Subject Section

Statistical inference of the rate of RNA polymerase II elongation by total RNA sequencing

Yumi Kawamura 1, Shinsuke Koyama 1,2 and Ryo Yoshida 1,2,*

1Department of Statistical Science, The Graduate University for Advanced Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan and
2Department of Statistical Modeling, The Institute of Statistical Mathematics, Research Organization of Information and Systems, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan.

* To whom correspondence should be addressed.

Abstract

Motivation: Sequencing total RNA without poly-A selection enables us to obtain a transcriptomic profile of nascent RNAs undergoing transcription with co-transcriptional splicing. In general, the RNA-seq reads exhibit a sawtooth pattern in a gene, which is characterized by a monotonically decreasing gradient across introns in the 5' to 3' direction, and by substantially higher levels of RNA-seq reads present in exonic regions. Such patterns result from the process of underlying transcription elongation by RNA polymerase II, which traverses the DNA strand in a 5' to 3' direction as it performs a complex series of mRNA synthesis and processing. Therefore, data of sequenced total RNAs could be utilized to infer the rate of transcription elongation by solving the inverse problem.

Results: Though solving the inverse problem in total RNA-seq has the great potential, statistical methods have not yet been fully developed. We demonstrate what extent the newly developed method can be useful. The objective is to reconstruct the spatial distribution of transcription elongation rates in a gene from a given noisy, sawtooth-like profile. It is necessary to recover the signal source of the elongation rates separately from several types of nuisance factors, such as unobserved modes of co-transcriptionally occurring mRNA splicing, which exert significant influences on the sawtooth shape. The present method was tested using published total RNA-seq data derived from mouse embryonic stem (ES) cells. We investigated the spatial characteristics of the estimated elongation rates, focusing especially on the relation to promoter-proximal pausing of RNA polymerase II, nucleosome occupancy and histone modification patterns.

Availability and implementation: A C implementation of PolSter and sample data are available at https://github.com/yoshida-lab/PolSter

Contact: yoshidar@ism.ac.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Sequenced total RNAs without poly-A selection (total RNA-seq) consist of the pool of nascent transcripts and mature polyadenylated RNAs. RNA polymerase II (Pol II) traverses on the DNA strand from the 5’ to 3’ direction and generates nascent transcripts combined with co-transcriptional splicing (Brown et al., 2012). It has been reported that total RNA-seq exhibits a sawtooth pattern in the read density of a gene (Ameur et al., 2011) as characterized by a monotonically decreasing 5’-3’ slope in intronic regions and substantially higher levels of RNA present in exonic regions (Fig 1). One of the major determinants that influence the observed sawtooth pattern is the rate of transcription elongation by Pol II. For example, the faster Pol II elongation becomes, the steeper the decreasing gradient appears in introns, and vice versa. Hence, it has been argued that the total RNA-seq could potentially be utilized to obtain the relative measures of transcription elongation rates genome-wide (Bentley, 2014; Luco et al., 2011; Singh et al., 2009). However, the use of total
RNA-seq has been less widespread, possibly because of the difficulty in analyzing considerably noisy data with low read coverage.

Several types of experimental technologies have recently emerged for genome-wide measurements of Pol II elongation rates, such as global run-on and sequencing (GRO-seq) (Jonkers et al., 2015), native elongating transcript sequencing (NET-seq) (Churchman and Weissman, 2011), precision run on sequencing (PRO-seq) (Kwak et al., 2013), nascent RNA sequencing (Nascent-seq) (Rodriguez et al., 2012), and metabolic labeling of nascent RNA using microarrays (Radle et al., 2013). The objective common to these methods is to deeply sequence RNAs at the binding sites of transcriptionally active Pol II running on DNA strands in cells. Typically, elongation rates are measured by tracking a wave front of transcriptionally active Pol II traversing 5'→3' over time. The observed travelling distance of the wave fronts between two consecutive time points is used to calculate the velocity. Such methods operate with intractable drug-driven interventions to induce the Pol II wave, such as manipulations for halting and restarting transcriptions. Furthermore, the time progresses of induced waves are visually undistinguishable and often infeasible to track for most genes as will be shown later. In addition, the spatial resolution of observable elongation rates is dependent on the length of the time interval. It is difficult to acquire high frequency time course data because of intractability in the protocols of such nascent transcript sequencing.

