A Unified Probabilistic Modeling Framework for Eukaryotic Transcription Based on Nascent RNA Sequencing Data

Adam Siepel

Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY, USA

Abstract

Nascent RNA sequencing protocols, such as GRO-seq and PRO-seq, are now widely used in the study of eukaryotic transcription, and these experimental techniques have given rise to a variety of statistical and machine-learning methods for data analysis. These computational methods, however, are generally designed to address specialized signal-processing or prediction tasks, rather than directly describing the dynamics of RNA polymerases as they move along the DNA template. Here, I introduce a general probabilistic model that describes the kinetics of transcription initiation, elongation, pause release, and termination, as well as the generation of sequencing read counts. I show that this generative model enables estimation of separate rates of initiation, pause-release, and termination, up to a proportionality constant. Furthermore, if applied to time-course data in a nonequilibrium setting, the model can be used to estimate elongation rates. This model additionally leads naturally to likelihood ratio tests for differences between genes, conditions, or species in various rates of interest. A version of the model in which read counts are assumed to be Poisson-distributed leads to convenient, closed-form solutions for parameter estimates and likelihood ratio tests. I present extensions to Bayesian inference and to a generalized linear model that can be used to discover genomic features associated with rates of elongation. Finally, I address technicalities concerning estimation of library size, normalization and sequencing replicates. Altogether, this modeling framework enables a unified treatment of many common tasks in the analysis of nascent RNA sequencing data.

Introduction

In recent years, several experimental protocols have emerged for measuring newly produced RNAs on a genome-wide scale [1–5]. These *nascent RNA sequencing* methods can be thought of as adaptations of traditional nuclear run-on assays that exploit modern massively-parallel sequencing technologies to simultaneously map the positions of engaged RNA polymerases across an entire genome. They have proven useful in a wide variety of applications, including measurement of transcription levels independent of RNA decay [1], measurement of rates of elongation [6], mapping of transcription start sites [7], identification of active enhancers [8,9], and estimation of relative RNA half-lives [10]. These measurements have led, in turn, to a number of new insights into the biology of transcription, such as the widespread incidence of promoter-proximal pausing and divergent transcription [1, 2], the remarkably similar architecture of transcription initiation at promoters and enhancers [7], pause-release as a crucial rate-limiting step in transcription [11], and distinct waves of transcriptional responses to the induction of transcription [9].

A variety of computational and statistical methods have been used to analyze these data, ranging from relatively simple count- and ratio-based tests to hidden Markov models and discriminative machine learning approaches (e.g., [6, 8, 9, 12, 13]). Some of these methods are sophisticated, powerful, and widely used. For

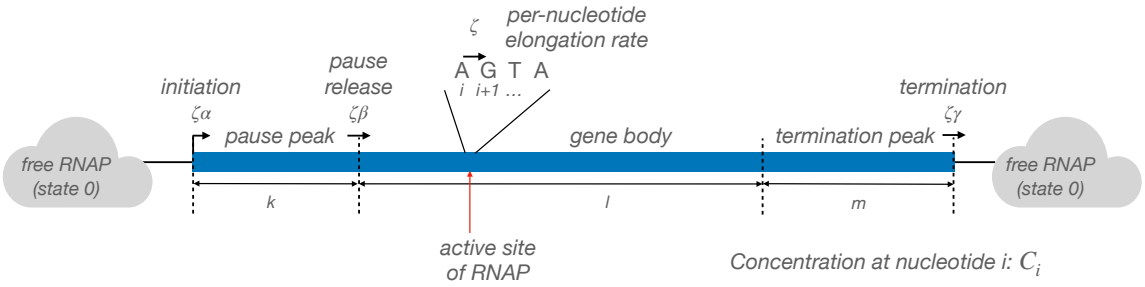

Figure 1: General kinetic model for transcription initiation, elongation, pause release, and termination along a transcription unit.

the most part, however, they have focused fairly narrowly on specific prediction tasks (such as elongation rate estimation or TSS prediction), rather than on more generally modeling the biophysical processes underlying transcription. As a result, with a few partial exceptions [14, 15], the existing computational methods for nascent RNA sequence data, or closely related data types, generally do not permit direct estimation of biophysical parameters of interest, such as rates of initiation or termination. In addition, owing to their heuristic description of the underlying processes, they can be difficult to adapt for nuances in the process, such as promoter-proximal pausing or variation in elongation rate along the gene body.

In this article, I introduce a unified framework that combines a simple kinetic model of transcription with a generative model for nascent RNA sequence data, and permits direct inference of kinetic parameters from sequence data. Separate layers in the model account for stochasticity in the process and noise in the sequencing data. I show that this model can be applied to a number of problems of interest, including estimation of initiation and elongation rates. It also leads directly to powerful statistical tests for differences in rates of interest, which naturally consider both the information in the read count data and the underlying physical process. The basic modeling framework is simple in structure and straightforward to apply but it has natural extensions to Bayesian inference and to a generalized linear model that accommodates covariates of elongation rate. Finally, I show how to address technical issues concerning estimation of library size, normalization, and sequencing replicates within this modeling framework. I conclude with a discussion of some current limitations and possible extensions of the current framework.

General Kinetic Model for Transcriptional Initiation, Elongation, Pause-Release, and Termination

Let us begin with a general kinetic model for the process by which RNA polymerase (RNAP) enzymes initiate transcription at the transcription start site (TSS) of a transcription unit (TU) and move along the DNA template, synthesizing an elongating nascent RNA molecule. This model is defined by a series of states corresponding to each nucleotide position of RNAP as it moves along the template (Figure 1). In particular, state i represents the positioning of the active site of RNAP at nucleotide i of the TU and corresponds to a nascent RNA of length $i-1$. We will assume that each RNAP eventually traverses the entire DNA template, borrowing from evidence suggesting that premature termination occurs at relatively low rates across most transcription units [16] (but see Discussion).

I will distinguish among three segments of the TU: (1) the first k nucleotides, known as the *pause peak*, where RNAP tends to accumulate owing to promoter-proximal pausing (typically $k \approx 50$) [1]; and (2) the subsequent l nucleotides, where RNAP tends to be relatively unimpeded, which is typically referred to as the *gene body*; and the final m nucleotides, where RNAP accumulates once again owing to delays from termination of transcription, which I will call the *termination peak* (see Figure 2). The TU is of length

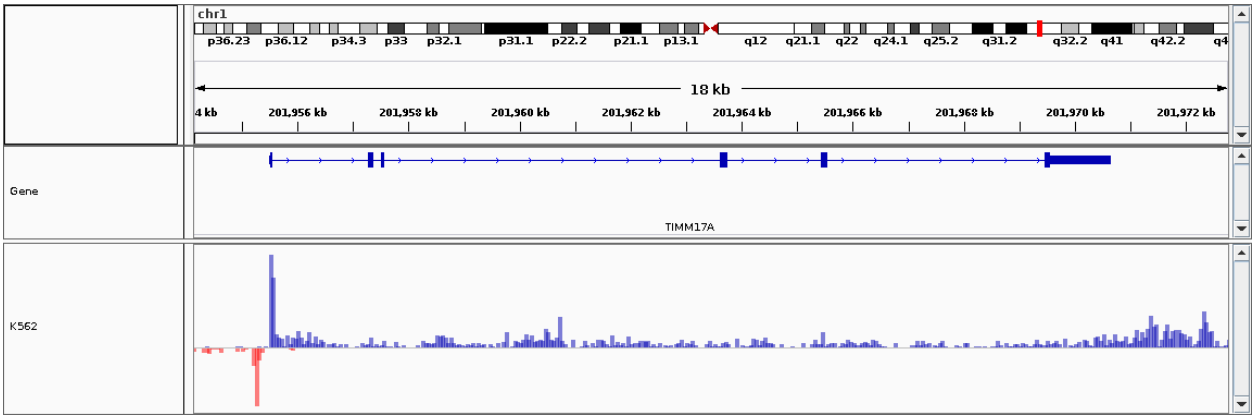

Figure 2: Example of real nascent RNA sequencing data in the region of the human *TIMM17A* gene. The data were obtained by applying PRO-seq to K562 cells, as described in ref. [9]. The pause peak and termination peak are clearly visible. Notice, however, that the read counts are quite noisy along the gene body. In addition, transcription in this case seems to have run past the annotated transcription termination site, a phenomonen that occurs frequently. As a result, the methods described in this paper would likely require some adjustment of annotated gene coordinates based on the raw data itself (e.g., see [13]).

$N = k + l + m$. The model has $N + 1$ states, corresponding to the N nucleotide positions of the TU plus an additional state (labeled 0) that abstractly represents *free RNAP*, that is, RNAP that is not currently engaged in transcription and is available for a new initiation.

I will assume, as with most existing protocols, that the data summarize a relatively large population of cells, and denote by C_i the concentration of cells in the sampled population having RNAP positioned at nucleotide i. As will be seen below, the sequencing reads from a PRO-seq experiment that map to position i at their $3'$ ends should have a depth roughly proportional to C_i. The situation is similar for other protocols.

