TSSer: An automated method to identify transcription start sites in prokaryotic genomes from dRNA-seq data

Hadi Jorjani and Mihaela Zavolan

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland.

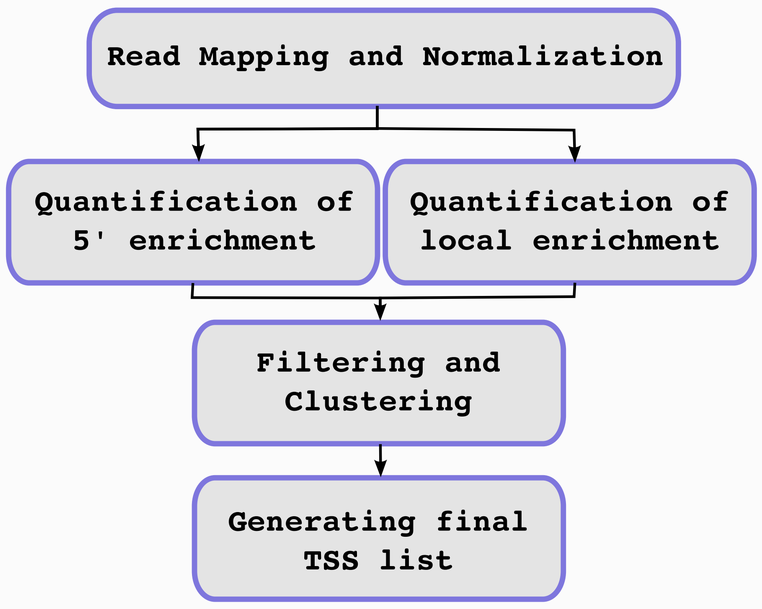
Swiss Institute of Bioinformatics

BIOZENTRUM

Universität Basel
The Center for
Molecular Life Sciences

1. Introduction: Promoter identification is a key step in studying the regulation of gene expression. Recently, a novel differential RNA sequencing (dRNA-seq) method was developed to discover bacterial transcription start sites (TSSs) at a genome wide scale. It uses 5' mono-phosphate-dependent terminator exonuclease (TEX) that specifically degrades 5' mono-phosphorylated RNA species such as processed RNA, mature rRNAs and tRNAs whereas primary transcripts remain intact. This approach results in an enrichment of primary transcripts, allowing TSSs to be identified by comparison of the TEX-treated libraries to control untreated ones. So far, an automated computational method to identify TSSs based on dRNA-seq data has not been available, and the TSS identification has been done to a great extent manually. To support future analyses of dRNA-seq data, we here introduce a rigorous computational method that helps identifying a large proportion of bona fide TSSs with relative ease. Our method is based on quantifying 5' enrichment of transcription start sites and also the significance of their expression relative to nearby putative TSSs. We have benchmarked our method on several recently published data sets and demonstrated that it enables accurate and automated TSS identification.

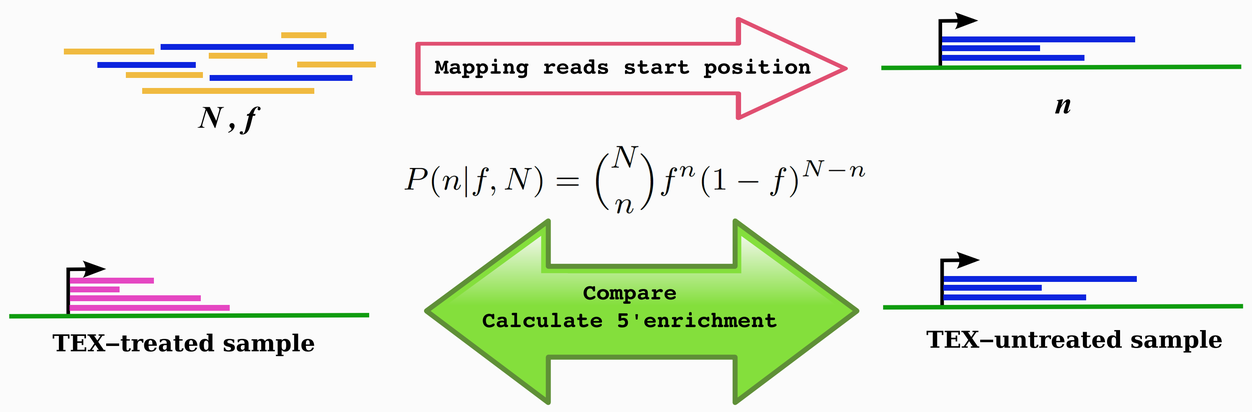
2. Schematic view of TSS identification pipeline



3.1 Quantification of 5' enrichment in TEX-treated compared to TEX-untreated sample

Basic Model:

Binomial distribution of reads starting at a given genomic position



$$P(n|f,N) = \binom{N}{n} f^n (1-f)^{N-n}$$

Quantification of enrichment based on a single paired sample

$$P(f_+ - f_- > 0 | n_+, N_+, n_-, N_-) = \phi\left(\frac{x_+ - x_-}{\sqrt{\frac{x_+(1-x_+)}{N_+} + \frac{x_-(1-x_-)}{N_-}}}\right)$$