Well-established total RNA sequencing has great promise as a tool to elucidate genome-wide transcription elongation rates. We focused on the use of total RNA-seq. The proposed method relies on a state space representation that describes a mathematical relationship between the observed read density and the spatially varying elongation rates. A prior distribution is placed on the elongation rates and splicing patterns, then followed by Bayesian inference by performing sequential Monte Carlo calculations (Bolíć et al., 2004). The data capture the pool of different kinds of source signals associated with spatial dynamics on elongation rates and co-transcriptionally occurring mRNAs splicing such as exon skipping, intron retention, recursive splicing (RS) (Duff et al., 2015; Sibley et al., 2015), and so on. The problem is a kind of blind source separation in which unobserved splicing patterns influence the observed sawtooth as a secondary signal to be decoded, and the data contain a considerably high level of noise because of the low read depth, especially in short introns. We have also investigated some important characteristics of the data and described the advantages and disadvantages over GRO-seq. We explored the Pol II elongation rates in 659 genes in mouse ES cells (Sigova et al., 2013). The estimated elongation rates were compared with some epigenetic observations on nucleosome occupancy and histone modification patterns in mouse ES cells that have been reported in different studies (Teif et al., 2012; Maeson et al., 2008; Creyghton et al., 2010). We found that position-specific variations in the elongation rates agree to some extent with the observed epigenetic landscape.

2 Methods
2.1 Sawtooth observation in total RNA-seq

Transcription elongation is coupled to splicing. In the process of Pol II running through a gene from the 5' to 3' end, a nascent transcript gets elongated successively and an intron is removed, conventionally when the Pol II reaches the 3' end of the intron. In addition to mature mRNAs, there exist in cells nascent transcripts at different stages of the elongation process coupled with co-transcriptional splicing. It was first found by Amour et al., 2011 that a sawtooth shape appears in the read density since the sequenced reads capture the pool of mature and immature RNAs in the cells as schematically shown in Fig. 1.

Let \( z(t) \) be the probability of existence of Pol II instantly occurring at nucleotide position \( t \) on the DNA strand \( t = 1, \ldots, T \). The 5' and 3' ends of the gene correspond to \( t = 1 \) and \( t = T \), respectively. The existence probability is inversely proportional to the elongation rate \( v(t) \). The 5th nucleotide is spliced out when Pol II reaches the position \( s(t) \leq u(t) \leq t \). Then, the expected read density \( r(t) \) is expressed by the integral of \( z(t) \) over the interval between its transcribed position \( t \) and the splice site \( s(t) \):

\[
r(t) = \int_{t}^{s(t)} z(u) du. \tag{1}
\]

The conversion between the read density \( r(t) \) and the Pol II density \( z(t) \) could be carried out by taking the integral or differentiation. If the splicing mode is conventional, that is, all exons are retained in the final product and introns are removed when Pol II reaches the 3' end, the expected read density becomes

\[
r(t) = \begin{cases} z(t) & t \in I_k \\ \frac{T_k(t)}{N_k} z(u) du & t \in E_k \end{cases}
\]

where \( I_k \) and \( E_k \) denote sets of nucleotide positions for the \( k \)th intron and the 5th exon, respectively, and \( T_k(t) \) denotes the 3' end in \( I_k \). It is assumed that, for each gene, \( K \) exons and \( K - 1 \) introns are arranged as \( E_1, E_2, \ldots, E_{K-1}, E_K \) from the 5' to 3' direction.

In this case, the sawtooth pattern has the following characteristics.