The kinetic model is defined by four rates: an initiation rate α, a pause-release rate β, a termination rate γ, and a constant per-nucleotide elongation rate ζ. For mathematical convenience, I assume that the initiation, pause-release, and termination steps are coupled with single-nucleotide elongation steps and occur at rates $\zeta\alpha$, $\zeta\beta$, and $\zeta\gamma$, respectively. As a result, as long as ζ is the same across nucleotides, it can be considered a scaling factor that applies equally to all steps in the process. (Later we will consider a generalization that allows this rate to vary across nucleotides.) Finally, I will make the simplifying assumption that elongation upstream of the pause site occurs at rate $\zeta\beta$, rather than rate ζ, to capture the effect of RNAP "piling up" behind the pause site. In effect, polymerase pausing is assumed to be rate limiting throughout the entire pause peak, not just at the site of pausing. Similarly, elongation is assumed to occur at a rate of $\zeta\gamma$ throughout the termination peak.

With these assumptions, the concentrations C_i, for $i \in \{0, \ldots, N\}$, are governed by the following

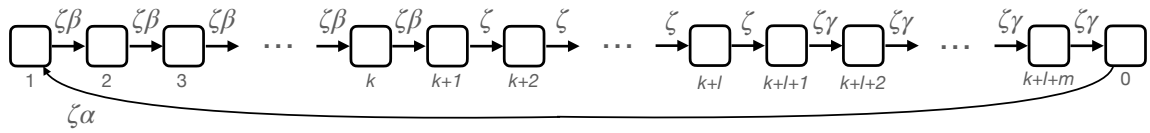

Figure 3: State-transition diagram for continuous-time Markov model for a single RNAP molecule moving along the DNA template.

system of $N+1$ differential equations,

$$
\begin{aligned}
\frac{d}{dt}C_1(t) &= \zeta\alpha C_0(t) - \zeta\beta C_1(t) \\
\frac{d}{dt}C_2(t) &= \zeta\beta C_1(t) - \zeta\beta C_2(t) \\
&\vdots \\
\frac{d}{dt}C_k(t) &= \zeta\beta C_{k-1}(t) - \zeta\beta C_k(t) \\
\frac{d}{dt}C_{k+1}(t) &= \zeta\beta C_k(t) - \zeta C_{k+1}(t) \\
\frac{d}{dt}C_{k+2}(t) &= \zeta C_{k+1}(t) - \zeta C_{k+2}(t) \\
&\vdots \\
\frac{d}{dt}C_{k+l}(t) &= \zeta C_{k+l-1}(t) - \zeta C_{k+l}(t) \\
\frac{d}{dt}C_{k+l+1}(t) &= \zeta C_{k+l}(t) - \zeta\gamma C_{k+l+1}(t) \\
\frac{d}{dt}C_{k+l+2}(t) &= \zeta\gamma C_{k+l+1}(t) - \zeta\gamma C_{k+l+2}(t) \\
&\vdots \\
\frac{d}{dt}C_{k+l+m}(t) &= \zeta\gamma C_{k+l+m-1}(t) - \zeta\gamma C_{k+l+m}(t) \\
\frac{d}{dt}C_0(t) &= \zeta\gamma C_{k+l+m}(t) - \zeta\alpha C_0(t).
\end{aligned}
$$

We will consider both the time evolution of the C_i values as well as their values at steady-state.

Continuous-time Markov Model

If instead of thinking in terms of a population of cells, we consider the probability distribution over states (nucleotide-specific positions) for a given RNAP, the same model can be represented using the convenient and flexible formalism of continuous-time Markov models. The continuous-time Markov version of the model consists of $N+1$ states and transition rates exactly analogous to the ones above (Figure 3). In this case, however, each state i corresponds to a binary random variable Z_i, indicating whether a given RNAP is ($Z_i = 1$) or is not ($Z_i = 0$) at a particular position at a particular time t.

It is easy to see that this model is identical to the previous one up to a normalization term. Given a sufficiently large collection of independent RNAP molecules, the overall concentration $C_i(t)$ is expected to

be proportional to the probability of occupancy $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ for any individual RNAP. Furthermore, for $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ to remain a proper probability distribution, the constant of proportionality must be given by $\frac{1}{\mathcal{Z}(t)}$ such that $\mathcal{Z}(t) = \sum_{i=0}^{N} C_i(t)$.

This Markov chain is defined by an $(N+1) \times (N+1)$ infinitesimal generator matrix, $\mathbf{Q} = \zeta \mathbf{R}$, such that,

$$\mathbf{R} = \begin{pmatrix} -\alpha & \alpha & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -\beta & \beta & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -\beta & \beta & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ & \vdots & & & & & \vdots & & & & & \vdots & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\beta & \beta & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ & \vdots & & & & & \vdots & & & & & \vdots & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -\gamma & \gamma & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -\gamma & \gamma & \cdots & 0 \\ & \vdots & & & & & \vdots & & & & & \vdots & & & & \vdots \\ \gamma & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -\gamma \end{pmatrix}.$$

The element at row j and column k of $\mathbf{Q}$ indicates the instantaneous rate at which transitions occur from state j to state k. By convention, the values along the diagonal are set such that the rows sum to zero. Because the states must be visited in a sequence, this matrix has a particularly simple form: it has nonzero terms only on the main diagonal, the diagonal immediately above it, and the single element at bottom left (which allows RNAP to "wrap around" from the last state to the first one).

The probability of transitioning from any state j to any other state k over a period of time $t \geq 0$ can be computed in the usual way as,

$$\mathbf{P}(t, \alpha, \beta, \gamma, \zeta) = e^{\mathbf{Q}t} = e^{\mathbf{R}\zeta t}, \tag{1}$$

where the element at row j and column k of the matrix $\mathbf{P}$ indicates the conditional probability of occupancy of state k at time t given occupancy of state j at time 0, and the matrix exponential is given by the Taylor expansion $e^{\mathbf{R}\zeta t} = \mathbf{I} + \mathbf{R}\zeta t + \frac{(\mathbf{R}\zeta t)^2}{2!} + \frac{(\mathbf{R}\zeta t)^3}{3!} + \ldots$ [17]. If transcription is initiated at time 0 (meaning that the initial state is state 0), then the distribution over states at a later time t is simply given by row $j = 0$ of $\mathbf{P}$. In particular, the element at the ith column of row $j = 0$ indicates the quantity I have denoted $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$.

Notice that the elongation rate ζ and the time t are nonidentifiable under these assumptions. As will be discussed below, one of these quantities must be known to estimate the other.

Generative Model for Sequence Data

In the analysis of nascent RNA sequencing data, we have the additional complication that the positions of RNAP molecules are not directly observed, but instead are indirectly indicated by aligned sequence reads. The read counts reflect a sampling process that probes only a relatively small subset of cells in the starting material. Moreover, this process can be biased by factors such as base composition or RNA secondary structure that can lead to subtle differences in the efficiency of various steps in library preparation, including DNA fragmentation, purification, and PCR amplification.

I will address this problem by adding a second layer to the model that describes the probabilistic process by which sequencing read counts are generated conditional on an underlying "true" density of RNAP at each

nucleotide. In this way, we obtain a full generative model for the observed sequence data that is defined by the parameters of both the kinetic model described above and these conditional distributions for sequencing reads, enabling inference of both sets of parameters from data.

Let μ_i be the expected read depth at position i. We will assume that μ_i is proportional to $P(Z_i|t, \alpha, \beta, \gamma, \zeta)$ (as computed from the continuous-time Markov chain), with a proportionality constant that depend on the sequencing depth, that is, with $\mu_i = \lambda P(Z_i|t, \alpha, \beta, \gamma, \zeta)$, where λ is a measure of sequencing depth. In addition, we will abstractly assume a general generating distribution, ψ, for read counts given μ_i and any other relevant parameters, which will be cumulatively denoted θ. Let the data be denoted $\mathbf{X} = (X_1, \ldots, X_N)$, where X_i representes the number of sequencing reads having their $3'$ end aligned to position i. Assuming independence of sequencing reads, the likelihood for the data in terms of the time t and the free parameters of the model can be defined as,

$$P(\mathbf{X} \,|\, t, \alpha, \beta, \gamma, \zeta, \theta) = \prod_{i=1}^{N} \psi(X_i \,|\, \mu_i, \theta) \tag{2}$$

where $\mu_i = \lambda P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ and $P(Z_i \,|\, t, \alpha, \beta, \gamma, \zeta)$ is obtained from the matrix $\mathbf{P}(t) = e^{\mathbf{Q}t}$, as described above. As will be seen below, λ can generally be preestimated from global properties of a sequencing data set. Notice that typically there will be a data point corresponding to each state of the model, with the exception of state 0, which is unobservable.

Nonequilibrium Case and Application to Time-Course Data

Suppose nascent RNA sequencing data have been collected in a time course, with read counts $\mathbf{X}^{(t)}$ for $t \in \{t_1, t_2, \ldots, t_M\}$. Typically $t_1 = 0$ and represents the case prior to induction of transcription (Figure 4) [6]. Assuming independence of the data for the time points, a joint likelihood for all data can be defined as,

$$P(\mathbf{X}^{(t_1)}, \ldots, \mathbf{X}^{(t_M)} \,|\, \alpha, \beta, \gamma, \zeta, \theta) = \prod_{j=1}^{M} \prod_{i=1}^{N} \psi(X_i^{(j)} \,|\, \mu_i^{(j)}, \theta), \tag{3}$$

where $\mu_i^{(j)} = \lambda^{(j)} P(Z_i \,|\, t_j, \alpha, \beta, \gamma, \zeta)$, analogous to the case above. Notice that we allow here for a time-point-specific sequencing depth, $\lambda^{(j)}$.

