Bioinformatics, 35(11), 2019, 1877–1884
doi: 10.1093/bioinformatics/bty886
Advance Access Publication Date: 30 October 2018
Original Paper

OXFORD

Gene expression

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

Yumi Kawamura[1], Shinsuke Koyama[1,2] and Ryo Yoshida[1,3,*]

[1]Department of Statistical Science, The Graduate University for Advanced Studies (SOKENDAI), Tachikawa 190-8562, Japan, [2]Department of Statistical Modeling, The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa 190-8562, Japan and [3]Department of Statistical Data Science, The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa 190-8562, Japan

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on March 9, 2018; revised on August 3, 2018; editorial decision on October 6, 2018; accepted on October 29, 2018

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′–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′–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 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

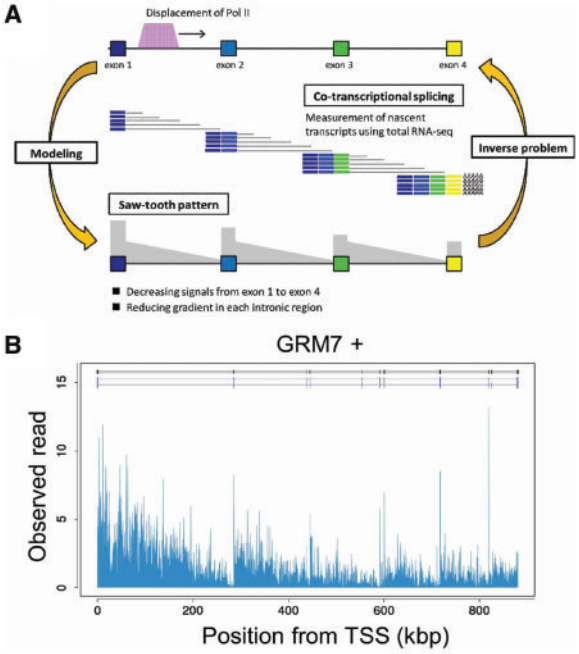

Fig. 1. Inverse problem of the transcription elongation rate. (**A**) Total RNA-seq captures a mixture of matured and nascent transcripts in a pool of cells. During the displacement of Pol II from 5′ to 3′, elongating and co-transcriptionally spliced RNAs can take various states as shown in the middle. The sawtooth pattern of sequenced RNA-seq reads shown in the bottom results from the expected frequency of nucleotides included in those transcripts at various stages. This figure was created by referring to Figure 2 of Ameur *et al.* (2011). (**B**) Total RNA-seq reads of a gene (GRM7) in human fetal brain (Ameur *et al.*, 2011). Splice variants reported in hg19, GRCh37 (Genome Reference Consortium Human Reference 37) are shown in the upper side

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 and Padgett, 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 and Lis, 2015), native elongating transcript sequencing (Churchman and Weissman, 2011), precision run-on sequencing (Kwak *et al.*, 2013), nascent RNA sequencing (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 traveling 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 progressions 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 (SMC) calculations (Bolić *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 mRNA 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 decoupled, 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 embryonic stem (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 (Creyghton *et al.*, 2010; Marson *et al.*, 2008; Teif *et al.*, 2012). We found that position-specific variations in the elongation rates agree to some extent with the observed epigenetic landscape.

2 Materials and 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 Ameur *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 Figure 1.

Let $x(t)$ be the probability of existence of Pol II instantly occurring at nucleotide position t on the DNA strand $t \in \{1, \ldots, T\}$. 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 $v(t) \propto 1/x(t)$. The tth nucleotide is spliced out when Pol II reaches the position $s(t)(t \leq s(t) \leq T)$. Then, the expected read density $r(t)$ is expressed by the integral of $x(t)$ over the interval between its transcribed position t and the splice site $s(t)$:

$$r(t) = \int_t^{s(t)} x(u)\mathrm{d}u. \tag{1}$$

The conversion between the read density $r(t)$ and the Pol II density $x(t)$ could be carried out by taking the integral or differentiation.

If the splicing mode is conventional, i.e. all exons are retained in the final product and introns are removed when Pol II reaches the 3' ends, the expected read density becomes

$$r(t) = \begin{cases} \int_t^{T(I_k)} x(u)\mathrm{d}u & t \in I_k \\ \int_t^{T} x(u)\mathrm{d}u & t \in E_k \end{cases}$$

where I_k and E_k denote sets of nucleotide positions for the kth intron and the kth exon, respectively, and $T(I_k)$ denotes the 3' end in I_k.