- **Non-monotonic increasing gradient in an intron**: \( \forall t \geq s\) and \( (t, s) \in I_k \times E_k \), \( r(t) \leq r(s) \).
- **Non-monotonic increasing gradient in exons**: \( \forall t \geq s\) and \( (t, s) \in E_k \times E_0 \), such that \( k \leq h, r(t) \leq r(s) \).
- **Higher read density in an exon than in subsequent introns**: \( \forall t \geq s\) and \( (t, s) \in E_k \times I_k \), such that \( k \leq h, r(t) \geq r(s) \).

These characteristics are retained only for the given splicing mode, but the statements imply an important feature of the data: shorter introns or exons closer to the 3' end of a gene exhibit lower read counts. As shown later, read depths indeed correlate negatively with intron lengths, and sawtooth patterns become less clear in shorter introns because of the lack of sufficient amounts of reads. In other words, the inference of elongation rates is feasible only to a small subset of longer genes without performing deep sequencing.

2.2 State space representation

Each intron was divided into bins with intervals equal to 400 bp. An exonic region was treated positionally as a single point. Accordingly, the Pol II density is discretized into the corresponding \( N \) grid points as \( \{x_n|n = 1, \ldots, N\} \), and the read counts were averaged within each range giving the dataset \( \{y_n|n = 1, \ldots, N\} \). It is assumed here that \( n = 1 \) and \( n = N \) denote the 5' and 3' ends of a gene, respectively. The state variables to be inferred from the data comprise the Pol II existence probability \( \{x_n|n = 1, \ldots, N\} \) and the splice site \( s_n (\geq n) \) of the \( n \)th position in a transcribed RNA. The grid points \( \{1, \ldots, N\} \) consist of \( K \) exonic regions, \( E_1, \ldots, E_K \), and \( K - 1 \) introns, \( I_1, \ldots, I_{K-1} \). Note that, by definition, the first and last exonic regions become \( E_1 = \{1\} \) and \( E_K = \{N\} \). The 5' and 3' ends of a reduced intronic region \( I_k \) are denoted by \( S(I_k) \) and \( T(I_k) \), respectively.
The state space representation is then
\[
\log y_n = \log r_n + \eta_n, \quad \eta_n \sim N(\mu, \sigma),
\]
\[
x_n = \sum_{i=1}^n x_i,
\]
\[
\log x_n = \log x_{n+1} + \nu_n, \quad \nu_n \sim N(0, \gamma),
\]
\[
s_n = p(s_n|s_{n+1}, s_{n+2}, \ldots, s_k),
\]  
with the initial distributions on the state variables, \( \log y_0 \sim N(\mu_0, \sigma_0) \) and \( s_0 = N \). As in the first equation, referred to as the measurement model, the read count is subject to the expected read count \( r_n \) corrupted by the multiplicative measurement noise \( \eta_n \), of the log-normal with mean \( \mu \) and variance \( \sigma \). In the second line, the expected read count is represented by the sum of the PoI II existence probabilities over the interval between \( n \) and \( s_n \), which corresponds to a discretization of the integral in Eq. 1. The last two equations, referred to as the system model, describe the state transition processes; a first-order random walk is imposed on the transition of \( x_n \) to induce spatially smooth estimates on the PoI II existence probabilities. The splice sites following the conditional distribution will be detailed in the next subsection. Note that the PoI II existence probabilities and the splice sites \( s_n \) and the configured preceding sequence \( s_0 \) are removed out from the transcribed RNA. The difficulty is to avoid the occurrence of infeasible splicing patterns during the random generation.

