OXFORD

Model-based characterization of the equilibrium dynamics of transcription initiation and promoter-proximal pausing in human cells

Yixin Zhao [1], **Lingjie Liu** [1,2], **Rebecca Hassett**[1] **and Adam Siepel** [1,2,*]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Graduate Program in Genetics, Stony Brook University, Stony Brook, NY, USA

[*]To whom correspondence should be addressed. Tel: +1 516 367 6922; Fax: +1 516 367 5501; Email: asiepel@cshl.edu

Abstract

In metazoans, both transcription initiation and the escape of RNA polymerase (RNAP) from promoter-proximal pausing are key rate-limiting steps in gene expression. These processes play out at physically proximal sites on the DNA template and appear to influence one another through steric interactions. Here, we examine the dynamics of these processes using a combination of statistical modeling, simulation, and analysis of real nascent RNA sequencing data. We develop a simple probabilistic model that jointly describes the kinetics of transcription initiation, pause-escape, and elongation, and the generation of nascent RNA sequencing read counts under steady-state conditions. We then extend this initial model to allow for variability across cells in promoter-proximal pause site locations and steric hindrance of transcription initiation from paused RNAPs. In an extensive series of simulations, we show that this model enables accurate estimation of initiation and pause-escape rates. Furthermore, we show by simulation and analysis of real data that pause-escape is often strongly rate-limiting and that steric hindrance can dramatically reduce initiation rates. Our modeling framework is applicable to a variety of inference problems, and our software for estimation and simulation is freely available.

Graphical abstract
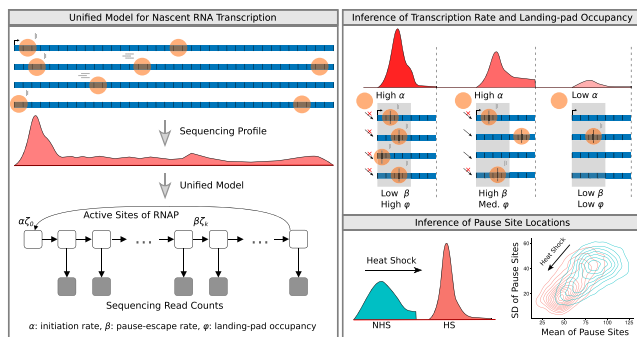
Introduction

Across all branches of life, homeostatic control of gene expression arises from a series of dynamic equilibria among competing processes. For example, cellular RNA concentrations reflect an equilibrium between RNA production and decay; RNA polymerase (RNAP) occupancy reflects an equilibrium among transcription initiation, elongation, and termination; and protein concentrations reflect an equilibrium between protein synthesis and degradation. When such concentrations change, say, across conditions, cell types, or developmental stages, it is typically because the relative rates of competing processes are altered and a new equilibrium is reached, rather than because any individual process is wholly enabled or disabled.

For decades, research on the regulation of eukaryotic gene expression focused on RNAP recruitment and transcription initiation, which were thought to be the rate-limiting steps under most circumstances (1). In recent years, however, it has become clear that many downstream steps in transcription can be regulated. One striking observation, first noted at particular loci in *Drosophila* and mammals (2–6), is that RNAPs are frequently held in a 'paused' position about 18–60 nt downstream of the transcription start site, before escaping into productive elongation (7–16). Later studies showed that such promoter-proximal pausing is widespread across metazoans (14). Several lines of evidence indicate that the escape from promoter-proximal pausing is frequently a regulated step in gene expression (17). It has been argued that regulation at the

pause-escape stage may be particularly advantageous when a rapid transcriptional response is required, as in heat shock or other stimulus-controlled pathways (11,17–22).

It has been noted that promoter-proximal pausing not only has a direct effect on the rate at which RNAPs proceed into productive elongation, but may also have an indirect influence on the rate of productive initiation, owing to steric hindrance between paused and initiating RNAPs. The key observation is that the physical space immediately downstream of the transcription start site (TSS) is limited: each RNAP has a 'footprint' of ~33–35 nt (23,24) and the pause site is typically no more than 60 nt downstream of the TSS. Thus, depending on precisely how much space is required between two adjacent RNAPs, there is typically room for only one, or perhaps two or three, RNAPs in the pause region before new initiation events begin to be blocked. Indeed, computer simulations of the movement of RNAPs along the DNA template have suggested that such steric hindrance could substantially reduce initiation rates (24). More recently, genome-wide studies of human (25,26) and *Drosophila* (27) cells found strong evidence that initiation rates were restricted by pause-escape rates at many genes. Thus, it appears that rates of productive initiation are often governed by a dynamic equilibrium between transcription initiation and pause-escape.

In recent years, two types of approaches have dominated in the study of transcriptional dynamics: (1) high-resolution imaging approaches based on fluorescence *in situ* hybridization, photobleaching, or electron micrography in single cells (e.g. (28–32)); and (2) genomic approaches based on nascent RNA sequencing (NRS) or chromatin immunoprecipitation (ChIP) and sequencing across populations of cells (e.g. (17,25,27)). The imaging approaches allow for more direct characterization of the dynamics of individual RNAPs, but as yet, they cannot be carried out at scale; instead, they are typically applied to one or a few loci. The genomic approaches are more indirect, requiring tagging and pull-down of sequences such as newly synthesized RNA or RNAP-associated DNA. In addition, because they produce summaries for large populations of cells, they typically require fairly complex statistical analyses for interpretation. Nevertheless, these methods have the crucial advantage of being applicable at genome-wide scale for modest cost, by exploiting the many recent advances in genome-sequencing and related technologies.

For these reasons, we focus in this article on the study of transcriptional dynamics using genomic data. We focus in particular on the use of data from the 'RO-seq' family of NRS protocols—including GRO-seq (9), PRO-seq (15,33) and ChRO-seq (34)—which specifically detect active RNAPs in transcriptionally-competent complexes (35). These protocols have matured dramatically in recent years, with major improvements in resolution, background signal, ease-of-use and cost. They can be thought of as producing 'snapshots' of the positions of engaged RNAPs across a population of cells either at steady-state or in a time course after a stimulus is applied. We develop a general statistical modeling framework that allows interpretation of these snapshots in a manner that reveals relative rates of transcriptional initiation, promoter-proximal pause escape, and elongation, as well as interrelationships among these processes.

Notably, we focus on modeling transcriptional dynamics under equilibrium conditions, which, in comparison to studies of time courses following transcriptional stimuli, allows us to examine larger sets of genes and avoid the potential off-target effects of commonly used drugs such as triptolide. The focus on steady-state conditions also leads to relatively simple and interpretable mathematical results. Importantly, our methods are applicable not only to newly produced NRS data sets, but to the thousands of sequenced samples that are already publicly available in databases such as the Gene Expression Omnibus (36).

We apply these new methods to both simulated and real NRS data, and refine our model to account for variable pause sites across cells and steric hindrance of transcription initiation from paused RNAP. To support our experiments, we have developed a fast and flexible simulator for NRS data, called SimPol, which is freely available along with our software for parameter estimation (see 'Code Availability'). Altogether, we find strong quantitative evidence that promoter-proximal pausing has major importance in the dynamics of transcription in human cells, that RNAP occupancy in the pause region tends to be high at many genes, and that initiation rates are frequently limited by paused RNAPs. We discuss various implications of these findings in detail.

Materials and methods

A simple probabilistic model for transcription initiation, promoter-proximal pausing, and elongation

Our initial model consists of two layers: a continuous-time Markov model for the movement of individual RNA polymerases (RNAPs) along a transcription unit (TU), and a conditionally independent generating process for the read counts at each nucleotide site (Figure 1A, B, see also bioRxiv: https://doi.org/10.1101/2021.01.12.426408). Together, these components produce a full generative model for NRS read counts along the TU, permitting inference of transcriptional rate parameters from the raw data. There is a long history of application of similar quantitative models to the processes of transcription and translation (37–43), but to our knowledge, they have not been used to study promoter-proximal pausing or its feedback on transcription initiation rates.

Our Markov model consists of a state Z_i for each nucleotide position $i \in \{1, \ldots, N\}$ of the RNAP plus an additional state, Z_0, that represents free RNAPs. It is parameterized by a transcription-initiation rate α, a rate of promoter-proximal pause escape β, and a termination rate γ (see Supplementary Table S1 for a summary of modeling notation). In addition, it includes an elongation rate ζ_i for each position i, which can either be assumed constant across sites (with $\zeta_i = \zeta$ as throughout this manuscript) or allowed to vary. For mathematical convenience (see below), the parameters α, β and γ are multiplied by the corresponding ζ_i parameters. In this manuscript, we focus on inference of α and β and largely ignore γ, because the 3′ ends of TUs are difficult to characterize.

