BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 10 2014, pages 1409–1416
doi:10.1093/bioinformatics/btu042

Gene expression

Advance Access publication January 27, 2014

Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models

Bogdan Mirauta[1,*], Pierre Nicolas[2,†] and Hugues Richard[1,†]

[1]Biologie Computationnelle et Quantitative, UPMC and CNRS UMR7238, Paris, France and [2]Mathématique Informatique et Génome, INRA UR1077, Jouy-en-Josas, France

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: The most common RNA-Seq strategy consists of random shearing, amplification and high-throughput sequencing of the RNA fraction. Methods to analyze transcription level variations along the genome from the read count profiles generated by the RNA-Seq protocol are needed.

Results: We developed a statistical approach to estimate the local transcription levels and to identify transcript borders. This transcriptional landscape reconstruction relies on a state-space model to describe transcription level variations in terms of abrupt shifts and more progressive drifts. A new emission model is introduced to capture not only the read count variance inside a transcript but also its short-range autocorrelation and the fraction of positions with zero counts. The estimation relies on a particle Gibbs algorithm whose running time makes it more suited to microbial genomes. The approach outperformed read-overlapping strategies on synthetic and real microbial datasets.

Availability: A program named Parseq is available at: http://www.lgm.upmc.fr/parseq/.

Contact: bodgan.mirauta@upmc.fr

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

Received on May 3, 2013; revised on November 30, 2013; accepted on January 21, 2014

1 INTRODUCTION

Sequencing technologies play an increasing role in the investigation of gene expression [RNA sequencing (RNA-Seq)]. The most common RNA-Seq strategy is based on random shearing, amplification and high-throughput sequencing of the RNAs, yielding millions of sequence reads that serve to characterize whole-genome transcriptional profiles (Holt and Jones, 2008; Marioni *et al*., 2008). Current protocols provide strand-specific data (Levin *et al*., 2010). After mapping onto a reference genome sequence, the number of reads found at each position of the genome is recorded and those counts can be used to derive estimates of gene expression up to the isoform level (Jiang and Wong, 2009; Mortazavi *et al*., 2008;Richard *et al*., 2010) under the assumption that read counts are proportional to transcript length and transcription level. The read coverage along the genome also provides a rich information that is often used to map new transcriptionally active regions (Nagalakshmi *et al*., 2008; van Dijk *et al*., 2011; Yassour *et al*., 2009).

Despite the amount of data collected in the past decade, initially with microarrays and now with sequencing, for most of the organisms, there are still no or incomplete annotations of their transcripts. Even the most studied model organisms are lacking a full characterization of their transcriptome architecture. Not only the complete condition-dependent repertoire of transcripts proves difficult to establish but also the biological meaning of important transcripts' categories, such as pervasive transcription in eukaryotes (Consortium *et al*., 2012; Graur *et al*., 2013; van Bakel *et al*., 2010) and antisense RNAs in bacteria (Nicolas *et al*., 2012; Raghavan *et al*., 2012; Thomason and Storz, 2010), remains elusive—hence, the importance of developing new computational approaches that could help to extract more information from RNA-Seq datasets.

A major research direction toward the identification of transcript structures is based on read assembling (Martin and Wang, 2011). Reference-based methods (Guttman *et al*., 2010; Trapnell *et al*., 2010) begin with the alignment of reads on the genome. Fragments are constructed by joining reads on the basis of paired-end or fragment length information. Fragment overlapping is then examined to build a connection graph. At the end, the connected reads are predicted to belong to the same transcript. Isoform structure can be inferred from the path of fragment contigs (Trapnell *et al*., 2010), and expression levels can be estimated after allocation of the reads to the inferred transcripts. Although this approach provides insightful results at a computationally affordable cost and can use reads overlapping exon junctions as direct evidence for splicing (Wang *et al*., 2010), it has also some intrinsic limitations. The most obvious is that limited depth of sequencing combined with technical biases may cause gaps that lead to artificial splits in transcript structure. Irrespective of the sequencing depth, this approach is also unable to point to overlapping transcripts caused by promoter multiplicity and incomplete termination. However, these two mechanisms contribute substantially to the transcriptome's complexity in organisms with compact-sized genomes (Nicolas *et al*., 2012).

Our aim in this study is to develop a principled strategy for analyzing changes in expression levels whose output could help to identify the variety of mechanisms shaping the transcriptional landscape. The task is complicated because of the existence of several types of protocol-induced biases that cause longitudinal variability of coverage along the chromosome. Part of these artifacts can be explained by the influence of the local nucleotide

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

composition on the priming step (Li *et al.*, 2010) and by other pre-sequencing procedures that can introduce biases in read coverage (Griebel *et al.*, 2012; Wu *et al.*, 2010;). To tackle these issues, we present a probabilistic model of RNA-Seq count data, which integrates transcription level variation as well as a generic description of the longitudinal variability induced by the sequencing protocol.

This modeling approach builds on previous works, originally motivated by the analysis of comparative genomic hybridization and transcription tiling array data that aimed at segmenting the signal into regions of piecewise constant expression. In this context, two major issues are the choice of the correct number of breakpoints (Picard *et al.*, 2005) and the assessment of uncertainty on breakpoint position (Huber *et al.*, 2006). The alternative adopted here consists of extending the probabilistic model to account for the full dynamics of the transcription signal (Nicolas *et al.*, 2009). Transcriptional landscape reconstruction is then conducted in the framework of hidden Markov models (HMMs) with hidden process in continuous state space, also known as state-space models (SSMs). We developed here procedures to estimate the model parameters, reconstruct local transcription levels, call transcribed regions and identify coverage breakpoints based on this framework.