As illustrated in Fig 2, we modeled three models of splicing events: (i) exon skipping, (ii) intron retention, and (iii) RS of an intron. The occurrence of alternative donor/acceptor sites is not taken into consideration because of the reduction of exonic regions into single points. RS is a stepwise removal process of an intron that has more often been observed in exceptionally long introns (Sibley et al., 2015). The occurrence of splicing in the middle of an intron brings a valley in the sawtooth shape of total RNA-seq reads at the RS site (Duff et al., 2015; Sibley et al., 2015). Deviation from the monotonic decreasing gradient in the RNA-seq density of an intron could be indicative of RS. As reported in previous studies, there are also a large number of apparent RS sites in the data that we analyzed as shown in Fig 3.

2.3 Prior distribution of unknown splice variants

One difficulty of the inverse problem lies in the fact that splicing variations cause significant deviations from the expected sawtooth pattern as previously shown. Hence, it is essential to infer the splicing patterns simultaneously with the elongation rate through analysis of a given read density. The prior distribution \( p(s_n|s_{n+1}, s_{n+2}, \ldots, s_k) \) is used in the SMC calculation to sequentially produce unknown splicing sites for which the sites \( n \) are removed out from the transcribed RNA. The difficulty is to avoid the occurrence of infeasible splicing patterns during the random generation.

As illustrated in Fig 2, we modeled three modes of splicing events: (i) exon skipping, (ii) intron retention, and (iii) RS of an intron. The occurrence of alternative donor/acceptor sites is not taken into consideration because of the reduction of exonic regions into single points. RS is a stepwise removal process of an intron that has more often been observed in exceptionally long introns (Sibley et al., 2015). The occurrence of splicing in the middle of an intron brings a valley in the sawtooth shape of total RNA-seq reads at the RS site (Duff et al., 2015; Sibley et al., 2015). Deviation from the monotonic decreasing gradient in the RNA-seq density of an intron could be indicative of RS. As reported in previous studies, there are also a large number of apparent RS sites in the data that we analyzed as shown in Fig 3.

The prior distribution describes the dependence of the splicing site \( s_n \) of the position \( n \) on the preceding ones, \( s_{n+1}, s_{n+2}, \ldots, s_k \). Adjacent \( s_n \) and \( s_{n+1} \) in the same intron should be more likely to take the same value; for example, they would be the 3’ end of the intron, conventionally. However, if the ith position is an RS site, it then holds that \( s_n = n \) while the neighboring \( s_{n+1} \) turns out to be the 3’ end of the intron with high probability. On the other hand, \( s_n \) for an exonic region tends to take the 3’ end of the gene if no skipping occurs, but the intronic \( s_{n+1} \) is likely to be the 3’ end of the intron. In this way, a sequence \( s_0 \) is not smoothly evolved, and the prior probability of \( s_0 \) should be dependent on whether or not \( n \) is an exon or an intron as well as the configuration of \( s_{n+1}, \ldots, s_k \).

The procedure for successively constructing such a sequence is summarized in quasi-code Algorithm 1. Several generators are switched into the active or inactive mode according to the if statements that classify the current position \( n \) and the configured preceding sequence \( s_{n+1}, \ldots, s_k \) into several conditions. This classification is employed to exclude the emergence of unlikely occurring splice variants as illustrated in Fig 2. For example, consider that a gene consists of \( E_1, E_2, E_3 \) with the three exons \( E_k \ (k = 1, 2, 3) \) and the two introns \( I_k \ (k = 1, 2) \). Conventionally, when the second exon \( E_2 \) is skipped out, it temporally forms with the previous and next introns, \( I_1 \) and \( I_2 \), a nascent dangling from the DNA strand, and they are removed out together at the same time, possibly when the 3’ end of \( I_2 \) is transcribed and isolated. This splicing mode is represented as \( E_1 (E_2 I_2 E_3) \), where the unit in the parentheses is isolated simultaneously. On the other hand, \( E_1 (I_1 E_2 I_2 E_3) \) would be unlikely to occur. This mode describes a nascent transcript comprised of \( E_1 E_2 I_2 \) dangling from the DNA.
strand temporarily, and its subunit $E_2 I_2$ is removed while only $E_1$ is retained in the transcript when Pol II reaches the 3' end of $I_2$. Such an unrealistic splicing mode should not be allowed to emerge. Meanwhile, $(E_1 I_1)(E_2 I_2)E_3$ could realistically happen as the first exon is spliced out together with the first intron, and then a nascent transcript consisting of the second exon and the second intron disappears simultaneously. Consequently, our generator follows the statements shown below:

- **Rule 1.** Let $s_n$ be a splice site of the exonic nucleotide in $E_i$, and then $s_{n-1}$ and $s_{n+1}$ be its nearest neighbors in the 5' and 3' directions, respectively. If $s_{n-1} = s_{n+1}$ but $s_n \neq s_{n+1}$, all upstream exonic nucleotides closer to the 5' end, i.e., any $m \leq E_i$, satisfy $s_{m+1} = s_n$.

- **Rule 2.** Whenever being skipped out, the exonic nucleotide $n = E_i$ is removed together with the neighboring intronic nucleotide (i.e., $s_{n-1} = s_{n+1}$) or the most surviving exon $s_n = s_k$ where $s_k = \min\{s_m|m \in E_{k+1}, \ldots, E_N\}$.

### 2.4 Hyperparameters

For each gene, the hyperparameters on the log-normal measurement noise, $\mu$ and $\sigma$, were determined as follows: (i) a smoothing spline $f(n)$ was fitted to the logarithmically transformed read counts, which provides an initial guess on the expected reads, i.e., $\log \hat{r}_n = \log \sum_{m=0}^{n} r_i$ (see the measurement equation in Eq. 2), and then (ii) the mean and the variance of the residuals were given to $\mu$ and $\sigma$, respectively. Using the estimated expected reads, we could derive the estimates on the state variables as $x_{n+1} = \exp f(n) - \exp f(n+1)$ ($n = 1, \ldots, N - 1$). The variance of the first-order differences $log(x_{n+1} - log(x_n))$ ($n = 1, \ldots, N - 1$) was given to $\eta$, and the mean of $x_{n+1}$ was given to $\mu_0$.

### 2.5 Total RNA-seq data

Total RNA-seq that we used was derived from mouse ES cells (Sigova et al., 2013). As already discussed, the RNA-seq reads were considerably sparse, especially in shorter genes, hence we began by selecting genes analyzable. The objective was to identify introns in which almost monotonically decreasing slopes were observed in the 5'–3' direction. To assess the monotonicity of an intron, we used Pearson's correlation coefficients between intronic read counts and their positions. Supplementary Materials F1 shows the relationship between the lengths of introns and the correlation coefficients. We then selected introns with lengths ≥ 5000 bp and with correlation coefficients ≥ 0.5, providing 665 genes that contain one or more such selected introns.
3 Results

For each gene, we calculated the Pol II density, the splicing sites, and the expected reads by taking the averages of 10^5 particles generated from the posterior distribution, which could be summarized with known splice variants as in Fig 4. The reconstructed elongation rates of the 653 genes are displayed by a heatmap in Fig 5.

First, we compared the estimated Pol II densities and two ChIP-seq profiles of Pol II (GSM1865697, GSM1865698), which were generated from mouse ES cells in a different study (Flynn et al., 2016). As shown in Fig 6(iv), the Pol II densities obtained by the different experimental methods exhibited a significantly strong correlation; the number of genes exhibiting significant positive correlations was nearly 11 times larger than significantly negative genes at the 5% significance level (Supplementary Materials F3 (iii)).