It is assumed that, for each gene, K exons and K–1 introns are arranged as $E_1 I_1 E_2 I_2 \ldots I_{K-1} E_K$ 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 $(t,s) \in I_k \times I_k$, $r(t) \leq r(s)$.
- *Non-monotonic increasing gradient in exons:* $\forall t \geq s$ and $(t,s) \in E_k \times E_h$ 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 $(t,s) \in E_k \times I_h$ 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 =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 nth 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

$$\begin{aligned} \log y_n &= \log r_n + \eta_n,\ \eta_n \sim N(\mu, \sigma), \\ r_n &= \sum_{i=n}^{s_n} x_n, \\ \log x_n &= \log x_{n+1} + \nu_n,\ \nu_n \sim N(0, \gamma), \\ s_n &\sim p(s_n | s_{n+1}, s_{n+2}, \ldots, s_N), \end{aligned} \tag{2}$$

with the initial distributions on the state variables, $\log x_N \sim N(\mu_0, \tau_0)$ and $s_N = N$. 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 η_n of the log-normal with mean μ and variance σ. In the second line, the expected read count is represented by the sum of the Pol II existence probabilities over the interval between n and s_n, which corresponds to a discretization of the integral in Equation (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 Pol II existence probabilities. The splice sites following the conditional distribution will be detailed in the next subsection. Note that the Pol II existence probabilities and the splice sites are sequentially generated in the 3'–5' direction ($n = N, N-1, \ldots, 1$) since the expected read r_n at the nth position could be calculated with the given $\{x_n, x_{n+1}, \ldots, x_N\}$ and $\{s_n, s_{n+1}, \ldots, s_N\}$.

The estimated values of x_n and s_n are calculated through a SMC method that draws a set of samples from the posterior distribution $(X, S) \sim p(X, S | Y)$ to derive estimates such as the posterior mean. A class of SMC methods provides rather easy-to-implement algorithms to produce Monte Carlo samples from analytically intractable posteriors. The standard reference is (Doucet and Johansen, 2011). The methods share a common algorithmic structure with genetic algorithms. The system model in Equation (2) is used to generate samples of (x_n, s_n) with given history, $\{x_{n+1}, \ldots, x_N\}$ and $\{s_{n+1}, \ldots, s_N\}$. Fitness scores of the generated samples are assessed based on the measurement model with respect to given y_n. Samples having better fitness have a better chance at surviving in the next generation. This process keeps on iterating from N to 1 and at the end, samples from the targeted posterior will be produced. The algorithmic details are shown in Supplementary Material M1.

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 $p(s_n | s_{n+1}, s_{n+2}, \ldots, s_N)$ 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 Figure 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 Figure 3.

The prior distribution describes the dependence of the splicing site s_n of the position n on the preceding ones, $s_{n+1}, s_{n+2}, \ldots, s_N$.

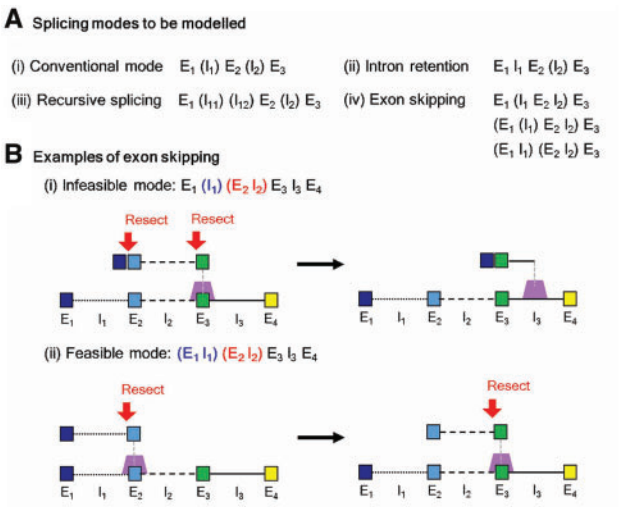

Fig. 2. (**A**) Four splicing modes to be modeled in the system with illustrative examples: (i) conventional mode, (ii) intron retention, (iii) RS of introns and (iv) exon skipping. (**B**) Infeasible and feasible modes of exon skipping are exemplified in (i) and (ii), respectively

