BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 971–974
doi:10.1093/bioinformatics/btt752

Gene expression

Advance Access publication December 25, 2013

TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data

Hadi Jorjani and Mihaela Zavolan*

Computational and Systems Biology, Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Accurate identification of transcription start sites (TSSs) is an essential step in the analysis of transcription regulatory networks. In higher eukaryotes, the capped analysis of gene expression technology enabled comprehensive annotation of TSSs in genomes such as those of mice and humans. In bacteria, an equivalent approach, termed differential RNA sequencing (dRNA-seq), has recently been proposed, but the application of this approach to a large number of genomes is hindered by the paucity of computational analysis methods. With few exceptions, when the method has been used, annotation of TSSs has been largely done manually.

Results: In this work, we present a computational method called 'TSSer' that enables the automatic inference of TSSs from dRNA-seq data. The method rests on a probabilistic framework for identifying both genomic positions that are preferentially enriched in the dRNA-seq data as well as preferentially captured relative to neighboring genomic regions. Evaluating our approach for TSS calling on several publicly available datasets, we find that TSSer achieves high consistency with the curated lists of annotated TSSs, but identifies many additional TSSs. Therefore, TSSer can accelerate genome-wide identification of TSSs in bacterial genomes and can aid in further characterization of bacterial transcription regulatory networks.

Availability: TSSer is freely available under GPL license at http://www.clipz.unibas.ch/TSSer/index.php

Contact: mihaela.zavolan@unibas.ch

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

Received on June 27, 2013; revised on December 16, 2013; accepted on December 21, 2013

1 INTRODUCTION

Identification of transcription start sites (TSSs) is a key step in the study of transcription regulatory networks. It enables identification of promoter regions, and thereby the focused search for binding sites of transcription factors. Although for species such as mouse and human, methods to capture TSSs have been developed >10 years ago (Shiraki *et al.*, 2003), owing to differences in messenger RNA (mRNA) processing, these methods cannot be applied to bacteria. Recently, however, a method for genome-wide identification of bacterial TSSs has been proposed (Sharma *et al.*, 2010). The method, called differential RNA sequencing (dRNA-seq), uses the 5′ mono-phosphate-dependent terminator exonuclease (TEX) that specifically degrades 5′ mono-phosphorylated RNA species such as processed RNA, mature ribosomal RNAs and transfer RNAs, whereas primary mRNA transcripts that carry a 5′ triphosphate remain intact. This approach results in an enrichment of primary transcripts, allowing TSSs to be identified by comparison of the TEX-treated samples to control untreated ones. As an automated computational method to identify TSSs based on dRNA-seq data has not been available, TSS annotation based on dRNA-seq data required substantial effort on the part of the curators. The aim of our work was to develop an automated analysis method to support future analyses of dRNA-seq data. We here introduce a rigorous computational method that enables identification of a large proportion of *bona fide* TSSs with relative ease. The method is based on quantifying 5′ enrichment of TSSs and also the significance of their expression relative to nearby putative TSSs. Benchmarking our method on several recently published datasets, we find that the identified TSSs are in good agreement with those annotated manually, and that a relatively large number of additional TSSs that also have the expected transcription regulatory signals are identified. TSSer is freely available at http://www.clipz.unibas.ch/TSSer/index.php.

2 APPROACH

The input to TSSer is dRNA-seq data, consisting of one or more pairs of TSS-enriched (TEX-treated) and TSS-not-enriched samples. There are two main criteria that we use to define TSSs. The first criterion stems from the obvious expectation that TSSs are enriched in the TEX-treated compared with the TEX-untreated samples (Sharma *et al.*, 2010). To quantify the enrichment, we explored two methods. In one approach we calculated, for each genomic position, a '*z*-score' of the observed number of reads in the TEX-treated sample compared with number of reads in the TEX-untreated sample. The second method aims to take advantage of the information from multiple replicates: we use a Bayesian framework to quantify the probability that a genomic position is overrepresented across a number of TEX-treated samples. The second main criterion that we use to pinpoint reliable TSSs rests on the observation that in bacteria, the majority of genes have a single TSS (Cho *et al.*, 2009). Thus, we expect that in a specific sample, for each transcribed gene, there will typically be one main TSS, as opposed to multiple TSSs in relatively close vicinity. In other words, *bona fide* TSSs should exhibit a 'local enrichment' in reads compared with neighboring genomic positions. We will now describe the computation of different measures of TSS enrichment.