In this version of the model, promoter-proximal pausing is assumed to occur at a fixed position k along the DNA template, where k can be pre-estimated from the data. This assumption will be relaxed in subsequent sections. We further assume that the read counts at each position i are Poisson-distributed with mean μ_i, where μ_i is a scaled version of the probability density of Z_i (and, hence, the RNAP density at nucleotide i) that reflects the initiation and elongation rates, as well as the sequencing depth. Finally, in order to make use of a time-homogeneous Markov chain, we assume that RNAPs
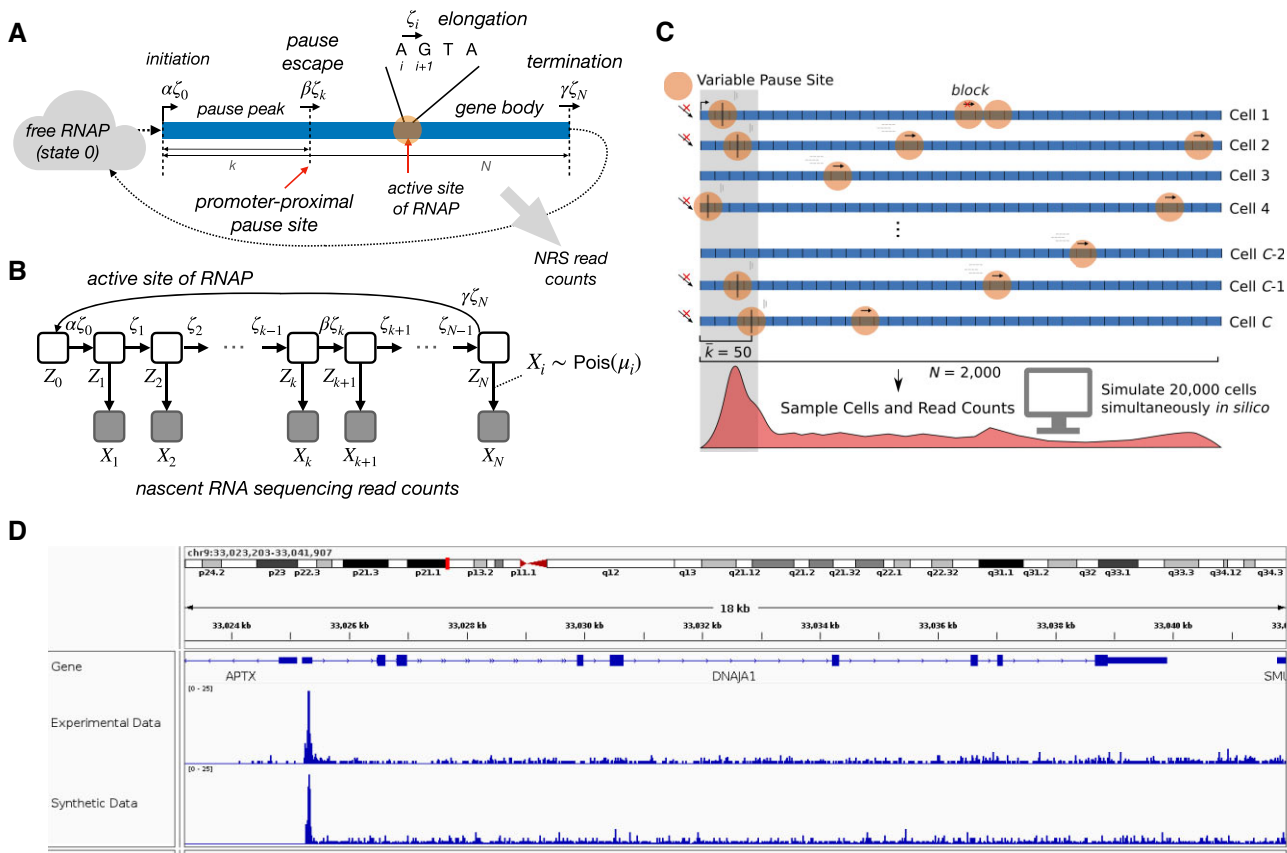
Figure 1. (**A**) Conceptual illustration of model, focusing on the kinetic model for RNAP movement on the DNA template. Gray arrow indicates that a second layer of the model describes generation of nascent RNA sequencing (NRS) read counts based on the distribution of RNAP positions across cells. (**B**) Graphical model representation with unobserved continuous-time Markov chain (Z_i) and observed read counts (X_i). Read counts at each site X_i are conditionally independent and Poisson-distributed given mean μ_i, which reflects both the density $P(Z_i)$ and the sequencing depth λ. (**C**) Design of SimPol ('Simulator of Polymerases'). Based on user-defined initiation, pause-escape, and elongation rates, SimPol tracks the movement *in silico* of RNAPs across N-bp DNA templates in C cells, then samples synthetic read counts based on RNAP positions. SimPol identifies collisions and prohibits RNAPs from passing one another. It also models variable pause sites and elongation rates. (**D**) Example of synthetic nascent RNA sequencing data from SimPol, shown in IGV (74) alongside matched real PRO-seq data from (19) for the *DNAJA1* gene on chromosome 9 of the human genome.

are sufficiently sparse along the DNA template that collisions between them are rare—another assumption that will be relaxed later. Notably, the model can be applied either in a nonequilibrium setting based on time-course data or to a single data set at steady state. We focus in this manuscript on the more limited steady-state case, because it is more mathematically convenient and interpretable, yet, as we show below, still allows for a quantitative characterization of the dynamic equilibrium between transcription initiation and pause-escape. In order to ensure identifiability of initiation and pause-escape rates at steady state, we assume that initiating RNAPs remain on the DNA template for its entire length, with negligible rates of premature termination (see Discussion).

This simple version of the model results in convenient, closed-form maximum-likelihood estimators for β and a parameter closely related to α. Because the initiation rate α is confounded at steady-state with the elongation rate ζ and the sequencing depth λ, we instead work with a compound parameter $\chi = \lambda \frac{\alpha}{\zeta}$ representing the read-depth-scaled ratio of the initiation rate to the elongation rate. It turns out that the MLE for χ is simply equal to the average read depth along the gene body and the MLE for β is given by the ratio of the average read depth in the gene body to that in the pause peak (see below). The estimates of both α and β can be considered *relative* to ζ, based on the multiplicative parameterization of the

model. Notably, similar average-read-depth estimators have been widely used in the analysis of nascent RNA-sequencing data, typically with more heuristic justifications. For example, the inverse of the estimator for β is commonly known as the 'pausing index' (14,44), and the estimator for χ is often used as a general measure of transcription output (19). In our case, these estimators emerge as MLEs under a generative probabilistic model, making it possible to characterize the dynamics of transcription from raw NRS read counts at steady state.

Continuous-time Markov model

The probabilistic model (Figure 1A, B) consists of a continuous-time Markov model that describes the stochastic movement of individual RNAPs along the DNA template, and a conditional generating process by which read counts arise independently at each site in proportion to RNAP occupancy (defined below). The Markov model consists of N + 1 states corresponding to the N possible nucleotide positions of the active site of an RNAP as it moves along an N-nucleotide DNA template, plus an additional state (labeled 0) that abstractly represents 'free' RNAPs, not currently engaged in transcription and available for new initiation events. Each state i in the Markov model corresponds to a binary random variable Z_i, indicating whether the RNAP is ($Z_i = 1$) or is not

($Z_i = 0$) at position i at a particular time t (Figure 1B). In our setting, the use of this time-homogeneous model depends on two key assumptions: (i) that collisions between RNAPs are rare, allowing the movement of each RNAP to be considered independently of the others and (ii) that premature termination of transcription is sufficiently rare that each RNAP can be assumed to traverse the entire DNA template if it is given enough time (see Discussion for limitations).

The model distinguishes between two segments of each transcription unit: (i) the first k nucleotides, known as the *pause peak*, where RNAP tends to accumulate owing to promoter-proximal pausing (typically $k \approx 50$) (9); and (ii) the subsequent $N - k$ nucleotides, where RNAP tends to be relatively unimpeded, which is typically referred to as the *gene body*. Movement of the RNAP is defined by four rate parameters: an initiation rate α (from state 0 to state 1), a pause-escape rate β (from state k to state $k + 1$), a termination rate γ (from state N to state 0), and a constant per-nucleotide elongation rate ζ (for all other allowable transitions). Because the states must be visited in a sequence, the infinitesimal generator matrix for the Markov chain $\mathbf{Q} = \{q_{ij}\}$ has a simple form, with positive terms only on the diagonal q_{ij} such that $j = i + 1$, negative terms on the main diagonal, and zeroes elsewhere. For mathematical convenience, we assume that the initiation, pause-escape, 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. Specifically,

$$q_{ij} = \begin{cases} \alpha\zeta & i = 0, j = 1 \\ \beta\zeta & i = k, j = k + 1 \\ \gamma\zeta & i = N, j = 0 \\ \zeta & \text{all other } i, j \text{ such that } j = i + 1 \\ 0 & \text{all other } i, j \text{ such that } i \neq j \\ -\sum_{i':i' \neq j} q_{i'j} & i = j, \end{cases} \tag{1}$$

where element q_{ij} indicates the instantaneous rate at which an RNAP transitions from state i to state j, and by convention, the values along the main diagonal are set such that the rows sum to zero.

Stationary distribution

This continuous-time Markov model allows for calculation of the probability of transitioning from any state i to any other state j over a period of time $t \geq 0$, but in this article we are interested in steady-state conditions and therefore focus on the limiting distribution for ending states as $t \to \infty$. This stationary distribution, denoted $\boldsymbol{\pi}$, is invariant to ζ and can be found easily by solving the equation $\boldsymbol{\pi}\mathbf{Q} = 0$. Because state 0 is simply an abstraction to allow for the recirculation of RNAP, we omit it and describe the stationary distribution conditional on RNAP occupancy along the DNA template. This conditional stationary distribution can be expressed as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ such that,

$$\pi_i = \frac{1}{\mathcal{Z}} \cdot \begin{cases} \frac{1}{\beta} & i = k \\ \frac{1}{\gamma} & i = N \\ 1 & i \in \{1, \ldots, N - 1\}, i \neq k, \end{cases} \tag{2}$$

with normalization constant $\mathcal{Z} = N - 2 + \frac{1}{\beta} + \frac{1}{\gamma}$.

