use of the Poisson distribution which fails to account for overdispersion of the read counts.

Figure 1 displays the segmentations obtained with each of five methods for gene YAL030W. In this particular example, postCP, EBS, and PDPA recover the segmentation in Nagalakshmi et al. (2008), while PELT largely overestimates the number of exons and CART misses the $3'$ boundary of the first exon and erroneously splits the second exon into three.

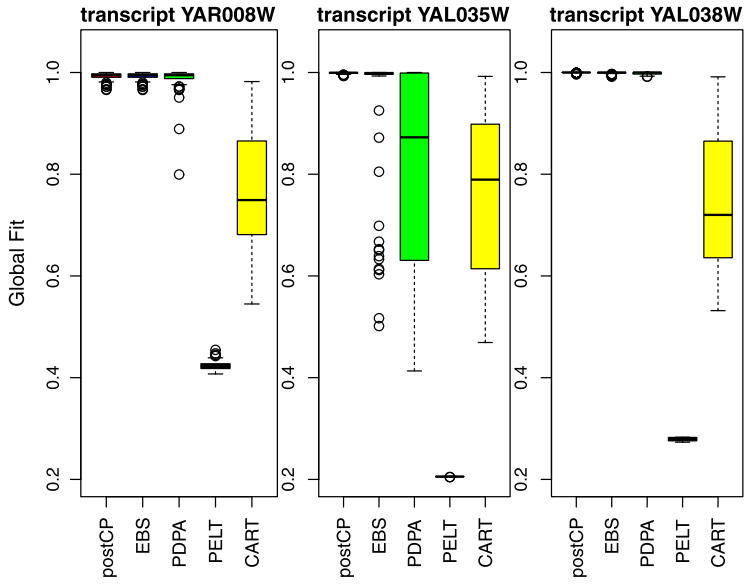
### 3.2. QUALITY, PRECISION, AND CONFIDENCE IN THE CHANGE-POINT LOCALIZATION

For both the negative binomial and mixture of uniforms simulation scenarios, methods implemented with the negative binomial distribution performed better than others according to the global fit criterion. As expected, we observed a general trend of slight improvement as the signal intensity $\mu$ increased. Figure 2 illustrates the performance of each method according to the global fit index for a particular value of $\mu$ corresponding to a moderate level of expression.

On the data sets simulated by resampling, however, we noticed that the effect of length was significant (Figure 2). For instance, for the long gene YAL035W, methods PDPA and




(a) Global fit for NB and MU simulations




(b) Global fit for RS simulations

Figure 2. Global fit. (a) Box plots of the global fit index by segmentation method for data sets simulated with the NB (left) and MU (right) models, for moderate expression levels of $\mu = 2.2$. (b) Box plots of the global fit index by segmentation method for data sets simulated by resampling for three different genes: YAR008W (left), which is short and lowly-expressed, YAL035W (middle), which is very long and more highly-expressed, and YAL038W (right), which is of average size but of very high expression level. See Supplementary Materials for more detail on the three genes. Methods implemented with the negative binomial distribution (EBS, PDPA, and postCP) outperform those based on the Poisson distribution (CART and PELT).