2 A PROBABILISTIC MODEL FOR TRANSCRIPTION LEVELS AND READ COUNTS

2.1 The SSM framework

Throughout this work, we refer to the transcription level at position t of the genome as u_t. This level is scaled such that it corresponds to the expectation of the count y_t of reads whose 5′-ends map at position t: it is thus also proportional to the total number of reads sequenced. It cannot be directly equated to the read count y_t because of the randomness of the selection of the sequenced reads and to local variability artifacts. Our aim is to reconstruct the trajectory $\mathbf{u} = (u_t)_{t\geq 1}$ from the sequence of read counts $\mathbf{y} = (y_t)_{t\geq 1}$. For this purpose, we consider an SSM where u_t is a hidden variable taking values on the real half line $[0; +\infty)$ whose distribution depends on u_{t-1} via a Markov transition kernel and y_t is an observation whose emission distribution depends on u_t. This framework allows accounting for the longitudinal dependency between the u_t's and provides great flexibility in the modeling of y_t given u_t.

2.2 Longitudinal model of transcriptional level

The Markov transition kernel $k(u_{t+1}; u_t)$ that we use distinguishes expressed ($u_t>0$) and non-expressed ($u_t = 0$) regions and assigns a positive probability for unchanged transcription level between t and $t + 1$. The allowed changes of transcription levels between t and $t + 1$ breaks down into distinct types: jump from between expressed and non-expressed regions as well as changes of transcription level within transcribed regions—accounting for transcription initiation and termination sites in presence of overlapping transcription units. Following the work on tiling array data (Nicolas *et al.*, 2009), changes within transcribed regions further subdivide into two types that differ by their amplitudes and are referred as shifts (large amplitude) and drifts (small amplitude). Coexistence of shifts and drifts is designed to pull apart well-defined initiation or termination sites internal to transcribed regions from smoother changes in measured transcriptional levels that can have a biological origin (e.g. random termination events) or can reflect technical artifacts (e.g. longitudinal bias caused by messenger RNA capture and fragmentation protocols).

The Markov transition kernel $k(u_t; u_{t-1})$ for transcriptional level writes

$$\begin{aligned}&\mathbf{1}_{\{u_{t-1}=0\}}[(1-\eta)\delta_0(u_t)+\eta f(u_t)]\\&\quad+\mathbf{1}_{\{u_{t-1}>0\}}[\alpha\delta_{u_{t-1}}(u_t)+\beta f(u_t)+\beta_0\delta_0(u_t)\\&\quad+\gamma_u g_u(u_t; u_{t-1},\lambda)+\gamma_d g_d(u_t; u_{t-1},\lambda)],\end{aligned}$$

where $\mathbf{1}$ denotes the indicator function that serves to indicate whether $t - 1$ is an expressed or non-expressed position, and δ_x denotes the Dirac delta function with mass at point x that gives a non-zero probability for unchanged transcription level and for jumping to 0 between $t - 1$ and t. The parameters $\eta \in (0, 1)$ and $(\alpha, \beta, \beta_0, \gamma_u, \gamma_d) \in (0, 1)^5$ with $\alpha + \beta + \beta_0 + \gamma_u + \gamma_d = 1$ define the probabilities of the different types of moves. The terms $f(u_t; \zeta)$, $g_u(u_t; u_{t-1}, \lambda)$ and $g_d(u_t; u_{t-1}, \lambda)$ are probability densities for the transcription level u_t, at the beginning of a transcribed region (occurring with probability η when $u_{t-1} = 0$) or after a shift (probability β when $u_{t-1}>0$), after an upward drift (probability γ_u when $u_{t-1}>0$) and after a downward drift (probability γ_d when $u_{t-1}>0$), respectively. The density $f(u_t; \zeta)$ corresponds to an exponential distribution of rate ζ (mean $1/\zeta$) and the parameter $\lambda>0$ defines the average relative change caused by drifts: $(u_t - u_{t-1})/u_{t-1}$ if upward drift or $(u_{t-1} - u_t)/u_t$ if downward drift.

2.3 Distribution of read counts in real datasets

The variability of read counts observed when resequencing the same library has been described as almost compatible with a Poisson distribution (Marioni *et al.*, 2008). However, when compared between samples (or even replicate libraries), it exhibits overdispersion, and the negative binomial (NB) distribution is often used to accommodate this behavior (Anders and Huber, 2010; Robinson *et al.*, 2010). Initially, we planned to rely also on the NB to account for read counts overdispersion between positions inside each transcript. It seems required to involve a mixed Poisson distribution to account simultaneously for the incompressible variance of the final sampling by sequencing (Poisson) and for the extra-variability introduced by randomness in library preparation and by position-specific biases that can be introduced at all steps of the protocols. In this context, the NB is viewed as a gamma–Poisson mixture [$y_t \sim$; Poisson($u_t z_t$), where z_t follows a gamma distribution with mean 1 and variance ϕ] stands as the most tractable model (Karlis and Xekalaki, 2005).