This distribution has an intuitive interpretation. First, it is natural that, conditional on RNAP occupancy, the steady-state distribution is invariant to both α (which defines the rate at which occupancy is initiated but has no effect thereafter) and ζ (which defines the 'flow' along the DNA template but does not favor one nucleotide position over another). In addition, as a result of local slowdowns in elongation, π_i is elevated relative to the gene body by factors of $\frac{1}{\beta}$ and $\frac{1}{\gamma}$ at the pause peak and termination peak, respectively. Notice that both peaks take the form of 'spikes' at single nucleotide positions under this model; in later sections we will generalize the model to allow for a broader pause peak.

When comparing different transcription units, we have to allow for differences in TU-specific initiation and elongation rates. In particular, in addition to obeying equation 2, the relative RNAP densities at TU j will be proportional to $\frac{\alpha_j}{\zeta_j}$, where α_j and ζ_j are the TU's initiation and elongation rates, respectively. Furthermore, as detailed below, estimation of these rates is confounded by the sequencing depth, λ. Because these parameters are not identifiable at steady state, we represent them by the compound parameter $\chi_j = \frac{\lambda\alpha_j}{\zeta_j}$.

In addition, the termination peak tends to be difficult to characterize with real nascent RNA sequencing data, owing to transcriptional run-on, poorly characterized 3' ends of genes, and other factors. Therefore, from this point on, we omit the parameter γ and assume N is defined such that ambiguities at the 3' ends of TUs are excluded.

With these assumptions, when comparing the RNAP densities at various sites i along various TUs j, we expect them to be proportional to,

$$\mu_{i,j} = \begin{cases} \frac{\chi_i}{\beta} & i = k \\ \chi_j & i \in \{1, \ldots, N\}, i \neq k. \end{cases} \tag{3}$$

Generative model for sequence data

To allow for the fact that the RNAP positions are only indirectly observed through the sequencing process, we add a second layer to the model that describes the probabilistic process by which sequencing read counts are generated conditional on an underlying density of RNAP at each nucleotide, as defined by the continuous-time Markov model. In this way, we obtain a full generative model for the observed sequence data that is defined by the model parameters, enabling inference of all parameters from the data.

Because we have freedom in how to set the read-depth scaling parameter λ (see below), we simply take μ_{ij} (as defined in equation 3) to be the expected read depth at position i of TU j. We then assume that the read count $X_{i,j}$ is Poisson-distributed with this mean. It is possible to use other generating distributions to allow for overdispersion of the read counts (see bioRxiv: https://doi.org/10.1101/2021.01.12.426408), but the Poisson assumption seems to be adequate in our case and it is particularly convenient for parameter inference.

With these assumptions, let the data for a single TU be denoted $\mathbf{X} = (X_1, \ldots, X_N)$, where X_i represents the number of sequencing reads having their 3' end aligned to position i. (We omit the j index in this case to simplify the notation.) Assuming conditionally independent Poisson distributions at each site, the steady-state log likelihood function for a single TU

is given by,

$$\ell(\mathbf{X};\ \chi,\beta) = \log P(\mathbf{X}\,|\,\chi,\beta)$$

$$= \log\left\{\left[\prod_{i=1}^{N}\text{Pois}\left(X_i\,|\,\chi\right)\right]\times\frac{\text{Pois}\left(X_k\,|\,\chi/\beta\right)}{\text{Pois}\left(X_k\,|\,\chi\right)}\right\}$$

$$= \log\left\{\left[\prod_{i=1}^{N}\frac{\chi^{X_i}e^{-\chi}}{X_i!}\right]\times\beta^{-X_k}e^{-\chi\left(\frac{1}{\beta}-1\right)}\right\}$$

$$= s\log\chi - X_k\log\beta - \chi\left(N+\frac{1}{\beta}-1\right) - \log\mathcal{Z},\tag{4}$$

where $s = \sum_{i=1}^{N}X_i$ is a statistic equal to the sum of all read counts 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.

Inference at steady state

Under this likelihood function (equation 4), the maximum-likelihood estimators for χ and β have simple closed-form solutions:

$$\hat{\chi} = \frac{s - X_k}{N-1},\tag{5}$$

$$\hat{\beta} = \frac{s - X_k}{X_k(N-1)}.\tag{6}$$

Notice that the estimator for χ is simply the average read depth, excluding the pause peak, and the estimator for β is the ratio of that same average read depth to the read depth in the pause peak. These are estimators that have been widely used in the analysis of nascent RNA sequencing data, with more heuristic justifications (14,19,44).

In practice, it tends to be better to avoid the complex signal in the pause region in estimating χ and instead estimate it from a downstream portion of the gene body. We define $s' = \sum_{i=j}^{j+M-1}X_i$, where M is the length of the interval considered, and estimate χ and β as,

$$\hat{\chi} = \frac{s'}{M}$$

$$\hat{\beta} = \frac{s'}{X_k M}.\tag{7}$$

For the remainder of the paper, we assume the use of these simpler, more robust estimators for χ and β.

Allowing for variation in the pause site

In real samples, the pause site tends to vary across cells, leading to a broad pause peak in nascent RNA sequencing data. We address this complication by allowing the location of the pause peak, k, to vary between a $k_{\min}$ and a $k_{\max}$ according to an appropriate distribution, and then assuming each read count X_k within this range reflects a mixture of cells that do and do not have their pause site at position k. Assume that, for $k \in \{k_{\min},...,k_{\max}\}$, f_k represents the fraction of cells with pause site k, and that the read count X_k derives from a mixture of one Poisson distribution with rate $\chi/\beta \cdot f_k$ and a second Poisson distribution with rate $\chi \cdot (1 - f_k)$. If we denote by Y_k the (unknown) portion of the read count that derives from the

first process, then the log likelihood function can be expressed as (cf. equation 4),

$$\ell(\mathbf{X};\ \chi,\beta) = \sum_{i=1}^{k_{\min}-1}\log\left[\text{Pois}\left(X_i\,|\,\chi\right)\right]$$

$$+ \sum_{k=k_{\min}}^{k_{\max}}\log\left[\sum_{Y_k=0}^{X_k}\text{Pois}\left(Y_k\,|\,\chi/\beta\cdot f_k\right)\right.$$

$$\left.\times\text{Pois}\left(X_k - Y_k\,|\,\chi\cdot(1-f_k)\right)\right]$$

$$+ \sum_{i=k_{\max}+1}^{N}\log\left[\text{Pois}\left(X_i\,|\,\chi\right)\right].\tag{8}$$

This log likelihood function does not have a closed-form solution but it is straightforward to maximize by expectation maximization. The complete-data log likelihood function, with known values of Y_k, can be expressed in terms of compact sufficient statistics (analogous to equation 4) as,

$$\ell_c(\mathbf{X},\mathbf{Y};\ \chi,\beta) = s\log\chi - t\log\beta - \chi\left(N+\frac{1}{\beta}-1\right)$$

$$+ \sum_{k=k_{\min}}^{k_{\max}}Y_k\log f_k + (X_k - Y_k)\log(1-f_k)$$

$$- \log Z,\tag{9}$$

where $s = \sum_i X_i$ and $t = \sum_k Y_k$. The expected value of this function, averaging over the latent variables Y_k—the quantity to maximize in EM—is simply,

$$\langle\ell_c(\mathbf{X},\mathbf{Y};\ \chi,\beta)\rangle = s\log\chi - \langle t\rangle\log\beta - \chi\left(N+\frac{1}{\beta}-1\right)$$

$$+ \sum_{k=k_{\min}}^{k_{\max}}\langle Y_k\rangle\log f_k$$

$$+(X_k - \langle Y_k\rangle)\log(1-f_k) - \log Z,\tag{10}$$

where $\langle Y_k\rangle$ and $\langle t\rangle = \sum_k\langle Y_k\rangle$ denote the posterior expected values of Y_k and t, respectively. For simplicity, assume that χ is pre-estimated for a portion of the gene body downstream of $k_{\max}$ using equation 7. If in addition the values of f_k are fixed, then β can be simply estimated as,

$$\hat{\beta} = \frac{\hat{\chi}}{\langle t\rangle},\tag{11}$$

The values $\langle Y_k\rangle$ can be computed by observing that, because X_k is the sum of two Poisson-distributed variables, $Y_k\,|\,X_k$ is binomially distributed with probability,

$$p_k = \frac{f_k\chi/\beta}{f_k\chi/\beta + (1-f_k)\chi} = \begin{cases}\frac{1}{1-\beta+\beta/f_k} & f_k\neq 0\\0 & f_k = 0\end{cases}\tag{12}$$

Therefore, $\langle Y_k\rangle = X_k\cdot p_k$, and

$$\langle t\rangle = \sum_k X_k\cdot p_k = \sum_k\frac{X_k}{1-\beta+\beta/f_k}I\quad(f_k\neq 0).\tag{13}$$

Thus, an EM algorithm can be implemented by iteratively applying equations 11 and 13, in turn, until convergence.

