Bayesian inference for transcription elongation rates by using total RNA sequencing

河村　優美

博士（統計科学）

総合研究大学院大学
複合科学研究科
統計科学専攻

平成３０（２０１８）年度

Bayesian inference for transcription elongation rates by using total RNA sequencing

Yumi Kawamura

Department of Statistical Science
School of Multidisciplinary Sciences SOKENDAI (The Graduate University for Advanced Studies)

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2019

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 presented in exonic regions. Such patterns result from the process of underlying transcription elongation by RNA polymerase II, which traverses the DNA strand in the 5' to 3' direction as it performs a complex series of mRNA syntheses and processing. Therefore, data of sequenced total RNAs could be used to infer the rate of transcription elongation by solving the inverse problem. We addressed this issue by using a signal reconstruction technique based on a sequential Monte Carlo algorithm. The objective was to reconstruct the spatial distribution of transcription elongation rates from a given noisy, sawtooth-like profile.

Results: 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 influence on the sawtooth shape. The present method was applied using published total RNA-seq data derived from mouse embryonic stem (ES) cells. We describe the spatial characteristics of the estimated elongation rates, focusing especially on promoter-proximal sites, exons, and introns. We found that the predicted elongation rates are highly correlated with the epigenetic landscape of nucleosome occupancy and histone modification patterns.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Transcription elongation rates are known to play an important role in co-transcriptional events such as splicing, termination, and genome stability [66]. However, how to measure genome-wide elongation rates and processing is poorly understood and controversial. RNA polymerase II (Pol II) is the producer of RNA. A better understanding of Pol II transcription elongation rates is important to understand co-transcriptional processing and its mechanism.

Common approaches to studying the rates of transcription elongation rely on advanced experimental technologies specifically designed to measure the moving distance of Pol II by conducting time-course experiments. 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) [30], native elongating transcript sequencing (NET-seq) [14], precision run-on sequencing (PRO-seq) [38], nascent RNA sequencing (Nascent-seq) [50], and metabolic labeling of nascent RNA using microarrays [49]. 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 progressions of induced waves are visually indistinguishable, and it is 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 the intractability in the protocols of such nascent transcript sequencing.

We have developed a simple statistical scheme that estimates the rates of transcription elongation using a widely used, well-established experimental technique called total RNA

sequencing (total RNA-seq). Experimental protocols of total RNA-seq are much easier to perform than the other existing methods for measuring transcription elongation rates. A statistical method based on Bayesian inference is the key to address this issue.

As will be described, in a given profile of total RNA-seq, a specific pattern of data that we call the sawtooth-like profile is often observed, which is caused by an unobserved process of transcription elongation on Pol II traversing the DNA strand in the 5' to 3' direction. We propose a Bayesian method that predicts the transcription elongation rates from observed total RNA-seq reads by solving an inverse problem. To be specific, after forwardly modeling the given sequenced RNA-seq reads for unknown rates of elongating Pol II and unknown modes of splicing, the backward prediction is performed according to Bayes' law to inversely predict the unknowns. As a proof of principle, we applied our approach on the total RNA-seq data derived from mouse ES cells. We identified some spatial features of elongation rates such as slow down of transcription at exons and promoter-proximal regions. In addition, spatial features of the predicted elongation rates were comprehensively investigated in relation to some epigenetic observations, *i.e.*, nucleosome positioning and histone methylation. This association study has revealed some previously unknown mechanisms.

Until now, genome-wide measurements for the rates of transcription elongation have required us to obtain time-course data in order to capture moving Pol II. In general, such methods involve tedious operation times and high costs for repeated RNA sequencing. However, our proposed method enables us to estimate transcription elongation rates using a one-time operation of total RNA-seq without time-course experiments. This will be a great contribution to the study of the mechanisms of transcription.

Chapter 2

Background

2.1 Transcription elongation

Here, we briefly describe the basic biology relevant to transcription elongation.

2.1.1 Elongation stage by the Pol II and molecular mechanisms

For many genes, Pol II pauses at $\sim$30-50 base pairs (bp) downstream of the transcription start site (TSS). This is called the promoter proximal pausing stage. Pol II passes the pausing phase, and then enters a productive elongation stage. At the termination stage, reaching the 3' end and the RNA chain stabilized by the addition of a poly(A) tail, it releases from the DNA [60].

Fig 2.1 illustrates the following mechanism. After Pol II initiates, Pol II enters the promoter proximal pausing stage under the control of the negative elongation factor (NELF) and 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole sensitivity-inducing factor (DSIF), which is a prevalent feature across metazoan genomes. Before the transition into productive elongation, NELF and DSIF form an elongation complex that is responsible for transcription inhibition. The positive transcription elongation factor b (P-TEFb) mediates the transition into the productive elongation stage. DSIF is phosphorylated by P-TEFb, and NELF is eliminated from the elongation complex, whereas DSIF remains in the complex. Other elongation factors come into the complex, such as the polymerase-associated factor complex (PAFc) and the super elongation complex (SEC). Splicing and polyadenylation machinery to promote a properly processed messenger RNA (mRNA) are coupled with transcription elongation [73].

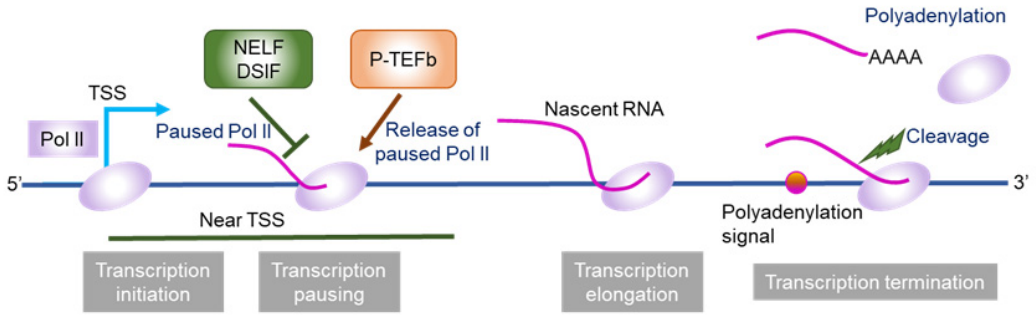

Fig. 2.1 After Pol II initiates, Pol II pauses at a promoter-proxy site under the control of NELF and DSIF. Pol II begins to elongate by P-TEFb. In transcription termination, Pol II is released from the DNA after the poly(A) tail is added to transcripts [13, 30].

2.1.2 Elongation rate controls alternative exons

Transcription elongation rates by Pol II influence pre-mRNA processing such as splicing, termination, and genome stability. Numerous factors and regulatory mechanisms are involved in the stages of transcription elongation by Pol II, suggesting that the elongation process is highly complex. Elongation rates play an important role in the regulation of transcription by Pol II.

It has also been observed that transcription elongation rates are highly related to the determination of splice patterns. As illustrated in Fig 2.2, Pol II elongation rates control exon inclusion and skipping. Faster Pol II elongation, passing an exon, does not have time to recognize it. As a consequence, the alternative exon is often skipped. Slower Pol II elongation as it goes through an exon has enough time for the splicing machinery to recognize and splice out the exon [9, 22].

2.2 Total RNA sequencing

Total RNA sequencing has been widely used to investigate different populations of RNAs, including pre-mRNA and non-coding RNA species in pooled cells. Here, we provide an overview of total RNA-seq experiments with RNA selection independent of the poly(A) tails, which were used in our analysis as shown in later chapters.

2.2.1 Total RNA-seq with and without poly(A) selection

In the gene expression process, RNA synthesis and processing such as 5' end capping, pre-mRNA splicing, RNA editing, and 3' end cleavage and polyadenylation occur. Splicing takes place co-transcriptionally, and pre-mRNA splicing conducts the removal of non-

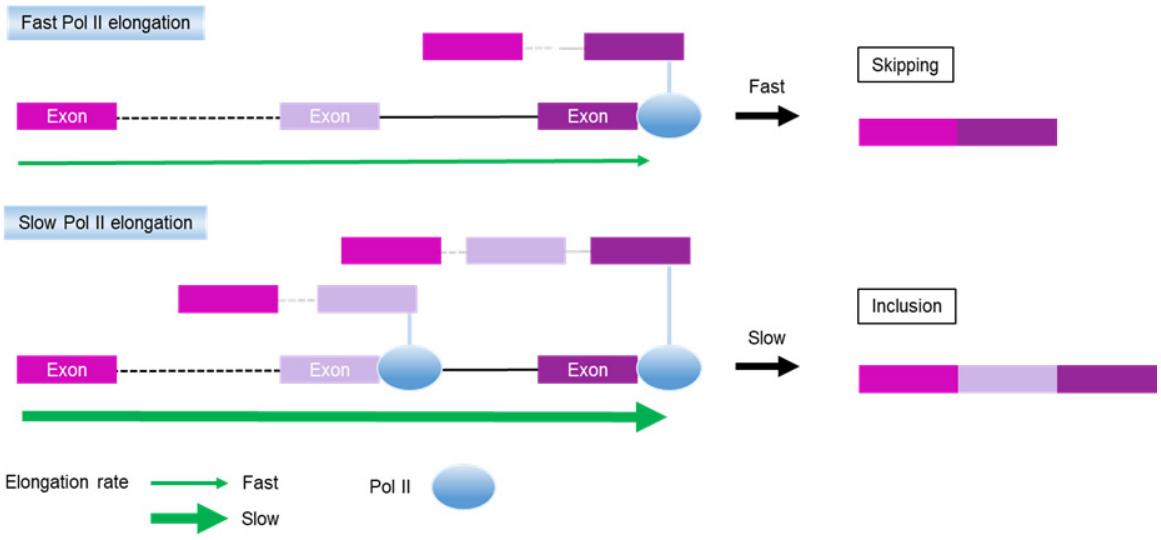

Fig. 2.2 Co-transcriptional splicing controls alternative splicing by Pol II elongation rates. Fast elongation rates tend to lead to the skipping of alternative exons (top). Slow elongation rates tend to lead to the inclusion of exons (bottom). This figure is modified from Ref [9].

coding RNA sequences (introns) and the ligation of coding RNA sequences (exons) to form the mature mRNA. Pol II functions in RNA synthesis and is responsible for the production of nascent RNAs. The use of nascent RNAs and deep sequencing techniques such as with next-generation sequencing (NGS) make it possible to track the process of transcription elongation. In particular, non-coding RNA sequences in intronic regions provide important statistical information to infer the transcription elongation process as detailed in Section 2.2.2.

Currently, NGS technology is an indispensable tool for various studies of the transcriptome, such as RNA-seq [48]. The complete set of transcripts is referred to as the transcriptome and includes protein-coding mRNA (coding RNA) and non-coding RNAs (ribosomal RNA (rRNA), transfer RNA (tRNA), and other non-coding RNAs) [40].

In principle, total RNA sequencing can preferentially detect more of a specific type of transcript by performing RNA selection by poly(A) tailing [69]. Sequenced ribosomal RNA-depleted (ribo-minus or rm) transcripts from total RNAs have been used as a purification method that enriches the RNA transcripts by the selective depletion of rRNAs from total RNAs. Ribo-minus RNA leads to be the absence of poly(A) tailing on transcripts (poly A(-)). The resulting RNA-seq without poly(A) selection can capture both the coding and non-coding RNA transcripts. The data that we analyzed were taken from the total RNA-seq without poly(A) selection. Conversely, RNA-seq with poly(A) selection (RNA-seq) detects only mature RNA transcripts with poly(A) tails (poly A(+)).