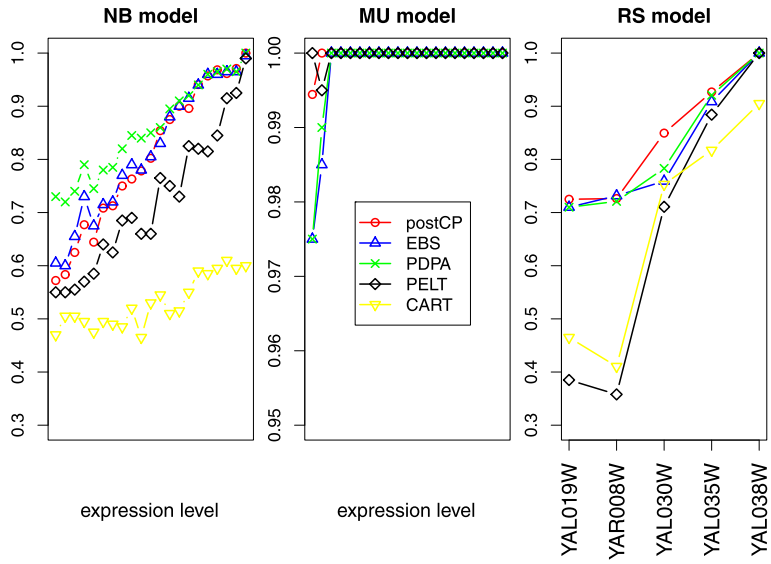



Figure 3. Local fit. Average of the proportions of simulations for which a method finds a change-point within three bases of the first and last change-points, respectively (i.e., $lf(\tau_1) > 0$ and $lf(\tau_{K-1}) > 0$) vs. expression level $\mu$. Left: Negative binomial. Middle: Mixture of uniforms. Right: Resampled data for the five yeast genes (ordered by increasing $\mu$). Increasing the level of expression improves the ability to identify change-points. Once again, methods implemented with the negative binomial distribution (EBS, PDPA, and postCP) outperform those based on the Poisson distribution (CART and PELT).

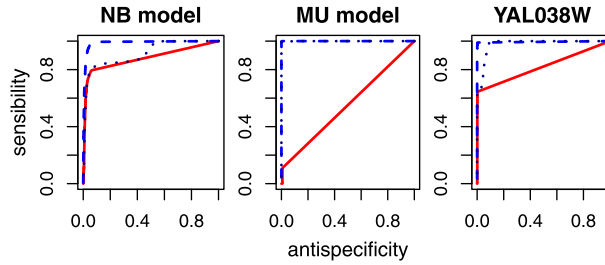



Figure 4. ROC curves. ROC curves for comparing the performance of EBS and postCP under different simulation scenarios (averaged over 100 simulations): postCP (—), EBS (- - -), and EBS-a (···). Left: Negative binomial ($\mu = 4.2$). Center: Mixture of uniforms ($\mu = 4.2$). Right: Resampling for gene YAL038W.

PELT drastically worsened. EBS and postCP consistently showed satisfying results and CART and PELT remained the least accurate methods.

Figure 3 displays the performance of each method in terms of local fit averaged over the first and last change-points, which are of particular interest in the context of UTR annotation. We observe that local fit improves for all methods as the expression level $\mu$ increases, although the methods tend to overestimate the number of segments when $\mu$ is high (see Figure 7 of Supplementary Materials).

Methods EBS and postCP yield posterior change-point probabilities for any given genomic location (see Equation (2.1)). An example is given in Figure 3 of the Supplementary Materials. This can be used to evaluate false positive and false negative rates for, say, the

Table 3. Credibility intervals. Median length of the 95 % credibility intervals and percentage of simulations for which the intervals covered the true change-point (out of 100).

| | Interval length | | | Coverage | | |
|---|---|---|---|---|---|---|
| Gene | postCP | EBS | EBS-a | postCP | EBS | EBS-a |
| YAL019W | 18 | 20 | 393 | 0.35 | 0.95 | 1 |
| YAR008W | 16 | 17 | 354 | 0.2 | 0.98 | 1 |
| YAL030W | 10 | 37 | 398 | 0.12 | 0.99 | 1 |
| YAL035W | 8 | 10 | 322 | 0.1 | 0.97 | 1 |
| YAL038W | 4 | 7 | 198 | 0.1 | 0.99 | 1 |

first change-point $\tau_1$. Specifically, for a given simulation and threshold $s$, a position $t$ is declared as first change-point if $\mathbf{P}\{\tau_1 = t|y, K\} \geq s$. Averaging the resulting proportions of false positives and false negatives over simulations and varying $s$ leads to the ROC-like curves of Figure 4.

The postCP's performance is acceptable when the data are simulated according to its negative binomial model, but very poor for the mixture of uniforms. Furthermore, performance deteriorates with increasing expression level $\mu$ (results not shown). A possible explanation is the very sharp aspect of the posterior distribution for the change-point location, which leads to false positives as soon as the mode is not equal to the true change-point. For the Resampling scenario, postCP's performance is good, but again worsens as the level of expression increases (see Figure 4 of the Supplementary Materials). Method EBS has good overall performance, with nearly perfect ROC for each model. Averaging over the number of segments $K$ (EBS-a), as in Equation (2.2), does not seem to improve the results.