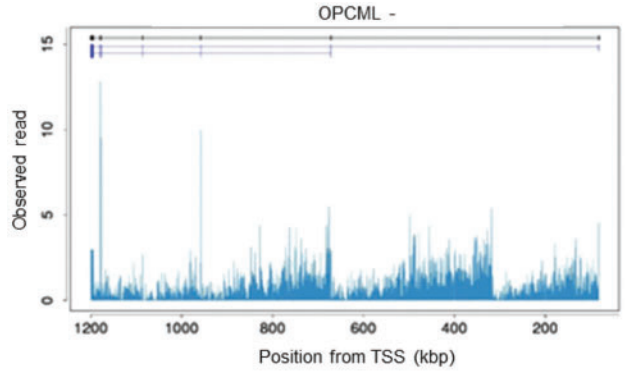

Fig. 3. Read density of the OPCML gene in human fetal brain (Ameur *et al.*, 2011). The observed valley in the intron implies the occurrence of RS

Adjacent s_n and s_{n+1} in the same intron should be more likely to take the same value; e.g. they would be the 3' end of the intron, conventionally. However, if the nth 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 $\{s_1, \ldots, s_N\}$ is not smoothly evolved, and the prior probability of s_n should be dependent on whether or not n is an exon or an intron as well as the configuration of $s_{n+1}, \ldots, s_N$.

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 $s_{n+1}, \ldots, s_N$ into several conditions. This classification is employed to exclude the emergence of unlikely occurring splice variants as illustrated in Figure 2. For example, consider that a gene consists of $E_1 I_1 E_2 I_2 E_3$ 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 transcript 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 $E_1(I_1 E_2 I_2)E_3$, 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_k, and then s_{n-1} and s_{n+1} be its nearest neighbors in the 5' and 3' directions, respectively. If $s_n = s_{n+1}$ but $s_n \neq s_{n-1}$, all upstream exonic nucleotides closer to the 5' end, i.e. any $m \in E_h \; \forall h \; <= k$, satisfy $s_m \leq s_n$.
- Rule 2. Whenever being skipped out, the exonic nucleotide $n \in E_k$ is removed together with the neighboring intronic nucleotide (i.e. $s_n = s_{n+1}$) or the most surviving exon $s_n = s_*$ where $s_* = \min\{s_m | m \in E_{k+1}, \ldots, E_K\}$.

Algorithm 1 Generator for splice sites $p(s_n | s_{n+1}, \ldots, s_N)$

Input: $s_{n+1}, \ldots, N, \alpha, \beta, \delta, \epsilon, \phi_{\text{thr}}$
Output: s_n, ϕ_{thr}
$t_1 \ldots t_p = \text{unique.intron}(s_{n+1}, \ldots, s_N)$ (# get unique values from the given splice sites of only intronic regions)
Remove from $\{t_1 \ldots t_p\}$ N and those less than ϕ_{thr}, and then we have $u_1 \ldots u_q$.
if $n \in \{T(I_1), \ldots, T(I_{K-1})\}$ **then** (# 3' end of the intron)

$$s_n = \begin{cases} n & \text{with probability } \alpha \\ s_{n+1} & \text{otherwise} \end{cases}$$

if $s_n \neq s_{n+1}$ and $s_{n+1} \in \{u_1 \ldots u_q\}$ **then**

$\phi_{\text{thr}} = n$

end if
end if
if $n \in I_1 \backslash \{t(I_1)\} \cup \ldots \cup I_{K-1} \backslash \{t(I_{K-1})\}$ **then** (# intronic region other than the 3' end)

$$s_n = \begin{cases} n & \text{with probability } \beta \text{ (RS)} \\ s_{n+1} & \text{otherwise} \end{cases}$$

end if
if $n \in E_1 \cup E_2 \ldots \cup E_{K-1}$ **then**

$$s_n = \begin{cases} N & \text{with probability } \delta \\ s_{n+1} & \text{with probability } (1-\delta)\epsilon \\ u_i & \text{with probability } (1-\delta)(1-\epsilon)/q \text{ for } i = 1, \ldots, q \end{cases}$$