Quantification of enrichment based on replicate paired samples

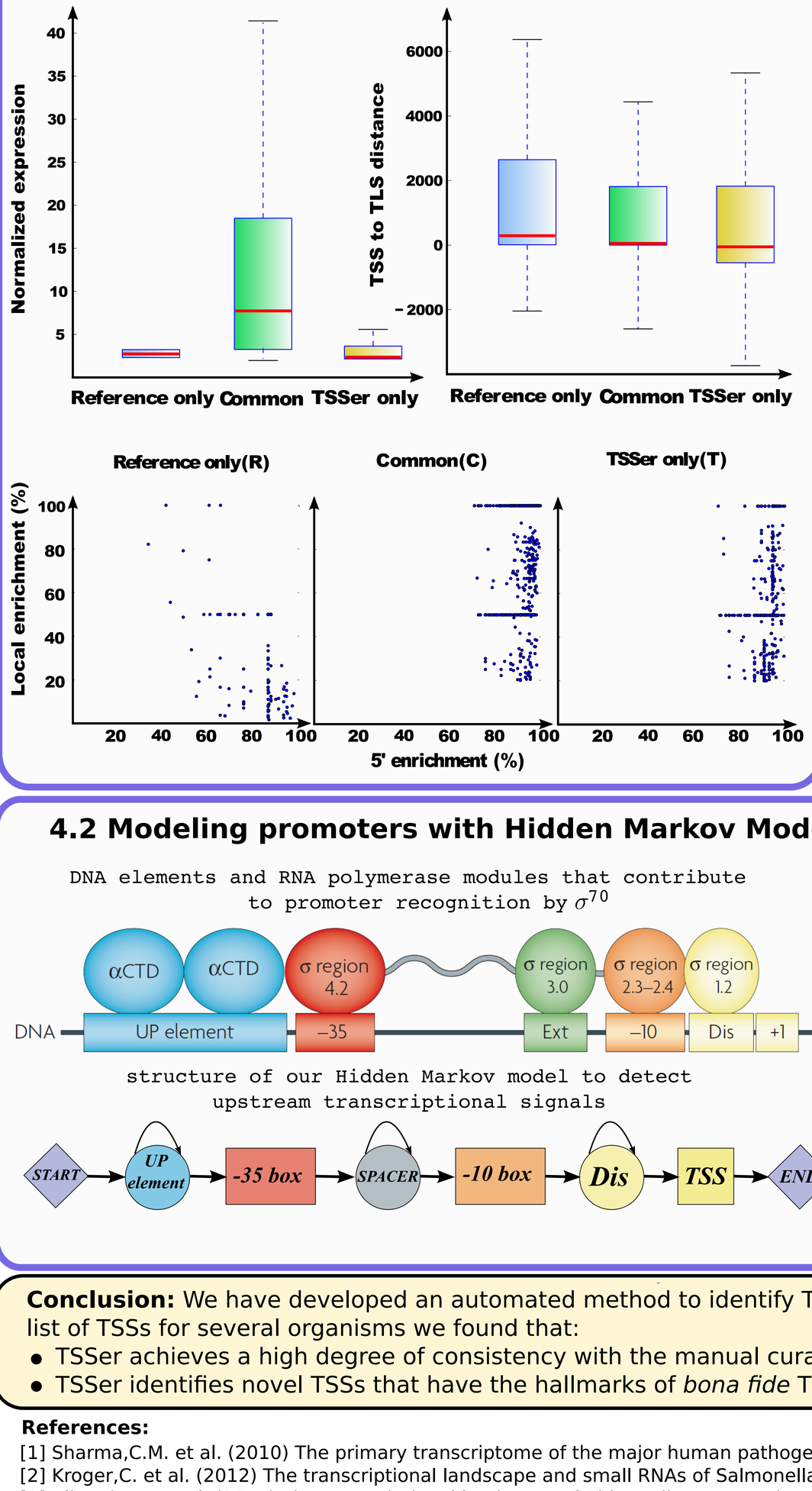
$$P(\mu > 1|\lambda) = \frac{\int_1^\infty \left(\frac{1}{(\mu-\mu_*)^2+\sigma_*^2}\right)^{\frac{n-1}{2}} d\mu}{\int_0^\infty \left(\frac{1}{(\mu-\mu_*)^2+\sigma_*^2}\right)^{\frac{n-1}{2}} d\mu}$$