Based on two real datasets, we examined the distribution of read counts inside regions expected to be homogeneous in terms of expression level. Namely, we asked whether the NB could capture the relationships between mean and variance and simultaneously account for the fraction of positions with zero-counts (Fig. 1). Both the characteristics are expected to impact directly on the decision to predict read counts at distant positions as originating from the same transcript. The most obvious

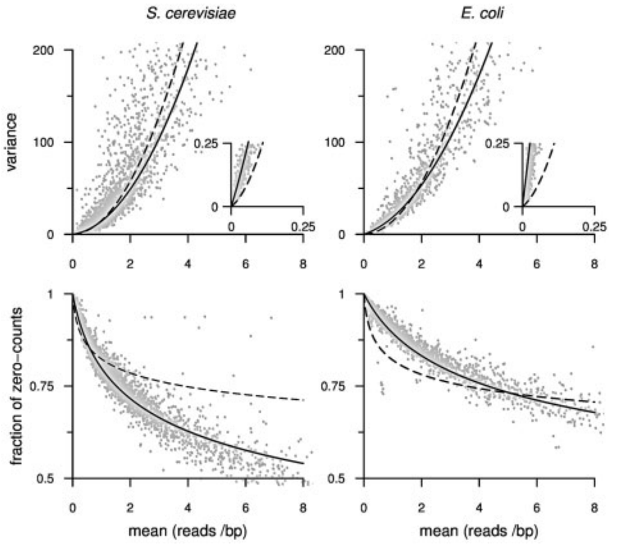

Fig. 1. Distribution of read counts inside regions of homogeneous expression. *S.cerevisiae* dataset SRR121907 (left); *E.coli* dataset SRR794838 (right). Each long open-reading frame (region without in-frame stop codon) identified on the genome is represented by a dot. Dashed lines show the fit of the NB model with overdispersion parameter estimated via variance (reads^2/bp^2) versus mean (reads/bp) regression; plain lines show the fit with the Parseq model

discrepancy between the data and the NB is with respect to the zero counts: given the mean and the variance of the empirical distribution, the fraction of positions with zero counts under the NB assumption tends to be too low for low expression levels and too high for high expression levels.

The usual parametrization of the NB with overdispersion parameter ϕ mentioned above is also contradicted by the data. The variance increases markedly faster than the mean u even for low expression level, in sharp contrast with the prediction that the variance should write $u + \phi u^2$. In the Poisson-mixture context, breaking these relationships that arise from law of total variance implies that the relationship between the mixing distribution and u is more subtle than a simple scaling. This prompted us to search for a more accurate model that would make sense from a mechanistic perspective.

2.4 Read count emission model

We developed a new RNA-Seq read count emission model that fits much better the characteristics of the real data than the simple NB (Fig. 1). Its construction intends to account for the three main steps of the experimental protocol: (i) initial molecule sampling and fragmentation, (ii) amplification and (iii) final sampling by sequencing. Namely, we write $y_t \sim$; Poisson$(x_t a_t)$, where a_t is distributed over $[0, +\infty)$ with mean μ_a and x_t follows a discrete distribution over $\{0, 1, \ldots, +\infty\}$ with mean u_t/μ_a; hence $E(y_t) = u_t$. The term x_t is aimed at representing the number of molecules with 5′-end mapping to position t after initial sampling; a_t wishes to capture the effect of randomness in amplification and position-specific biases in amplification and sequencing, μ_a should be interpreted as an amplification coefficient corresponding to the average number of reads per initial molecule sampled; the Poisson distribution accounts for the final sampling. For simplicity, we choose a gamma distribution for a_t and an NB distribution for x_t.

In practice, the parameters of the gamma distribution for a_t (size κ, scale θ, $\mu_a = \kappa\theta$) are obtained by examining the distribution of counts in regions of low expression, i.e. where x_t is expected to be 1 if a count is observed. The NB for x_t is obtained by writing $x_t \sim$; Poisson$(u_t s_t/\kappa\theta)$ where s_t follows a gamma distribution with mean 1 and variance κ_s (i.e. size κ_s and scale $1/\kappa_s$), the parameter κ_s is estimated on the variance versus mean and fraction of zero counts versus mean plots (Fig. 1). Decomposing the NB for x_t as Poisson-mixture allows to account for the pattern of short-range autocorrelation between counts (Supplementary Fig. S2) by making $\mathbf{s} = (s_t)_{t\geq 0}$ a piecewise constant Markov chain.

Integrating out these three sources of variability and the possibility of outliers, the density $\pi(y_t; u_t, s_t)$ of our complete read count emission model writes the following:

$$(1-\varepsilon_b-\varepsilon_o)\sum_{x_t=0}^{\infty}\text{Poisson}\left(x_t;\frac{u_t s_t}{\kappa\theta}\right)\cdot\text{NB}\left(y_t;\kappa,\frac{x_t\theta}{x_t\theta+1}\right)$$
$$+\,\varepsilon_b\text{NB}_{-\{0\}}\left(y_t;\kappa,\frac{\theta}{\theta+1}\right)+\varepsilon_o\text{Uniform}(y_t;0\ldots b),$$

The parameters $(\varepsilon_b, \varepsilon_0) \in (0, 1)^2$ account for the possibility of background noise outside transcribed regions and outliers $(\varepsilon_b + \varepsilon_0 \leq 1)$, respectively. The NB density term within the sum arises after integration over all possible values of a_t. A complete description of the relationships between the variables y, u, s and the parameters (hereafter referred collectively as Θ) is found in Supplementary Sections S1 and S3.