Both postCP and EBS also provide posterior credibility intervals. Table 3 presents the average width and the percentage of nominal 95 % credibility intervals covering the true change-point $\tau_1$ (over 100 simulations). We display the results for the first change-point $\tau_1$ for ease of comparison with the Bayesian aggregation method EBS-a (indeed, studying the $k$th change-point would require a segmentation into at least $k + 1$ segments and thus the modification of the prior used for $K$), but results are similar for other change-points. The empirical coverage of EBS is close to the nominal credibility of 95 %, with reasonably narrow intervals. The empirical coverage of EBS-a is 100 %, at the price of huge credibility intervals, precluding its use in practice. This observation and the ROC curves of Figure 4 suggest that EBS-a yields a well-located posterior mode, but too large a posterior variance. postCP showed very poor coverage (fewer than 40 % of the simulations had a 95 % credibility interval covering the true location), due to its small credibility intervals that do not account for uncertainty in the estimation of the parameters $\theta_k$ and $\phi$. Results of the comparison are similar across expression levels $\mu$.

### 3.3. NUMBER OF CHANGE-POINTS

All results presented up to now are based on methods that involve estimating the number of segments $K$. The accuracy of the resulting segmentation could therefore be affected by a poor choice of $K$. Figures 5–9 of the Supplementary Materials show the distribution

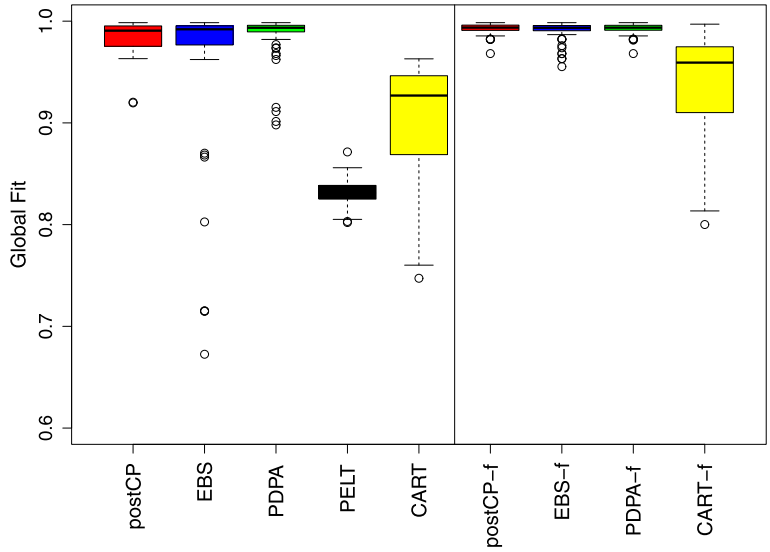



Figure 5. Global fit and estimation of $K$. Box plots of the global fit index by segmentation method, for estimated (left) and true (right, suffix '-f' added to method name) number of segments $K$. Data sets simulated from negative binomial model with $K = 9$ and $\mu = 1.1$. Knowing the number of segments improves the ability to recover the optimal segmentation. PELT cannot benefit from this advantage.

of estimates of $K$. In the context of genome re-annotation, where it is reasonable to assume that the number of segments is known (for instance, a gene with $K_e$ exons will have $K = 2 \times K_e + 1$ segments), it is of interest to compare segmentations based on the true and estimated $K$. Note that PELT cannot take advantage of the knowledge of $K$, as the estimation of $K$ is embedded in the algorithm.

The box plots in Figure 5 illustrate the advantage of providing the true $K$ to a segmentation method: for methods postCP, EBS, and CART, the global fit index $gf$ is less variable and higher with the true $K$ (right) than with estimated $K$ (left). This trend is mostly observed on data sets simulated with the negative binomial model. As expected, as the expression level $\mu$ increases and the segmentation becomes more obvious, the impact of the choice of $K$ for methods postCP and EBS lessens, as the ICL criterion becomes more accurate. In the case of PDPA, the model selection criterion already provides the true value of $K$ in more than 90 % of the simulations, thus the knowledge of $K$ does not yield a noticeable gain.