*To whom correspondence should be addressed.

3 METHODS

3.1 Quantifying 5′ enrichment in a TEX-treated compared with a TEX-untreated sample

In preparing the dRNA-seq sample, one captures mRNAs from bacterial cells and sequences their 5′-ends. The capture of the mRNAs could be viewed as a sampling process that gives rise to hypergeometrically distributed counts of reads from individual positions in the genome. However, given that the number of reads originating at a given genomic position is small relative to the total number of obtained reads, we can approximate the hypergeometric distribution by a binomial distribution. That is, if the total number of reads in the sample is N, and the fraction f of these corresponds to a given TSS of interest, then the probability to observe the TSS represented by n of the N reads in the sample follows a binomial distribution:

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

Letting f_+ and f_- denote the frequency of reads derived from a given genomic position in the TEX-treated (TSS-enriched) and TEX-untreated (non-enriched) samples, respectively, what we would like to determine is the enrichment defined as follows:

$$P(f_+ > f_- | n_+, N_+, n_-, N_-) = P(f_+ - f_- > 0 | n_+, N_+, n_-, N_-).$$

We do not know the underlying frequencies f_+ and f_-. Rather, we approximate the probability of enrichment based on observed counts as explained in the Supplementary Material. With x being the observed frequency of reads derived from a given position (i.e. $x_+ = \frac{n_+}{N_+}$ and $x_- = \frac{n_-}{N_-}$ for the TEX-enriched and not enriched samples, respectively), the probability that a genomic position has a higher expression in the TEX-treated compared with the untreated sample is given by the following equation:

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

where ϕ is the cumulative of Gaussian distribution (error function). In case of having multiple paired samples, the average value of $\phi(t)$ for a given genomic position would quantify the 5′ enrichment probability. We call this measure 'z-score'. Alternatively, when we have replicates of paired (TEX-treated and untreated) samples, we can calculate the 5′ enrichment λ_s for each pair separately:

$$\lambda_s = \langle \frac{f_+}{f_-} \rangle$$

Assuming that λ_s follows a normal distribution with mean μ and variance σ^2, we can calculate the probability that a TSS is enriched across a panel of k replicate paired samples:

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

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)$ and μ_* and σ_* are mean and variance of λ, respectively, and k is the number of replicates (details of the derivation are given in the Supplementary Material).

3.2 Quantifying local enrichment

To quantify the local enrichment of a putative TSS, we examine the frequencies of sequenced reads in a region of length 2l centered on the putative TSS ($[x-l, x+l]$). That is, we define the local enrichment L as follows:

$$L = \frac{\sum_{i \in [x-l, x+l], n_{+,i} \le n_{+,x}} n_{+,i}}{\sum_{j \in [x-l, x+l]} n_{+,j}} \quad (1)$$

where $n_{+,i}$ is number of reads derived from position i in the TEX-treated sample. The value of L would be 1 for the position with maximum expression in the interval, corresponding to a perfect local enrichment. When replicates are available, we compute the average local enrichment over these samples. We chose l such that it covers typical 5′ UTR lengths and intergenic regions, i.e. 300 nt. This value is of course somewhat arbitrary, but we found that it allows a good selection of TSSs in practice.

3.3 Identification of TSSs

To identify TSSs, we compute these measures based on all available samples. Because we observed that the precision of start sites is not perfect but there are small variations in the position used to initiate transcription, we also apply single linkage clustering to select the representative among closely spaced (up to 10 nt) TSSs. We then select the parameters that give us the maximum number of annotated genes being associated with TSSs, restricting the total number of predicted TSSs to be in within a narrow range, ±50% of the number of annotated genes in the genome.