3 TRANSCRIPTIONAL LANDSCAPE RECONSTRUCTION

3.1 Markov chain Monte Carlo with particle Gibbs

In SSMs, the reconstruction of the hidden trajectory, given the parameter values and the observed data (here the characterization of $\mathbf{u}|\mathbf{y}, \Theta$), is more challenging than in a classical HMM where only discrete values are considered for the hidden variable. The forward–backward recursions that provide exact answers in the context of classical HMMs need to be substituted by particle filtering algorithms build on sequential Monte Carlo (SMC) principles whose results are only approximate for finite numbers of particles (Doucet and Johansen, 2008). Parameter inference that relies heavily on hidden trajectory reconstruction in this category of models is also directly impacted. Here, the existence of a second hidden variable s_t and the sequence lengths ranging in millions of base pairs increase the difficulty.

To circumvent these problems, we used a recently described SMC method known as particle Gibbs (PG) that makes it possible to obtain exact (but correlated) joint samples of the hidden trajectory and parameters, given the data (Andrieu *et al.*, 2010). PG is based on a modified SMC step, the conditional SMC (CSMC), that is integrated into more general Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of the parameters. In this setup, this also allowed to combine the reconstructions of $\mathbf{u}|\mathbf{s}, \mathbf{y}, \Theta$ and $\mathbf{s}|\mathbf{u}, \mathbf{y}, \Theta$ to obtain a joint reconstruction of $\mathbf{u}, \mathbf{s}|\mathbf{y}, \Theta$ and to extract the marginal of interest $\mathbf{u}|\mathbf{y}, \Theta$. We also implemented an additional PG step intended to preserve $\mathbf{s}\times\mathbf{u}$ by rescaling $\mathbf{s}$ when updating $\mathbf{u}$. The problem posed by sequence length was properly handled within the PG framework by successive partial (block) CSMC updates of the hidden trajectories. To validate the implementation of our PG algorithm, we extended the algorithm to sample the joint $(\mathbf{s}, \mathbf{u}, \mathbf{y}, \Theta)$ distribution and verified that we could retrieve the priors. Detailed descriptions of the parameter priors, MCMC and validation procedures used in this work are provided in Supplementary Sections S2 and S3.

3.2 The Parseq workflow

In theory, our PG algorithm permits to tackle parameter estimation and transcriptional landscape reconstruction simultaneously, but our software Parseq subdivides the problem in three successive steps for practical reasons (Fig. 2). The parameters of the read count emission model are estimated, and the emission density corresponding to the different values of $u_t s_t$ are tabulated (step 1). PG iterations are too time-consuming to be performed on a single CPU for genomes of moderate sizes such as the yeast *Saccharomyces cerevisiae* (~12 Mb). To distribute computation on independent CPUs, we decided to subdivide each chromosome in fragments (~1 Mb each), to perform parameter estimation separately on these fragments, and then to select a common set of parameters based of the obtained results (step 2). Posterior sampling of transcriptional landscape trajectories **u** is then carried out on a different CPU for each genome fragment, but with common parameters (step 3). With an Intel Core i7-3610QM CPU @ 2.30 GHz, each complete sweep of the MCMC algorithm was recorded to take ~1 min for 1 Mb using 150 particles in each CSMC update. In this study, we used 2200 sweeps, including 200 burn-in sweeps, for parameter estimation (step 2), and 2200 sweeps for making predictions at fixed parameters (step 3). Thus, on multi-CPU computers, the complete procedure takes slightly <3 days for each dataset with this algorithm setup, which we currently recommend for applications.

The output of the algorithm is a sample of transcriptional landscape trajectories drawn from $\mathbf{u}$—$\mathbf{y}$, Θ that conveys rich information about the actual transcriptional landscape. Here, these trajectories served to estimate the expected value of u_t, the 95% credibility interval of u_t and the probability of $u_t > 0$ (transcribed position), together with the probability of the different types of breakpoints along the sequence. Because of the posterior uncertainty on the exact position of each breakpoint, we further aggregated the breakpoint probabilities at adjacent positions into small regions with high cumulative probabilities using a local-score approach. The weighted center of each small region was taken as a point estimate of the position of the breakpoint and the cumulative probability served as a confidence measure. According to the direction of the change in expression level, the breakpoints were identified as upshifts or downshifts. To better distinguish genuinely expressed regions from (biological or technological) background noise, we also realized the relevance of computing the probability for u_t to be above a selected cutoff and to predict the breakpoints that lead the trajectory **u** above this cutoff. Details on the workflow, including parameter estimation and post-processing, are provided in Supplementary Section S4. Transcriptional landscape reconstruction is illustrated on Figure 3.