Furthermore, to estimate the distribution of pause sites from the data at each TU, we assume the f_k values have a

truncated Gaussian distribution with mean μ and variance σ^2,

$$f_k = \frac{1}{\mathcal{Z}} \times \begin{cases} \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{k-\mu}{\sigma}\right)^2} & k \in \{k_{\min}, \dots, k_{\max}\} \\ 0 & \text{otherwise,} \end{cases} \qquad (14)$$

where the explicit normalization constant $\mathcal{Z}$ is needed because the distribution is applied to a bounded interval and is defined at integer values only.

In this case, the EM updates for μ and σ^2 are simply,

$$\hat{\mu} = \frac{\langle u \rangle - \langle z \rangle}{\langle t \rangle - \langle w \rangle}, \qquad \hat{\sigma}^2 = \frac{\langle v \rangle - \langle r \rangle}{\langle t \rangle - \langle w \rangle} - \hat{\mu}^2, \qquad (15)$$

where,

$$\langle t \rangle = \sum_k \langle Y_k \rangle, \quad \langle w \rangle = \sum_k \frac{f_k}{1-f_k}(X_k - \langle Y_k \rangle),$$

$$\langle u \rangle = \sum_k \langle Y_k \rangle k, \quad \langle z \rangle = \sum_k \frac{f_k}{1-f_k}(X_k - \langle Y_k \rangle)k,$$

$$\langle v \rangle = \sum_k \langle Y_k \rangle k^2, \quad \langle r \rangle = \sum_k \frac{f_k}{1-f_k}(X_k - \langle Y_k \rangle)k^2.$$

Notice that all of these quantities are easily computed from the $\langle Y_k \rangle$ values together with f_k and X_k.

Approximate relationship between β and the pausing index I_P

The model above requires iterative estimation, but an approximate closed-form expression for β in terms of the pausing index I_P can be obtained by noting that the excess reads in the pause region, t, can be estimated reasonably well by multiplying the difference between the average read depths in the pause region and gene body by the length of the pause region, $L = k_{\max} - k_{\min} + 1$:

$$t \approx L\left(\hat{\chi}_P - \hat{\chi}\right) = L\hat{\chi}\left(I_P - 1\right), \qquad (16)$$

where $\hat{\chi}_P$ is the average read depth in the pause region and $I_P = \frac{\hat{\chi}_P}{\hat{\chi}}$ is the pausing index as computed by averaging across the pause region. Therefore,

$$\hat{\beta} = \frac{\hat{\chi}}{t} \approx \frac{1}{L(I_P - 1)}. \qquad (17)$$

Thus, an interpretable pause-escape rate parameter can be estimated approximately from the pausing index I_P. This equation provides a physical interpretation for I_P and also explains why our naive initial estimation strategy tended to overestimate β by a factor of approximately L. This strategy, however, does not allow estimation of the mean or variance of the pause site at each gene.

Allowing for steric hindrance of initiation

To accommodate steric hindrance of initiation at steady state, we introduce a distinction between a *potential* rate of initiation in the absence of occlusion of the initiation site, α, and the *effective* rate of initiation after a portion of initiation events are blocked by an existing RNAP molecule, which we denote ω. We assume $\omega = (1 - \phi)\alpha$, where φ is the probability of that the 'landing pad' required for a new initiation event is already occupied by an RNAP. Thus, $\omega \leq \alpha$. Notice that any estimation of initiation rates based on the density of RNAPs in the gene-body will be representative of ω, not α; a correction may be required to estimate α accurately.

We first assume that the 'footprint' of an engaged polymerase, ℓ, is sufficiently large that at most one RNAP can be present in this region at a time. (We relax this assumption below.) We further assume that elongation up to position k occurs much faster than the initiation rate αζ or the pause-escape rate βζ, so that the dynamics of elongation through the pause peak can be ignored. In this case, occupancy of the landing pad can be described using a simple two-state continuous-time Markov model (Figure 2A). Here, either the landing pad is unoccupied (state 0) and is therefore available for new initiation events, which occur at rate αζ; or the landing pad is already occupied by an RNAP (state 1) and no new initiation events can occur until that RNAP escapes from the pause site, which occurs at rate βζ. Thus, at steady state, the landing-pad occupancy φ is simply given by the stationary distribution of the occupied state,

$$\phi = \frac{\alpha\zeta}{\alpha\zeta + \beta\zeta} = \frac{\alpha}{\alpha + \beta}, \qquad (18)$$

and the effective initiation rate, allowing for steric hindrance, is given by,

$$\omega = \alpha(1 - \phi) = \frac{\alpha\beta}{\alpha + \beta}. \qquad (19)$$

Notice that these equations also imply that $\omega = \phi\beta$, meaning that, at steady state, the effective initiation rate ω must always be less than or equal to the pause-escape rate β, with ω approaching β as the landing-pad occupancy φ approaches unity. Therefore, if one estimates an effective initiation rate $\hat{\omega}$ and a pause-escape rate $\hat{\beta}$ from the data, as described above, then one can obtain estimates of φ and α as follows,

$$\hat{\phi} = \frac{\hat{\omega}}{\hat{\beta}}, \qquad \hat{\alpha} = \frac{\hat{\omega}}{1 - \hat{\phi}} = \frac{\hat{\omega}\hat{\beta}}{\hat{\beta} - \hat{\omega}}. \qquad (20)$$

Notice that the estimator for φ will be proportional to the read-counts in the pause peak (see, e.g., equation 7).

Steric hindrance with multiple RNAPs

As it turns out, the assumption of ≤1 RNAPs per pause region is often too restrictive, and the presence of more than one RNAP in this region can have a substantial impact on the landing-pad occupancy φ. In this section, we generalize the model for steric hindrance to allow for any number r of RNAPs in the pause region, focusing in particular on the case of $r \leq 3$, which we expect to cover essentially all plausible scenarios in human cells.

Let s_p be the minimum center-to-center spacing, in nucleotides, between adjacent RNAPs on the DNA template. As noted in the main text, structural data suggests s_p is at least 33 nt but more plausibly $s_p \approx 50$ nt (23,24,45,46); we also consider the case of $s_p = 70$ nt for comparison. We further assume that a new initiation event can successfully occur if, and only if, the previous RNAP has advanced to a position $i > s_p$ (in other words, the 'landing pad' for new initiation events has size $\ell = s_p$). Consequently, the maximum possible number of RNAPs in the pause region is $r = 1$ if the pause site $k \leq s_p$, $r = 2$ if $s_p < k \leq 2s_p$, $r = 3$ if $2s_p < k \leq 3s_p$, and so on (see Figure 2B). In general, $r = \lceil \frac{k}{s_p} \rceil$.

In addition, we assume a probability mass function f_k for the pause site k across cells, with cumulative distribution function $F(k) = \sum_{i=0}^{k} f_i$. Let q_1 be the density associated with $r = 1$; that is, $q_1 = F(k = s_p)$. Similarly, q_2 is the density associated
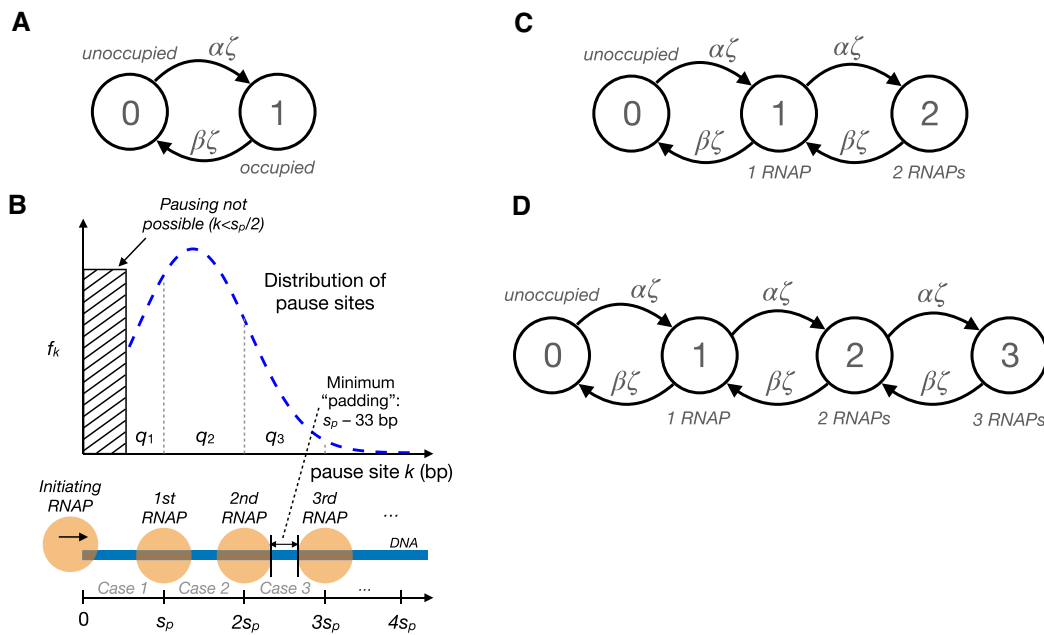
Figure 2. (**A**) Two-state continuous-time Markov model for steric hindrance of transcriptional initiation, assuming at most one RNAP at a time in the pause region. The pause region must be either unoccupied (state 0) or already occupied by another RNAP (state 1). Transitions from state 0 to state 1 occur at the (unimpeded) initiation rate, $\alpha\zeta$, and transitions from state 1 to state 0 occur at the pause-escape rate, $\beta\zeta$. The stationary frequency of state 1 defines the landing-pad occupancy ϕ and is given by $\frac{\alpha}{\alpha+\beta}$. (**B**) Illustration showing a hypothetical distribution of pause sites k and its implications for the number of RNAPs that can simultaneously occupy the pause region. When $k \leq s_p$, where s_p is the minimum center-to-center spacing between adjacent RNAPs, only one RNAP is possible (Case 1 in the text); when $s_p < k \leq 2s_p$, up to two are possible (Case 2); and when $2s_p < k \leq 3s_p$, up to three are possible (Case 3). Notice that the portion of the density corresponding to each Case r is given by q_r. (**C**) Generalization of Markov model to accommodate up to two RNAPs in the pause region (Case 2). (**D**) Further generalization to accommodate up to three RNAPs (Case 3). The equation for ϕ can be generalized to account for these cases (see text).

with $r = 2$, $q_2 = F(k = 2s_p) - F(k = s_p)$; and $q_3 = F(k = 3s_p) - F(k = 2s_p)$. In general, $q_r = F(k = rs_p) - F(k = (r-1)s_p)$ (Figure 2B).