The ROC curves in Figure 6 illustrate the impact of the estimation of $K$ for methods EBS and postCP in terms of false positive and false negative rates. The gain from using the true $K$ lessens as the level of expression increases, regardless of the performance of the methods. Moreover, while performance improves for method EBS, postCP worsens with expression level.

### 3.4. EXTENSION TO MORE COMPLEX ORGANISMS

A natural question is how the methods would compare for an organism with a more complex gene structure than *S. cerevisiae*. We have therefore considered the artificial sce-

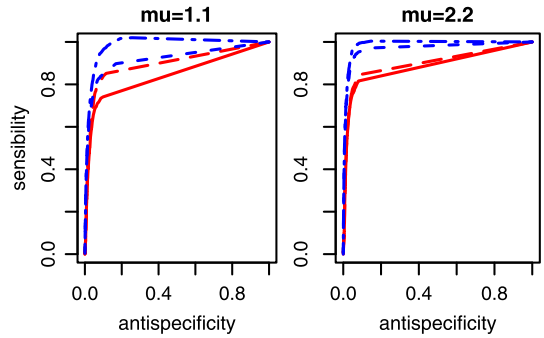



Figure 6. ROC curves and estimation of $K$ ROC curves for comparing the performance of EBS and postCP with known and estimated number of segments $K$: postCP with estimation of $K$ (—), postCP with known $K$ (– –), EBS with estimation of $K$ (- - -), and EBS with known $K$ (·-·). Data simulated from negative binomial model with $\mu = 1.1$ and $\mu = 2.2$. Once again, the knowledge of $K$ improves the performance of the algorithms.

nario in which two of the isoforms of the *Drosophila melanogaster* gene Inr-a, Inr-a-RB (six exons) and Inr-a-RC (two exons), are expressed at different levels (Figure 10 of the Supplementary Materials illustrates the gene and its isoforms).

In our simulation, we used the annotation of the Inr-a gene from FlyBase (*http://www.flybase.org*) to define the true segmentation. To simulate read counts, we pooled the observed counts from several yeast genes to create three classes of expression: intronic, low, and medium. Then, on each segment, the read counts were obtained by re-sampling at random, with replacement from the three groups. Specifically, we used the intronic class for segments corresponding to intronic regions of the two isoforms, the medium expression class for exons of Inr-a-RB, and the low expression class for exons of Inr-a-RC. Thus, for exons shared by the two isoforms, read counts are sums of counts from the low and medium classes. This created a synthetic signal for a gene of length 5000 nucleotides with 14 segments.

Figure 11(a) of the Supplementary Materials shows that all methods but PDPA fail to recover the right number of segments. This leads to poor results in terms of local and global fit. However, the prior information on $K$ allows method EBS to yield almost perfect ROC curves (see Figure 11(b) of Supplementary Materials), while methods PELT and CART still fail to retrieve an acceptable number of true change-points.

The segmentation methods considered in this article can be applied to organisms of a large range of complexity (in the number of exons, isoforms, etc.), provided that the exons of different genes of interest do not overlap, in which case it would not be possible to assign a segment to a specific gene. Fast methods such as CART, PELT or PDPA can be applied regardless of gene length. However, the EBS algorithm is restricted to sequences no longer than $10^4$ bases.

### 3.5. TRANSFORMATION OF THE DATA

One might be interested in transforming the discrete RNA-Seq read counts to allow the use of a wider range of methods, for instance, continuous data segmentation methods developed for microarrays. Because of encouraging results with EBS, we compared its

performance on the resampled data sets, with the negative binomial distribution, as above, and with the Gaussian distribution applied to log-transformed counts ($\tilde{y}_t = \log(y_t + 1)$). We also applied the variance-stabilization transformation corresponding to the negative binomial distribution (which involves the $arsinh$ function), but the results were similar to the widely-used and dispersion-independent log-transformation and thus are not presented here.

We observed that on the RS simulations, the two approaches yield very similar results. However, as illustrated by the ROC curves in Figure 12 of the Supplementary Materials, the negative binomial distribution is better for more complex scenarios, where some segments can be very small, as is the case with *D. melanogaster*'s introns.

## 4. CONCLUSION

This simulation study showed that each method is adapted to a different type of problem. CART and PELT perform worse in all situations for distinct reasons: CART is a heuristic that is most appropriate when the signal is long and segments are well-delimited (for example, large changes in the mean), while PELT fails because of its inability to choose an appropriate number of segments. Indeed, PELT was designed to segment profiles in which the number of segments increases with the length of the signal, which is not the case in our framework.