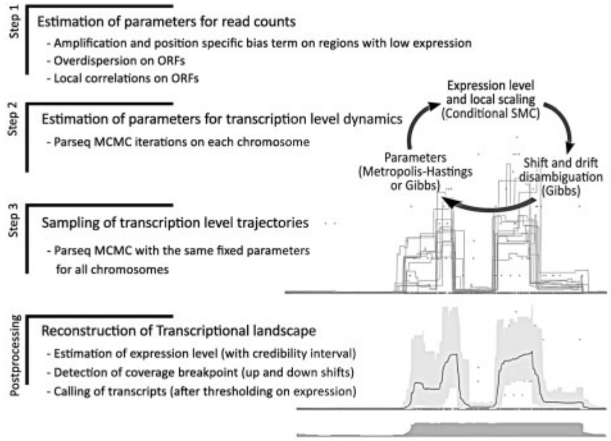
Step 1
Estimation of parameters for read counts
- Amplification and position specific bias term on regions with low expression
- Overdispersion on ORFs
- Local correlations on ORFs
Step 2
Estimation of parameters for transcription level dynamics
- Parseq MCMC iterations on each chromosome
Expression level and local scaling (Conditional SMC)
Shift and drift disambiguation (Gibbs)
Parameters (Metropolis-Hastings or Gibbs)
Step 3
Sampling of transcription level trajectories
- Parseq MCMC with the same fixed parameters for all chromosomes
Postprocessing
Reconstruction of Transcriptional landscape
- Estimation of expression level (with credibility interval)
- Detection of coverage breakpoint (up and down shifts)
- Calling of transcripts (after thresholding on expression)

Fig. 2. The Parseq workflow: from parameter estimation to reconstruction of transcriptional landscape

4 RESULTS AND DISCUSSION

4.1 Evaluation on synthetic data

The difficulty to find a reference annotation that could be considered as a gold standard motivated the idea of starting our analysis with a synthetic dataset. Strand-specific datasets of increasing sequencing depth (between 0.025 and 0.4 reads/bp after mapping) were simulated with the Flux simulator v1.2 (Griebel *et al.*, 2012) using the sequence and annotation of the *S.cerevisiae* S288C chromosome IV (Supplementary Section S4). The 50 bp-long reads were aligned on the reference sequence with Bowtie 1 v0.12.7 (Langmead *et al.*, 2009), allowing only one mismatch in a 5 bp seed (`-n 1`), and discarding multiple alignments (`-m 1`).

The accuracy of transcriptional landscape reconstruction was assessed from two different standpoints: the number of transcribed positions that can be correctly called based on the estimated value of u_t, and the number of transcript 5′- and 3′-ends at <50 bp of an identified upshift and downshift, respectively. To establish the lists of predictions, we used a probability cutoff set to 0.5 for both the probability of $u_t > 0$ and the cumulative probability of shift in the small region delineated by local-score approach. When comparing the predictions with a reference annotation, we needed to take into account that Parseq models the distribution of the 5′-end of the reads. For this reason, the regions predicted as transcribed by Parseq were extended of l_3 bp on their 3′-ends, and the same correction needs to be applied to the predicted downshifts before comparing with transcript 3′-ends (adjusted to 50 bp for the simulated dataset). To report results in terms of sensitivity and positive predictive values (PPV) we computed the fraction of the true positives that could be matched to a prediction and the fraction of the predictions that could be matched to a true positive. Parseq predictions were systematically compared with the results of Cufflinks v2.1.1 (Trapnell *et al.*, 2010), a method for transcript assembly that is based on read overlapping.