Total RNA-seq generates millions of reads and has the potential to determine the full range of the abundance of RNA transcripts. Conventionally, total RNA-seq has been used to quantify just the abundance of RNA molecules originating from both coding and noncoding regions. This study provides another way to use total RNA-seq in genome-wide studies of transcription elongation.

2.2.2 Sawtooth pattern in total RNA-seq

Sequenced total RNAs without poly-A selection (total poly-A(-) RNA-seq) consist of the pool of nascent transcripts and mature polyadenylated RNAs. Pol II traverses on the DNA strand from the 5' to 3' direction and generates nascent transcripts combined with co-transcriptional splicing [9]. It has been reported that total poly-A(-) RNA-seq exhibits a sawtooth pattern in the read density of a gene [2] as characterized by a monotonically decreasing 5'-3' slope in the intronic regions and substantially higher levels of RNA present in the exonic regions (Fig 3.1). One of the major determinants that influences 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 total poly-A(-) RNA-seq could potentially be utilized to obtain relative measures of transcription elongation rates across the genome [4, 42, 57]. However, the use of total poly-A(-) RNA-seq for such purposes has not been widespread, possibly because of the difficulty in analyzing considerably noisy data with low read counts. The underlying mechanisms behind the presence of sawtooth patterns and the statistical methods for inversely inferring the elongation rates from such observations will be described in a later chapter.

2.3 Experimental technologies of measuring transcription elongation rates

Here, we review existing technologies that have been used to measure the rate of transcription elongation at individual-gene and whole-genome levels. Advantages and disadvantages of each method are summarized in Tables 2.1 and 2.2.

2.3.1 Measurement on individual genes

Fluorescence recovery after photobleaching (FRAP) [70] is an *in vivo* imaging method to detect Pol II elongation during steady state transcription. This technique enables us to monitor the recovery rate of fluorescent-tagged transcription factors in living cells. Upon completing transcription, the recovery rate depends on the amount of time that it takes for the photobleached elongating Pol II. The fluorescence recovery rate after photobleaching at a single locus or at multiple loci gives information on transcription elongation rates.

Quantitative RT-PCR (qRT-PCR) coupled with using the compound 5,6-dichlorobenzimidazole 1-β-d-ribofuranoside (DRB), which reversibly prevents gene transcription *in vivo*, can be used to analyze transcription and RNA processing. DRB inhibits the P-TEFb-dependent Ser2 phosphorylation of Pol II. As a consequence, qRT-PCR measures the time during which it fails to progress from the initiation to the elongation step of transcription [57].

The advantages and disadvantages of such gene-by-gene based experimental methods are summarized in Table 2.2.

2.3.2 ChIP-seq of Pol II

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) of Pol II can be used to directly observe the spatial distribution of nascent transcripts over the genome that undergo transcription [15]. Observing Pol II transiting between two consecutive time points provides the information required to infer the velocity (see Table 2.2 for summary).

2.3.3 Nascent RNA-seq

Several experimental technologies, generally referred to as nascent RNA-seq, have recently emerged for genome-wide measurements of Pol II elongation rates, such as global run-on and sequencing (GRO-seq) [30], native elongating transcript sequencing (NET-seq) [14], precision run-on sequencing (PRO-seq) [38], nascent RNA sequencing (Nascent-seq) [50], and metabolic labeling of nascent RNA using microarrays [49]. The objective common to

Table 2.1 Comparison between nascent RNA-seq and total RNA-seq

Method	Disadvantages	Advantages
Nascent RNA-seq and others	Costly to acquire time-course data (possibly high frequency). Difficulty in the wave front identification (Bioconductor 'groHMM'). Long genes only.	Absolute elongation rates.
Total RNA-seq	Relative elongation rates only. Long genes only.	Widely used, well-established.

these methods is to deeply sequence RNAs at the binding sites of transcriptionally active Pol II running on DNA strands in cells. Typically, the 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 transcription. Furthermore, the time progressions of induced waves are visually indistinguishable, and it is often infeasible to track for most genes as will be shown in later chapters. In addition, the spatial resolution of the observed elongation rates depends on the length of the time interval. It is difficult to acquire high frequency time-course data because of the intractability in the protocols of such nascent transcript sequencing (Table 2.2).

2.3.4 Estimation of elongation rates using groHMM

The groHMM package is an R library that can be used to detect the boundaries of transcription waves induced from GRO-seq experiments [12]. After measuring GRO-seq data that shows the leading edge of the Pol II wave by short 17β-estradiol (E2) [27] treatments with inducers and inhibitors of gene activation, a hidden Markov model (HMM) is used to estimate three regions: (1) upstream of the wave, (2) the Pol II wave, and (3) downstream of the wave. GRO-seq is used to detect the elongation rates that are the moving distance of Pol II divided by a defined time period [17]. In real applications, which will be shown later, the proposed method will be compared with a method based on GRO-seq plus groHMM.

Table 2.2 Advantages and disadvantages of various experimental methods used to measure transcription elongation rates [30]

Method	Principle	Advantages	Disadvantages
In vivo imaging by FRAP [10, 44, 3, 70, 6, 8, 18]	Quantifies GFP-tagged Pol II at specific loci following photobleaching in real-time	Performed in vivo. Allows single cell measurements.	Limited to multiple aligned genes or gene arrays. Uses ectopically (over)expressed protein. Low resolution.
RT-qPCR [57, 63]	Intron detection after release from DRB-mediated elongation block	Can be performed at different locations within a gene. Can measure co-transcriptional splicing.	Individual gene experiments only. Can be performed only on long genes, insufficient resolution on short genes.
ChIP-seq (genome-wide exploration) and qPCR (gene-specific) [1, 45]	Immunoprecipitation of Pol II	Can be performed at different locations within a gene. Potentially genome-wide exploration.	Restricted to highly expressed genes; high background noise.
GRO-seq (<50 bp resolution) and PRO-seq (single bp resolution) [29, 17, 51]	Directly measures the nascent transcription at specific sites in the genome after treatment with elongation blockers with small inhibitory drugs like DRB	Can be performed at different locations within a gene. Potentially genome-wide exploration. Highly sensitive.	Can be performed only on long genes, insufficient resolution on short genes.
BruDRB-seq and 4sUDRB-seq [66, 25]	Similar to GRO-seq, measuring the transcription "*wave*" with bromo-labelled UTP or 4sU combined with a DRB wash-out	Can be performed at different locations within a gene. Potentially genome-wide exploration. Highly sensitive.	Can be performed only on long genes, insufficient resolution on short genes.
Total poly(A)$(-)$ RNA-seq [29, 2]	Uses the read density slope of intronic regions as the relative measure of elongation rates	Utilizing many existing datasets has no need for individual experimental design.	Poor sensitivity. Measures only relative elongation rates. High deep sequencing needed. Long introns only.

2.4 Determinants that control the rates of transcription elongation

In Chapter 4.3, gene-to-gene variations in the estimated elongation rates will be associated with some epigenetic observations obtained from independent analyses. Here, we describe the biological implications of chromatin modifications and nucleosomes as major determinants of elongation rates.

2.4.1 Histone modifications

The fundamental unit of chromatin is the nucleosome, which consists of 147 bp of DNA wrapped around a core histone protein octamer made of two dimers of H2A and H2B and a tetramer of H3 and H4. All histones can have post-translational modifications, including acetylation, phosphorylation, methylation, and ubiquitylation, which regulate the chromatin structure. As a result, the structures of the chromatin state change in relation to histone modifiers [71] (Fig 2.4).

In general, active transcription states are due to high levels of lysine acetylation on the H3 and H4 tails, trimethylation of H3 at lysine 4, trimethylation of H3 at lysine 79, ubiquitylation of H2B, and trimethylation of H3 at lysine 36 (Fig 2.3). Silent and repressed transcription is caused by trimethylation of lysine 27, ubiquitylation of H2A on lysine 119, and trimethylation of H3 at lysine 9 (Fig 2.3).

One of the major determinants of transcription elongation is histone modification. There is some evidence that the histone modifications along gene bodies are associated with transcription elongation. A subset of histone modifications includes histone H3 methylated at lysines 4, 36, and 79 (H3K4me, H3K36me, and H3K79me), and histone H2B monoubiquitylated on lysine 120 (H2Bub1) [61].

Set1 enzymes are the catalytic subunits of Set1/COMPASS that are H3K4 methyltransferase complexes [54]. Methyltransferase of the COMPASS complex is responsible for all H3K4 methylations in yeast. H3K4 di- and tri-methylation by Set1/COMPASS is a highly regulated process that depends on mainly monoubiquitination of H2B on lysine 123 (H2Bub) by the Rad6/Bre1 complex. H2Bub is recognized by COMPASS component Cps35 (Swd2) and then recruits the other subunits of COMPASS to conduct H3K4 di- and tri-methylation. H2B ubiquitination is the product of a complex regulatory cascade for which Pol II functions as a central platform. Methylation of H3K4 requires H2B ubiquitination, but not vice versa [59].

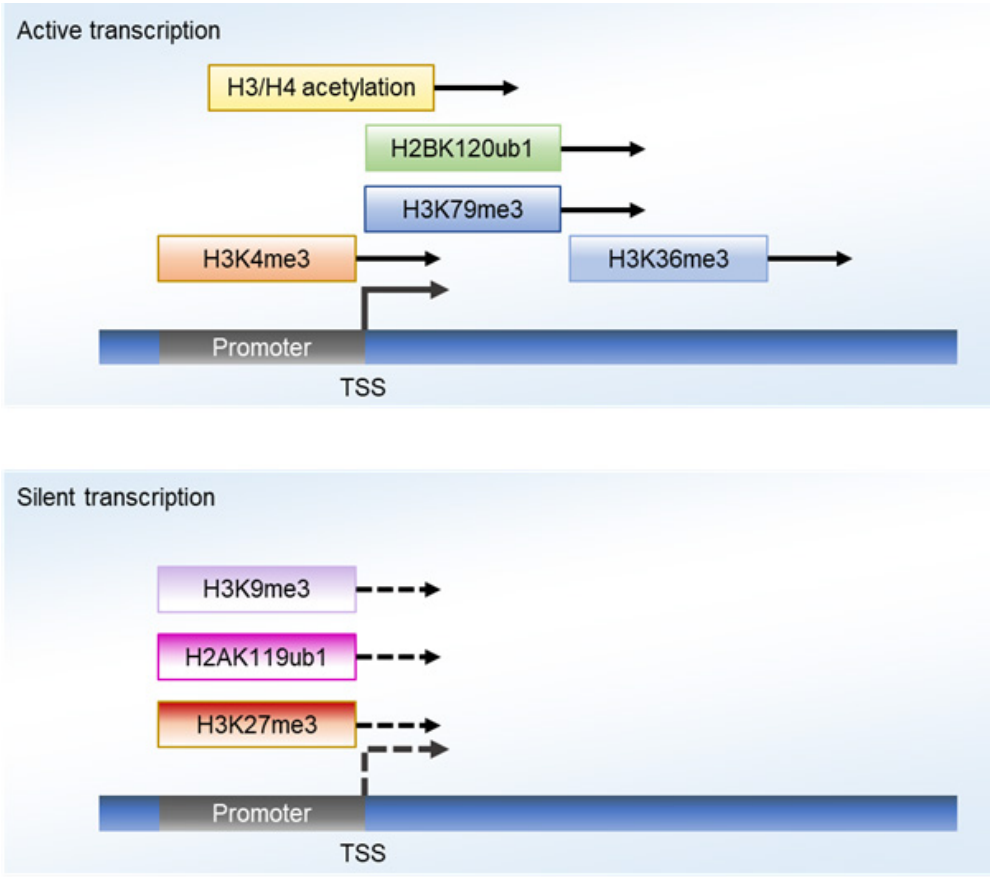

Fig. 2.3 Role of histone modifications in active or silent transcription. This figure is modified from Ref [71].