where

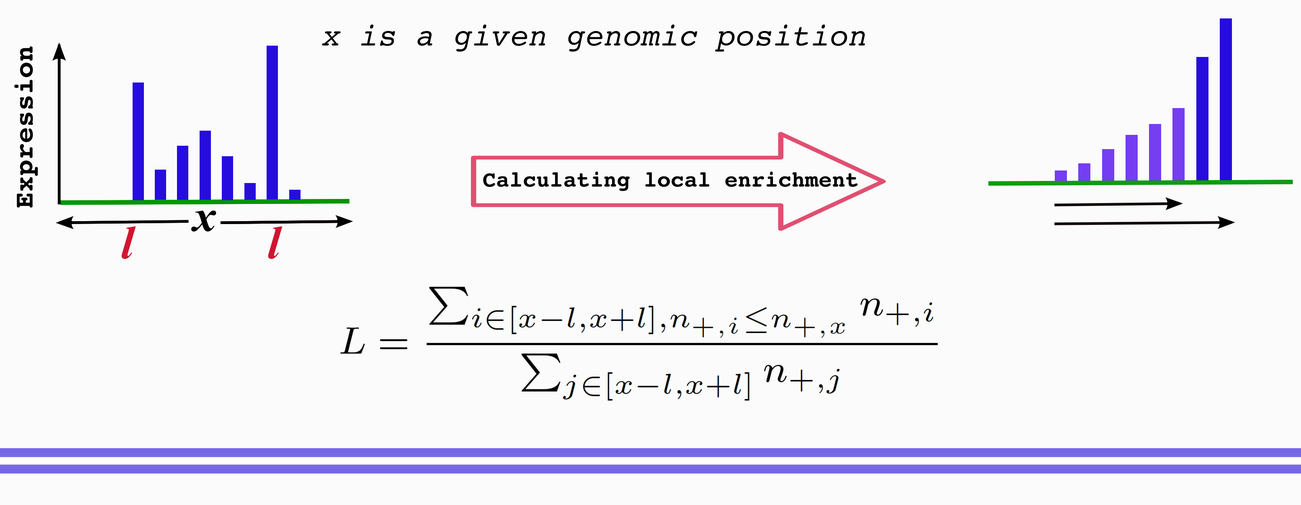
$$\lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \text{ , } \lambda_s = \langle \frac{f_+}{f_-} \rangle$$

4.1 Evaluation of TSSer

Comparison of TSSer-identified promoters to the reference promoter set of Helicobacter Pylori



3.2 Quantification of local enrichment in TEX-treated samples



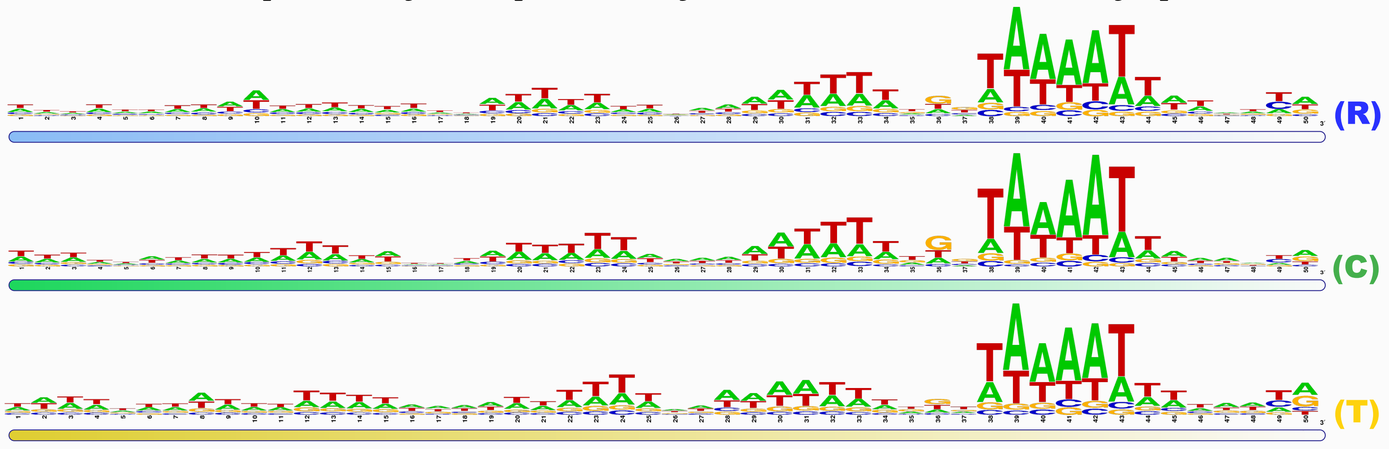
$$L = \frac{\sum_{i \in [x-l, x+l], n_{+,i} \leq n_{+,x}} n_{+,i}}{\sum_{j \in [x-l, x+l]} n_{+,j}}$$

4.2 Modeling promoters with Hidden Markov Model

DNA elements and RNA polymerase modules that contribute to promoter recognition by σ^{70}

structure of our Hidden Markov model to detect upstream transcriptional signals

Sequence logo of upstream regions for TSSs in each category



Conclusion: We have developed an automated method to identify TSSs given dRNA-seq data. Benchmarking TSSer on recently published reference list of TSSs for several organisms we found that:

- TSSer achieves a high degree of consistency with the manual curation
- TSSer identifies novel TSSs that have the hallmarks of *bona fide* TSSs

References:

[1] Sharma,C.M. et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori , Nature, 464, 250-255

[2] Kroger,C. et al. (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium , Proc Natl Acad Sci U S A, 109, 1277-1286

[3] Albrecht,M. et al. (2011) The transcriptional landscape of Chlamydia pneumoniae , Genome Biol, 12