Now, in each cell, k has a single value and therefore one of a series of mutually exclusive cases must apply. Let us denote by Case r (for $r \in \{1, 2, 3, ...\}$) that the maximum possible number of RNAPs is r. Thus, in Case 1, $r = 1$ and $k \leq s_p$; in Case 2, $r = 2$ and $s_p < k \leq 2s_p$; and so on. Given f_k, we know that Case r occurs with probability q_r. Therefore, we can calculate ϕ as a mixture of case-specific landing-pad probabilities, $\phi = \sum_{r=1}^{\infty} q_r \phi_r$, where ϕ_r is the probability that the landing pad (first s_p nucleotides) is occupied in case r. In practice, we approximate this quantity as $\phi \approx \sum_{r=1}^{R} q'_r \phi_r$ where R is the maximum plausible value of r (here, $R = 3$) and,

$$q'_r = \begin{cases} q_r & r \in \{1, 2, \ldots, R-1\} \\ 1 - \sum_{r=1}^{R-1} q_r & r = R, \end{cases} \quad (21)$$

where the last term, q'_R, is 'rounded up' to account for the remaining tail of the distribution.

The case of $r = 1$ has already been described in the previous section. It can be captured by a two-state model, where the landing-pad is either unoccupied (state 0) or occupied by a single RNAP (state 1; Figure 2A). Therefore,

$$\phi_1 = \frac{\alpha}{\alpha + \beta}. \quad (22)$$

It turns out that the cases of larger values of r follow naturally through the addition of more states. For example, when $r = 2$, the landing pad is either unoccupied (state 0), occupied by one RNAP (state 1), or occupied by two RNAPs (state 2). Assuming no two events can occur simultaneously, state 2 can

be reached only when a new initiation event occurs while state 1 is occupied (at rate $\alpha\zeta$), and a pause-release event in state 2 causes a return to state 1 (at rate $\beta\zeta$). The result is a chain of states as shown in Figure 2C. In addition, under the assumption that elongation through the pause peak is instantaneous, the landing pad is occupied if, and only if, state 2 is occupied. Thus, ϕ_2 is given by the stationary frequency of state 2 in this model, which can be shown to be,

$$\phi_2 = \frac{\alpha^2}{\alpha^2 + \beta^2 + \alpha\beta}. \quad (23)$$

Similarly, the case of $r = 3$ can be addressed by extending the chain further with a state 3, and setting ϕ_3 equal to the stationary frequency of that state (Figure 2D),

$$\phi_3 = \frac{\alpha^3}{\alpha^3 + \beta^3 + \alpha^2\beta + \alpha\beta^2}. \quad (24)$$

Therefore, assuming $R = 3$, we can estimate ϕ as,

$$\begin{aligned} \phi &= q_1\phi_1 + q_2\phi_2 + (1 - q_1 - q_2)\phi_3 \\ &= \frac{q_1\alpha}{\alpha+\beta} + \frac{q_2\alpha^2}{\alpha^2+\beta^2+\alpha\beta} + \frac{(1-q_1-q_2)\alpha^3}{\alpha^3+\beta^3+\alpha^2\beta+\alpha\beta^2}, \end{aligned} \quad (25)$$

allowing for the possibilities of one, two, or three RNAPs in the pause region of each cell. As s_p grows larger and/or f_k shifts toward the TSS, the fraction q_1 approaches 1, and Equation (25) approaches Equation (18).

As it turns out, substitution of $\omega/(1 - \phi)$ in place of α in this more general expression for ϕ leads to a complex polynomial that cannot be easily solved for ω. Instead, we solve this equation numerically to obtain ϕ and α from ω and β (analogous to equation 20).

Fitting the steric-hindrance model to data

The multi-RNAP steric hindrance model can be combined with the model for variable pause sites across cells and fitted to the data by a relatively straightforward extension of the EM algorithm described above. As in that case, we assume that the unscaled initiation rate, χ, is pre-estimated from data in the gene body. In addition, however, this case requires pre-estimation of the scale-factor λ and the elongation rate ζ, so that a scaled estimate of the effective initiation rate, $\omega = \frac{\chi \zeta}{\lambda}$, can be obtained from each estimate of χ. In practice, we obtain these scale factors by calibrating with respect to other studies (see 'Calibrating the initiation rate,' below).

In addition, we must address the problem that the landing-pad occupancy ϕ is constrained to fall between 0 and 1, but the relationships above do not enforce such a constraint. For example, in the single-RNAP case, where $\phi = \frac{\omega}{\beta}$ (equation 20), ϕ will be undefined whenever $\omega > \beta$. A similar (but more complicated) relationship holds in the multi-RNAP case.

To address this problem, we take a Bayesian approach and assume a (weakly) informative prior distribution for ϕ. This strategy not only restricts ϕ to the allowable range but has the benefit of regularizing the model when information in the data about ϕ is weak. With the assumption of a Beta$(\phi \mid a, b)$ prior for ϕ, with shape parameters a and b (we assume $a = b = 2$ throughout), the likelihood (cf. equation 8) becomes,

$$
\begin{aligned}
\mathcal{L}(\mathbf{X};\, \chi, \beta) = \int & \left[\prod_{i=1}^{k_{\min}-1} \text{Pois}\,(X_i \mid \chi) \right] \\
\times & \left[\prod_{k=k_{\min}}^{k_{\max}} \sum_{Y_k=0}^{X_k} \text{Pois}\,(Y_k \mid \chi/\beta \cdot f_k) \right. \\
& \left. \times \text{Pois}\,(X_k - Y_k \mid \chi \cdot (1 - f_k)) \right] \\
\times & \left[\prod_{i=k_{\max}+1}^{N} \text{Pois}\,(X_i \mid \chi) \right] \times \text{Beta}\,(\phi \mid a, b)\ d\phi,
\end{aligned}
$$

$$(26)$$

and the expected complete-data log likelihood (cf. equation 10) becomes,

$$
\begin{aligned}
\langle \ell_c(\mathbf{X}, \mathbf{Y};\, \chi, \beta, \phi) \rangle & \\
= s \log \chi & - \langle t \rangle \log \beta - \chi \left(N + \frac{1}{\beta} - 1 \right) \\
+ & \left[\sum_{k=k_{\min}}^{k_{\max}} \langle Y_k \rangle \log f_k + (X_k - \langle Y_k \rangle) \log(1 - f_k) \right] \\
+ & (a - 1) \log \phi + (b - 1) \log(1 - \phi) - \log Z. \quad (27)
\end{aligned}
$$

From a comparison of equations 10 and 27, it is evident that the calculation of the summary statistics $\langle t \rangle$, $\langle u \rangle$, $\langle v \rangle$, $\langle w \rangle$, $\langle z \rangle$, and $\langle r \rangle$ and the updates for μ and σ^2 will all remain unchanged (equations 12–15). However, this simplified presentation obscures that the parameters ϕ and β are implicitly linked by a function, $\phi = g(\beta; \chi)$, which is indirectly defined by equation 25 as well as by the relationship $\omega = (1 - \phi)\alpha$. Thus, the update for β in this case is no longer that shown in equation 11 but now must also consider the terms that depend on ϕ. As it turns out, in the full multi-RNAP model, rewriting equa-

tion 27 in terms of β only (i.e., by substitution for ϕ) leads to rather unwieldy polynomial expressions. Nevertheless, the M-step in the EM algorithm can be performed numerically without much trouble. In particular, on each iteration of the algorithm, we calculate all expected sufficient statistics as before (E-step), but then, for the M-step, we estimate β by numerically maximizing the portion of equation 27 that depends on β and ϕ, that is,

$$
\hat{\beta} = \arg\max_{\beta} \left[- \langle t \rangle \log \beta - \frac{\chi}{\beta} + (a - 1) \log \phi \right.
$$
$$
\left. + (b - 1) \log(1 - \phi) \right] \qquad (28)
$$

subject to the constraints of equation 25 and the relationship $\omega = (1 - \phi)\alpha$. Thus, the previous EM algorithm can be adapted to the multi-RNAP steric hindrance case by simply replacing the M-step for β with this numerical optimization. No other changes are required.