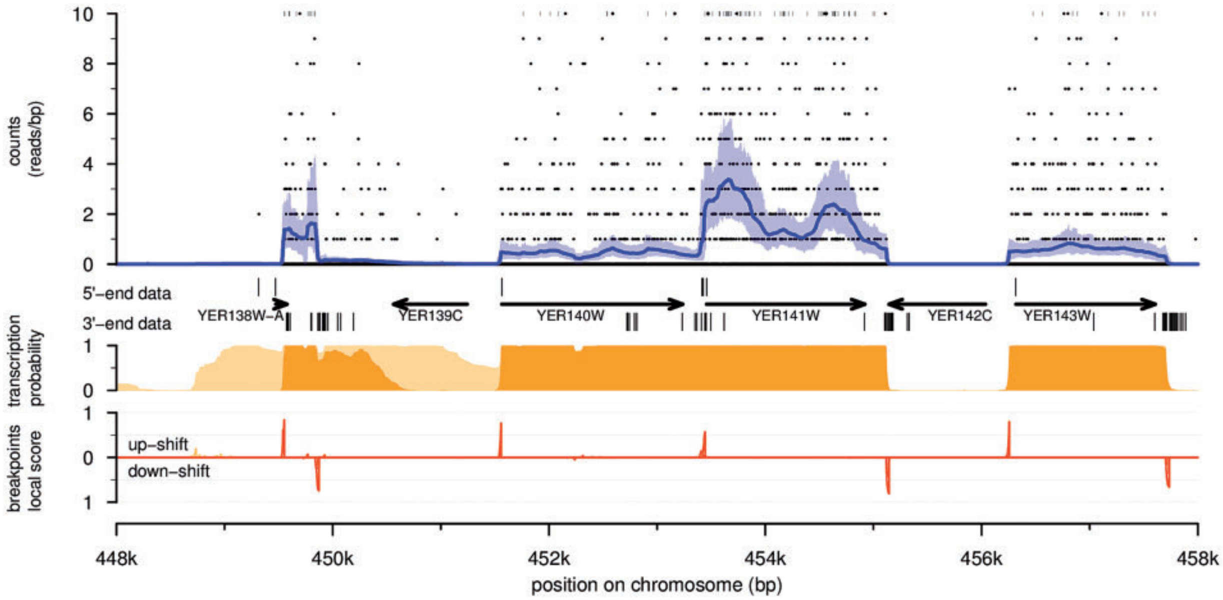

Fig. 3. Transcriptional landscape reconstruction with Parseq. Example of results on a 10 kb region of the first strand of *S.cerevisiae* chromosome V (dataset SRR121907). From top to bottom: read counts (dots) and the estimated expression profile (blue line) with its 95% credibility interval (light blue area); annotated CDSs (arrows) complemented with specific datasets of 5′- and 3′-ends (brown); probability of transcription with a cutoff on expression level set to 0^+ (light orange) or 0.1 reads/bp (orange); local score in high-scoring segments for the detection of breakpoints associated with upshifts and downshifts (red). This example illustrates the detection of overlapping transcription units (upshifts before YER140W and YER141W) and incomplete termination sites (downshift after YER138W-A)

The results obtained on synthetic data are summarized in Figure 4. While both Parseq and Cufflinks perform well when the depth of sequencing exceeds an average of 0.12 reads/bp, below this level the differences between the two methods become evident. Even though they do not have the same sensitivity-specificity trade-off, it appears clearly that the results obtained by Parseq are better. The model-based approach adopted in Parseq makes it possible to extrapolate transcription across coverage gaps and this results in a better calling of transcribed positions (not shown) and transcript borders. The mechanistic interpretation of our new emission model is also well supported by the results: Parseq estimation of the amplification coefficient (μ_a) distinguishes remarkably well the two scenarios considered in our simulations where sequencing depth increases either as a consequence of higher amplification or as a consequence of higher number of initial molecules sampled.

4.2 Evaluation on real data

On synthetic data, both the model-based approach of Parseq and the read-overlapping approach of Cufflinks perform well at detecting transcribed positions and transcript borders once the sequencing depth becomes high enough (0.12 reads/bp in our simulations). However, despite the efforts made on the simulation pipeline to mimic the different types of artifacts, the synthetic data do not have the complexity of a real dataset.

For evaluation on real data, we chose strand-specific single-end datasets from two major model microorganisms: the yeast *S.cerevisiae* and the bacterium *Escherichia coli*. The *S.cerevisiae* dataset was sequenced on a SOLiD platform (Short Read Archive identifier SRR121907) and published in a study on regulatory non-coding RNAs (van Dijk *et al.*, 2011). It has a read length of 50 bp and a sequencing depth of 1.6 reads/bp after mapping. The *E.coli* dataset (SRR794838) was sequenced on an Illumina platform and published together with the

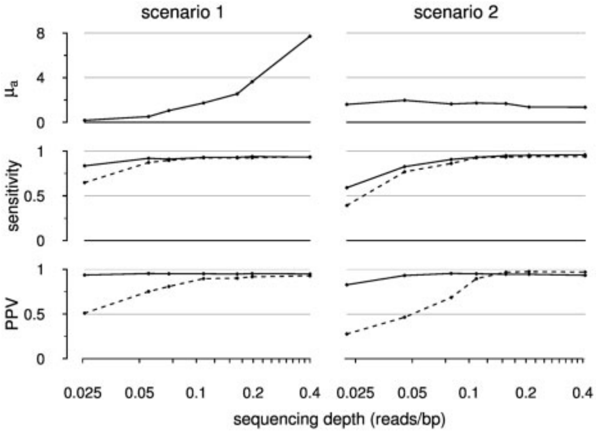

Fig. 4. Impact of sequencing depth on transcript borders prediction in synthetic data. Two scenarios were considered to achieve higher sequencing depth: increasing the amount of amplification (left column) or increasing the number of initial molecules before amplification (right column). The evolution of the amplification coefficient μ_a estimated by Parseq distinguishes the two scenarios (top row). The results of Parseq and Cufflinks (default parameters) are represented by continuous and dashed lines, respectively (middle and bottom rows). The results were similar for 5′- and 3′-ends and were pooled here

Table 1. Detection of transcribed positions and transcript borders on *S.cerevisiae* (SRR121907) and *E.coli* (SRR794838) datasets

Features	*S.cerevisiae*			*E.coli*			
	Reference	Parseq	Cufflinks	Reference	Parseq	Cufflinks	Rockhopper
Transcripts							
Sensitivity	CDSs and UTRs	0.83 (0.91)	0.83 (0.87)	Operons	0.56 (0.81)	0.60 (0.75)	0.21 (0.39)
PPV	CDSs and UTRs	0.90 (0.68)	0.90 (0.81)	Operons	0.76 (0.57)	0.72 (0.61)	0.91 (0.86)
5′ End							
Number		6689 (8353)	5484 (13 622)		1846 (2193)	1577 (7962)	2949 (4401)
Sensitivity	TSSs	0.64 (0.65)	0.43 (0.45)	Promoters	0.24 (0.25)	0.15 (0.23)	0.12 (0.19)
PPV	TSSs and 5′UTRs	0.48 (0.4)	0.49 (0.22)	Promoter and operon 5′-ends	0.49 (0.42)	0.34 (0.11)	0.24 (0.23)
3′ End							
Number		6287 (7440)	5484 (13 622)		1327 (1342)	1577 (7962)	2949 (4401)
Sensitivity	pAs	0.60 (0.62)	0.43 (0.44)	Terminators	0.12 (0.11)	0.08 (0.13)	0.03 (0.08)
PPV	pAs and 3′UTRs	0.57 (0.51)	0.51 (0.22)	Terminator and operon 3′-ends	0.35 (0.32)	0.24 (0.08)	0.07 (0.11)