In this case, α, β, γ, and ζ can all be estimated from the data, at least in principle, although the information about some of them may be weak. However, there is a practical problem relating to polymerase positions that are impossible under the model (for example, when $t = 0$ and transcription has not yet commenced) but for which the corresponding read count X_i is nonzero simply owing to the background noise of the assay. This problem can be addressed by defining $\psi(X_i|\mu_i, \theta)$ by a suitable background distribution whenever $\mu_i = 0$ (or perhaps whenever $\mu_i < T$, where T is a background-derived threshold). In this way, it should be possible to estimate an elongation rate ζ for each transcription unit directly from time-course data by numerical optimization of equation 3. This approach would provide an alternative to current hidden Markov model-based inference methods and would have the advantage of allowing estimation of α and β as well. The termination rate γ may be more difficult to estimate, but it may be possible to do so with a sufficiently long time course.

As observed above, ζ and t cannot be separately estimated from data, because they appear in the likelihood only as the product ζt. However, in a time-course setting, t is known and thus ζ can be estimated.

Stationary Distribution and Inference at Steady-State

This Markov chain is ergodic and, therefore, will eventually reach a steady-state equilbrium defined by a unique stationary distribution over states. This stationary distribution, denoted $\boldsymbol{\pi}$, is invariant to ζ and can

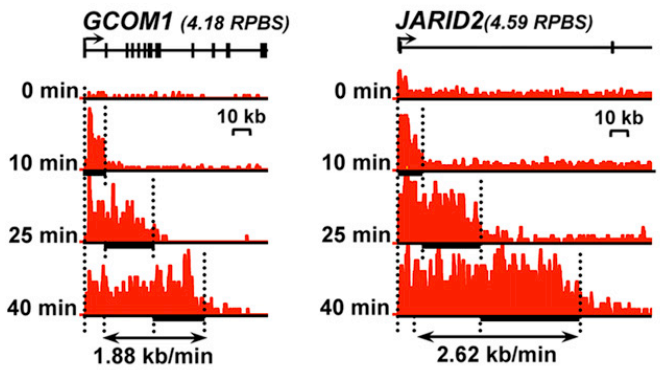

Figure 4: Examples of nascent RNA sequencing data collected in a time-course and used to estimate rates of elongation. Data were collected by GRO-seq at 0, 10, 25, and 40 minutes in MCF-7 cells [6]. Results are shown for the *GCOM1* (left) and *JARID2* (right) genes. In both cases, the left edges of the *x*-axes correspond to the annotated TSS. The RNAP "wave" can be seen to move rightward as time advances. Figure adapted from reference [6].

be found by solving the equation $\boldsymbol{\pi}(\mathbf{R}+\mathbf{I}) = \boldsymbol{\pi}$, or equivalently, $\boldsymbol{\pi}\mathbf{R} = \mathbf{0}$. Owing to the simple structure of $\mathbf{R}$, the solution for $\boldsymbol{\pi}$ can easily be determined by substitution.

Because state 0 is simply an abstraction to allow for the recirculation of RNAP and cannot be measured, we will ignore its stationary probability and instead describe the stationary distribution conditional on RNAP occupancy along the DNA template. This conditional stationary distribution is given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ such that,

$$\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{\alpha}{\beta} & i \in \{1, \ldots, k\} \\ \alpha & i \in \{k+1, \ldots, k+l\} \\ \frac{\alpha}{\gamma} & i \in \{k+l+1, \ldots, N\} \end{cases} \tag{4}$$

with normalization constant $\mathcal{Z} = k\frac{\alpha}{\beta} + l\alpha + m\frac{\alpha}{\gamma}$.

This distribution has an intuitive interpretation. First, note that it will typically be the case that $\beta, \gamma < 1$, owing to slowdowns in elongation from pausing and termination. By contrast, RNAP should proceed unimpeded in the gene body, and therefore its density at steady state should reflect the initiation rate. Accordingly, π_i is proportional to α in the gene body, but elevated by factors of $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ in the pause peak and termination peak, respectively (Figure 5). The quantity $\frac{1}{\beta}$ is sometimes estimated from nascent RNA sequencing data and described as the *pause index*. By analogy, the quantity $\frac{1}{\gamma}$ can be described as a *termination index*.

Notice that I have chosen to write the elements of $\boldsymbol{\pi}$ such that α appears in each numerator, despite that these α terms would cancel with the denominator and hence could be omitted. The reason for this choice will become clear later in the article when multiple genes are considered, each with its own version of α. The inclusion of α in the stationary distribution will allow these values to be compared across genes.

The stationary distribution $\boldsymbol{\pi}$ can be used to define a likelihood function for a system in steady state, described by a single data set (as opposed to a time course), denoted $\mathbf{X} = (X_1, \ldots, X_N)$. In this case, we can set the expected read depth at each site i to $\mu_i = \lambda\pi_i$, and calculate the likelihood of the data as (cf.

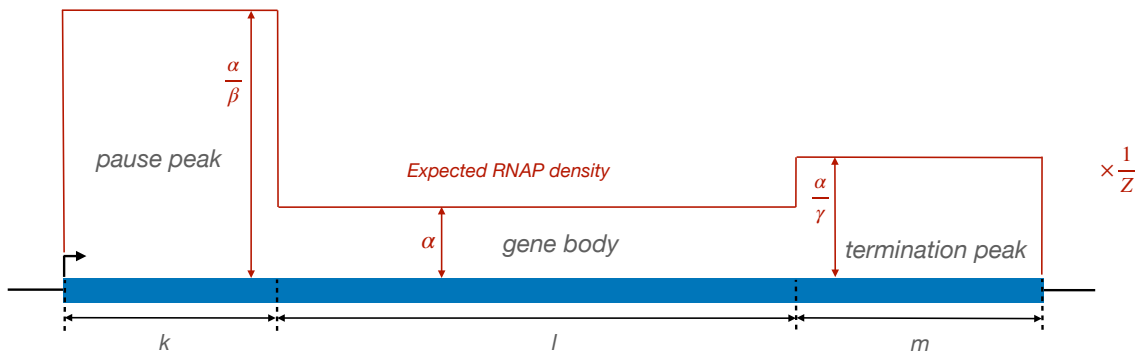

Figure 5: Expected density of RNAP at steady-state under the full continuous-time Markov model. The simplified four-state model has the same expectation but with a different normalization constant $\left(\frac{1}{Z}\right)$.

equation 2),

$$
\begin{aligned}
P(\mathbf{X}|\alpha,\beta,\gamma,\theta,\lambda) &= \prod_{i=1}^{N} \psi(X_i \,|\, \mu_i, \theta) \\
&= \prod_{i=1}^{k} \psi\left(X_i \,\middle|\, \frac{\lambda\alpha}{\beta}, \theta\right) \times \prod_{i=k+1}^{k+l} \psi\left(X_i \,|\, \lambda\alpha, \theta\right) \times \prod_{i=k+l+1}^{N} \psi\left(X_i \,\middle|\, \frac{\lambda\alpha}{\gamma}, \theta\right)
\end{aligned} \tag{5}
$$

where the normalization constant for $\boldsymbol{\pi}$ is assumed to be absorbed by λ.

Reduced Four-State Markov Model

It is evident from the preceding section that, in the steady-state case, the model makes no distinction among the states corresponding to each of the three major segments of the transcription unit—the pause peak, gene body, and termination peak. Therefore, in this case it is natural to work instead with a much simpler four-state model, consisting of a single state for each of these segments and a state for free RNAP. The transitions among these states will again be described using parameters labeled α, β, and γ, but notice that the new parameters are not directly comparable to the previous ones, because they represent rates of transitioning between segments of a TU (including elongation within each segment), not nucleotide-specific rates (see Figure 6).

This model has transition matrix,

$$
Q_{\text{reduced}} = \begin{pmatrix} -\alpha & \alpha & 0 & 0 \\ 0 & -\beta & \beta & 0 \\ 0 & 0 & -1 & 1 \\ \gamma & 0 & 0 & -\gamma \end{pmatrix}. \tag{6}
$$

Its conditional stationary distribution is given by $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ such that

$$
\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{\alpha}{\beta} & i = 1 \\ \alpha & i = 2 \\ \frac{\alpha}{\gamma} & i = 3, \end{cases} \tag{7}
$$

with normalization constant $\mathcal{Z} = \frac{\alpha}{\beta} + \alpha + \frac{\alpha}{\gamma}$ (see Figure 5).

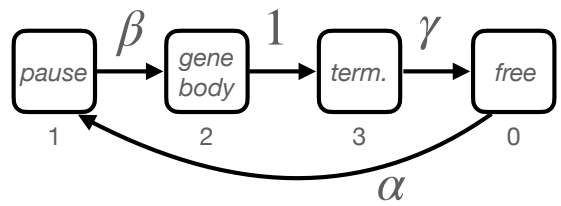

Figure 6: State-transition diagram for the reduced four-state model.