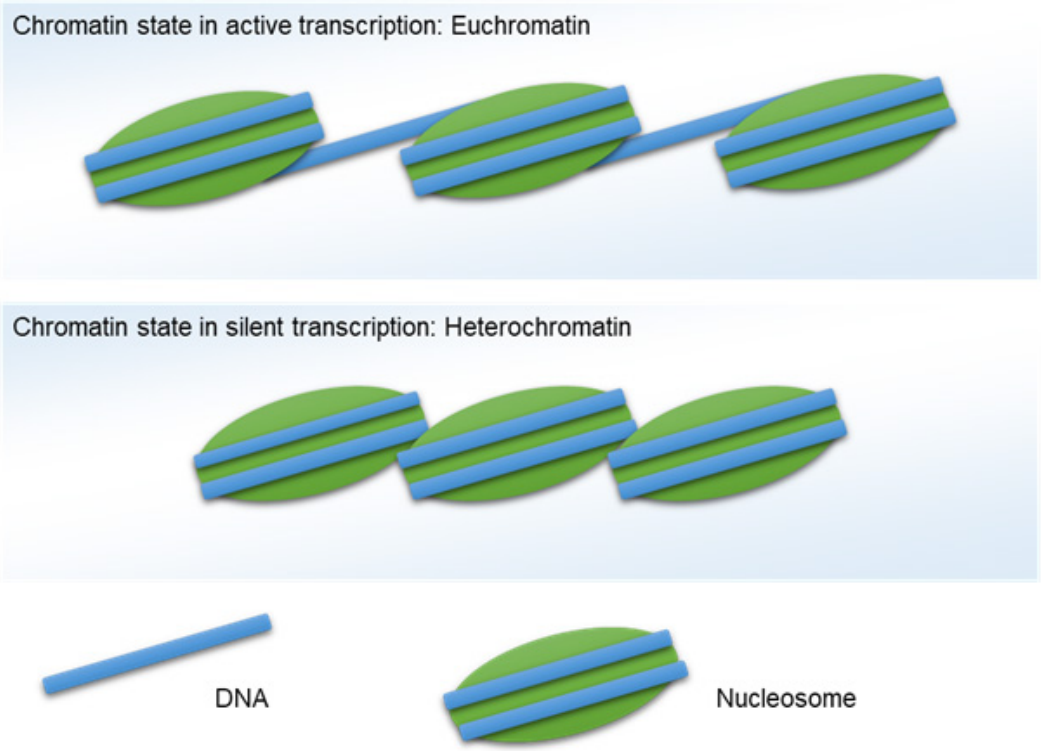

Fig. 2.4 Active and inactive transcription states are regulated via the modification of chromatin states, which are called euchromatin and heterochromatin, respectively.

Dot1/DOT1L (disruptor of telomeric silencing-1) catalyzes mono-, di-, and tri-methylation of histone H3 at lysine 79 via a non-processive mechanism using S-adenosylmethionine (SAM) as a cofactor. Efficient H3K79 trimethylation mainly requires ubiquitination of H2B, such as with H3K4. The Paf1 complex, associated with elongating Pol II, enhances the recruitment of Rad6 and Bre1 to chromatin, linking this modification to transcription elongation [68]. Some DOT1L-associated complexes have been identified in mammals that contain the Pol II Ser2-specific CTD kinase P-TEFb, implicating the involvement of Dot1 in transcription elongation. However, whether this mechanism relates H3K79 methylation to transcription activation and elongation is still not clear [59].

H3K36 methylation is catalyzed by Set2 family enzymes within the coding regions of transcribed genes. H3K36me2 and H3K36me3 are generally associated with transcription activation, and H3K36me3 levels in particular are known to correlate with transcription elongation rates. H3K36 methylation is associated with transcribed genes, and therefore it is used as a reference of active transcription [59, 67].

2.4.2 Transcription regulation through nucleosomes

Eukaryotic genomes are packaged into nucleosomes, which are chromatin composed of repeating units of 147 bp of DNA wrapped around eight histone proteins. Nucleosome assembly prevents Pol II access to DNA because Pol II has to cross the nucleosome barrier to get access to DNA and transcribe genes efficiently [64, 37]. Nucleosome occupancy and positioning are measured by micrococcal nuclease sequencing (MNase-seq).

Most core histone proteins have a diversity of minor (e.g. H2B.1, H3.3) or major (e.g. H2A.Z, H2A.Bbd, centromere protein A [CENP-A]) modifications in their amino acid sequences. Nucleosomes containing histone variants might change their bonding and interactions with DNA, which can clarify the role of varying DNA sequence preferences and nucleosome positions [52].

The intrinsic DNA sequence and structural preferences of nucleosomes play a crucial role in nucleosome occupancy and positioning, and this chromatin landscape is further distributed by chromatin remodelers [52, 31]. Chromatin remodelers use the energy from ATP hydrolysis to evict, assemble, or slide nucleosomes. The main subfamilies of remodelers are divided into four classes: SWI/SNF, ISWI, INO80, and CHD [11].

The chromatin remodeler Chd1 evicts nucleosomes downstream of the promoter, and Chd1 is responsible for the vast majority of transcription-mediated nucleosome turnover. Chd1 is required to overcome the nucleosomal barrier and enables the Pol II promoter to escape in order to be transcribed. It was reported in a previous study that Chd1 plays a crucial role in breaking the nucleosome barrier for progressive transcription [58].

2.5 Recursive splicing is a stepwise removal event of a long intron

As a fortuitous byproduct, the current proposed method could be used to analyze a recently reported novel splicing event, referred to as recursive splicing (RS). It is commonly accepted that splicing of introns is a single removal process as one unit from mRNA transcripts. RS is a stepwise removal process of an intron that has most often been observed in exceptionally long introns [21, 55]. As mapped reads of total RNA-seq without poly(A) selection contain a mixture of mRNA, pre-mRNA, and nascent RNA transcripts, RNA-seq reads contain transcripts originating from introns that are non-coding sequences. Co-transcriptional splicing can be observed in total RNA-seq data that shows the sawtooth pattern of repeatedly decreasing gradient read densities across introns in the 5'-3' direction of transcription. Previous studies have identified the occurrence of RS using observations from total RNA-seq ([21, 55]). It has been observed that the read density of total RNA-seq shows a lot of valleys in intronic regions that can be used as signals for the occurrence of RS (Fig 2.5). RS depends on 3' and 5' splice-site sequences, called recursive splice sites, that are in the way of long introns. Many of the functional sites of ratchet points are conserved across *Drosophila* strains, indicating that RS is evolutionarily conserved [21]. A total of 197 ratchet points in 130 introns of a total of 115 *Drosophila* genes that have been known to show RS have been identified [21]. The technology of total RNA-seq has the potential to identify the unknown RS sites from the splicing patterns of read density. Some previous studies have suggested that many genes that have longer introns in which RS occurs are related to neurological diseases and autism [33, 39, 47].

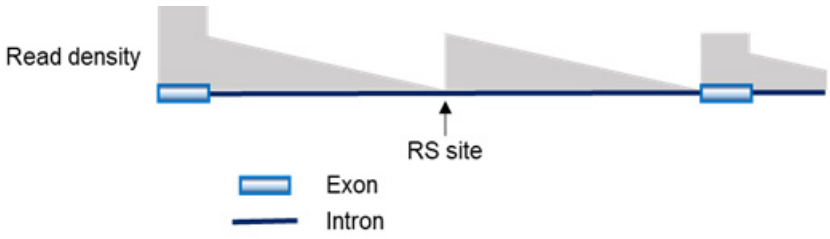

Fig. 2.5 Schematic diagram for the identification of RS sites using the observed read density of total RNA-seq. This figure is modified from Ref [21, 55].

2.6 Statistical inference

Well-established total RNA sequencing is the most promising tool for elucidating genome-wide transcription elongation rates. We focused on the use of total poly-A(-) RNA-seq. The proposed method relies on a state space representation [34–36] that describes a mathematical relationship between the observed read density and spatially varying elongation rates. A prior distribution is placed on the elongation rates and splicing patterns, which is then followed by Bayesian inference by performing sequential Monte Carlo calculations (SMC) [7, 20, 19]. The data captures the pool of different kinds of source signals associated with the spatial dynamics of elongation rates and co-transcriptionally occurring mRNA splicings such as exon skipping, intron retention, RS [21, 55], etc. 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 the disadvantages over GRO-seq. We explored the Pol II elongation rates in 659 genes in mouse ES cells [56]. The estimated elongation rates were compared with some epigenetic observations of nucleosome occupancy and histone modification patterns in mouse ES cells that have been reported in different studies [62, 43, 16]. We found that position-specific variations in the elongation rates agree to some extent with the observed epigenetic landscape.

Chapter 3

Forward modeling and backward prediction

3.1 Sawtooth observation in total poly-A(-) 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, typically when 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.* [2] that a saw-tooth shape appears in the read density, since the sequenced reads capture the pool of mature and immature RNAs in cells as schematically shown in Fig 3.1.

Let $x(t)$ be the probability of existence of Pol II instantly occurring at nucleotide position t on a 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 $s(t)$:

$$r(t) = \int_t^{s(t)} x(u)\mathrm{d}u. \tag{3.1}$$

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

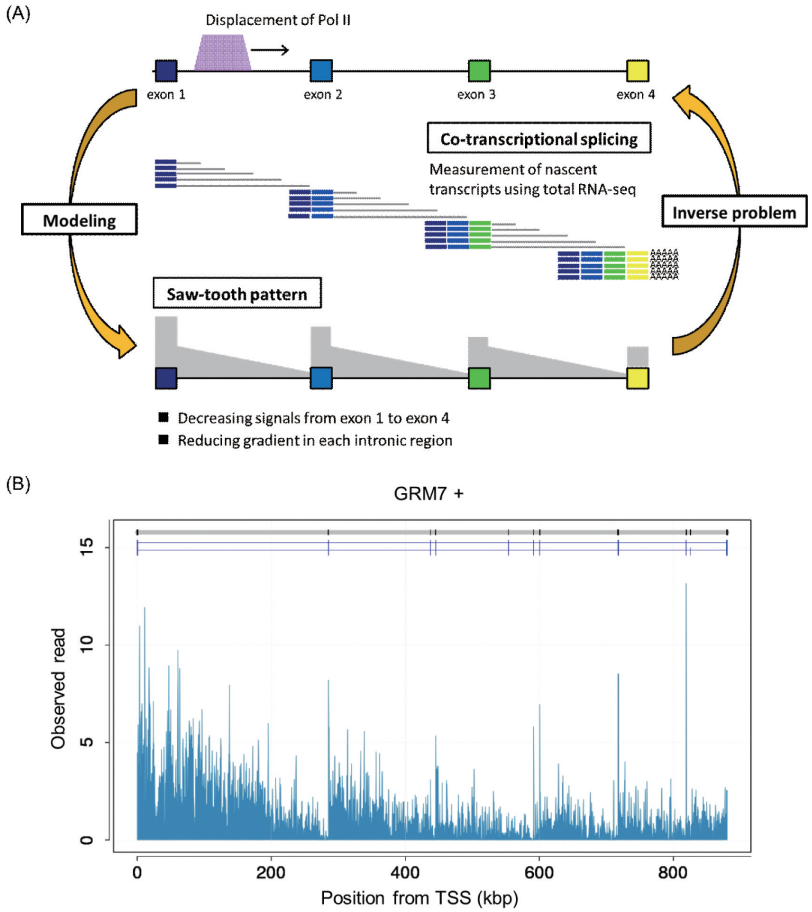

Fig. 3.1 The inverse problem of the transcription elongation rate. (A) Total poly-A(-) RNA-seq captures a mixture of mature and nascent transcripts in a pool of cells. During the displacement of Pol II from 5' to 3', elongating and co-transcriptionally spliced RNAs can take various states as shown in 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 Fig. 2 of Ameur *et al.* [2]. (B) Total poly-A(-) RNA-seq reads of a gene (GRM7) in human fetal brain [2]. Splice variants reported in hg19, GRCh37 (Genome Reference Consortium Human Reference 37) are shown in the upper side.