Predictions and reference data were matched based on a ±50 bp distance cutoff (for a ±25 bp distance cutoff, see Supplementary Table S3). Outside parentheses: results obtained after applying a stricter expression cutoff. *S.cerevisiae*: 0.1 reads/bp for Parseq, 100 fragments per transcript for Cufflinks. *Escherichia coli*: 0.25 reads/bp cutoff for Parseq, 200 fragments per transcript for Cufflinks, $z = 0.2$ for Rockhopper. Between parentheses: 0^+ reads/bp for Parseq, 5 fragments per transcript for Cufflinks and $z = 0.01$ for Rockhopper.

presentation of the Rockhopper workflow for bacterial RNA-Seq data processing (McClure *et al.*, 2013). It has a read length of 100 bp and a sequencing depth of 2.4 reads/bp after mapping.

As a reference annotation for the transcribed positions in *S.cerevisiae*, we relied on the 5874 coding sequences (CDSs) found in the *S.cerevisiae* database SGD (Cherry *et al.*, 2012) and lists of untranslated regions (UTRs) mapped from RNA-Seq experiments in Yassour *et al.* (2009) (5200 5′UTRs and 5295 3′UTRs). To better assess the accuracy of the prediction of transcripts 5′- and 3′-ends, we also included comparison with experimental data that aimed at mapping precisely these sites: 4393 transcriptional start sites (TSSs) (Zhang, 2005), and 7977 polyadenylation sites (pAs) (Ozsolak *et al.*, 2010). For *E.coli,* we used annotations available in the RegulonDB database (Salgado *et al.*, 2013) (2438 promoters and 2647 operons) and also the sequence-based predictions of 2260 rho-independent transcription terminators obtained with Petrin software (d'Aubenton Carafa *et al.*, 1990).

Table 1 presents a detailed breakdown of the results according to the different sets of reference annotations, which could be considered to assess accuracy. In this context, we found that the probability of $u_t > 0$ (expression cutoff 0^+) is not necessarily the most relevant to compare the prediction of transcribed positions with a reference annotation. The best trade-offs are obtained near 0.1 reads/bp on the *S.cerevisiae* dataset, and 0.25 reads/bp on the *E.coli* dataset. These values are in agreement with the presence of a large number of positions associated with low expression level, resembling a background noise (Supplementary Figs S6 and S7). The accuracy of the detection of transcribed position is remarkable (e.g. 83% sensitivity, 90% PPV with the 0.1 reads/bp expression cutoff on *S.cerevisiae*), but similar to Cufflinks (Table 1). In keeping with our observations on synthetic data, this suggests that detecting transcribed positions is easy at high sequencing depth, and consequently, the model-based approach implemented in Parseq provides only small benefits.

The accurate identification of transcript borders is by far more challenging. For instance, on *S.cerevisiae* 5′-ends, with the same 0.1 reads/bp expression cutoff, the sensitivity reaches 64% and the PPV 48%. On *E.coli*, PPVs remain acceptable, but sensitivity values are much lower. This could be due to a combination of the following: lower quality of the data (μ_a estimated to 6.15 in *E.coli* versus 1.18 in *S.cerevisiae*, adjusted l_3 is 50 bp for *S.cerevisiae* versus 160 bp for *E.coli*); lower quality of the annotation taken as reference (e.g. Petrin predictions are expected to contain substantial numbers of false positives and false negatives); higher proportion of genes with low or no expression and thus for which promoters and terminators cannot be detected (with the 0^+ expression cutoff, sensitivity for detection of transcribed regions is only 0.81 in *E.coli* versus 0.91 in *S.cerevisiae*). On both datasets and for 5′-ends and 3′-ends alike, Parseq results are consistently better than the ones obtained by Cufflinks, particularly in terms of sensitivity. This confirms our expectations, as Cufflinks reconstruction ignores the possibility of overlapping transcripts and thus overlooks transcript-ends in these configurations. We also included in our comparison the predictions made on *E.coli* by Rockhopper (Table 1). As we were interested here in *de novo* predictions but this software could not run without annotations, we discarded successively the annotation on one-tenth of the genome and recorded the predictions on it. Parseq and Cufflinks provide results markedly better than Rockhopper in this comparison setup.

4.3 Importance of drift and local scaling

Transcript borders are detected on the basis of significant changes in read counts. Therefore, high variability in read counts can lead to breakpoint overpredictions resulting in a loss of specificity when not properly incorporated in the model. We palliated this need by introducing two different components in our model: a drift term on the transition kernel for progressive variations as opposed to the abrupt changes modeled by shifts,

Table 2. Impact of drift and local scaling

Parseq components	Included in the model			
Drift[a]	+	+	−	−
Autocorrelation[b]	+	−	+	−
5′-ends number	6689	13 881	15 994	31 428
TSS sensitivity	64%	70%	74%	79%
TSS PPV	48%	28%	25%	15%
3′-ends number	6,287	11 880	16 613	32 357
pAs sensitivity	60%	63%	70%	74%
pAs and 3′UTR PPV	57%	34%	29%	17%
CV[c] within CDSs	0.37	0.57	0.43	0.59

Results obtained on *S.cerevisiae* (SRR121907) chromosome IV (both strands) with expression cutoff 0.1 reads/bp.