Next, we investigated the spatial features of the transcription elongation rates in neighboring regions of the transcription start sites (TSSs) as shown in Supplementary Materials F2 (i). The averaged elongation rates in 0-3 kb and 3-6 kb downstream from the TSSs were compared. Nearly 1.75-fold slower elongation was observed in the TSS adjacent regions than in the downstream regions. This is due to a widely known fact, i.e. the promoter-proximal pausing of Pol II at ~30-50 bp downstream of the TSS, which is mediated by negative elongation factors (Jonkers et al., 2015). In addition, as shown in Supplementary Materials F2 (ii), a comparison of the average elongation rates between exons and introns strongly suggests that Pol II slows down significantly at exons, presumably to facilitate splicing (Brown et al., 2012; Tanny et al., 2014). On the other hand, a lack of correlation was observed between the estimated Pol II densities and GC content in the DNA sequences (Supplementary Materials F2 (iii)), though several studies suggest that GC-richer sequences negatively influence elongation rates (Jonkers et al., 2014).

The effects of nucleosome occupancy and histone modification on elongation rates were investigated by assessing the correlation between the estimated Pol II densities and epigenetic-level profiles derived from mouse ES cells in independent studies (Teif et al., 2012; Marson et al., 2008; Creyghton et al., 2010). Pearson’s correlation coefficients were evaluated with respect to the nucleosome occupancies observed through MNase-seq from mouse ES cells (GSE40910: GSM1004652), neural progenitor cells (NPCs) derived from these ES cells (GSE40910: GSM1004653), and mouse embryonic fibroblasts (MEFs) from the corresponding mouse strain (GSE40910: GSM1004654) (Teif et al., 2012). Nucleosomes form barriers against Pol II elongation as nucleosome-depleted regions become more accessible by Pol II (Teves et al., 2014). Indeed, the correlation coefficients indicated negative relationships between the estimated Pol II densities and the nucleosome positioning patterns (Kulaeva et al., 2013) in many genes (Fig 6(ii) and Supplementary Materials F3 (ii)).

For the association with histone modification patterns, we used the ChIP-seq profiles of histone modifiers involved in epigenetic silencing (histone H3 lysine 79 di-methylation (H3K79me2)) and activation (histone H3 lysine 4 tri-methylation (H3K4me3), histone H3 lysine 4 mono-methylation (H3K4me1), histone H3 lysine 36 tri-methylation (H3K36me3), histone H3 lysine 27 acetylation (H3K27ac)) (GSE11724, GSE24165) (Marson et al., 2008; Creyghton et al., 2010). For many genes, the estimated Pol II densities seem to be positively related to the histone modification marks associated with transcriptional activation (Fig 6(i) and Supplementary Materials F3 (iii)). The number of genes exhibiting significant positive correlations was more than eight times larger than those with negative correlations at the 5% significance level. On the other hand, the histone modification patterns of the silencer groups tend to correlate negatively with the Pol II densities within the gene bodies (Fig 6(i) and Supplementary Materials F3 (iii)). The number of genes exhibiting statistically significant negative correlations was nearly 1.5 times larger than those with positive correlations. Even though these epigenetic data are derived from different laboratories, we found that the estimated Pol II densities are highly consistent in pattern with the observed epigenetic landscape.
5. The estimated elongation rates of the 653 genes are arranged on the vertical axis. The horizontal axis denotes the relative position from TSS. The color scale chart shown on the side denotes the estimated values normalized to [0, 1].