The likelihood of a data set reflecting a system at steady state is simply,

$$P(\mathbf{X}|\alpha,\beta,\gamma,\theta,\lambda) = \psi\left(X_1,\ldots,X_k \,|\, \frac{\lambda\alpha}{\beta},\theta\right) \times \psi\left(X_{k+1},\ldots,X_{k+l} \,|\, \lambda\alpha,\theta\right) \times \psi\left(X_{k+l+1},\ldots,X_N \,|\, \frac{\lambda\alpha}{\gamma},\theta\right), \tag{8}$$

where I implicitly assume ψ has been modified to compute a density for a sequence of read counts, rather than just a single nucleotide count.

The Poisson Case

In some applications, the generating distribution ψ may need to be complex to capture realistic patterns in sequencing reads, but in others, it may be sufficient to assume a simple form for this distribution. The assumption of a Poisson distribution at each nucleotide for ψ leads to particularly straightforward closed-form estimators in many applications of interest.

If a Poisson generating distribution is combined with the four-state model at steady state, the likelihood becomes simply,

$$\begin{aligned} P(\mathbf{X}|\alpha,\beta,\gamma,\lambda) &= \prod_{i=1}^{k} \text{Pois}\left(X_i \,|\, \frac{\lambda\alpha}{\beta}\right) \times \prod_{i=k+1}^{k+l} \text{Pois}\left(X_i \,|\, \lambda\alpha\right) \times \prod_{i=k+l+1}^{N} \text{Pois}\left(X_i \,|\, \frac{\lambda\alpha}{\gamma}\right) \\ &= \prod_{i=1}^{k} \frac{\left(\frac{\lambda\alpha}{\beta}\right)^{X_i} e^{-\frac{\lambda\alpha}{\beta}}}{X_i!} \times \prod_{i=k+1}^{k+l} \frac{(\lambda\alpha)^{X_i} e^{-\lambda\alpha}}{X_i!} \times \prod_{i=k+l+1}^{N} \frac{\left(\frac{\lambda\alpha}{\gamma}\right)^{X_i} e^{-\frac{\lambda\alpha}{\gamma}}}{X_i!} \\ &= \frac{1}{\mathcal{Z}} \left(\frac{\lambda\alpha}{\beta}\right)^{s_P} (\lambda\alpha)^{s_B} \left(\frac{\lambda\alpha}{\gamma}\right)^{s_T} \exp\left[-\lambda\alpha\left(\frac{k}{\beta}+l+\frac{m}{\gamma}\right)\right] \end{aligned} \tag{9}$$

where $s_P = \sum_{i=1}^{k} X_i$, $s_B = \sum_{i=k+1}^{k+l} X_i$, and $s_T = \sum_{i=k+l+1}^{N} X_i$ are sufficient statistics for the pause peak (P), the gene body (B), and the termination peak (T), respectively, and $\mathcal{Z} = \prod_{i=1}^{N} X_i!$ is a normalization term that does not depend on the model parameters and can be ignored during optimization. The log likelihood is given by,

$$\ell(\mathbf{X};\alpha,\beta,\gamma,\lambda) = s_P \log\left(\frac{\lambda\alpha}{\beta}\right) + s_B \log(\lambda\alpha) + s_T \log\left(\frac{\lambda\alpha}{\gamma}\right) - \lambda\alpha\left(\frac{k}{\beta}+l+\frac{m}{\gamma}\right) - \log\mathcal{Z}. \tag{10}$$

Notice that the terms λ and α always occur as a product in equations 9 and 10. However, as will be shown below, the read-depth parameter λ can be estimated in a preprocessing step, allowing separate estimation

of α. With this observation, the maximum likelihood estimators for α, β, and γ have simple closed-form solutions:

$$\hat{\alpha} = \frac{s_B}{l\lambda} \tag{11}$$

$$\hat{\beta} = \frac{s_B/l}{s_P/k} \tag{12}$$

$$\hat{\gamma} = \frac{s_B/l}{s_T/m}. \tag{13}$$

Thus, in this case, the estimator for α is simply the average read depth in the gene body divided by λ, and the estimators for β and γ are ratios of average read depths of the gene body to the pause peak and the termination peak, respectively. Notably, these are estimators that have been widely used in the analysis of nascent RNA sequencing data, with more heuristic justifications.

While the nucleotide-specific counts in real data sets tend to be overdispersed, the simplicity and speed of the Poisson model may still make it an attractive option in many analyses. For example, as will be sees below, it may be desirable to perform statistical tests based on the Poisson model and then adjust *p*-values using an empirical calibration.

Allowing for Overdispersion

In general, read counts in nascent RNA sequencing data tend to be overdispersed (with variance exceeding the mean), just as in RNA-seq data. A relatively straightforward way to allow for overdispersion is to assume a mixture of Poisson rates within each region while maintaining the assumption of independence across nucleotide sites. In this way, some sites are allowed to accumulate reads at higher-than-average rates, and others at lower-than-average rates. If the mixture over rates is assumed to be Gamma-distributed, then this assumption simply implies that the site-specific Poisson distributions for read counts are replaced with negative binomial distributions.

In particular, if we assume that the read depth at each nucleotide i is scaled by a random variable ρ_i, which is Gamma-distributed with shape parameter A and rate parameter $\frac{1}{A}$ (so that $E[\rho_i] = 1$), and we maintain the assumption of independent Poisson read counts as in equation 9, then we obtain a likelihood function of,

$$\begin{aligned}
P(\mathbf{X} \mid \alpha, \beta, \gamma, \lambda, A) = &\prod_{i=1}^{k} \left[\int \text{Gamma}\left(\rho_i \mid A, \frac{1}{A}\right) \text{Pois}\left(X_i \mid \frac{\lambda\alpha}{\beta}\right) d\rho_i \right] \\
&\times \prod_{i=k+1}^{k+l} \left[\int \text{Gamma}\left(\rho_i \mid A, \frac{1}{A}\right) \text{Pois}\left(X_i \mid \lambda\alpha\right) d\rho_i \right] \\
&\times \prod_{i=k+l+1}^{N} \left[\int \text{Gamma}\left(\rho_i \mid A, \frac{1}{A}\right) \text{Pois}\left(X_i \mid \frac{\lambda\alpha}{\gamma}\right) d\rho_i \right] \\
= &\prod_{i=1}^{k} \text{NB}\left(X_i \mid A, \frac{A\lambda\alpha}{A\lambda\alpha + \beta}\right) \prod_{i=k+1}^{k+l} \text{NB}\left(X_i \mid A, \frac{A\lambda\alpha}{A\lambda\alpha + 1}\right) \\
&\times \prod_{i=k+l+1}^{N} \text{NB}\left(X_i \mid A, \frac{A\lambda\alpha}{A\lambda\alpha + \gamma}\right),
\end{aligned} \tag{14}$$

where $\text{NB}(X \mid r, p)$ indicates that X has a negative binomial distribution with parameters r and p, that is,

such that,

$$p(X = x \mid r, p) = \binom{x + r - 1}{r - 1}(1 - p)^x p^r. \tag{15}$$

In this case, however, there are no closed-form expressions for the maximum likelihood estimates of the parameters, and they would need to be estimated numerically or by expectation maximization. For this reason, I will focus on the Poisson model in many of the sections below (but see Discussion).

Likelihood Ratio Tests for Differences Between Transcription Units, Conditions, or Species

These likelihood functions lead naturally to a class of likelihood ratio tests (LRTs) for comparing aspects of the transcriptional dynamics of different TUs, or of the same TU under different conditions. For example, one might test whether two genes have different initiation rates, different pause-release rates, or different termination rates. In each case, it might be desirable to allow some parameters to vary freely, while testing for evidence in the data of a difference in another parameter; for example, one might test for a difference in the pause-release rate allowing for differences in the initiation rate. I will focus on tests based on steady-state data, but one could equally well conduct a test for a difference in elongation rate based on time courses representing different conditions.

These LRTs all have the same general form. Consider read counts captured at steady-state for two TUs, $\mathbf{X}^{(1)} = (X_1^{(1)}, \ldots, X_{N_1}^{(1)})$ and $\mathbf{X}^{(2)} = (X_1^{(1)}, \ldots, X_{N_2}^{(2)})$. As noted, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ could represent the same TU under different conditions, different TUs under the same conditions, or even different TUs under different conditions. In a comparative genomic analysis, they may represent orthologous TUs in different species (say, human and chimpanzee). Let the free parameters for the two genes be denoted $\Theta_1 = (\alpha_1, \beta_1, \gamma_1, \lambda_1, \theta_1)$ and $\Theta_2 = (\alpha_2, \beta_2, \gamma_2, \lambda_2, \theta_2)$, respectively (ignoring the ζ parameters for now, and focusing on the steady-state). Finally, let the log likelihood of a particular data set be denoted $\ell(X; \Theta)$, and let the maximized value of that likelihood be denoted $\hat{\ell}(X; \Theta) = \max_\Theta \ell(X; \Theta)$.