[a] Drift is removed by setting $\gamma_u = \gamma_d = 0$.

[b] Short-range autocorrelation is removed by setting $\alpha_s = 0$, overdispersion is preserved by writing x_t as drawn from a NB instead of a Poisson–gamma mixture.

[c] Coefficient of variation.

and a local scaling Markov-dependent variable **s** intended to capture short-range autocorrelations. By monitoring the accuracy in terms 5′- and 3′-ends detection, we assessed the effect of these two model components on the quality of the inference. The results are reported in Table 2 and confirm that taken individually the drift and the local scaling improve the results. Moreover, the results also demonstrate that the two terms are complementary rather than redundant, as their combination leads to further improvements.

5 CONCLUSION

This article presents a model-based approach for analyzing the RNA-Seq read count profiles along the genome. The model aims for an account of artifactual longitudinal variability's sources, via a new model of overdispersion able to capture not only the variance versus mean relationships but also the fraction of zero counts and the short-range autocorrelations. From a methodological standpoint, our work also demonstrates the feasibility of analyzing genome-scale data within the framework of SSMs. The recently described PG algorithm (Andrieu *et al.*, 2010) was instrumental in this success. Running time does not depend on the depth of the sequencing, but is proportional to genome length, which makes it more suited to microbial genomes.

The method outperforms a read assembly approach at low sequencing depth, and shows a clear improvement on real data even for high sequencing depth. We believe that the availability of such a tool will become increasingly useful as the use of RNA-Seq becomes more popular. In particular, the availability of confidence scores and credibility intervals will be relevant to build a reference annotation from a compendium of experiments as done from tiling array data (Nicolas *et al.*, 2012) and also to compare global RNA-Seq profiles with the results of protocols targeting more specifically the sequencing of transcript ends (Lin *et al.*, 2013; Pelechano *et al.*, 2013). In principle, one could also envision to extend the model for situations where data would be collected with distinct RNA-Seq protocols on the same biological sample.

ACKNOWLEDGEMENTS

The authors are grateful to Adeline Samson and Claude Thermes for initial discussions, respectively, on the inference framework and on the RNA-Seq technology. The authors also thank Alessandra Carbone for her support and the reviewers for their stimulating comments.

Funding: Ministère de l'Enseignement Supérieure et de la Recherche, doctoral fellowship and teaching assistantship (B.M.).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome. Biol.*, **11**, R106.

Andrieu,C. *et al.* (2010) Particle markov chain monte carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **72**, 269–342.

Cherry,J.M. *et al.* (2012) *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.

Consortium,E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

d'Aubenton Carafa,Y. *et al.* (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their rna stem-loop structures. *J. Mol. Biol.*, **216**, 835–858.

Doucet,A. and Johansen,A.M. (2008) A tutorial on particle filtering and smoothing: fifteen years later. *Technical report, Department of Statistics, University of British Columbia*.

Graur,D. *et al.* (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome. Biol. Evol.*, **5**, 578–590.

Griebel,T. *et al.* (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, **40**, 10073–10083.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Holt,R.A. and Jones,S.J. (2008) The new paradigm of flow cell sequencing. *Genome res.*, **18**, 839–846.

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

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Karlis,D. and Xekalaki,E. (2005) Mixed poisson distributions. *Int. Stat. Rev.*, **73**, 35–58.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Levin,J.Z. *et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**, 709–715.

Li,J. *et al.* (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R25.

Lin,Y.F. *et al.* (2013) A combination of improved differential and global rna-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional rnas in propionibacterium acnes, a major contributor to wide-spread human disease. *BMC Genomics*, **14**, 620.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

McClure,R. *et al.* (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.*, **41**, e140–e140.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Nicolas,P. *et al.* (2009) Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics*, **25**, 2341–2347.

Nicolas,P. *et al.* (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, **335**, 1103–1106.

Ozsolak,F. *et al.* (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**, 1018–1029.

Pelechano,V. *et al.* (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.

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

Raghavan,R. *et al.* (2012) Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio*, **3**, e00156–12.

Richard,H. *et al.* (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.*, **38**, e112.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Salgado,H. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.

Thomason,M.K. and Storz,G. (2010) Bacterial antisense RNAs: how many are there, and what are they doing? *Annu. Rev. Genet.*, **44**, 167–188.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, **28**, 511–515.

van Bakel,H. *et al.* (2010) Most Dark matter transcripts are associated with known genes. *PLoS Biol.*, **8**, e1000371.

van Dijk,E.L. *et al.* (2011) XUTs are a class of xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, **475**, 114–117.

Wang,K. *et al.* (2010) Mapsplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Wu,Z. *et al.* (2010) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, **27**, 502–508.

Yassour,M. *et al.* (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 3264.

Zhang,Z. (2005) Mapping of transcription start sites in saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.*, **33**, 2838–2851.