PDPA showed excellent results in proposing a segmentation close to the true one, especially when the signal was not both very long and high. The criterion used for the choice of the number of segments yielded good performances even when other methods failed. Its use is promising in a range of biological settings, such as transcript discovery or assessment of which genes are expressed.

Finally, postCP and EBS demonstrated the ability to both propose a segmentation that is very close to the true one and return distributions for the location of change-points, thereby allowing precise and confident re-annotation. Both methods showed equivalent results for their optimal segmentation, but EBS had better results in terms of ROC curves on the true data sets and showed a clear improvement when the number of segments was known. Figure 7 illustrates the results of method EBS on actual RNA-Seq data for the five yeast genes of interest.

Segmentation methods are of interest in related contexts such as whole-genome (re-)annotation or copy-number estimation using DNA-Seq data. However, in practice, few methods are fast enough to be applied in those frameworks. Indeed, in our simulation study, performed on a standard computer (Intel-Core2 Duo CPU P8400, 2.26 GHz $\times$ 2 with 3 Gio of RAM), the average run-times were very different among the methods. PELT and CART were almost instantaneous, while PDPA (respectively postCP) needed a few seconds per simulation (about 4 s (resp. 10 s) on model-simulated data sets, up to 20 s (resp. 50 s) on the longer genes). EBS was by far the slowest, needing about 15 seconds for model-simulated data sets and up to 6.5 minutes on the longer genes. While we would recommend using method EBS (with prior information on $K$ when available) for targeted transcript re-annotation, its run-time prohibits its use for larger segmentation problems.