In general, we will compare a null hypothesis H_0 that the two data sets can be jointly described by (at least some) shared parameters against an alternative hypothesis H_A that they require different parameters. The parameter can be partitioned into three mutually exclusive and exhaustive categories:

- *Fully tied* parameters (Θ_{T}). These parameters are shared by both TUs under the null and alternative hypotheses.
- *Fully free* parameters (Θ_{F1} and Θ_{F2}). These parameters vary freely in the two data sets under both the null and alternative hypotheses.
- *Tested* parameters (Θ_* or Θ_{*1} and Θ_{*2}). These parameters vary freely under the alternative hypothesis but are tied under the null hypothesis.

With these assumptions, the test statistic for an LRT for a difference in the tested parameters can be defined generally as,

$$T = \hat{\ell}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}; \Theta_{\text{T}}, \Theta_{\text{F1}}, \Theta_{\text{F2}}, \Theta_{*1}, \Theta_{*2}) - \hat{\ell}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}; \Theta_{\text{T}}, \Theta_{\text{F1}}, \Theta_{\text{F2}}, \Theta_*) \tag{16}$$

Here, each term represents the joint likelihood of the two data sets. Generally, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ will be assumed to be conditionally independent given the parameters, so that their joint likelihood is a product of their individual likelihoods; however, I present them in this more general form to emphasize the need to estimate tied parameters jointly. In the usual manner, the first term represents the maximized likelihood under the alternative hypothesis and the second term represents the maximized likelihood under the null hypothesis. If the tested parameters, Θ_{*1} and Θ_{*2}, are completely unconstrainted under the alternative hypothesis, T

represents a two-sided test. However, they can be constrained, if desired, to achieve a one-sided test (for example, allowing only $\alpha_2 \geq \alpha_1$ to test for an increased initiation rate in the second TU).

To make this framework concrete, let us consider some specific examples of tests. First, suppose we wish to test for a difference in initiation rate between two TUs from the same data set. In this case, $\Theta_{\text{T}} = \{\lambda, \theta\}$, $\Theta_{\text{F1}} = \{\beta_1, \gamma_1\}$, $\Theta_{\text{F2}} = \{\beta_2, \gamma_2\}$, and $\Theta_* = \{\alpha\}$.

Next, suppose we wish to test for a difference in the pause-release rate between two TUs same data set. In this case, $\Theta_{\text{T}} = \{\lambda, \theta\}$, $\Theta_{\text{F1}} = \{\alpha_1, \gamma_1\}$, $\Theta_{\text{F2}} = \{\alpha_2, \gamma_2\}$, and $\Theta_* = \{\beta\}$.

Next, suppose we wish to test for a difference in initiation rate in the same TU under two conditions, represented by separately sequenced nascent RNA sequencing data sets. Here, $\Theta_{\text{T}} = \{\theta\}$, $\Theta_{\text{F1}} = \{\beta_1, \gamma_1, \lambda_1\}$, $\Theta_{\text{F2}} = \{\beta_2, \gamma_2, \lambda_2\}$, and $\Theta_* = \{\alpha\}$, where the data sets are assumed to be sufficiently similar that the same generating distribution for read counts can be used for both (therefore θ can be tied).

Finally, suppose we wish to test for a difference in termination rate in the same TU under two conditions. In this case, $\Theta_{\text{T}} = \{\theta\}$, $\Theta_{\text{F1}} = \{\alpha_1, \beta_1, \lambda_1\}$, $\Theta_{\text{F2}} = \{\alpha_2, \beta_2, \gamma_2, \lambda_2\}$, and $\Theta_* = \{\gamma\}$.

In general, for two-sided tests, the asympotic distribution of $2T$ under the null hypothesis will be a chi-square distribution with $|\Theta_*|$ degrees of freedom, enabling convenient calculation of nominal *p*-values. In many cases, however, it may be preferred to use empirical *p*-values instead, for example, by computing T for a collection of TUs believed to be representative of the null hypothesis, and using this empirical distribution in place of the null.

Poisson-based Likelihood Ratio Tests

In the case of the Poisson generating distribution and the four-state model at steady state, it is often possible to find closed-form solutions for the test statistic T. For example, in the test for differences in initiation rate ($H_0 : \alpha_1 = \alpha_2$ vs. $H_A : \alpha_1 \neq \alpha_2$), T can be expressed as,

$$\begin{aligned} T = s_{B,1} &\left[\log\left(\frac{s_{B,1}}{l_1}\right) - \log\left(\frac{\lambda_1(s_{B,1} + s_{B,2})}{l_1\lambda_1 + l_2\lambda_2}\right)\right] \\ &+ s_{B,2}\left[\log\left(\frac{s_{B,2}}{l_2}\right) - \log\left(\frac{\lambda_2(s_{B,1} + s_{B,2})}{l_1\lambda_1 + l_2\lambda_2}\right)\right], \end{aligned} \tag{17}$$

where $s_{B,i}$ denotes the sum of read counts and l_i denotes the length of the gene body in TU i. Thus, the LRT for a difference in α depends only on the data in the gene body, not on the pause peak or the termination peak.

When $\lambda_1 = \lambda_2$, as when comparing the read counts for two different TUs from the same data set, this expression reduces to,

$$\begin{aligned} T = s_{B,1} &\left[\log\left(\frac{s_{B,1}}{l_1}\right) - \log\left(\frac{s_{B,1} + s_{B,2}}{l_1 + l_2}\right)\right] \\ &+ s_{B,2}\left[\log\left(\frac{s_{B,2}}{l_2}\right) - \log\left(\frac{s_{B,1} + s_{B,2}}{l_1 + l_2}\right)\right]. \end{aligned} \tag{18}$$

Furthermore, when in addition $l_1 = l_2$, as when considering the reads only within a fixed interval in the gene body, T reduces further to an expression that depends only on $s_{B,1}$ and $s_{B,2}$:

$$T = s_{B,1}\log s_{B,1} + s_{B,2}\log s_{B,2} - (s_{B,1} + s_{B,2})\log\frac{s_{B,1} + s_{B,2}}{2}. \tag{19}$$

In this way, T can be seen to be a measure of discordance in read depth, normalized for differences in library size and gene-body length. Notice in particular that, when $\frac{s_{B,1}}{l_1} = \frac{s_{B,2}}{l_2}$ in equation 18, or $s_{B,1} = s_{B,2}$ in equation 19 (i.e., when the average read depths are the same in the two gene bodies), then $T = 0$ and

the null hypothesis cannot be rejected. T grows larger as these average read depths become more different from one another. Importantly, however, T depends on the raw read counts, not only on their ratios. Thus, T tends to be small when the $s_{B,i}$ values are small, and to grow larger as they increase, reflecting greater confidence in the rejection of the null hypothesis with more data.

The test statistic for a difference in pause-release rates β_1 and β_2, allowing for differences in initiation and termination rates, is messier but can still be expressed in closed form as,

$$\begin{aligned} T = {} & s_{P,1} \log s_{P,1} + s_{B,1} \log s_{B,1} + s_{P,2} \log s_{P,2} + s_{B,2} \log s_{B,2} \\ & - (s_{P,1} + s_{P,2}) \log (s_{P,1} + s_{P,2}) - (s_{B,1} + s_{B,2}) \log (s_{B,1} + s_{B,2}) \\ & - (s_{P,1} + s_{B,1}) \log (s_{P,1} + s_{B,1}) - (s_{P,2} + s_{B,2}) \log (s_{P,2} + s_{B,2}) \\ & + (s_{P,1} + s_{B,1} + s_{P,2} + s_{B,2}) \log (s_{P,1} + s_{B,1} + s_{P,2} + s_{B,2}) , \end{aligned} \tag{20}$$

where, for simplicity, I focus on the case of shared $\lambda_1 = \lambda_2$, $k_1 = k_2$, $l_1 = l_2$, and $m_1 = m_2$. Here, it can be shown easily that $T = 0$ if the ratios of read counts at the pause peak and the gene body are equal, i.e., if $\frac{s_{P,1}}{s_{B,1}} = \frac{s_{P,2}}{s_{B,2}}$, and that T grows larger as these ratios become more divergent, but in a manner that depends not only on the ratios but also on the absolute counts.

Analogously, the test statistic for a difference in termination rates γ_1 and γ_2, allowing for a difference in initiation and pause-release rates is,

$$\begin{aligned} T = {} & s_{B,1} \log s_{B,1} + s_{T,1} \log s_{T,1} + s_{B,2} \log s_{B,2} + s_{T,2} \log s_{T,2} \\ & - (s_{B,1} + s_{B,2}) \log (s_{B,1} + s_{B,2}) - (s_{T,1} + s_{T,2}) \log (s_{T,1} + s_{T,2}) \\ & - (s_{B,1} + s_{T,1}) \log (s_{B,1} + s_{T,1}) - (s_{B,2} + s_{T,2}) \log (s_{B,2} + s_{T,2}) \\ & + (s_{B,1} + s_{T,1} + s_{B,2} + s_{T,2}) \log (s_{B,1} + s_{T,1} + s_{B,2} + s_{T,2}) , \end{aligned} \tag{21}$$

again assuming $\lambda_1 = \lambda_2$, $k_1 = k_2$, $l_1 = l_2$, and $m_1 = m_2$. In this case, $T = 0$ if $\frac{s_{T,1}}{s_{B,1}} = \frac{s_{T,2}}{s_{B,2}}$.