SimPol simulator

SimPol ('Simulator of Polymerases') tracks the independent movement of RNAPs along the DNA templates of a large number of cells (Figure 1C). It accepts several user-specified parameters, including the initiation rate, pause-escape rate, a constant or variable elongation rate, the mean and variance of pause sites across cells, as well as the center-to-center spacing constraint between RNAPs (s_p), the number of cells being simulated, the gene length, and the total time of transcription (Supplementary Table S2). The simulator simply allows each RNAP to move forward or not, in time slices of 10^{-4} minutes, according to the specified position-specific rate parameters. It assumes that at most one movement of each RNAP can occur per time slice. The simulator monitors for collisions between adjacent RNAPs, prohibiting one RNAP to advance if it is at the boundary of the allowable distance from the next. After running for the specified time, SimPol outputs a file in csv format that summarizes all RNAP positions across cells. An R script (also provided) can then be used to simulate nascent RNA sequencing read counts based on this file. SimPol is written in C++ and makes use of multi-threading to improve performance. A typical simulation (e.g., transcribing a 2kb gene in 20,000 cells for 400,000 steps) requires only a few minutes on a desktop computer.

Generation of synthetic NRS data

Using SimPol, we simulated genes 2000 bp in length with initiation ($\alpha\zeta$) and pause-escape ($\beta\zeta$) rates that spanned two orders of magnitude, ranging from 0.1 to 10 events per min. per cell. Elongation rates at each nucleotide position were randomly sampled from a truncated normal distribution, with mean = 2000 nt/min, sd = 1000 nt/min, min = 1500 nt/min and max = 2500 nt/min. When a fixed pause site was assumed, it occurred at position $k = 50$ nt; variable pause sites assumed a truncated normal distribution for k, with mean = 50 nt, sd = 25 nt, min = 17 nt and max = 200 nt. Our main simulations assumed a center-to-center spacing $s_p = 50$ nt, but alternative simulations assumed $s_p = 33$ nt and $s_p = 70$ nt. For each parameter combination, we simulated 20 000 cells for the equivalent of 40 min (400 000 time slices), which appeared to be sufficient to reach equilibrium in all cases (see also Supplementary Table S2 for a summary).

Based on the output of each SimPol run, we randomly sampled 5000 of the 20 000 cells. This sampling step was performed 50 times for each run. To save in computation, the same source collection of 20 000 cells was used for each replicate, after verifying that it made little difference to rerun the full simulation to equilibrium each time. For each of these replicates, we then sampled a read count at each position i from a Poisson distribution with mean μ_i such that,

$$\mu_i = \lambda_c \cdot \frac{R_i}{C_s}, \tag{29}$$

where R_i is the number of sampled cells having an RNAP whose active site (center) is at position i and $C_s = 5000$ is the total number of sampled cells. The scale parameter λ_c was calibrated such that a typical choice of initiation ($\alpha\zeta = 1$ events/min) and pause-escape ($\beta\zeta = 1$ events/min) rate parameters resulted in an average read depth of 0.049 reads/bp in the gene body, as we observed in the real data from ref. (19) (median value).

For simplicity, we performed this step separately at each nucleotide only in the pause region (the first 200 nucleotides). Because the RNAP densities throughout the gene bodies are fairly homogeneous, and the corresponding read counts are averaged anyway, we sampled the total read count for each gene body in one final step by scaling the Poisson distribution appropriately. This strategy also allowed us to extrapolate from our 2,000-bp simulated genes to genes of more realistic length. Specifically, the total read count for the gene body was sampled from a Poisson distribution with mean μ_{GB} such that,

$$\mu_{GB} = \lambda_c \cdot \frac{R_{GB}}{C_s} \cdot \frac{l_{targ}}{l_{sim}}, \tag{30}$$

where R_{GB} is the total number of RNAPs across the simulated gene body, $C_s = 5000$ is the total number of cells sampled, $l_{targ} = 19\,800$ bp is the target gene-body length, and $l_{sim} = 1,800$ bp is the simulated length. Notice that the scale parameter λ_c was held fixed throughout so that average read-depths would increase or decrease appropriately as the rate parameters were altered.

Analysis of real data

We obtained published K562 PRO-seq libraries from the heat shock (20) and celastrol studies (19) and processed them using the PROseq2.0 pipeline (https://github.com/Danko-Lab/proseq2.0) in single-end mode (47). The 3′ ends of reads—which approximately represent the active sites of isolated RNAPs—were recorded in bigWig files and used for analysis. Mapping was performed with human genome assembly GRCh38.p13 and gene annotation were downloaded from Ensembl (release 99) in GTF (48). Annotations of protein-coding genes from the autosomes and sex chromosomes were used, excluding overlapping genes on the same strand. To improve TSS positioning, we augmented the gene annotations with CoPRO-cap (Coordinated Precision Run-On and sequencing with 5′ Capped RNA) data for K562 cells from ref. (16) (see (26,49) for similar uses of GRO-cap data). In particular, we used the position with highest CoPRO-cap signal within a 250 bp radius around each annotated TSS as a refined TSS and discarded genes for which no CoPRO-cap signal was found. We then considered the 200 bp starting at this refined TSS as the 'pause region,' and the region from 1250 bp to up to 90 kb downstream (but not past the annotated end of

the gene) of this TSS as the 'gene body.' We excluded any gene with fewer than 20 reads mapped to either the pause peak or the gene body. For the remaining genes, the read counts at each of 200 positions in the pause region, and the total read counts in the gene body, were summarized in a table and used for all downstream analyses.

Calculation of half-lives

RNAP half-lives were calculated from estimates of the pause-escape rate β and average values of the pause-site position k by the following equation,

$$T_{1/2} = \log 2 \cdot \frac{\bar{k} - 1 + 1/\hat{\beta}}{\zeta}, \tag{31}$$

where $\bar{k} = 50$ nt, $\zeta = 2000$ nt/min, and $\hat{\beta}$ is the estimated pause-escape rate. Here, the quantity $(\bar{k} - 1 + 1/\hat{\beta})/\zeta$ represents the expected time required for an RNAP to pass through the pause region and the pause site, and the factor $\log 2$ converts the mean of an exponential distribution to a half-life.

Discriminative motif finding

DNA sequences 200 bp downstream of the TSS were extracted, and STREME (50) was used to identify the motifs enriched in the 10% genes with narrowest pause peaks (smallest $\hat{\sigma}^2$) compared with the 10% of genes with broadest peaks (largest $\hat{\sigma}^2$). Tomtom (51) was then used to annotate the results with known motifs from HOCOMOCO v11 (52). To validate the sequence motif enrichment of TAF1, ChIP-seq signals for TAF1 in K562 cells downstream of the TSS were extracted separately for genes exhibiting narrow and broad peaks. These signals were plotted using the R package Genomation (53).

Calibrating the initiation rate

As described in the Results section, we made use of a 'low' (L) calibration of 0.2 initiation events per minute (54–57), and a 'high' (H) calibration based on ref. (26), who reported a median of 1.0 initiation events per minute for mRNAs in K562 cells. These calibrations were performed separately for each of our four analyzed data sets: the untreated and treated samples from the heat-shock (20) and celastrol (19) studies. We performed all calibrations using 'housekeeping' (HK) genes from ref. (58), to minimize sensitivity to differences between assays and other properties. We first identified a subset of genes that fell in the intersection of the HK set, the genes from (26), and each of our sets. This subset numbered between 1079 and 1082 genes for our four data sets. In each case, for the L calibration, we simply scaled our ω values so that the median value of $\omega\zeta$ within this set was equal to 0.2 events/min in the control samples. For the H calibration, we scaled our ω values so that the median value of $\omega\zeta$ within this set matched the median value reported by Gressel et al. (26) for the same subset of genes, which were 1.52 and 1.53 events/min in the control samples. The $\omega\zeta$ in treated samples were then scaled based on the ratio of spike-in reads (19) or total mappable reads (20) between the control and treated samples. Finally, to focus on genes exhibiting robust expression in both the treated and untreated samples, we identified a subset of genes that fell in the top 80% by $\hat{\chi}$ in both samples. This filter resulted in a total of 6,182 genes for the heat-shock data set and 5,964 genes for the celastrol dataset.

Results

An evaluation with simulated data reveals strengths and limitations of the initial model

To examine the quality of the maximum-likelihood estimators for the initial probabilistic model (Figure 1A, B, see Materials and Methods), we developed a fast, flexible computational simulator, called SimPol ('Simulator of Polymerases'), that tracks the progress over time of individual RNAPs across DNA templates in thousands of cells under user-defined initiation, pause-escape, and elongation rates (Figure 1C). Unlike our assumed model, SimPol tracks potential collisions between RNAPs and prohibits one RNAP from passing another along the template. In addition, it allows for variation across cells in pause-site location and variation across both cells and nucleotide positions in local elongation rate (see Materials & Methods for details). After running to equilibrium, the simulator generates synthetic read counts at each nucleotide position by counting the RNAPs across cells, and then sampling read counts conditional on local RNAP density (see Figure 1D for an example gene with synthetic data).