end if

2.4 Hyperparameters

For each gene, the hyperparameters on the log-normal measurement noise, μ and σ, 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. $\log r_n = \log \sum_{i=n}^{s_n} x_i$ [see the measurement equation in Equation (2)], and then (ii) the mean and the variance of the residuals were given to μ and σ, respectively. Using the estimated expected reads, we could derive the estimates on the state variables as $x_n = \exp f(n) - \exp f(n+1)$ $(n = 1, \ldots, N-1)$. The variance of the first-order differences $\log x_n - \log x_{n+1}$ $(n = 1, \ldots, N-1)$ was given to η, and the mean of x_n was given to μ_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 Material F1 shows the relationship between the lengths of introns and the correlation coefficients. We then selected introns with lengths $\geq$5000bp and with correlation coefficients $\geq$0.5, providing 653 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 Figure 4. The reconstructed elongation rates of the 653 genes are displayed by a heatmap in Figure 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 Figure 6D, 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 Material 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 Material 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 and Lis, 2015). In addition, as shown in Supplementary Material 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, 2014)<AQ12/>. On the other hand, a lack of correlation was observed between the estimated Pol II densities and GC content in the DNA sequences [Supplementary Material 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 (Creyghton *et al.*, 2010; Marson *et al.*, 2008; Teif *et al.*, 2012). 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 derived from these ES cells (GSE40910: GSM1004653) and mouse embryonic fibroblasts 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. 6B and Supplementary Material 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 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) (Creyghton *et al.*, 2010; Marson *et al.*, 2008). For many genes, the estimated Pol II densities seem to be positively related to the histone modification marks associated with transcriptional activation [Fig. 6A and Supplementary Material F3(i)]. 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. 6A and Supplementary Material F3(i)]. 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.

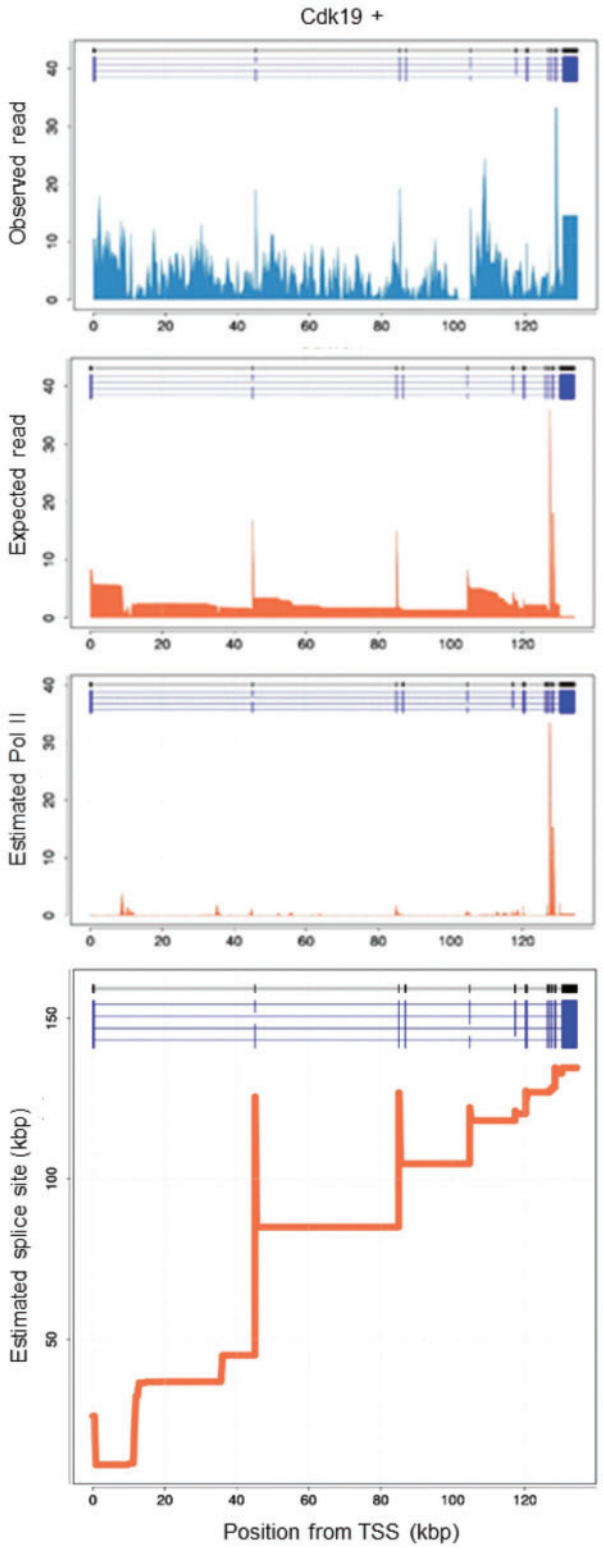

Fig. 4. Estimated Pol II density, expected read density and splicing patterns are shown on the DNA coordinates of the Cdk19 gene in the 5′–3′ direction. The observed read counts are shown in the top panel