Bayesian Inference Under the Poisson Model

Another advantage of the Poisson model at steady-state is that it leads to relatively straightforward Bayesian inference of the key model parameters, α, β, and γ. To simplify the mathematics, I will introduce a change of variables from β to $\omega = \frac{1}{\beta}$ and from γ to $\nu = \frac{1}{\gamma}$, and then assume Gamma priors for α, ω, and ν:

$$\begin{aligned} \alpha &\sim \text{Gamma}(a, b), \\ \omega &\sim \text{Gamma}(A, B), \\ \nu &\sim \text{Gamma}(\mathcal{A}, \mathcal{B}), \end{aligned}$$

where the first parameter of each Gamma distribution represents its shape and the second parameter represents its rate (inverse scale).

With these assumptions, the joint posterior distribution of the parameters is given, up to a normalization constant, by:

$$\begin{aligned} P(\alpha, \omega, \nu \,|\, \mathbf{X}, \lambda) &\propto P(\alpha \,|\, a, b)\, P(\omega \,|\, A, B)\, P(\nu \,|\, \mathcal{A}, \mathcal{B})\, P(\mathbf{X} \,|\, \alpha, \omega, \nu, \lambda) \\ &\propto \alpha^{a-1} e^{-b\alpha} \omega^{A-1} e^{-B\omega} \eta^{\mathcal{A}-1} e^{-\mathcal{B}\eta} \left(\lambda \alpha k \omega\right)^{s_P} \left(\lambda \alpha l\right)^{s_B} \left(\lambda \alpha m \eta\right)^{s_T} e^{-\lambda\alpha(k\omega + l + m\eta)} \\ &= \alpha^{a+s_P+s_B+s_T-1} \omega^{A+s_P-1} \eta^{\mathcal{A}+s_T-1} e^{-\lambda\alpha(k\omega+l+m\eta)}. \end{aligned} \tag{22}$$

Because α is entangled with ω and ν in the last term of equation 22, the marginal posterior distributions for the individual parameters do not have standard forms. However, one can sample from the joint posterior

distribution by rejection sampling or sampling-importance resampling using the unnormalized density in equation 22.

In addition, it can be shown that the conditional distribution of each parameter given the others reduces to a standard Gamma distribution:

$$P(\alpha \,|\, \mathbf{X}, \lambda, \omega, \eta) = \text{Gamma}(a + s_P + s_B + s_T, b + \lambda(k\omega + l + m\eta)) \tag{23}$$

$$P(\omega \,|\, \mathbf{X}, \lambda, \alpha, \eta) = \text{Gamma}(A + s_P, B + \lambda\alpha k) \tag{24}$$

$$P(\eta \,|\, \mathbf{X}, \lambda, \alpha, \omega) = \text{Gamma}(\mathcal{A} + s_T, \mathcal{B} + \lambda\alpha m) \tag{25}$$

As a result, it is straightforward to implement a Gibbs sampling algorithm to sample from the joint posterior, from which marginal posterior distributions can easily be approximated.

Using this framework, it would be possible to calculate Bayes factors for model comparison, analogous to the LRTs introduced above.

Generalization to Site-Specific Elongation Rates

The full N-state continuous-time Markov model introduced at the beginning of this article can easily be generalized to allow for a different elongation rate ζ_i at each nucleotide position i, instead of a constant rate across all nucleotides. In this case, the elongation rate no longer serves as a simple scaling factor for the infinitesimal generator matrix $\mathbf{Q}$, and there is potentially information in the data about site-specific elongation rates, even at steady-state.

With site-specific elongation rates ζ_i, the conditional stationary distribution becomes (cf. equation 4):

$$\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{\alpha}{\beta\zeta_i} & i \in \{1, \ldots, k\} \\ \frac{\alpha}{\zeta_i} & i \in \{k+1, \ldots, k+l\} \\ \frac{\alpha}{\gamma\zeta_i} & i \in \{k+l+1, \ldots, N\} \end{cases} \tag{26}$$

with normalization constant,

$$\mathcal{Z} = \sum_{i=1}^{k} \frac{\alpha}{\beta\zeta_i} + \sum_{i=k+1}^{k+l} \frac{\alpha}{\zeta_i} + \sum_{i=k+l+1}^{N} \frac{\alpha}{\gamma\zeta_i}. \tag{27}$$

Notice that I have maintained the parameters α, β, and γ here, despite that they appear only as compound parameters together with the site-specific elongation rates and hence are nonidentifiable. The reason will become clear in the next section.

A Generalized Linear Model for Discovering Features Associated with Elongation

In practice, nascent RNA sequencing data tends to be quite noisy at individual nucleotides, and it is not likely to be useful for direct inference of each ζ_i value. However, it may be possible to learn something about the site-specific elongation process by pooling information across many sites and many TUs.

In particular, let us define a generalized linear model to make use of a collection of generic features at each transcribed nucleotide across the genome. These could could include any property known or suspected to influence elongation rates, including, for example, the degree of chromatin accessibility in a cell type of interest (assayed by DNase-seq or ATAC-seq), proximity to a splice site, presence of nucleosomes (MNase-seq), presence of stem-loops or other secondary structures of RNAs, or local G+C composition. Suppose there are D such features, summarized by a vector $\mathbf{Y}_i$ at each site i. Let us define ζ_i as an exponentiated linear function of these features,

$$\zeta_i = e^{\boldsymbol{\kappa} \cdot \mathbf{Y}_i}, \tag{28}$$

where $\boldsymbol{\kappa}$ is a vector of D real-valued coefficients shared across all sites, and the function e^x ensures that ζ_i takes a positive value. It may be useful to define the first element of $\mathbf{Y}_i$ to be a constant of 1 to accommodate an intercept for the linear function at the corresponding position in $\boldsymbol{\kappa}$ (implying $D-1$ true features).

In order to separate global aspects of the elongation process from TU-specific (and highly variable) rates of initiation, pause-release, and termination, we will maintain separate instances of the parameters α, β, and γ for each TU, but share the coefficient vector κ across all TUs. Thus, for M (independent) TUs indexed by j, the joint log likelihood function under the Poisson model is (cf. equations 9 & 10),

$$\begin{aligned}\ell(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\kappa},\lambda,\mathbf{Y}) &= \sum_{j=1}^{M}\left[\sum_{i=1}^{k_j} X_{j,i}\log\left(\frac{\lambda\alpha_j}{\beta_j\zeta_{j,i}}\right)-\frac{\lambda\alpha_j}{\beta_j\zeta_{j,i}}\right]+\left[\sum_{i=k_j+1}^{k_j+l_j} X_{j,i}\log\left(\frac{\lambda\alpha_j}{\zeta_{j,i}}\right)-\frac{\lambda\alpha_j}{\zeta_{j,i}}\right] \\ &\quad+\left[\sum_{i=k_j+l_j+1}^{N_j} X_{j,i}\log\left(\frac{\lambda\alpha_j}{\gamma_j\zeta_{j,i}}\right)-\frac{\lambda\alpha_j}{\gamma_j\zeta_{j,i}}\right]-\log\mathcal{Z} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left[X_{j,i}\log\left(\frac{\lambda\alpha_j}{c_{j,i}}\right)-\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}-\frac{\lambda\alpha_j}{c_{j,i}e^{\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}}\right]-\log\mathcal{Z},\end{aligned} \tag{29}$$

where α_j β_j, and γ_j are the initiation, pause-release, and termination rates, respectively, for TU j; $X_{j,i}$ is the read-count, $\mathbf{Y}_{j,i}$ is the feature vector, and $\zeta_{j,i}$ is the elongation rate at the ith nucleotide in TU j; k_j, l_j, and m_j are the lengths of the pause peak, gene body, and termination peaks, respectively, for TU j; and, for simplicity of notation, $c_{j,i}$ is defined as follows:

$$c_{j,i}=\begin{cases}\beta_j & i\in\{1,\ldots,k_j\}\\ 1 & i\in\{k_j+1,\ldots,k_j+l_j\}\\ \gamma_j & i\in\{k_j+l_j+1,\ldots,N_j\}.\end{cases} \tag{30}$$

This joint log likelihood can be maximized by gradient ascent. Making use of the chain rule, the partial derivative with respect to the nth component of $\boldsymbol{\kappa}$ is simply,

$$\frac{\partial}{\partial\kappa_n}\ell(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\kappa},\lambda,\mathbf{Y})=\sum_{j=1}^{M}\sum_{i=1}^{N_j}Y_{j,i,n}\left(\frac{\lambda\alpha_j}{c_{j,i}e^{\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}}-X_{j,i}\right), \tag{31}$$

where $Y_{j,i,n}$ represents the nth element in the feature vector at nucleotide i of TU j. For a given value of $\boldsymbol{\kappa}$ and the other parameters, the maxima for each of α_j, β_j, and γ_j can be determined analytically as follows:

$$\hat{\alpha}_j=\frac{s_{P,j}+s_{B,j}+s_{T,j}}{\sum_{i=1}^{N_j}\frac{\lambda}{c_{j,i}e^{\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}}}, \tag{32}$$

$$\hat{\beta}_j=\frac{\sum_{i=1}^{k_j}\frac{\lambda\alpha_j}{e^{\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}}}{s_{P,j}}, \tag{33}$$

$$\hat{\gamma}_j=\frac{\sum_{i=k_j+l+1}^{N_j}\frac{\lambda\alpha_j}{e^{\boldsymbol{\kappa}\cdot\mathbf{Y}_{j,i}}}}{s_{T,j}}, \tag{34}$$

where $s_{P,j}$, $s_{B,j}$, and $s_{T,j}$ are the now-familiar sufficient statistics, but restricted to the jth TU. Thus, a gradient ascent algorithm can be implemented to iteratively improve estimates of $\boldsymbol{\kappa}$ and, on each iteration, to fully optimize each of the other parameters conditional on the previous parameter values.