Using SimPol, we generated synthetic data sets for a range of plausible initiation and pause-escape rates, with read depths corresponding to those of genes with median expression levels from real data (see Materials & Methods). We performed separate sets of simulations for cases where the pause site is fixed at a known value k and for cases where k is allowed to vary across cells according to a truncated Gaussian distribution. We assumed a mean value of k equal to 50 nt and a standard deviation of 25 nt, approximately as we observe in real data (see Materials & Methods), and required a minimum center-to-center distance between adjacent RNAPs of 50 nt (as in (25)). For each simulated data set, we estimated χ and β from the synthetic data and compared the estimated and true rates. In cases where the pause site was variable, we naively estimated β by using the average read depth across a 200 bp pause peak region, reasoning that such an averaging approach is typically used in computing a pausing index (e.g. (19)). However, we considered other estimators as well (see below). In addition, for β, we evaluated the product βζ—which represents the absolute rate of pause-escape—assuming the mean value of ζ = 2 kb/min that was used for simulation. For α, we calibrated our estimates relative to a baseline case simulated with αζ = 1 and no pausing and again assumed ζ = 2 kb/min. In this way, we were able to account for both the sequencing read depth and the elongation rate, and express both the initiation and pause-escape rates in absolute units of events per minute, for ease of interpretation.

We found that these initial estimates of α and β were accurate in some cases but significantly biased in others (Figure 3, Supplementary Figure S1). In particular, estimates of the initiation rate α were close to the truth when α was not too large and β was not too small; but either a moderately high α or a moderately low β led to a notable downward bias in the α estimates (Figure 3A). This bias ranged from ∼50% when α was high or β was low but the other parameter was in the favorable regime, to more than an order of magnitude when both parameters were unfavorable. As we explore below, these biases suggest an influence from steric hindrance of new initiation events owing to RNAPs in the pause peak, which would be expected to lead to under-estimation of α precisely when pause-escape rates are low and/or initiation rates are high. Importantly, these biases can be pronounced for plausible values

of α and β, suggesting that it can be misleading to treat average read counts in the gene body as a measure of the initiation rate (see Discussion).

By contrast, in the case where the pause site was held fixed and assumed known during estimation, estimates of β were accurate across a range of true α and β values (Figure 3B; Supplementary Figure S1), indicating that the model describes the dynamics of pause-escape well. When the distance to the pause site, k, was allowed to vary across cells, however, and the read-depth in the pause peak was estimated by averaging across sites, the increased read density owing to pausing was dramatically under-estimated, resulting in a strong upward bias in estimates of β (Figure 3C; Supplementary Figure S1). In this case, the 'spike' of read counts becomes a rounded 'peak', and using its average height across the region naturally over-estimates β. As it turns out, this simple source of bias is deceptively difficult to eliminate. For example, if the maximum, rather than the average, read-depth in the pause peak is used in the estimator for β, the bias is somewhat reduced but remains pronounced (Supplementary Figure S1). A better approach is to use the sum of read counts across the 200 bp pause-peak region (Figure 3D). Even in this case, however, a downward bias in β is observed when the initiation rate equals or exceeds the pause-escape rate. To remedy this problem, we develop an extension to the model in the next section.

Explicitly modeling pause-site variability across cells improves estimates of pause-escape rates

To address the bias in estimation of β in the presence of variable pause sites, we extended the model to allow for a distribution of pause sites across cells (see Materials & Methods). In this version of the model, the read counts at each site in the pause peak are assumed to arise from a mixture of cells in which an RNAP is and is not paused at that position. If a Gaussian distribution of pause sites is assumed, the model can be fitted to the data relatively simply and efficiently by expectation maximization (EM). This procedure results in maximum-likelihood estimates not only of χ and β but also of the mean and variance of the pause-site position across cells. Thus, this version of the model no longer requires prior information about the pause-site k, but instead allows its mean and variance to be estimated separately at each gene from the raw data.

We reanalyzed our simulated data using this version of the model and found that it was highly effective at correcting the bias in estimated values of β. In this case, the same model addresses the fixed and variable pause-site scenarios equally well across a range of values of α and β (Figure 4A; Supplementary Figure S2). Whereas the averaging approach described in the previous section resulted in over-estimates of β by more than two orders of magnitude, that bias was completely eliminated when variability in the pause site was modeled directly. Thus, typical average-based estimates of a 'pausing index' (I_P) for real data may substantially under-estimate the prominence of promoter-proximal pausing in RNAP dynamics (see Discussion). The model addresses this problem. Only when α is quite small and β is quite large (bottom of Supplementary Figure S2), resulting in few excess reads in the pause peak, do the model-based estimates sometimes exhibit a substantial downward bias. Notably, while the full model requires an iterative method such as EM, it is possible to derive an approximate closed-form expression for β in terms of the pausing index I_P,
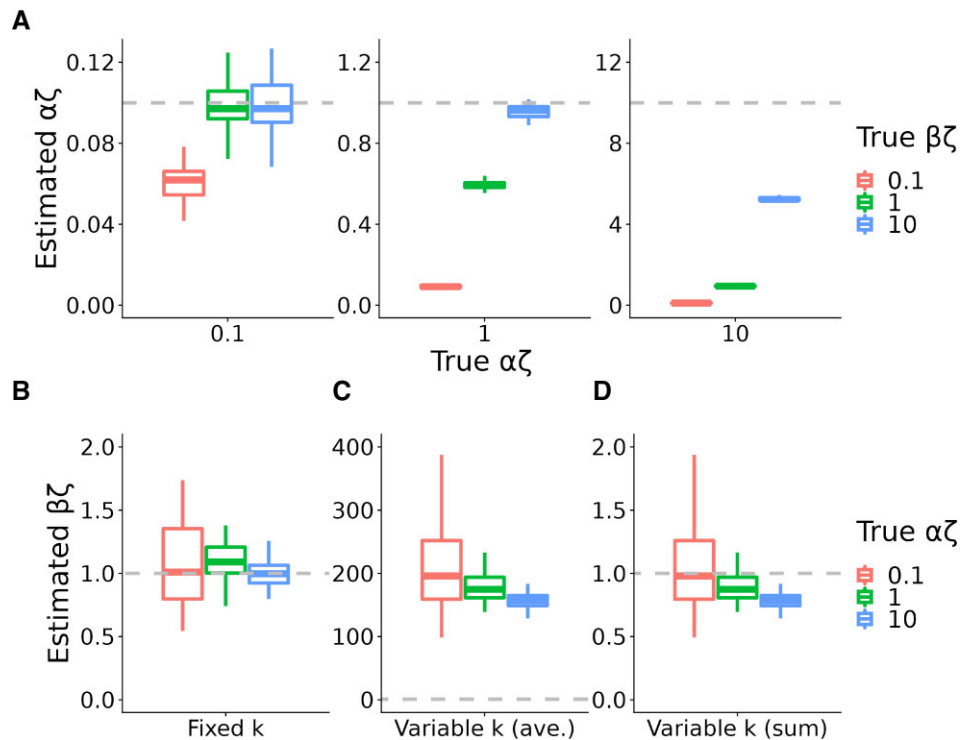
Figure 3. Accuracy of estimated values of the transcription initiation rate α and pause-escape rate β under the initial version of the model. Estimates are expressed as products with the elongation rate ζ ($\alpha\zeta$ and $\beta\zeta$). (**A**) Simulated true vs. estimated values of $\alpha\zeta$, for $\alpha\zeta \in \{0.1, 1, 10\}$ (left to right) and $\beta\zeta \in \{0.1, 1, 10\}$ (see key). Dashed lines indicate the ground truth. (**B–D**). Estimated values of $\beta\zeta$ for simulated true values of $\beta\zeta = 1$ and $\alpha\zeta \in \{0.1, 1, 10\}$ (see key), when the pause-site k is fixed (**B**) or variable across cells (**C, D**) in simulation. In panel C, β is estimated using the average read-depth in the pause peak, and in panel D it is estimated using the sum of read counts across the region. Dashed lines indicate the ground truth. Results for other values of $\beta\zeta$ are shown in Supplementary Figure S1. All boxplots summarize 50 replicates of the simulation; box boundaries indicate 1st and 3rd quartiles, and horizontal line indicates median. A value of $\zeta = 2$ kb/min is assumed; $\alpha\zeta$ and $\beta\zeta$ can be assumed to have units of events per minute. Pause sites occur at a mean position of $k = 50$ nt. In the variable case, we assume a Gaussian distribution with a standard deviation of 25 nt.

which clarifies the physical meaning of I_P (see Materials & Methods and Discussion).

To validate the ability of our model to capture differences in the pause-site distributions across cells, we applied it to synthetic data sets with larger and smaller variances in pause-site locations. We found that the model was able to recover the correct distributions fairly accurately, as long as the read counts in the pause peak were not too sparse (Figure 4B). Notably, the model is provided with no prior information about the mean or variance in the pause-site location, but is only constrained to consider a range of possible values (here from $k_{min} = 1$ to $k_{max} = 200$). When applied to real data, the model also appears to do well at identifying plausible distributions across cells in the locations of pause sites (Figure 4C).