4 EVALUATION OF THE TSS IDENTIFICATION METHOD

To evaluate our method and verify its accuracy, we applied it to several recently published datasets [*Helicobacter pylori*, *Salmonella enterica serovar Typhimurium* (Kröger *et al.*, 2012) and *Chlamydia pneumoniae* (Albrecht *et al.*, 2009)] for which a mixture of computational analysis and manual curation was used to annotate TSSs. We here present an in-depth analysis of the TSS identification approaches for *H.pylori*. Similar analyses for the other species are given in the Supplementary Tables S4–S6.

In the *H.pylori* genome, our method identified 2366 TSSs. Of these, 1306 (55%) TSSs are in the reference set of 1893 curated TSSs reported by Sharma *et al.*, 2010, which we refer to them as 'Common' TSSs. Thus, 69% of the curated sites are included in our TSS list. A number of reasons contributed to our method failing to identify another 31% curated TSSs, which we refer to them as 'Reference only'.

- In our approach, we only use reads that were at least 18 nt in length and mapped with at most 10% error to the genome. This selection appears to have led to the loss of 187 (32%) of the 587 curated TSSs in the mapping process, before applying the TSSer inference.
- The majority of the curated sites that we did not retrieve appear to have been supported by a small number of reads. Two hundred twenty-six (38%) of the 587 curated TSSs that we did not identify were supported by less than a single read per 100 000 on average and we required that a TSS is supported by at least 1 read (see Fig. 1a).
- Finally, 174 (30% of the curated TSSs that we did not retrieve) did not pass our enrichment criteria (see Fig.1c). Accepting these TSSs as putative TSSs would have to be accompanied by the inclusion of many false positives.

In summary, 70% of the manually curated TSSs that are not in the 'TSSer' prediction set were not lost due to TSSer scoring but rather before because they had little evidence of expression, even though we mapped 70.43% of the reads to the genome, compared with 80.86% in the original analysis (Sharma *et al.*, 2010). Only 30% of the TSSs that were in the reference list were not