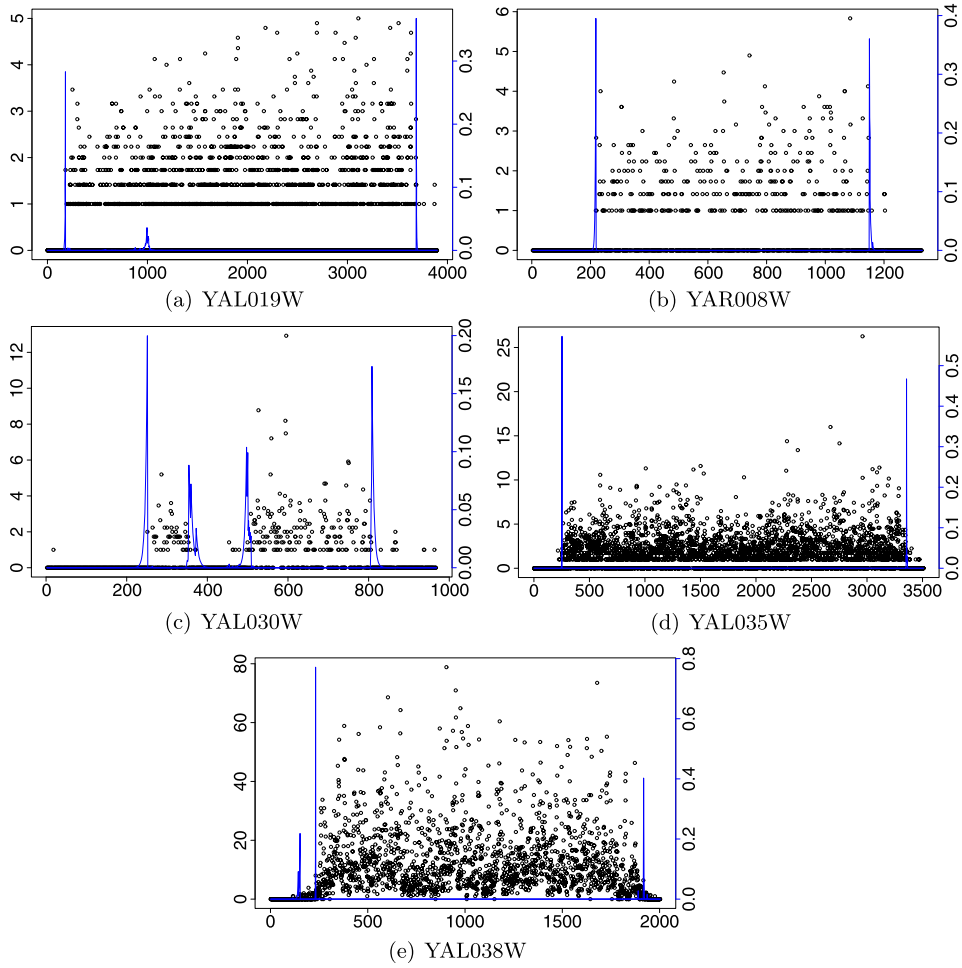



Figure 7. EBS segmentation of five yeast genes. Segmentation based on actual RNA-Seq read counts. Read counts were squared on the y-axis for more visibility.

A possible strategy would consist in first applying PDPA to large regions in order to delimit smaller regions of interest and then using EBS to obtain confidence intervals on the change-point locations within the smaller regions.

## 5. AVAILABILITY OF SUPPORTING DATA

The data set supporting the results of this article is available in the Sequence Read Archive repository, *http://www.ncbi.nlm.nih.gov/sra*, with the accession number SRA048710.

*[Received April 2013. Accepted September 2013. Published Online October 2013.]*

## REFERENCES

Arlot, S., and Celisse, A. (2010), "Segmentation of the Mean of Heteroscedastic Data via Cross-Validation," *Statistics and Computing*, 1–20.

Bai, J., and Perron, P. (2003), "Computation and Analysis of Multiple Structural Change Models," *Journal of Applied Econometrics*, 18, 1–22.

Barry, D., and Hartigan, J. (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*, 88 (421), 309–319.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011), "Control-Free Calling of Copy Number Alterations in Deep-Sequencing Data Using GC-Content Normalization," *Bioinformatics (Oxford, England)*, 27, 268–269.

Breiman, Friedman, Olshen, and Stone (1984), *Classification and Regression Trees*, Belmont: Wadsworth and Brooks.

Cleynen, A., Koskas, M., and Rigaill, G. (under review), "A Generic Implementation of the Pruned Dynamic Programing Algorithm," *arXiv:1204.5564*.

Cleynen, A., and Lebarbier, E. (under review), "Segmentation of the Poisson and Negative Binomial Rate Models: A Penalized Estimator," *arXiv:1301.2534*.

Guthery, S. B. (1974), "Partition Regression," *Journal of the American Statistical Association*, 69 (348), 945–947.

Hsu, L., Self, S., Grove, D., Randolph, T., Wang, K., Delrow, J., Loo, L., and Porter, P. (2005), "Denoising Array-Based Comparative Genomic Hybridization Data Using Wavelets," *Biostatistics*, 6, 211–226.

Hupé, P., Stransky, N., Thiery, J., Radvanyi, F., and Barillot, E. (2004), "Analysis of Array CGH Data: From Signal Ratio to Gain and Loss of DNA Regions," *Bioinformatics*, 20(18), 3413–3422.

Johnson, N., Kemp, A., and Kotz, S. (2005), *Univariate Discrete Distributions*, New York: Wiley.

Killick, R., and Eckley, I. (2011), changepoint: An R Package for Changepoint Analysis.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005), "Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data," *Bioinformatics (Oxford, England)*, 21 (19), 3763–3770.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2008), "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome," *Genome Biology*, 10.

Luong, T. M., Rozenholc, Y., and Nuel, G. (2013), "Fast Estimation of Posterior Probabilities in Change-Point Models Through a Constrained Hidden Markov Model," *Computational Statistics & Data Analysis*. *arXiv:1203.4394*.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008), "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing," *Science*, 320 (5881), 1344–1349.

Rigaill, G., Lebarbier, E., and Robin, S. (2012), "Exact Posterior Distributions and Model Selection Criteria for Multiple Change-Point Detection Problems," *Statistics and Computing*, 22, 917–929.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011), "GC-Content Normalization for RNA-Seq Data," *BMC Bioinformatics*, 12 (1), 480.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data," *Bioinformatics*, 26 (1), 139–140.

Scott, A., and Knott, M. (1974), "A Cluster Analysis Method for Grouping Means in the Analysis of Variance," *Biometrics*, 30, 507–512.