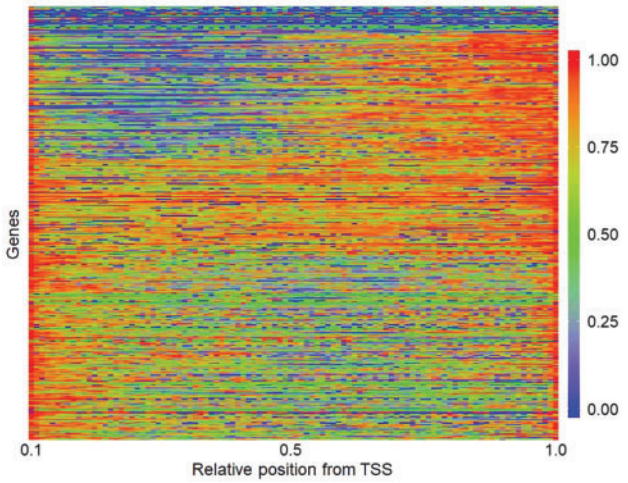

Fig. 5. The estimated elongation rates of the 653 genes are arranged on the vertical axis. The horizontal axis denotes the relative position from TSS. The color scale chart shown on the side denotes the estimated values normalized to $[0, 1]$

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 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 Figure 6C, it was found that some chromatin states, e.g. '*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 (Hah *et al.*, 2011; Jonkers and Lis, 2015). Using a hidden Markov model with the groHMM package (Chae *et al.*, 2015; Danko *et al.*, 2013) 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