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

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

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 a sufficient amount of reads. In other words, the inference of elongation rates is feasible only to a small subset of longer genes without performing deep sequencing.

3.2 Existing method

Despite the great potential to utilize total RNA-seq to study transcription elongation rates, there has been considerably less progress made in statistical methods. In a previous study, the slope of the read density in an intron was estimated using a linear regression model, and the estimated coefficient was used as an estimate of the relative elongation speed [2]. However, the estimated slope is just a measure of the average elongation rate in the intron. Obviously, the transcription elongation rates vary from one place to another in the intron. Furthermore, since the estimate of the slope is affected by the baseline expression level of the intron, it is infeasible to compare such estimates between different introns. In addition, different splicing modes bring different slopes to the read density. Therefore, any statistical estimation of elongation rates should be coupled with the identification of splicing variations.

One contribution of this study is to provide a method of estimating unobserved states of transcription elongation rates and splicing modes simultaneously.

3.3 State space representation

Each intron is divided into bins with intervals equal to 400 bp. An exonic region is 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 are 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

$$\begin{aligned}
\log y_n &= \log r_n + \eta_n, \quad \eta_n \sim N(\mu,\sigma), \\
r_n &= \sum_{i=n}^{s_n} x_i, \\
\log x_n &= \log x_{n+1} + \nu_n, \quad \nu_n \sim N(0,\gamma), \\
s_n &\sim p(s_n|s_{n+1},s_{n+2},\ldots,s_N),
\end{aligned} \tag{3.2}$$

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 Eq. 3.1. 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 an algorithm based on the SMC method [20, 41] 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 [20]. These methods share a common algorithmic structure with genetic algorithms. The system model in Eq. 3.2 is used to generate samples of (x_n, s_n) with a 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 a given y_n. Samples having better fitness have a better chance at surviving in the next generation. This process keeps iterating from N to 1, and at the end, samples from the targeted posterior will be produced. The algorithmic details are shown in Algorithm 1.

$$\text{Posterior distribution:} \quad p(x_{1:N}, s_{1:N}|y_{1:N}) \sim p(x_1)p(s_1)\prod_{n=2}^{N} p(y_n|x_{n+1:N}, s_n)p(x_n|x_{n+1})p(s_n|s_{n+1:N}) \tag{3.3}$$

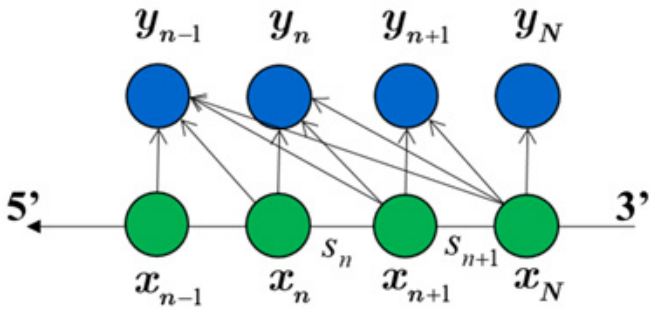

Fig. 3.2 Graphical model of a state space representation in this work

3.4 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 shown in previous studies. 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 challenge is to avoid the occurrence of infeasible splicing patterns during random generation.

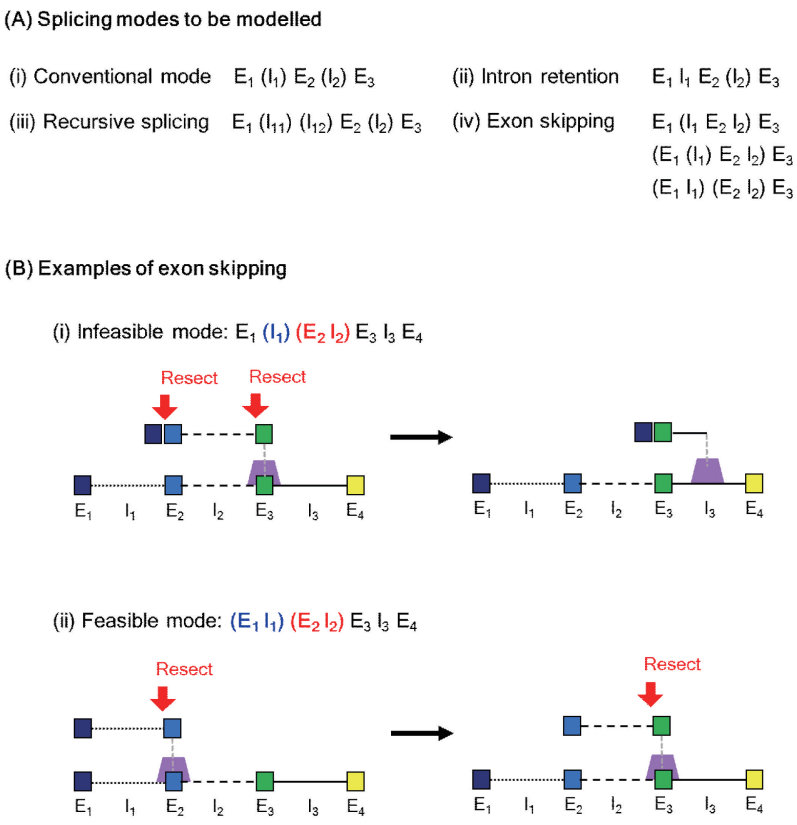

Fig. 3.3 (A) Four splicing modes to be modeled in the system with illustrative examples: (i) conventional mode, (ii) intron retention, (iii) RS of introns, and (iv) exon skipping. (B) Infeasible and feasible modes of exon skipping are exemplified in (i) and (ii), respectively.

Algorithm 1 sequential Monte Carlo

Input: $\{y_n\}_n$, μ, σ, γ, splice site generator $p(s_n|s_{n+1},\ldots,s_N)$, the number of particles M, μ_0, τ_0
Output: $\{x_n^m\}_{n,m}$, $\{s_n^m\}_{n,m}$
Initialize:
for $m = 1,\ldots,M$ **do**
  $x_N^m \sim N(\mu_0, \tau_0)$
  $s_N^m \to N$
end for
for $n = N-1,\ldots,1$ **do**
  for $m = 1,\ldots,M$ **do**
    Renewal of x_n**:** $x_n^m = x_{n+1}^m + \eta_n^m$ with $\eta_n^m \sim N(0,\gamma)$
    Renewal of s_n**:** $s_n^m \sim p(s_n^m|s_{n+1}^m,\ldots,s_N^m)$
    Expected read: $r_n^m = \sum_{i=n}^{s_n^m} x_i^m$
    Likelihood: $w^m = N(\log y_n - \log s_n^m|\mu,\sigma)$
  end for
  Resampling: Perform the resampling of $\{x_n^m, s_n^m\}_{(n,m)}$ with the selection probabilities proportional to $\{w^m\}_m$ (residual systematic resampling [7]).
  The resampled set forms a new ensemble set $\{x_n^m, s_n^m\}_{(n,m)}$.
end for

As illustrated in Fig 3.3, 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 most often been observed in exceptionally long introns [55]. The occurrence of splicing in the middle of an intron brings a valley in the sawtooth shape of total poly-A(-) RNA-seq reads at the RS site [21, 55]. 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 3.4.

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

The procedure for successively constructing such a sequence is summarized in quasi-code Algorithm 2. 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 Fig 3.3. For example, consider that a gene consists of $E_1I_1E_2I_2E_3$ 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 as 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 $E_1(I_1E_2I_2)E_3$, where the unit in the parentheses is isolated simultaneously. On the other hand, $E_1(I_1)(E_2I_2)E_3$ would be unlikely to occur. This mode describes a nascent transcript comprised of $E_1E_2I_2$ dangling from the DNA strand temporarily, and its subunit E_2I_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, $(E_1I_1)(E_2I_2)E_3$ 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 $s_* = \min\{s_m | m \in E_{k+1},\ldots,E_K\}$.

3.5 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 of the expected reads, *i.e.*, $\log r_n = \log\sum_{i=n}^{s_n} x_i$ (see the measurement equation in Eq. 3.2), 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 $x_n = \exp f(n) - \exp f(n+$

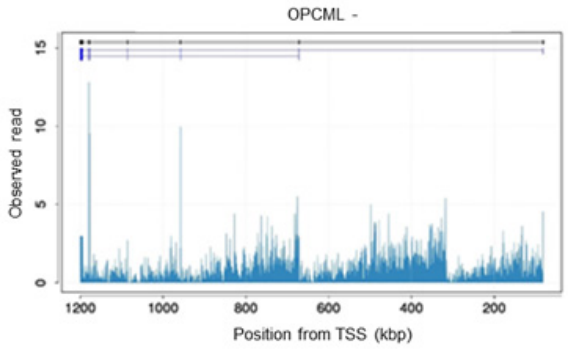

Fig. 3.4 Read density of the OPCML gene in human fetal brain [2]. The observed valley in the intron implies the occurrence of RS.

Algorithm 2 Generator for splice sites $p(s_n|s_{n+1},\ldots,s_N)$

Input: $s_{n+1},\ldots,N$, $\alpha,\beta,\delta,\varepsilon,\phi_{\text{thr}}$
Output: s_n, ϕ_{thr}

$t_1\ldots t_p =$ 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 $n \in I_1 \setminus \{t(I_1)\} \cup \ldots \cup I_{K-1} \setminus \{t(I_{K-1})\}$ **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**

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

end if

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.

3.5.1 Residual systematic resampling

We use the resampling algorithm residual systematic resampling (RSR), which is similar to residual resampling (RR) and systematic resampling (SR). The resampling process is to replicate particles with higher weights and discard particles with lower weights. In RR, the number of replications of a specific particle is determined in the loop by truncating the generation of the number of particles and the normalized weight using uniform distribution numbers. In RSR instead, the updated uniform distribution number is produced in a different procedure, which allows for only one iteration loop, and processing time is different from the distribution of the weights at the input. The RSR algorithm for N input and M output (resampled) particles is the following quasi-code Algorithm 3 [7].

