An RNA-Seq read count emission model for transcriptional landscape reconstruction with state-space models

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

[1] Génomique des Microorganismes, UPMC and CNRS UMR7238, Paris, France. [2] Mathématique Informatique et Génome, INRA UR1077, Jouy-en-Josas, France.
⋆ To whom correspondence should be addressed. † Contributed equally to this work.
`http://www.lgm.upmc.fr/parseq`

Motivation Sequencing technologies play an increasing role in the investigation of gene expression (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 which serve to characterize whole-genome transcriptional profiles [1]. We developed Parseq, a statistical method to estimate the local transcription levels and to identify transcript borders. Extending previous work on tilling array [2], this transcription landscape reconstruction relies on a State-Space Model (SSM) to describe transcription level variations in terms of abrupt shifts and more progressive drifts. The transcription level at the genome position t is denoted u_t and corresponds to the expectation of y_t, the count of reads with 5'-end mapping to position t. It cannot be directly equated to the read count y_t due to randomness and biases in library preparation and sequencing. We use Particle Gibbs, a Gibbs algorithm based on a conditional Sequential Monte Carlo, to estimate the SSM parameters and reconstruct the trajectory $\mathbf{u} = (u_t)_{t\geq 1}$ from the sequence of read counts $\mathbf{y} = (y_t)_{t\geq 1}$.

A key point to obtain good results with this framework is to incorporate a realistic emission model for the distribution of y_t given u_t.

Distribution of read counts in real data sets The variability of RNA-Seq read counts has been approximated by a Poisson ($\mathcal{P}$) distribution when re-sequencing of the same library [1] and by a Negative Binomial (NB) when comparing between samples [3]. For $y_t|u_t$, a mixed Poisson distribution seems unavoidable 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 occurring at all steps of the protocols. The Poisson-Gamma mixture is the simplest choice as it corresponds to the NB. Here, it leads to envision the relationships $y_t \sim P(u_t z_t)$ where z_t would follow a Gamma distribution with mean 1 and variance ϕ (variance $u_t + \phi u_t^2$).

We examined the distribution of read counts in regions of homogeneous expression level (*e. g.* ORFs) of two real data sets. In particular, 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). Marked discrepancies between the data and the NB are seen not only in the fraction

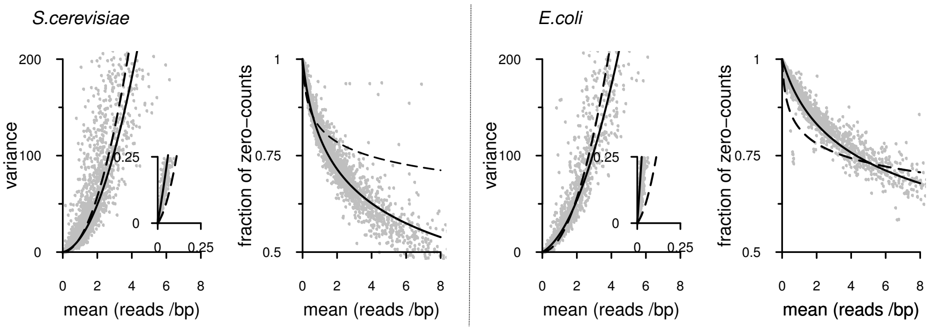

Fig. 1: Read counts distribution in regions of homogeneous expression. *S. cerevisiae* (SRR121907, SOLiD data, read coverage 1.6 reads/bp) - left; *E. coli* (SRR794838, Illumina, 2.4 reads/bp) - right. Dots: long open-reading frames (ORF, region without in-frame stop codon); Dashed lines: NB with overdispersion estimated via variance versus mean regression; Plain lines: Parseq model.

of zero-counts but also in the behaviour of the variance at low expression level where it exceeds $u_t + \phi u_t^2$. Breaking this relationship between mean and variance implies a relationships between the mixing distribution and u_t more subtle than a simple scaling.

A new read count model We adopted a mechanistic perspective to develop the more complex emission model incorporated in Parseq. Our aim is 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 $y_t \sim \mathcal{P}(x_t a_t)$ where a_t (mean μ_a) wishes to capture the effect of randomness in amplification, and x_t (mean u_t/μ_a) is aimed at representing the number of molecules after initial sampling. The Poisson distribution accounts for the final sampling. We capture the additional variability introduced by position-specific biases in library preparation with a supplementary random term s_t that impacts on initial molecule sampling : $x_t \sim \mathcal{P}(u_t s_t/\mu_a)$ with s_t having (for simplicity) a Gamma distribution of mean 1 an variance $1/\kappa_s$.

With a Gamma distribution (shape κ and scale θ) for the amplification term a_t and after integration over all a_t the density of our emission model writes

$$\pi(y_t; u_t, s_t) = \sum_{x_t=0}^{\infty} \mathcal{P}(x_t; \frac{u_t s_t}{\kappa\theta}) \cdot \mathcal{NB}(y_t; \kappa, \frac{x_t\theta}{x_t\theta + 1}).$$

This density could also be integrated over all s_t but in Parseq we introduced a Markov dependency between s_{t+1} and s_t to account for short-range autocorrelation between counts. This model makes it possible to capture the characteristics of the read count variability (fig. 1) and increases our ability to disentangle genuine transcription breakpoints from protocol induced variations.

1. J. C. Marioni, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.
2. P. Nicolas, et al. Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics*, 25(18):2341–2347, 2009.
3. Mark D. Robinson, et al. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.