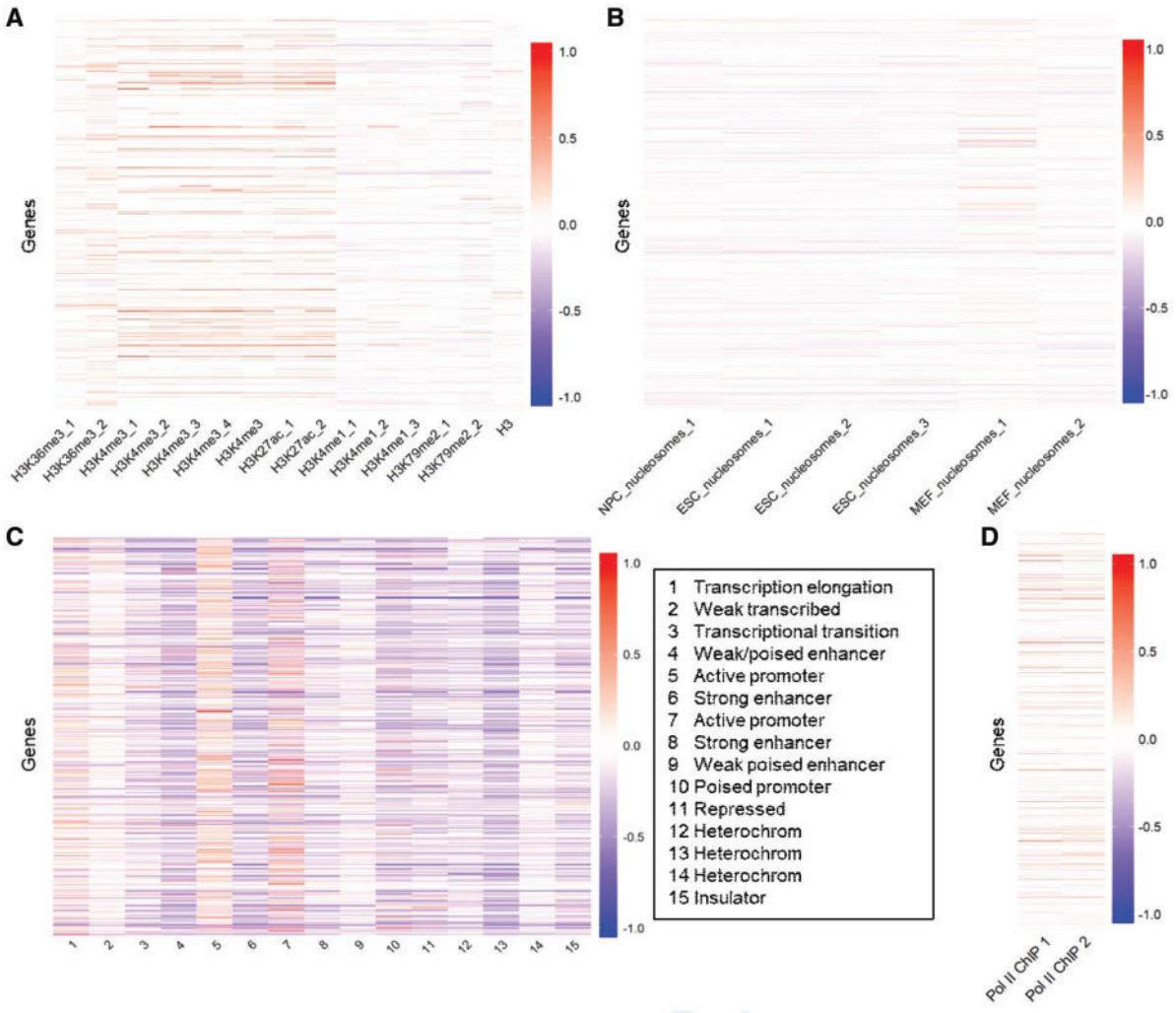

Fig. 6. Correlation coefficients between the estimated Pol II densities and (**A**) ChIP-seq profiles of histone modifiers, (**B**) nucleosome occupancies observed by MNase-seq from mouse ES cells. (**C**) Differences between the averages of the estimated Pol II densities in regions with and without 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. (**D**) Correlation coefficients between the estimated Pol II densities and two ChIP-sep profiles of Pol II. The color scale charts shown on the sides denote the given values in which the mean differences shown in (C) are scaled to $[-1, 1]$

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 Material F4). This inconsistency likely arises from the difficulty of identifying the induction waves of elongating Pol II with GRO-seq. As exemplified in Supplementary Material 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 considerably 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 Material 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 Material 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 Material 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 potenstially predictable with the well-established RNA-seq protocol.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI [grant number JP15K12145].

Conflict of Interest: none declared.

References

Ameur,A. *et al.* (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1440.

Bentley,D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.

Bolić,M. *et al.* (2004) Resampling algorithms for particle filters: a computational complexity perspective. *EURASIP J. Appl. Signal Process.*, **15**, 2267–2277.

Brown,S.J. *et al.* (2012) Chromatin and epigenetic regulation of pre-mRNA processing. *Hum. Mol. Genet.*, **21**, R90–R96.

Chae,M. *et al.* (2015) groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, **16**, 222.

Creyghton,M.P. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA*, **107**, 21931–21936.

Churchman,L.S. and Weissman,J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.

Danko,C.G. *et al.* (2013) Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell*, **50**, 212–222.

Doucet,A. and Johansen,A.M. (2011) A tutorial on particle filtering and smoothing: fifteen years later. In: *Handbook of Nonlinear Filtering*. Vol. 12, Oxford University Press, New York, pp. 656–704.

Duff,M.O. *et al.* (2015) Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, **521**, 376–379.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Flynn,R.A. *et al.* (2016) 7SK-BAF axis controls pervasive transcription at enhancers. *Nat. Struct. Mol. Biol.*, **23**, 231–238.

Hah,N. *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.

Jonkers,I. *et al.* (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*, **3**, e02407.

Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16**, 167–177.

Kulaeva,O.I. *et al.* (2013) Mechanism of transcription through a nucleosome by RNA polymerase II. *Biochim. Biophys. Acta*, **1829**, 76–83.

Kwak,H. *et al.* (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, **339**, 950–953.

Luco,R.F. *et al.* (2011) Epigenetics in alternative pre-mRNA splicing. *Cell,* **144**, 16–26.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.

Radle,B. *et al.* (2013) Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture. *J. Vis. Exp.*, **78**, e50195.

Rodriguez,J. *et al.* (2012) Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. *Mol. Cell*, **47**, 27–37.

Shen,Y. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 16–120.

Sibley,C.R. *et al*. (2015) Recursive splicing in long vertebrate genes. *Nature*, **521**, 371–375.

Sigova,A.A. *et al*. (2013) Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. USA*, **110**, 2876–2881.

Singh,J. and Padgett,R.A. (2009) Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.*, **16**, 1128–1133.

Tanny,J.C. (2014) Chromatin modification by the RNA Polymerase II elongation complex. *Transcription*, **5**, e988093.

Teif,V.B. *et al*. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, **19**, 1185–1192.

Teves,S.S. *et al*. (2014) Transcribing through the nucleosome. *Trends Biochem. Sci.*, **39**, 577–586.