Algorithm 3 residual systematic resampling [7]

Objective: product of an array of indices $\{i\}_1^N$ at time $n, n > 0$.
Input: an array of weights $\{w_n^{(m)}\}_1^N$, N input and M output number of particles.
Method:
$(i) = RSR(N, M, w)$
Generate a random number $\Delta U^0 \sim u[0, 1/M]$
for $m = 1 - N$ **do**
    $i^{(m)} = [w_n^{(m)} - \Delta U^{(m-1)} \cdot M] + 1$
    $\Delta U^{(m)} = \Delta U^{(m-1)} + i^{(m)}/M - w_n^m$
end for

Chapter 4

Bayesian inference of transcription elongation rates

4.1 Total poly-A(-) RNA-seq data

The total poly-A(-) RNA-seq data that we used was derived from mouse ES cells [56]. As already discussed, the RNA-seq reads were considerably sparse, especially in shorter genes, hence we began by selecting analyzable genes. 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. Fig 4.1 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 659 genes that contain one or more such selected introns.

4.2 Estimated Pol II density

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 4.2. The reconstructed elongation rates of the 653 genes are displayed by a heatmap in Fig 4.3. The results suggest that the estimated elongation rates of the 653 genes have roughly six patterns, such as faster or slower elongation rates across a gene.

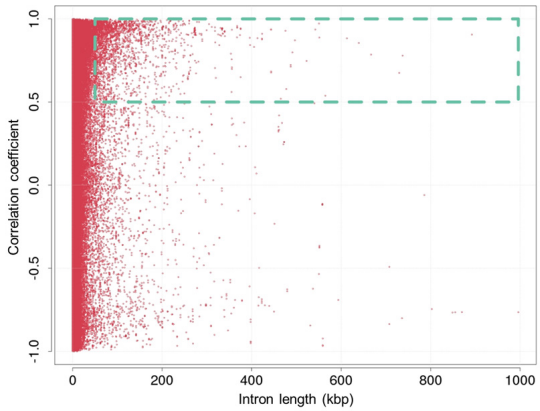

Fig. 4.1 Selection of introns to be analyzed that seem to exhibit monotonically decreasing gradients. Pearson's correlation coefficient between intronic read counts and their positions was used as the monotonicity measure. We selected introns in the rectangle with lengths ≥ 5000 bp and with correlation coefficients ≥ 0.5.

4.3 The spatial features of the transcription elongation rates

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 [24]. As shown in Fig 4.5(D), 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 that of significantly negative genes at the 5% significance level (Fig 4.6 (C)).

Next, we investigated the spatial features of the transcription elongation rates in neighboring regions of TSSs as shown in Fig 4.4 (A). The averaged elongation rates at 0-3 kb and 3-6 kb downstream from the TSSs were compared. A nearly 1.75-fold slower elongation was observed in the TSS-adjacent regions than in the downstream regions. This is due to the widely known promoter-proximal pausing of Pol II at $\sim$30-50 bp downstream of the TSS, which is mediated by negative elongation factors [30]. In addition, as shown in Fig 4.4 (B), 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 [9, 61]. On the other hand, a lack of correlation was observed between the estimated Pol II densities and the guanine-cytosine content (GC content) in the DNA sequences (Fig 4.4 (C)), though several studies suggest that GC-rich sequences negatively influence elongation rates [29].

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 [62, 43, 16]. Pearson's

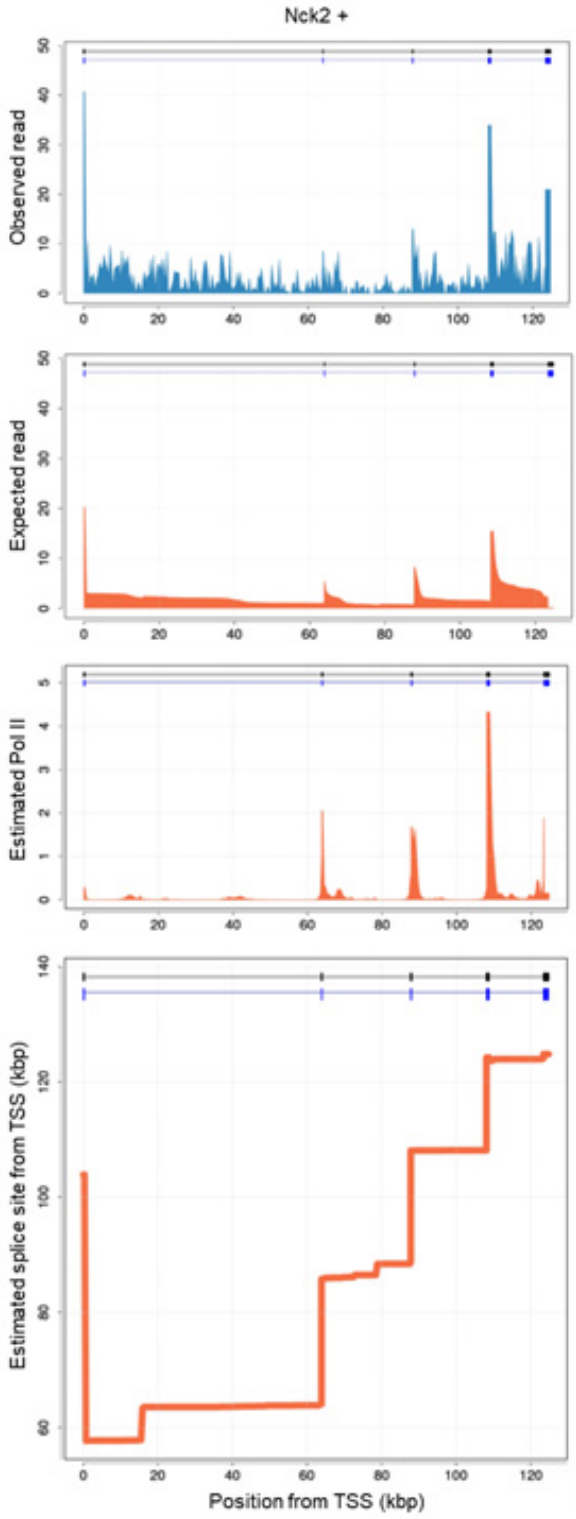

Fig. 4.2 Estimated Pol II density, expected read density, and splicing patterns are shown on the DNA coordinates of the Nck2 gene 5'-3' from the left to right. The observed read counts are shown in the top panel.

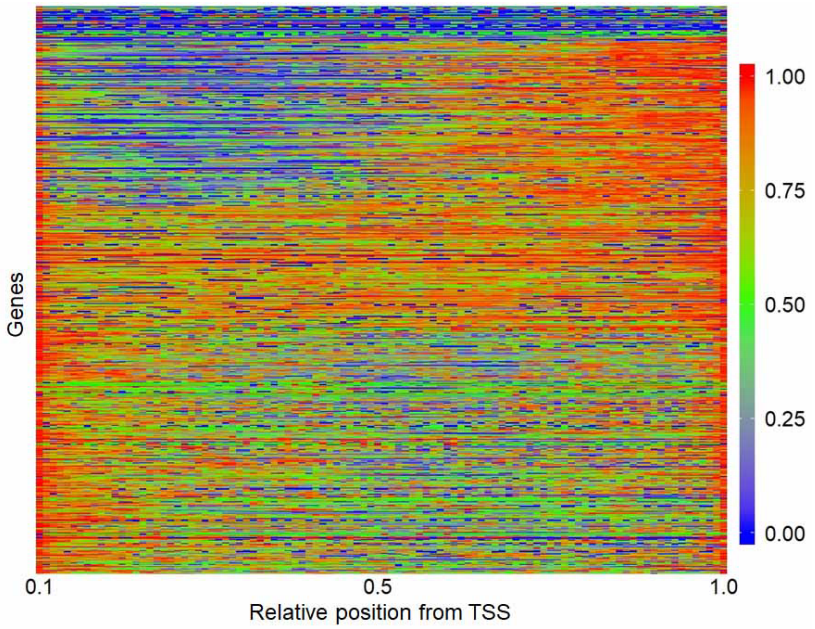

Fig. 4.3 The estimated elongation rates of 653 genes are arranged on the vertical axis. The horizontal axis denotes the relative position from the TSS. The color scale chart shown on the side denotes the estimated values normalized to [0; 1].

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) [62]. Nucleosomes form barriers against Pol II elongation, and nucleosome-depleted regions become more accessible by Pol II [64]. Indeed, the correlation coefficients indicated positive relationships between the estimated Pol II densities and the nucleosome positioning patterns [37] in many genes (Fig 4.5 (B) and Fig 4.6 (B)).

For the association with histone modification patterns, we used the ChIP-seq profiles of histone modifiers involved in epigenetic silencing (histone H3 lysine 79 dimethylation (H3K79me2)) and activation (histone H3 lysine 4 trimethylation (H3K4me3), histone H3 lysine 36 trimethylation (H3K36me3), and histone H3 lysine 27 acetylation (H3K27ac)) (GSE11724, GSE24165) [43, 16]. For many genes, the estimated Pol II densities seem to be positively related to the histone modification marks associated with transcriptional activation (Fig 4.5 (A) and Fig 4.6 (A)). The number of genes exhibiting significant positive correlations was more than eight times larger than those with negative correlations at the 5% significance level. However, the histone modification patterns of the silencer groups tended to correlate negatively with the Pol II densities within the gene bodies (Fig 4.5 (A) and Fig

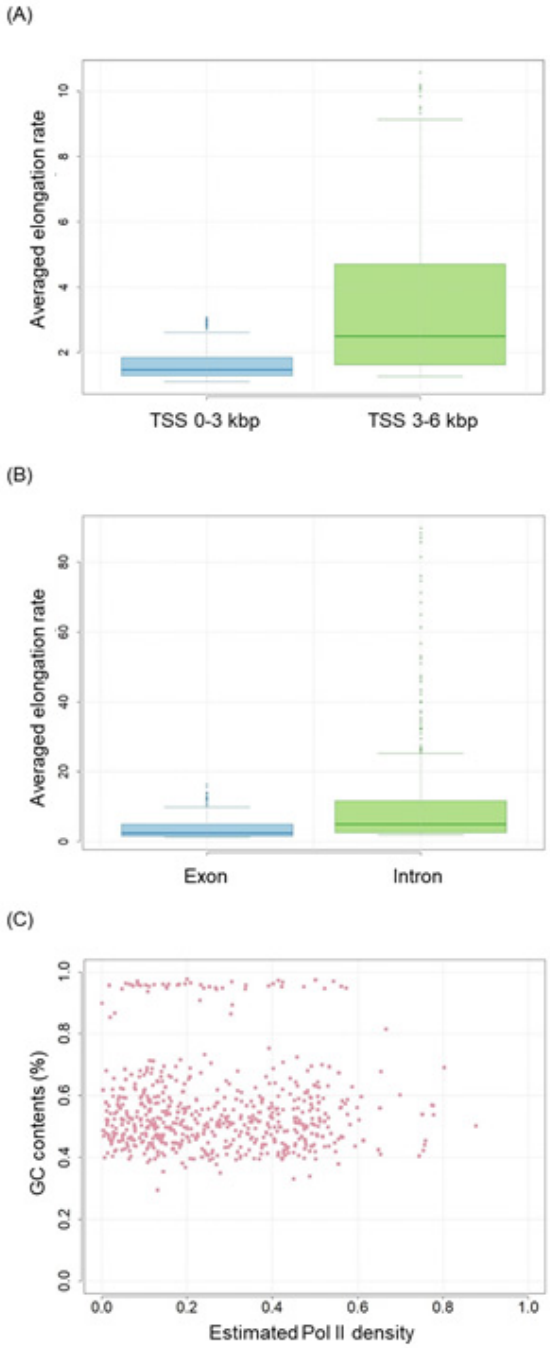

Fig. 4.4 Spatial features of the estimated elongation rates of 653 genes: (A) barplot describing the difference in the averaged elongation rates between the promoter-proximal regions 0-3 kb downstream from TSSs and the subsequent regions (3-6 kb from TSSs); (B) differences between exons and introns; and (C) the estimated elongation rates and GC contents in all the binned intronic regions.