In addition, the estimated Pol II densities were investigated in relation to computationally annotated chromatin states. We used 15 annotations of chromatin states (Shen et al., 2012), which were obtained by performing a Poisson-based multivariate hidden Markov model (ChromHMM) (Ernst and Kellis, 2012) on 7 ChIP-seq profiles of H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K27ac, the insulator-binding protein CCCTC-binding factor (CTCF) and Pol II in mouse ES cells (GSE29184). We then compared the averages of the estimated Pol II densities in regions with and without a given annotation. As shown in Fig 6(iii), it was found that some chromatin states, for example, ‘active promoter’ tend to show significant associations with high-density regions of Pol II in most genes. The estimated elongation rates of the 653 genes were compared to those estimated based on GRO-seq (Jonkers et al., 2015; Hah et al., 2011). Using a hidden Markov model with the groHMM package (Chae et al., 2015; Danko et al., 2015) of R language, we tracked the wave fronts of Pol II progression at 5, 12.5, 25, and 50 min after the release from the paused state of Pol II. The elongation rate was calculated by the moving distance of the adjacent wave fronts per minute. The Pol II densities obtained by our method were summed for each interval of the identified wave fronts at two consecutive times, and the relative elongation rate on each of the five intervals was calculated by dividing the inverse of the summed Pol II densities by the respective moving distance. Then, the correlation coefficients were calculated for each gene, showing a lack of agreement between the different estimates of elongation rates with total RNA-seq and GRO-seq (Supplementary Materials F4). This inconsistency likely arises from the difficulty of identifying the induction waves of elongating Pol II with GRO-seq. As exemplified in Supplementary Materials F5, it was quite hard in many genes even to recognize visually exact positions on the wave fronts of elongating Pol II. While induction waves should progress in time monotonically from 5’ to 3’, the tracked positions could take place in the reverse order across time points.

4 Discussion

In this study, we implemented a Bayesian framework for the reconstruction of transcription elongation rates from sawtooth-like observations derived from total RNA-seq. After forwardly modeling the given sequenced RNA-seq reads for unknown rates of elongating Pol II and unknown modes of splicing, the backward prediction was performed according to Bayes’law to inversely predict the unknowns. As a proof of principle, we tested our approach on the total RNA-seq data derived from mouse ES cells. We identified some spatial features of elongation rates such as the slowdown of transcription at exons and promoter-proximal regions. In addition, the predicted elongation rates were highly consistent spatially with epigenetic observations, i.e. nucleosome positioning and histone methylation, even though the data were acquired in different studies.

Despite the potentially great promise of utilizing total RNA-seq to study transcription elongation, there has been considerable less progress made in statistical methods. In some previous studies, the slope of the read density gradients, for instance, which is obtained using linear regression, was used as the relative elongation speed. However, as described in this study, different splicing modes could bring different slopes to the read density, thereby drawing the wrong conclusion in the absence of inferring the splicing variations. One contribution of this study is to provide a way to estimate unmeasured states of elongation rates and splicing modes simultaneously.

As a by-product of our method, the RS sites could be identified. Although details were not described, quite a lot of valleys, possibly indicating ratchet points of RS, were found in the intronic regions in addition to those shown in Supplementary Materials F6. For example, the luna gene in Drosophila melanogaster is known to contain a 108-kb intron with five ratchet points, such that the intron is removed in six stepwise RS events (Duff et al., 2015). As shown in Supplementary Materials F6, the splicing sites estimated by our method captured the five ratchet points reported in the previous study, though some seemingly false estimates of the splicing sites were also given.

This study focused only 653 genes since intronic reads were considerably sparse in most other genes. Supplementary Materials F7 shows an example of such data in which RNA-seq reads covered only 6.47% of the entire region. One difficulty is the infeasibility of inferring splicing sites from such data. The current method is applicable only for long introns. In our perspective, the currently achieved estimation accuracy might decline substantially for shorter introns, even for the selected 656 genes, where read coverages tend to be low. By performing deeper sequencing, a genome-wide elongation rate distribution is potentially predictable with the well-established RNA-seq protocol.

Funding

This work was supported by JSPS KAKENHI Grant Number JP15K12145.

References


Fig. 6. Correlation coefficients between the estimated Pol II densities and (i) ChIP-seq profiles of histone modifiers, (ii) nucleosome occupancies observed by MNase-seq from mouse ES cells. (iii) Differences between the averages of the estimated Pol II densities in regions with and without a chromatin state annotation. The 15 annotations shown in the right panel were obtained by performing ChromHMM on the ChIP-seq profiles of the histone modifiers. (iv) Correlation coefficients between the estimated Pol II densities and two ChIP-seq profiles of Pol II. The color scale charts shown on the sides denote the given values in which the mean differences shown in (iii) are scaled to \([-1, 1]\).