Notably, these results are closely related to those for standard Poisson regression. As in that case, the log likelihood function is guaranteed to be convex.

Technicalities

Estimation of λ and interpretation of rate parameters

As noted, the parameter for sequencing depth, λ, and the TU-specific initiation rates, α_j, are confounded in the sense that they appear only in the likelihood functions as a product. Because they both contribute linearly to increases in the expected read counts, they cannot be independently estimated without some exogenous calibration, which in most cases will not exist.

The most pragmatic path forward seems to be to estimate λ in a preprocessing step, hold it fixed throughout the analysis, and interpret estimates of α_j relative to each other and to the meaning of λ. Notably, β_j and γ_j must be interpreted in the same relative manner, since they are effectively tied to α_j through the structure of the continuous-time Markov model. In this sense, the general framework presented in this article should be regarded as a means for estimating *relative*, rather than absolute, rates of initiation, pause-release, termination, and, potentially, elongation. Nevertheless, these rates can effectively be compared with each other, across TUs, and—with appropriate care, as discussed below—across experiments.

Two possible ways to estimate λ are to make use of (1) the average read counts across all TUs, or (2) the read counts at one or more designated "reference" TUs. In either case, it would be sensible to make use of the gene bodies only of confidently annotated and likely expressed TUs. If the set of such TUs is denoted V (where V is a large, inclusive set in the first case or a smaller, more restricted one in the second case), then λ can be estimated in preprocessing simply as,

$$\hat{\lambda} = \frac{\sum_{j \in V} s_{B,j}}{\sum_{j \in V} l_j}. \tag{35}$$

(Notice that this formula will implicitly weight TUs by length; if desired, an unweighted average of averages could be used instead.)

The choice of V will determine the meaning of the α_j parameters, and in turn, the meaning of the β_j and γ_j parameters. If V is taken to include all (expressed) TUs, then α_j will be expressed relative to the average behavior across the genome, with $\alpha_j = 1$ indicating an average initiation rate, $\alpha_j < 1$ indicating a lower-than-average initiation rate, and $\alpha_j > 1$ indicating a higher-than-average initiation rate. If instead, V is a set of unusually highly or lowly expressed genes, α_j must be interpreted accordingly.

It is worth noting that if one were to have access to a set V of TUs with known absolute values of α_j, then these TUs could serve as a calibration for the α_j estimates, allowing them to be be assigned absolute, rather than relative, values. One possible experimental design that would enable such estimation would be to "spike" the sample with known concentrations of exogenous, distingishable RNAs (say, from a highly divergent species) and to treat these as the set V (see, e.g., [11]). For example, if both the RNA concentrations and the numbers of nuclei in a PRO-seq library preparation were known, one could, in principle, estimate α_j in numbers of events per nucleus per unit time, and then obtain corresponding estimates of β_j and γ_j.

Batch effects and normalization

A major practical problem in the analysis of all transcriptomic data, including nascent RNA sequencing data, is bulk differences in the distributions of values from separate experiments that are not biologically meaningful but instead reflect differences in sample or library preparation, sequencing, or some other technical aspect of data collection. In our case, such "batch effects" could be particularly problematic in tests of differences in initiation, pause-release, or termination rates across different conditions or species. If they are ignored, they will tend to produce many false-positive predictions of differences.

In many cases, batch effects can be effectively addressed in a preprocessing step that eliminates systematic differences across many TUs that are likely to reflect technical artifacts, leaving differences more

likely to be driven by the biological processes under study. I will focus here on quantile normalization [18], although the discussion will apply to other normalization methods as well, such as ones based on principal component analysis.

Let us assume quantile normalization is applied to the gene bodies of all TUs in each data set under study, such that the original value of $s_{B,j}$ is replaced by a new normalized value $s'_{B,j}$. For now, assume that a proportional correction is made to the read counts across the TUs (including in the pause and termination peaks), so that each count $X_{j,i}$ is replaced by $X'_{j,i} = \rho_j X_{j,i}$ where $\rho_j = s'_{B,j}/s_{B,j}$.

One possible strategy would be to apply the methods described in this paper directly to the normalized data set, without change. An issue with this approach is that the normalized read counts, in general, will no longer have integer values, and for some applications, they would either need to be rounded or a continuous distribution would need to be used in place of the Poisson or other choice of ψ (for example, a Gamma distribution).

An alternative strategy, which may be preferred in some cases, is to absorb the TU-specific rescaling induced by normalization into the λ constant. In particular, let each TU have its own read-depth parameter λ_j and set $\lambda_j = \hat{\lambda}/\rho_j$, where $\hat{\lambda}$ is the original (global) estimate of λ. In this way, the analysis can still make use of the original raw data $\mathbf{X}$, rather than the normalized data $\mathbf{X}'$, but the inferred rates α_j, β_j, and γ_j will effectively reflect the normalization transformation.

In tests of differences in the pause-release rate β, a second normalization may be needed. There are some indications that the degree to which RNAP accumulates in the promoter region may differ from one experiment to another, perhaps owing to differences in the concentration of factors that contribute to pause release. As a result, the data may display batch effects in pause peaks beyond those that can be explained by differences in the gene bodies. This problem could be addressed by a second round of normalization applied to the s_P values after having normalized for differences in s_B, as described above. As in the first case, this normalization could also be accommodated by altering the λ values per TU, but it would require the introduction of a second λ_j that applied only to the pause peaks.

Replicates

In many applications, it is desirable to perform multiple replicates (biological or technical) of each individual experiment, to mitigate the noisiness of the sequence data and increase confidence in parameter estimates or hypothesis tests. Replicates can easily be accommodated in the model by working with a joint likelihood function that assumes independence across replicates but shares parameters that represent an underlying biological "truth" expected to be the same in each replicate. Typically, the key rate parameters—α, β, γ and perhaps ζ (in the nonequilibrium case)—will be shared, whereas technical parameters—such as λ—will be separate for each replicate.

With R replicates per time point, the likelihood function for the nonequilibrium case (cf. equation 3) generalizes to,

$$P(\mathbf{X}^{(t_1)}, \ldots, \mathbf{X}^{(t_M)} \mid \alpha, \beta, \gamma, \zeta, \theta) = \prod_{j=1}^{M} \prod_{r=1}^{R} \prod_{i=1}^{N} \psi(X_{i,r}^{(j)} \mid \mu_i^{(j)}, \theta), \tag{36}$$

where $X_{i,r}^{(j)}$ represents the read count for nucleotide i and time-point j as measured in replicate r and $\mu_i^{(j)} = \lambda_r^{(k)} P(Z_i | t_k, \alpha, \beta, \gamma, \zeta)$ represents a shared "truth" across replicates for site i and time-point j that is separately normalized for each replicate r using read-depth parameter $\lambda_r^{(k)}$. Similarily, the likelihood

function for the steady-state case (cf. equation 8) becomes,

$$P(\mathbf{X}|\alpha,\beta,\gamma,\theta,\boldsymbol{\lambda}) = \prod_{r=1}^{R} \psi\left(X_{1,r},\ldots,X_{k,r} \mid \frac{\lambda_r\alpha}{\beta},\theta\right) \times \psi\left(X_{k+1,r},\ldots,X_{k+l,r} \mid \lambda_r\alpha,\theta\right) \times \psi\left(X_{k+l+1,r},\ldots,X_{N,r} \mid \frac{\lambda_r\alpha}{\gamma},\theta\right). \tag{37}$$

In the case of the Poisson model at steady-state, the likelihood function reduces to (cf. equation 9),

$$\begin{aligned} P(\mathbf{X}|\alpha,\beta,\gamma,\boldsymbol{\lambda}) &= \frac{1}{\mathcal{Z}} \prod_{r=1}^{R} \left(\frac{\lambda_r\alpha}{\beta}\right)^{s_{P,r}} (\lambda_r\alpha)^{s_{B,r}} \left(\frac{\lambda_r\alpha}{\gamma}\right)^{s_{T,r}} \exp\left[-\lambda_r\alpha\left(\frac{k}{\beta}+l+\frac{m}{\gamma}\right)\right] \\ &= \frac{1}{\mathcal{Z}'} \left(\frac{\alpha}{\beta}\right)^{s'_P} \alpha^{s'_B} \left(\frac{\alpha}{\gamma}\right)^{s'_T} \exp\left[-\lambda'\alpha\left(\frac{k}{\beta}+l+\frac{m}{\gamma}\right)\right] \end{aligned} \tag{38}$$

where $s'_P = \sum_r s_{P,r}$, $s'_B = \sum_r s_{B,r}$, $s'_T = \sum_r s_{T,r}$, $\lambda' = \sum_r \lambda_r$, and where we absorb the terms that depend only on λ_r in $\mathcal{Z}'$, assuming that the λ_r values will be pre-estimated. Thus, the likelihood function can be expressed in terms of sufficient statistics that are simply sums over replicates of the original sufficient statistics. Notice that this function would be identical if the read counts were simply pooled across replicates. Predictably, it leads to maximum-likelihood estimators of,

$$\hat{\alpha} = \frac{s'_{B,}}{\lambda'} \tag{39}$$

$$\hat{\beta} = \frac{s'_B/l}{s'_P/k} \tag{40}$$

$$\hat{\gamma} = \frac{s'_B/l}{s'_T/m}. \tag{41}$$

The likelihood ratio tests and generalized linear model under the Poisson model can similarly be expressed in terms of these new aggregate sufficient statistics.