4.6 (A)). 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 have a highly consistent pattern with the observed epigenetic landscape.

In addition, the estimated Pol II densities were investigated in relation to computationally annotated chromatin states. We used 15 annotations of chromatin states [53], which were obtained by performing a Poisson-based multivariate hidden Markov model (ChromHMM) [23] 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 4.5 (C), 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.

4.4 Comparison between the elongation rate of total poly-A(-) RNA-seq and GRO-seq

The estimated elongation rates of the 645 genes were compared to those estimated based on GRO-seq [30, 27]. Using a hidden Markov model with the groHMM package [12, 17] 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 Pol II paused state. 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 in each interval of the identified wave fronts at two consecutive times, and the relative elongation rate of 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, and they showed a lack of agreement between the different estimates of elongation rates with total RNA-seq and GRO-seq (Fig 4.7). This inconsistency likely arises from the difficulty in identifying the induction waves of elongating Pol II with GRO-seq. As exemplified in Fig 4.8, it was quite hard for many genes even to recognize the exact visual 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.

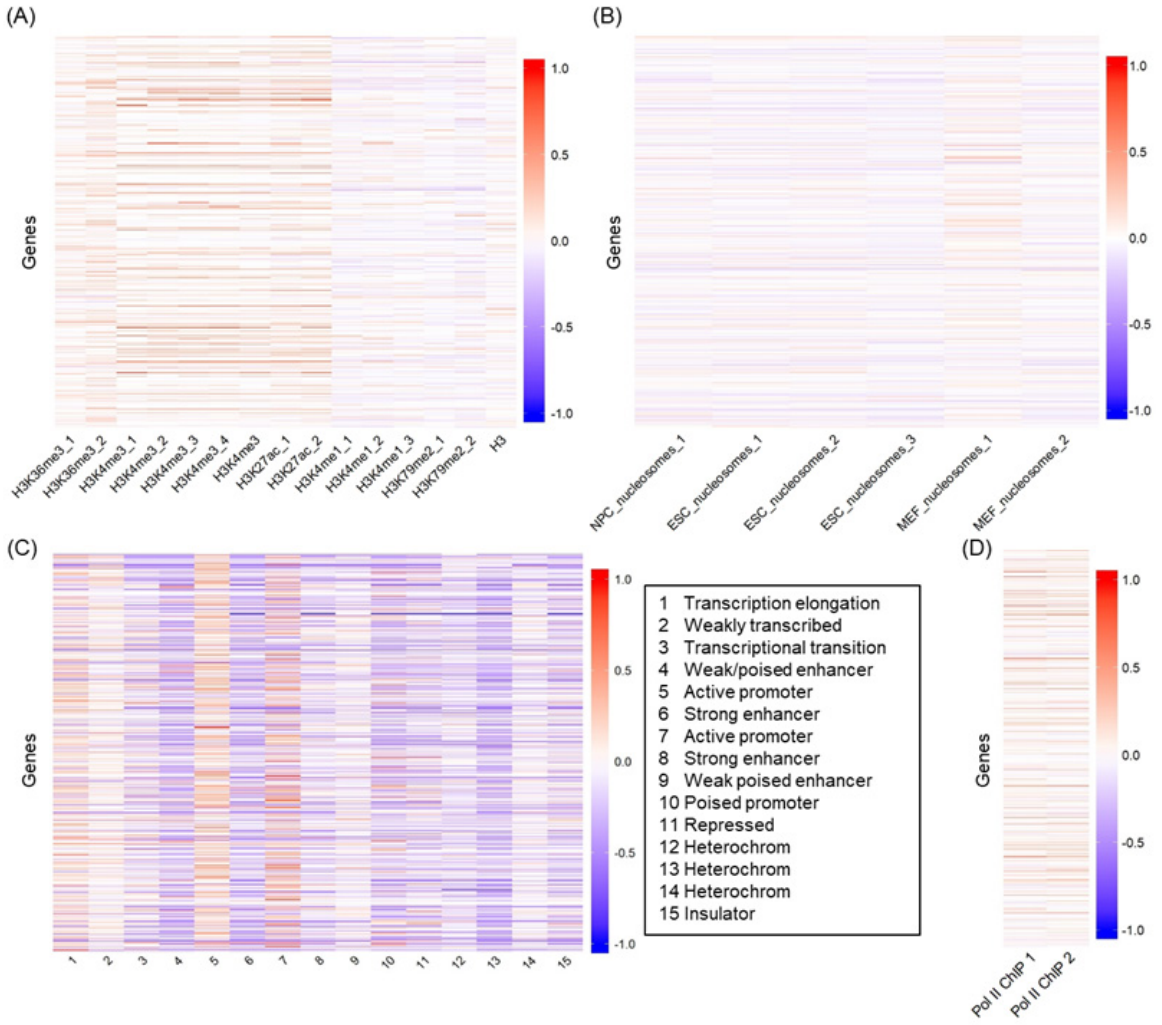

Fig. 4.5 Correlation coefficients between the estimated Pol II densities and (A) ChIP-seq profiles of histone modifiers and (B) nucleosome occupancies observed by MNase-seq from mouse ES cells. (C) Differences between the averages of the estimated Pol II densities in regions with and without the chromatin state annotated. 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-seq 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].

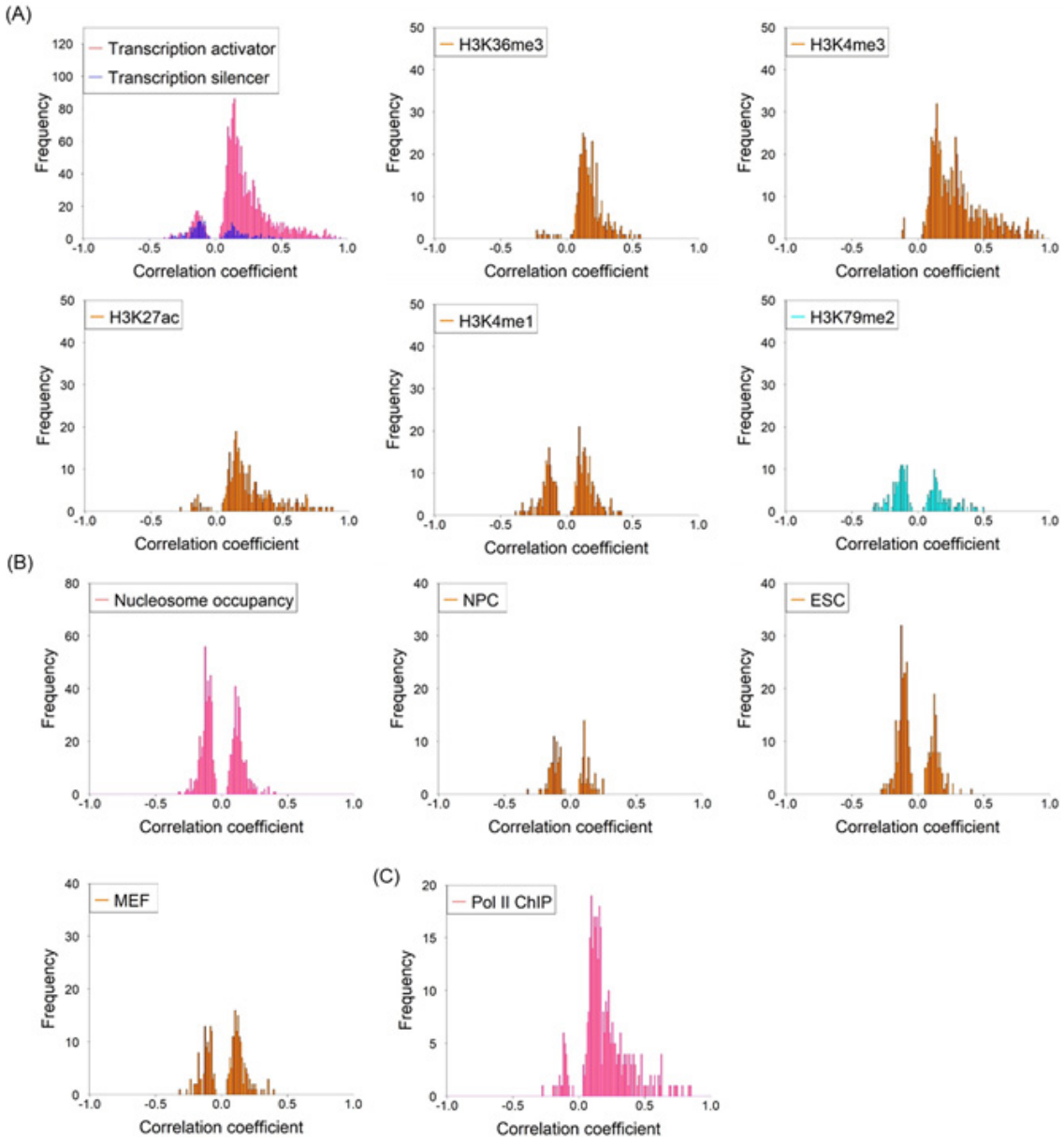

Fig. 4.6 Histograms of the statistically significant correlation coefficients (at a significance level 5%) between the estimated Pol II densities and (A) ChIP-seq profiles of histone modifiers, (B) nucleosome occupancies observed by MNase-seq, and (C) ChIP-seq profiles of Pol II from mouse ES cells.

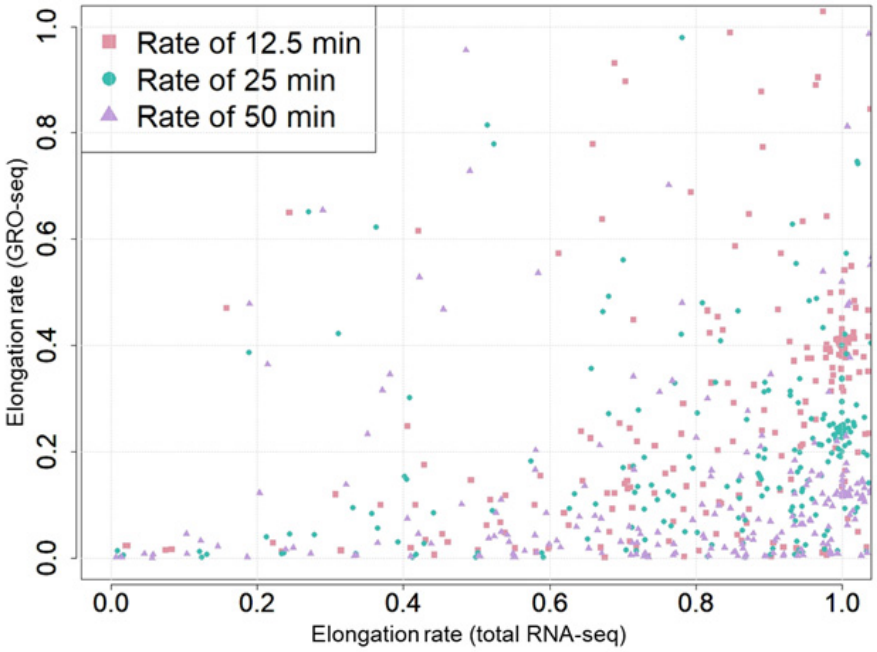

Fig. 4.7 Comparison between the estimated elongation rates using total poly-A(-) RNA-seq and GRO-seq with the hidden Markov model.