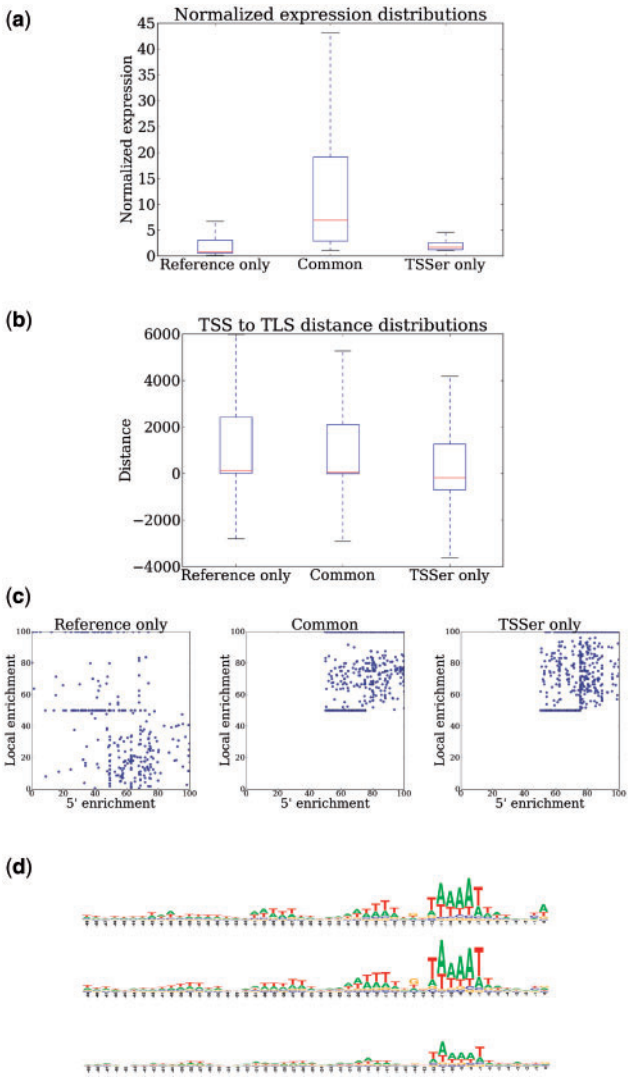

Fig. 1. Properties of TSSs that were present only in the reference list (left), both in the reference and the TSSer list (middle) or only in the TSSer list (right). (**a**) Box plot of averaged normalized expression (the boxes are drawn from the first to the third quantile and the median is shown with the red line). (**b**) Box plot of the displacement distribution relative to the start codon. (**c**) Scatterplots of 5′ versus local enrichment (both shown as percentage). (**d**) Sequence logos indicating the position-dependent (5′ → 3′ direction) frequencies of nucleotides upstream of the TSS (datasets are shown from top to bottom rather than from left to right)

present in the TSSer list because they did not satisfy our criteria for enrichment in reads. Further investigating the features [enrichment values, distance to start codon (TLS) and presence of transcriptional signals (see Supplementary Material)] of these TSSs that we did not identify, we found that a large proportion are likely to be *bona fide* TSSs, i.e. false negatives of our method.

On the other hand, we identified an even larger number of TSSs (1060) that were not present in the curated list. We refer to these as 'TSSer only'. Of these, 198 TSSs correspond to 142 genes that were not present in the reference list. Of the remaining 862 TSSs that are only identified by our method, 287 TSSs are 'Antisense' TSSs, 58 TSSs are 'Orphan' and 379 TSSs are alternative TSSs for genes that did have at least one annotated TSS in the reference set (the definition of these categories is given in Section 2.3 of Supplementary Material). These TSSs share the properties of TSSs jointly identified by our method and the manual curation (Fig. 1), indicating that they are also *bona fide* TSSs. To further support the TSSs that were identified by TSSer and were missing in the reference list, we compared these TSSs with the 'Common' category and also 'Reference only' category in the following aspects:

- Average normalized expression (Fig. 1a): 'TSSer only' TSSs have almost the same expression distribution as TSSs in 'Reference only' category and both have lower expression compared with the TSSs in the 'Common' set. This indicates that TSSs with high expression are equally well identified by the two methods, and that the difference between methods manifests itself at the level of TSSs with low expression.
- TSS to TLS distance: Figure 1b shows that TSSer identifies putative TSSs that are closer, on average, to the translation start, compared with the TSSs that were manually curated. The proportion of internal TSS identified by TSSer is also higher and it remains to be determined what proportion of these represents *bona fide* transcription initiation starts.
- Enrichment values: Figure 1c shows that TSSs identified by TSSer only have strong 5' and local enrichment, whereas those that are present in the 'Reference only' set have low local enrichment. This indicates that these sites are located in neighborhoods that give comparable initiation at spurious sites and thus these sites would be difficult to identify simply based on their expression parameters.
- Strength of transcriptional signals: Figure 1d shows that TSSs identified by TSSer share transcriptional signals such as the −10 box with the other categories of sites. The overall weaker sequence bias may indicate that a larger proportion of 'TSSer only' sites are false positives, consistent with the higher proportion of sites that TSSer identified downstream of start codons (Fig. 1a). To further investigate the transcription regulatory signals, we also implemented a hidden Markov model (HMM) that we trained on the 'Common' sites to find transcription regulatory motifs. We then applied this model to the sequences from each individual subset (see Supplementary Material for details). The results from the HMM further confirm that a large proportion of the 'TSSer only' sites have similar scores to the sites in the other two categories, indicating that TSSer captures a substantial number of *bona fide* TSSs that were not captured during manual curation.