Discussion

In this article, I have outlined a new probabilistic modeling framework for the analysis of nascent RNA sequencing data. This framework allows for a unified treatment of many problems of interest in the study of eukaryotic transcription, including estimation of the rates of initiation, elongation, pause-release, and termination. It leads naturally to likelihood ratio tests for differences in rates, and it has extensions to Bayesian inference and a generalized linear model for evaluating covariates of elongation rate. In addition, a number of issues of practical importance in the analysis of real data—such as accomodating differences in library size, batch effects, and experimental replicates—can be readily addressed in this framework.

A number of machine-learning and statistical methods already exist for analyzing nascent RNA sequencing data [6, 8, 12, 13], so it is reasonable to ask what is gained by addressing these problems using the generative probabilistic model described here. In my view, this generative model has several key advantages. First, the model does allow estimation of some quantities that have been out of reach of most previous methods, such as the relative rates of initiation, pause-release, and termination. Some of these quantities have been indirectly characterized by more heuristic methods—as when pausing is quantified using the pause index—but it has been difficult to know precisely how to interpret these measures. Second, the model can be fitted to the raw read counts in a completely unsupervised manner, with no need

for labeled training data. This property is important because, while high-quality, independent training data is available for some applications (e.g., [8]), no such data is available for others, including estimation of pause-release or elongation rates, or tests for differences in rates. Third, because the model allows several distinct problems to be addressed in a unified, coherent manner, it allows various parameter estimates and hypothesis tests to be interpreted jointly, without the concern that separate estimates might reflect different biases or different modeling assumptions—as when, say, applying one machine-learning method to estimate elongation rates and a second, completely separate method to characterize promoter-proximal pausing or correlates of elongation rate. Finally, by directly describing the actual process by which RNAP moves along the DNA template, the model potentially leads to new insights into this process, beyond simply providing predictions of quantities of interest, as a discriminative machine-learning method would. Altogether, the generative model both opens up new applications and provides a new perspective on the fundamental process of eukaryotic transcription.

The model does require a number of simplifying assumptions, some of which may not hold in all settings. For example, it assumes that RNAPs that initiate transcription gradually make their way all the way along the DNA template to the termination site, with negligible rates of premature termination. The frequency with which premature termination occurs is debated in the literature, with some studies suggesting it occurs at low rates across most genes [16] and others indicating higher rates [19, 20]. My hope is that this phenomenon occurs at low enough rates across most of the TU (with the possible exception of locations near the promoter) that it will have at most a minor impact on the rate estimators and hypothesis tests described here, but it could result in appreciable biases. The challenge in this setting is that, if premature termination were added to the model, its rate would be confounded with the initiation and elongation rates. I do not see a way of accommodating it in this framework except by conditioning on independent estimates of its rate.

Another imperfect simplifying assumption is that of constant reductions in elongation rate across the pause peak and the termination peak. A more realistic model would account for the interactions of individual RNAP molecules, allowing later ones to be slowed down by earlier ones as they back up like automobiles in a traffic jam. However, such a model would require a time-inhomogeneous rate matrix and would considerably complicate the mathematics. The assumption of a constant reduction in rate seems to be reasonable in the steady-state case, but in the nonequilibrium case, it will lead to an unrealistic reduction in rates at the pause peak at early time points, before RNAPs can back up. My hope is that, even in this case, the effect of this bias will be minor, because the pause peak tends to be quite short and emerges relatively quickly after induction of transcription. Nevertheless, more work will be needed to understand its impact.

The assumption of a Poisson distribution for read counts leads to a great deal of mathematical simplicity, and I have therefore relied heavily on this version of the model. I see three major points in favor of the Poisson model. First, it leads to closed-form expressions that are easy to interpet, and help to develop intuition for how the data drive parameter estimates and hypothesis tests. Second, it provides a convenient and rapid means for performing many computations of interest, without the need for complicated, slow, and error-prone software for numerical optimization. Third, it is likely that, even when the Poisson assumption is unrealistic (as when read counts are clearly overdispersed), it will nevertheless have limited impact on many of the estimators or tests of interest, particularly if care is taken in post-processing and interpretation. For example, in LRTs, one Poisson model is compared with another, and if log likelihood ratios are interpreted with respect to an empirical null distribution, the effects of overdispersion should be considerably mitigated. On the other hand, if it does prove necessary to discard the Poisson model, the negative binomial model is still relatively convenient to work with, despite requiring numerical optimization. Notably, the fully general framework would potentially allow the use of distributions for read counts that allowed not only for overdispersion but also for autocorrelation along the genome sequence, such as a Gaussian process model or hidden Markov model. More work will be needed to determine whether the use of more complex and computationally intensive generating distributions is justified.

It is worth noting that the analysis of real nascent RNA sequencing data would require addressing a

number of practical issues that have been glossed over in this article. Perhaps chief among these is the question of how to define the three intervals corresponding to k_j, l_j, and m_j at each TU j. For some analyses, it may be adequate to set, say, k_j and m_j to fixed values, and let l_j absorb the remainder of the TU. In real data, however, pause and termination peaks can be highly variable in both position and breadth, so it may be necessary to make use of data-driven methods to define the boundaries of interest. At the same time, using the data to define the intervals, and then using it a second time conditional on those intervals, could result in biases. A similar issue concerns the location of the transcription start and termination sites, which quite often appear to be discordant between the sequence data and the available gene annotations. These coordinates may also need to be estimated in a data-driven fashion [13]. Notably, the generative model could itself potentially be used as the basis for prediction of transcription start and termination sites, and/or the locations of the pause and termination peaks, for example, by relying on hidden Markov model or other changepoint prediction algorithms for segmentation. This approach, however, would likely be unwieldy and computational intensive, and more work would be needed to determine if it is practical.

Overall, I believe that the combination of a continuous-time Markov model and a generalized generating distribution for read counts, as outined in this article, provides a flexible and powerful modeling framework for nascent RNA sequencing data. The version of the model described here does make use of a number of simplifying assumptions, some of which may need to be relaxed over time. Nevertheless, the basic framework described here should, if nothing else, be useful as a starting point for further model and algorithm development.

Acknowledgments

I thank Charles Danko, Yifei Huang, Yixin Zhao, and Noah Dukler for helpful discussions, and Yixin Zhao for help with figure preparation.

References

[1] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–1848.

[2] Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature. 2011;469(7330):368–373.

[3] Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013;339(6122):950–953.

[4] Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell. 2015;161(3):541–554.

[5] Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016;11(8):1455–1476.

[6] Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell. 2013;50(2):212–222.

[7] Core LJ, Martins AL, Danko CG, Waters C, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of transcription initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46(12):1311–1320.

[8] Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015;12(5):433–438.

[9] Dukler N, Booth GT, Huang YF, Tippens N, Waters CT, Danko CG, et al. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Res. 2017;27(11):1816–1829.

[10] Blumberg A, Zhao Y, Huang YF, Dukler N, Rice EJ, Krumholz K, et al. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BioRxiv. 2019; p. 690644.

[11] Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife. 2014;3:e02407.

[12] Azofeifa JG, Allen MA, Lladser ME, Dowell RD. An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(5):1070–1081.

[13] Anderson WD, Duarte FM, Civelek M, Guertin MJ. Defining data-driven primary transcript annotations with primaryTranscriptAnnotation in R. Bioinformatics. 2020;36(9):2926–2928.

[14] wa Maina C, Honkela A, Matarese F, Grote K, Stunnenberg HG, Reid G, et al. Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data. PLoS Comput Biol. 2014;10(5):e1003598.

[15] Azofeifa JG, Dowell RD. A generative model for the behavior of RNA polymerase. Bioinformatics. 2017;33(2):227–234.

[16] Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, Wagner EJ, et al. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Mol Cell. 2019;76(5):738–752.

[17] Karlin S, Taylor HM. A First Course in Stochastic Processes. 2nd ed. Academic Press; 1975.

[18] Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. Journal of the American Statistical Association. 2001;96(456):1161–1170.

[19] Kamieniarz-Gdula K, Proudfoot NJ. Transcriptional Control by Premature Termination: A Forgotten Mechanism. Trends Genet. 2019;35(8):553–564.

[20] Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schubeler D. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Mol Cell. 2017;67(3):411–422.