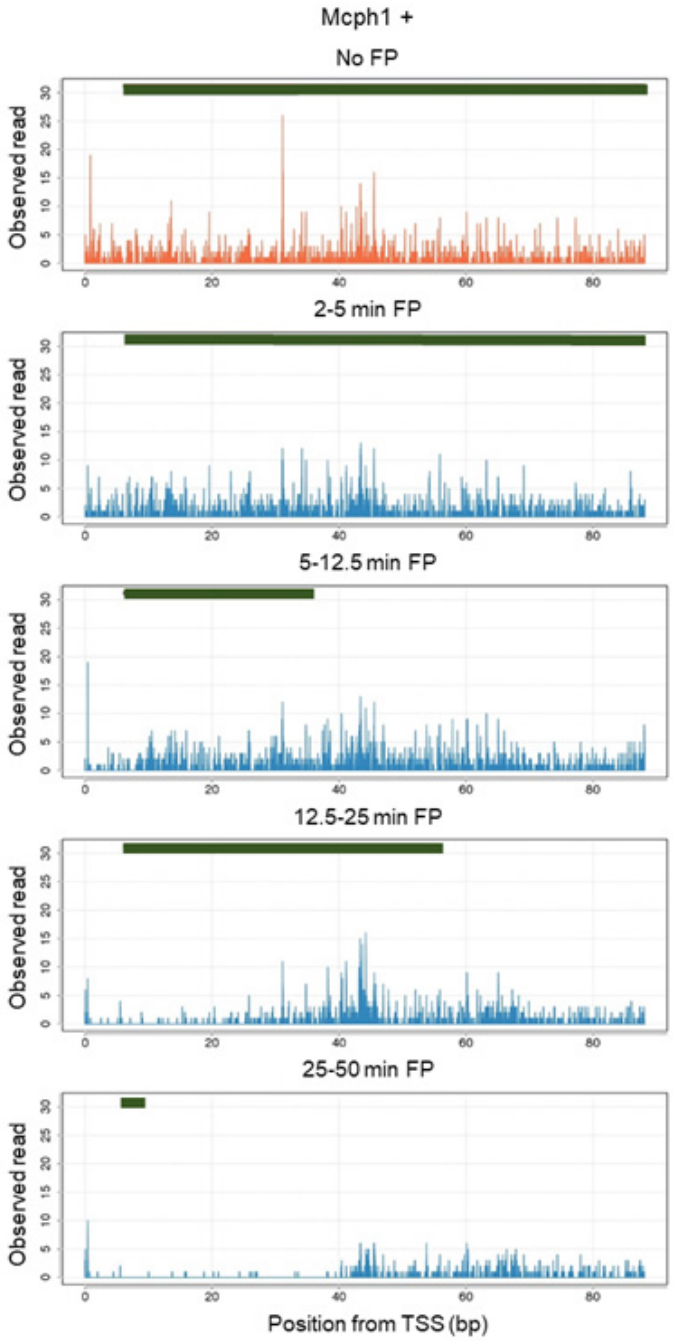

Fig. 4.8 Example of GRO-seq time-course data at $0, 2, 5, 12.5, 25$, and 50 min after release from the paused state of Pol II in the Mcph1 gene. The black rectangles shown at top denote the wave fronts of Pol II progression that were estimated by the hidden Markov model (the groHMM package in R).

4.5 Implementation of estimating Pol II density

An implementation using C software of "Pol II density estimated by the statistical inference of transcription elongation rates by total RNA-seq (PolSter)" and sample data are available at https://github.com/yoshida-lab/PolSter.

We demonstrate how to estimate the Pol II density with this software. At the outset, it needs to prepare an example total RNA-seq dataset [56] (mouse dataset) in which each intron is divided into bins with intervals equal to 400 bp, and an exonic region is treated positionally as a single point. Here's an example of a potential input file:

- $*\_read.txt$: read count data across genomic bin positions from TSS, * shows gene number.
- $*\_EI.txt$ exon or intron across genomic bin positions, * shows gene number.
- $Particle\_num.txt$: The number of particles for the particle filter: recommend more than 100,000.
- $SRR960177\_pickup\_gene\_num\_joint\_ei\_av\_mu\_tau\_sigma.txt$: hyper parameter for each gene, mu is the initial state of the variable of the state model, tau is noise of the system model, and sigma is noise of the measurement model (log). The details are described in our paper [32] and Section 3.5.
- $SRR960177\_pickup\_gene\_num\_joint\_ei\_av\_chrPosSE\_sig\_len5cor05\_cov01raw.txt$: gene number, chromosome number, gene name, strand, start, end. Here, one gene results in combining all isoforms in one.

It is possible to estimate the Pol II density in this mouse dataset by performing the following commands:

1. $gcc \quad -lm \quad Estimate\_Pol2.c$
2. $./a.out \quad /DIRECTORY\_input\_data/$

Chapter 5

Conclusion

We implemented a Bayesian framework for the reconstruction of transcription elongation rates from sawtooth-like observations derived from total RNA-seq. After forwardly modeling given 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 work, different splicing modes can bring different slopes to the read density, thereby drawing the wrong conclusion in the absence of inferring the splicing variations. One contribution of this work is to provide a way to estimate unmeasured states of elongation rates and splicing modes simultaneously.

As a byproduct of our method, RS sites could be identified. Quite a lot of valleys, possibly indicating ratchet points of RS, were found in the intronic regions in addition to those shown in Fig 5.1. 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 [21]. As shown in Fig 5.1, the splicing sites estimated by our method captured the five ratchet points reported in a previous study, though some seemingly false estimates of the splicing sites were also given.

This study focused on only 653 genes since intronic reads were considerably sparse in most other genes. Fig 5.2 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.

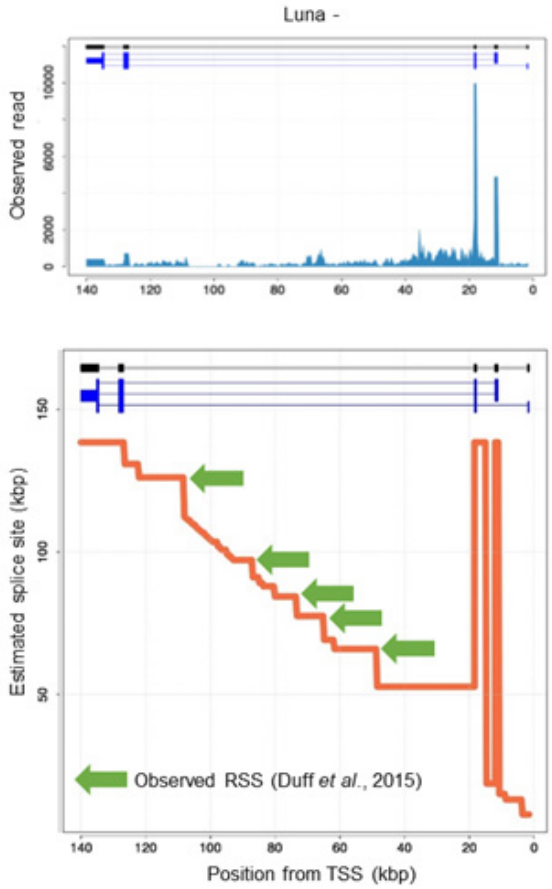

Fig. 5.1 The observed read density and splicing patterns are shown for the DNA coordinates of the Luna gene 5'-3' from the left to right.

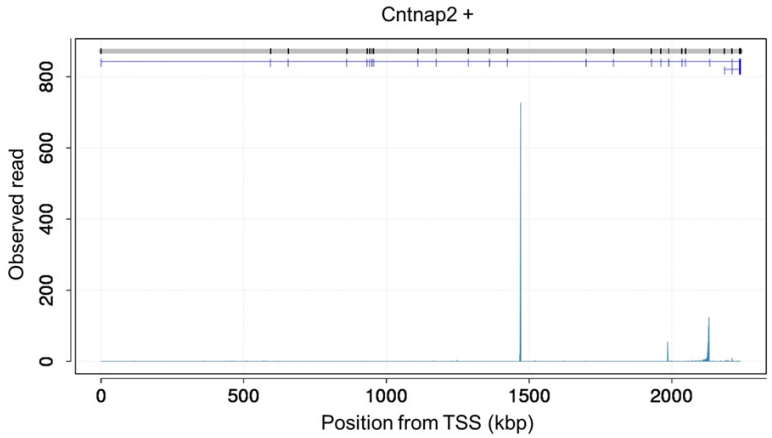

Fig. 5.2 The expected read density and splicing patterns are shown for the DNA coordinates of the Cntnap2 gene (mouse ES cells [56]) 5'-3' from the left to right.

Thoroughly as a byproduct, the inverse problem solution predicts RS sites from total RNA-seq reads represented as sawtooth patterns, where there are underlying transcription principles required to regulate the molecular mechanism. It is a worthwhile topic for exploring the theoretical side of a subject from real scientific data. It has actually been reported that many genes that have longer introns, in which RS occurs, are related to neurological diseases and autism [33, 39, 47].

However, the obtained data from total RNA-seq reads presents a challenge in analyzing clearly noisy data with low read coverage. At present, some false positive estimates of the splicing sites were also seen. To address this problem, the model must become more complex, and it clearly needs read density from deeper sequencing.

Acknowledgements

First, I would like to express my gratitude to my supervisor Ryo Yoshida, whose advice and suggestions were immensely valuable throughout the course of my study at The Graduate University for Advanced Studies (SOKENDAI) and The Institute of Statistical Mathematics (ISM). He led me to carry out a wonderful study that is a worthy accomplishment. Despite the various demands on his time and energy, he generously gave me his time and encouragement with our scientific endeavors. I hope that someday I will have the opportunity to do the same for future scientists.

I am indebted to the chair of my dissertation committee, Prof. Gen Ueno, whose comments made an enormous contribution to my study, and to the members of the committee, Prof. Rui Yamaguchi, Prof. Masayuki Henmi, and Prof. Shinsuke Koyama, who provided considerable and diverse feedback on my doctoral work.

I would like to thank Dr. Charles Boone, Prof. of the University of Toronto and the project leader of the Institute of Physical and Chemical Research (RIKEN), who gave me a chance to learn statistical science and machine learning as a Ph.D. candidate. I would also like to thank to my talented collaborator Dr. Chad Myers, Prof. of the University of Minnesota-Twin Cities (U of M) and a project member of RIKEN, who hosted my fruitful visit at U of M.

I am deeply grateful to the members of the Data Science Center for Creative Design and Manufacturing at ISM, who provided support and encouragement at various times in this study.

Finally, I would like to show my greatest appreciation to all my colleagues at SOKENDAI and ISM, people with whom I have spent so many joyful times throughout my life. My deepest gratitude goes to my family for their understanding while I pursued a Ph.D.

References

[1] Alexander, Ross, D., Innocente, Steven, A., Barrass, J, D., and Beggs, Jean, D. (2010). Splicing-dependent RNA polymerase pausing in yeast. *Molecular cell*, 40(4):582–593.

[2] Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol*, 18(12):1435–1440.

[3] Ardehali, M, B., Yao, J., Adelman, K., Fuda, N, J., Petesch, S, J., Webb, W, W., and Lis, J, T. (2009). Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J*, 28(8):1067–1077.