5 DISCUSSION

Deep sequencing has truly revolutionized molecular biology. It enabled not only the assembly of the genomes of thousands of species, but also annotation of transcribed regions in these genomes and the generation of a variety of maps for DNA-binding factors, non-coding RNAs and RNA-binding factors. High-throughput studies revealed that not only eukaryotic but also

prokaryotic genomes are more complex than initially thought. In particular, bacterial genomes encode relatively large numbers of non-coding RNAs with regulatory functions (Waters and Storz, 2009) and antisense transcripts (Georg and Hess, 2011). Such transcripts are of particular interest because they are frequently produced in response to and contribute to the adaptation to specific stimuli (Repoila and Darfeuille, 2009). The availability of a large number of bacterial genomes further enables identification of regulatory elements through comparative genomics-based approaches (Arnold *et al.*, 2012). However, these methods benefit from accurate annotation of TSSs that enables a focused search for transcription factor binding sites. Although the data supporting TSS identification can be obtained with relative ease (Sharma *et al.*, 2010), the annotation of TSSs has so far been carried out manually, which is tedious and likely leads to an incomplete set of TSSs. Only recently, as our manuscript was in the review process, methods for automated annotation of TSSs based on dRNA-seq data started to emerge (Dugar *et al.*, 2013) (see also http://www.tbi.univie.ac.at/newpapers/pdfs/TBI-p-2013-4.pdf). The method that we propose here is meant to provide a starting point into the process of TSS curation. Because it uses dRNA-Seq data, it is clear that only TSSs from which there is active transcription during the experiment can be annotated. As we have determined in the benchmark against the *H.pylori*, there remain TSSs for which the expression evidence is poor, yet have the properties of *bona fide* TSSs. Additional samples, covering conditions in which these TSSs are expected to be expressed are necessary to identify them. Alternatively, they can be brought in during the process of manual curation. Nonetheless, the advantage of an unbiased automated method such as the one we propose here is that it allows the discovery of TSSs that may not be expected or easily evaluated such as those of antisense transcripts, alternative TSSs and TSSs corresponding to novel genes. Furthermore, this method can provide an initial set of high-confidence TSSs that can be used to train more complex models of transcription regulation, which could be used to iteratively identify additional TSSs, that may be supported by a small number of reads. To illustrate this point, we here used an HMM, which we trained on high-confidence TSSs from the 'Common' category, to provide an additional list of putative TSSs that appear to have appropriate transcription regulatory signals but that were not captured with high abundance or enrichment in the experiment (Supplementary Table S8). Thirty-six percent of the TSSs that were only present in the reference annotation are part of this list. More sophisticated versions of this approach could be used toward comprehensive annotation of TSSs in bacterial genomes. Finally, the method can be applied to other systems in which genomic regions give rise to an increased number of transcripts in specific conditions.

6 CONCLUSION

We have proposed an approach for genome-wide identification of TSSs in bacteria, which uses dRNA-Seq data to quantify the 5′ and local enrichment in reads at putative TSSs and their corresponding significance. The method is implemented in an automated pipeline, which we applied to several recently published dRNA-Seq datasets. A thorough benchmarking of the TSSs proposed by our method relative to manual curation indicates that the method performs well in identifying known TSSs and is able to further detect novel TSSs that have the expected properties of *bona fide* TSS. Thus, our method should enable rapid identification of TSSs in bacterial genomes starting from dRNA-Seq data.

ACKNOWLEDGEMENTS

The authors thank A. R. Gruber and A. Rzepiela for critical reading of the manuscript.

Funding: Work in the Zavolan laboratory is supported by the University of Basel and the Swiss National Science Foundation (grant number 31003A_147013).

Conflict of Interest: none declared.

REFERENCES

Albrecht,M. *et al.* (2011) The transcriptional landscape of Chlamydia pneumoniae. *Genome Biol.*, **12**, R98.

Arnold,P. *et al.* (2012) MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, **28**, 487–494.

Cho,B.K. *et al.* (2003) The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.*, **27**, 1043–1049.

Dugar,G. *et al.* (2013) High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet.*, **9**, e1003495.

Georg,J. and Hess,W.R. (2011) cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.*, **75**, 286–300.

Kröger,C. *et al.* (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl Acad. Sci. USA*, **109**, 1277–1286.

Repoila,F. and Darfeuille,F. (2009) Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol. Cell*, **101**, 117–131.

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

Shiraki,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.

Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.