[4] Bentley, D, L. (2014). Coupling mRNA processing with transcription in time and space. *Nat Rev Genet*, 15(3):163–175.

[5] Bishop, C. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York.

[6] Boireau, S., Maiuri, P., Basyuk, E., de la Mata, M., Knezevich, A., Pradet-Balade, B., Backer, V., Kornblihtt, A., Marcello, A., and Bertrand, E. (2007). The transcriptional cycle of HIV-1 in real-time and live cells. *J Cell Biol*, 179(2):291–304.

[7] Bolić, M., Djurić P, M., and Hong, S. (2004). Resampling algorithms for particle filters : A computational complexity perspective. *EURASIP Journal on Applied Signal Processing*, 15:2267–2277.

[8] Brody, Y., Neufeld, N., Bieberstein, N., Causse, S, Z., Bohnlein, E, M., Neugebauer, K, M., Darzacq, X., and Shav-Tal, Y. (2011). The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol*, 9(1):e1000573.

[9] Brown, S, J., Stoilov, P., and Xing, Y. (2012). Chromatin and epigenetic regulation of pre-mRNA processing. *Hum Mol Genet*, 21(R1):R90–96.

[10] Buckley, M, S., Kwak, H., Zipfel, W, R., and Lis, J, T. (2014). Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev*, 28(1):14–19.

[11] Cairns, B. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261):193–198.

[12] Chae, M., Danko, C, G., and Kraus, W, L. (2015). groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, 16:222.

[13] Chiu, A., Suzuki, H., Wu, X., Mahat, D., Kriz, A., and Sharp, P. (2018). Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. *Mol Cell*, 69(4):648–663.e7.

[14] Churchman, L, S. and Weissman, J, S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373.

[15] Core, L., Waterfall, J., Gilchrist, D., Fargo, D., Kwak, H., Adelman, K., and Lis, J. (2012). Defining the status of rna polymerase at promoters. *Cell Rep*, 2(4):1025–1035.

[16] Creyghton, M, P., Cheng, A, W., Welstead, G, G., Kooistra, T., Carey, B, W., Steine, E, J., Hanna, J., Lodato, M, A., Frampton, G, M., Sharp, P, A., Boyer, L, A., Young, R, A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, 107(50):21931–21936.

[17] Danko, C, G., Hah, N., Luo, X., Martins, A, L., Core, L., Lis, J, T., Siepel, A., and Kraus, W, L. (2013). Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell*, 50(2):212–222.

[18] Darzacq, X., Shav-Tal, Y., de Turris, V., Brody, Y., Shenoy, S, M., Phair, R, D., and Singer, R, H. (2007). In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol*, 14(9):433–438.

[19] Doucet, A., De Freitas, N., and Gordon, N. (2001). *An introduction to sequential Monte Carlo methods*. Springer.

[20] Doucet, A. and Johansen, A, M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. *The Oxford Handbook of Nonlinear Filtering*, 12:656–704.

[21] Duff, M, O., Olson, S., Wei, X., Garrett, S, C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S, E., and Graveley, B, R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, 521(7552):376–379.

[22] Dujardin, G., Lafaille, C., de la Mata, M., Marasco, L., Munoz, M., Le Jossic-Corcos, C., Corcos, L., and Kornblihtt, A. (2014). How slow RNA polymerase II elongation favors alternative exon skipping. *Mol Cell*, 54(4):683–690.

[23] Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3):215–216.

[24] Flynn, R., Do, B., Rubin, A., Calo, E., Lee, B., Kuchelmeister, H., Rale, M., Chu, C., Kool, E., Wysocka, J., Khavari, P., and Chang, H. (2016). Sk-baf axis controls pervasive transcription at enhancers. *Nat Struct Mol Biol*, 23(3):231–238.

[25] Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome biology*, 15(5):R69.

[26] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[27] Hah, N., Danko, C, G., Core, L., Waterfall, J, J., Siepel, A., Lis, J, T., and Kraus, W, L. (2011). Examination of the possibility of a fluid-mechanics treatment of dense granular flows. *Cell*, 145(4):622–634.

[28] Ishihara, S. and Sugimura, K. (2012). Bayesian inference of force dynamics during morphogenesis. *J Theor Biol*, 313:201–211.

[29] Jonkers, I., Kwak, H., and Lis, J, T. (2014). Genome-wide dynamics of pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *elife*, 3:e02407.

[30] Jonkers, I. and Lis, J, T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol*, 16(3):167–177.

[31] Kaplan, N., Hughes, T., Lieb, J., Widom, J., and Segal, E. (2010). Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biol*, 11(11):140.

[32] Kawamura, Y., Koyama, S., and Yoshida, R. (2018). Statistical inference of the rate of rna polymerase ii elongation by total RNA sequencing. *Bioinformatics*, in press.

[33] King, I., Yandava, C., Mabb, A., Hsiao, J., Huang, H., Pearson, B., Calabrese, J., Starmer, J., Parker, J., Magnuson, T., Chamberlain, S., Philpot, B., and Zylka, M. (2013). Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 501(7465):58–62.

[34] Kitagawa, G. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association*, 82(400):1032–1041.

[35] Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, 46(4):605–623.

[36] Kitagawa, G. (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1):1–25.

[37] Kulaeva, O, I., Hsieh, F, K., Chang, H, W., Luse, D, S., and Studitsky, V. (2013). Mechanism of transcription through a nucleosome by RNA polymerase II. *Biochim Biophys Acta*, 1829(1):76–83.

[38] Kwak, H., Fuda, N, J., Core, L, J., and Lis, J, T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (New York, N.Y.)*, 339(6122):950–953.

[39] Lagier-Tourenne, C., Polymenidou, M., Hutt, K., Vu, A., Baughn, M., Huelga, S., Clutario, K., Ling, S., Liang, T., Mazur, C., Wancewicz, E., Kim, A. S., Watt, A., Freier, S., Hicks, G., Donohue, J., Shiue, L., Bennett, C., Ravits, J., Cleveland, D., and Yeo, G. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci*, 15(11):1488–1497.

[40] Lindberg, J. and Lundeberg, J. (2010). The plasticity of the mammalian transcriptome. *Genomics*, 95(1):1–6.

[41] Liu, J. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

[42] Luco, R, F., Allo, M., Schor, I, E., Kornblihtt, A, R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell*, 144(1):16–26.

[43] Marson, A., Levine, S, S., Cole, M, F., Frampton, G, M., Brambrink, T., Johnstone, S., Guenther, M, G., Johnston, W, K., Wernig, M., Newman, J., Calabrese, J, M., Dennis, L, M., Volkert, T, L., Gupta, S., Love, J., Hannett, N., Sharp, P, A., Bartel, D, P., Jaenisch, R., , and Young, R, A. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533.

[44] Martin, R, M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell Rep*, 134(3):521–533.

[45] Mason, P, B. and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol Cell*, 17(6):831–840.

[46] Nielsen, M. (2018). *Neural Networks and Deep Learning.[online]*. Neuralnetworksanddeeplearning.com.

[47] Polymenidou, M., Lagier-Tourenne, C., Hutt, K., Huelga, S., Moran, J., Liang, T., Ling, S., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J., Shiue, L., Bennett, C., Yeo, G., and Cleveland, D. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci*, 14(4):459–468.

[48] Qian, X., Ba, Y., Zhuang, Q., and Zhong, G. (2014). Rna-seq technology and its application in fish transcriptomics. *OMICS*, 18(2):98–110.

[49] Radle, B., Rutkowski, A, J., Ruzsics, Z., Friedel, C, C., Koszinowski, U, H., and Dolken, L. (2013). Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture. *Journal of visualized experiments : JoVE*, 78:e50195.

[50] Rodriguez, J., Menet, J, S., and Rosbash, M. (2012). Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture. *Mol Cell*, 47(1):27–37.

[51] Saponaro, M., Kantidakis, T., Mitter, R., Kelly, G, P., Heron, M., Williams, H., Soding, J., Stewart, A., and Svejstrup, J, Q. (2014). RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell*, 157(5):1037–1049.

[52] Segal, E. and Widom, J. (2009). What controls nucleosome positions? *Trends Genet*, 25(8):335–343.

[53] Shen, Y., Yue, F., McCleary, D., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V., and Ren, B. (2014). A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120.

[54] Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu Rev Biochem*, 81:65–95.

[55] Sibley, C, R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M, E., Hardy, J., Modic, M., Curk, T., Wilson, S, W., Plagnol, V., and Ule, J. (2015). Recursive splicing in long vertebrate genes. *Nature*, 7552(521):371–375.

[56] Sigova, A, A., Mullen, A, C., Molinie, B., Gupta, S., Orlando, D, A., Guenther, M, G., Almada, A, E., Lin, C., Sharp, P, A., Giallourakis, C, C., and Young, R, A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A*, 110(8):2876–2881.

[57] Singh, J. and Padgett, R, A. (2009). Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol*, 16(11):1128–1133.

[58] Skene, P., Hernandez, A., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. *elife*, 3:e02042.

[59] Smolle, M. and Workman, J. (2013). Transcription-associated histone modifications and cryptic transcription. *Biochim Biophys Acta*, 1829(1):84–97.

[60] Svetlov, V. and Nudler, E. (2013). Basic mechanism of transcription by RNA polymerase II. *Biochim Biophys Acta*, 1829(1):20–28.

[61] Tanny, J, C. (2014). Chromatin modification by the RNA polymerase II elongation complex. *Transcription*, 5(5):e988093.

[62] Teif, V, B., Vainshtein, Y., Caudron-Herger, M., Mallm, J, P., Marth, C., Hofer, T., and Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol*, 19(11):1185–1192.

[63] Tennyson, Christine, N., Klamut, Henry, J., and Worton, Ronald, G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature genetics*, 9(2):184.

[64] Teves, S, S., Weber, C, M., and Henikoff, S. (2014). Transcribing through the nucleosome. *Trends Biochem Sci*, 39(12):577–586.

[65] Tiberi, S., Walsh, M., Cavallaro, M., Hebenstreit, D., and Finkenstadt, B. (2018). Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics*, 34(17):i647–i655.

[66] Veloso, A., Kirkconnell, K, S., Magnuson, B., Biewen, B., Paulsen, M, T., Wilson, T, E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res*, 24(6):896–905.

[67] Wagner, E. and Carpenter, P. (2012). Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol*, 13(2):115–126.

[68] Wood, K., Tellier, M., and Murphy, S. (2018). DOT1L and H3K79 Methylation in Transcription and Genomic Stability. *Biomolecules*, 8(11):115–126.

[69] Wu, Q., Kim, Y., Lu, J., Xuan, Z., Chen, J., Zheng, Y., Zhou, T., Zhang, M., Wu, C., and Wang, S. (2008). Poly A- transcripts expressed in HeLa cells. *PLoS One*, 3(7):e2803.

[70] Yao, J., Munson, Katherine, M., Webb, Watt, W., and Lis, John, T. (2006). Dynamics of heat shock factor association with native gene loci in living cells. *Nature*, 442(7106):1050.

[71] Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. *EMBO Rep*, 16(11):1467–1481.

[72] Zhou, D. (2018). Deep distributed convolutional neural networks: Universality. *Analysis and Applications*, 16(6):895–919.

[73] Zhou, Q., Li, T., and Price, D. (2012). RNA polymerase II elongation control. *Annu Rev Biochem*, 81:119–143.