Estimates of pause-escape rates and pause-site distributions for real data

Having observed good performance on synthetic data, we applied the model to published PRO-seq data for K562 cells before and after heat shock (20), focusing at first on the untreated sample. In comparison to results from the previous version of the model, the distribution of β estimates was substantially shifted downward, as expected from simulations (Figure 4D). For comparison with experimental results (e.g. (25,26,59)), we converted these estimated pause-escape rates to half-lives for RNAP residence in the pause region, conditional on an assumed elongation rate of $\zeta = 2$ kb/min (18,25,26,59). For a direct comparison, we considered the

time required for elongation up to the pause site, as well as the waiting time for escape (see Materials & Methods). We estimated a median half-life of 1.4 min. and a mean of 2.8 min. (Supplementary Figure S3A). These estimates are fairly similar to ones obtained experimentally by Gressel et al. for human Raji B cells (25) (median pause duration of 1.4 min, corresponding to a median half-life of 1.0 min) and for K562 cells (26) (median pause duration of 1 min), but somewhat smaller than those from some other recent experimental studies (e.g. (59); see Discussion). We additionally examined PRO-seq data from another study of K562 cells (19) and estimated a similar distribution of half-lives, with a median of 1.2 min. and mean of 2.1 min. (Supplementary Figure S3B). Notably, induction of heat shock and treatment with the drug celastrol both increased half-lives by 2–3-fold (Supplementary Figure S3C, D), consistent with the observations of increased pause indices in the original reports (19,20).

Interestingly, the pause peaks estimated for both of these data sets varied substantially across genes in their breadth, as quantified by the estimated variance across cells in the pause-site position ($\hat{\sigma}^2$). Hypothesizing that the underlying DNA sequences might contribute to the precision of pausing, we searched for sequence motifs that distinguished the pause regions having the narrowest peaks (10% with smallest $\hat{\sigma}^2$) from those having the broadest peaks (10% with largest $\hat{\sigma}^2$), focusing again on the untreated samples (see Materials & Methods). In both the heat-shock (20) and celastrol (19) data sets, we found that the most strongly enriched motif in the narrow peaks closely matched the binding motif for
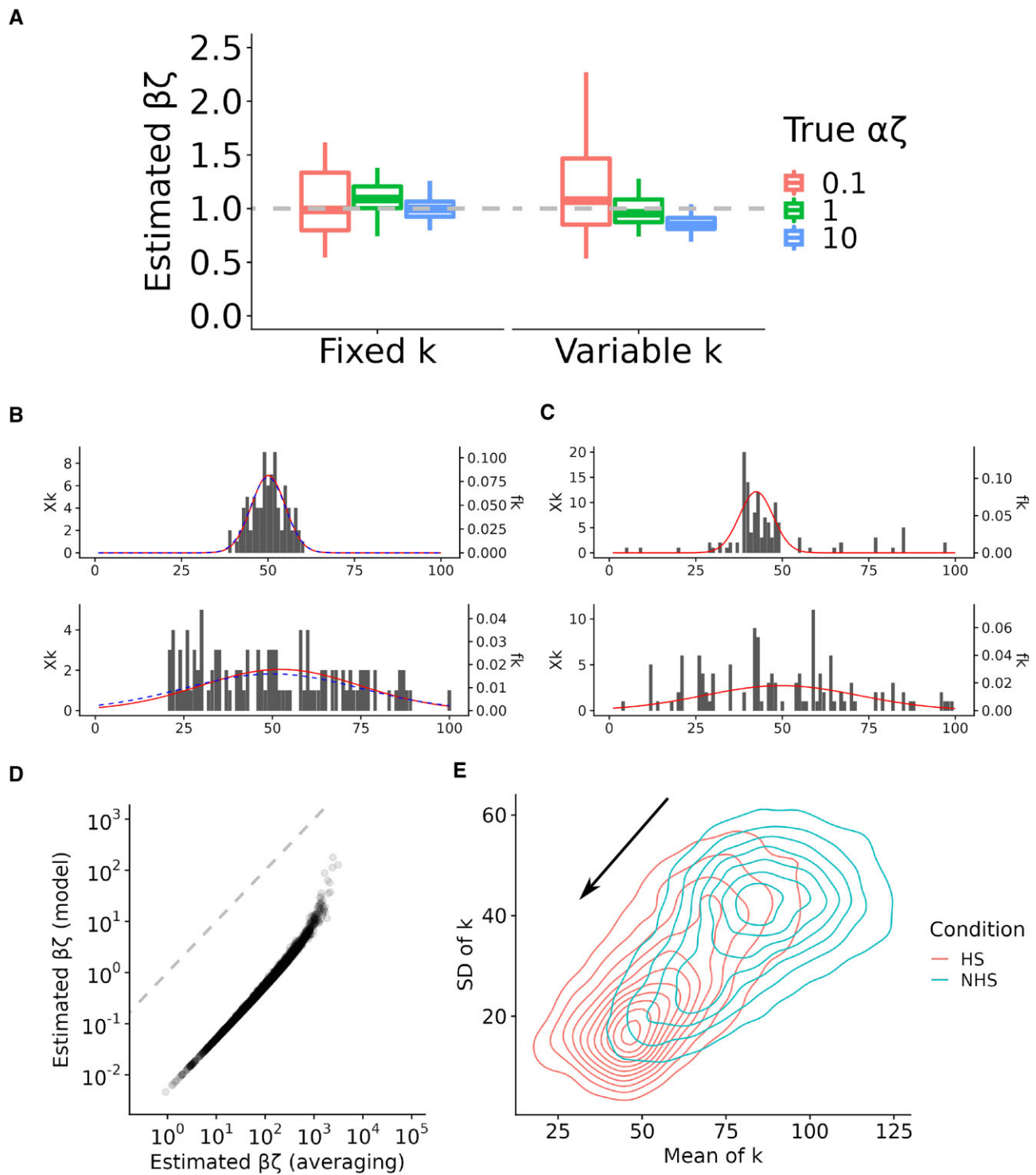
Figure 4. (**A**) Accuracy of estimated values of the pause-escape rate β under the version of the model that allows for a distribution of pause-sites k across cells. Shown are estimated values of βζ for simulated true values of βζ = 1 and αζ ∈ {0.1, 1, 10} (see key), when the pause-site k is fixed (left) or variable across cells (right) in simulation. Dashed lines indicate the ground truth. Results for other values of βζ are shown in Supplementary Figure S2. All boxplots summarize 50 replicates of the simulation; box boundaries indicate 1st and 3rd quartiles, and horizontal line indicates median. Simulated pause sites occurred at a mean position of k = 50 nt. In the variable case, we assumed a Gaussian distribution with a standard deviation of 25 nt. (**B**) Examples of pause peaks in simulated data, showing assumed distribution of pause sites (blue dashed line) and distribution inferred by expectation maximization (red solid line). (**C**) Similar examples from real data from (20). (**D**) Estimates of βζ under the original averaging approach (horizontal axis) vs. estimates of βζ under the model that allows for variable k across cells (vertical axis). (**E**) Contour plot showing the distribution of estimated means (horizontal axis) and standard deviations (vertical axis) of the pause peak position k, under the 'no heat shock' (NHS) and 'heat shock' (HS) conditions. Data from (20). In panels A and D, a value of ζ = 2 kb/min is assumed; thus, αζ and βζ can be assumed to have units of events per minute.

TATA-Box Binding Protein Associated Factor 1 (TAF1), the largest subunit of general transcription factor IID (TFIID) and a key component of the pre-initiation complex (PIC) ([60],[61]) (Supplementary Figures S4 and S5, panels A & B). Consistent with these motif enrichments, chromatin immunoprecipitation and sequencing (ChIP-seq) data for TAF1 in K562 cells from the ENCODE project ([62]) exhibits a substantially stronger signal at narrow peaks than at broad peaks, particularly near the TSS (Supplementary Figures S4C and S5C). Interestingly, it has recently been shown that the PIC alone is sufficient to establish RNAP pausing, and that rapid TAF1 depletion induces pause-release genome-wide ([46]). We found other enriched motifs, but they mostly consisted of repetitive, highly G+C-rich sequences, similar to observations in other recent studies ([25],[63],[64]).

We were also interested in possible effects of cellular stress on pause-site locations. Metaplots summarizing the accumulated PRO-seq signal across large classes of genes have suggested that pause peaks may tend to shift in location and become sharper following stress, particularly for downregulated genes (e.g., ([20],[33])). To test whether our model supported such a change, we compared our estimates of the pause site locations, before and after application of heat shock (Figure [4]E). Indeed, we observed a striking reduction in both the mean and variance of k after heat shock. This shift may reflect differences in activity of protein complexes required for pausing, including NELF, DSIF, P-TEFb, or the PIC (e.g. ([46],[65])). For comparison, we examined PRO-seq data from the celastrol study, which reported a stress response that resembles heat shock in some ways ([19]). In this case, however, we did not observe a clear change in the mean and variance of k (Supplementary Figure S6).

A modeling extension to accommodate steric hindrance in initiation

It seemed likely that the observed under-estimation of α was driven by steric hindrance of initiation from paused RNAPs (Figure [3]A), as has been noted in previous studies of both simulated and real data ([24]–[27]). To ensure that steric hindrance was indeed driving this phenomenon in our setting, we collected two types of auxiliary data from our simulation experiments. First, we tracked the fraction of cells in which the 'landing pad' for a potential new initiation event was already occupied by an RNAP ('landing-pad occupancy'), which is a close proxy in our simulations for the fraction of potential initiation events that were not allowed to occur owing to steric hindrance. We found that this fraction was frequently quite high (often 50% or more, and in some cases ≥90%), and tended to be highest when the bias in estimated α was most pronounced (Supplementary Figure S7A). Second, we measured the rate at which initiation events successfully occurred (not being blocked) and compared it with our estimates of α, finding much better agreement between this 'effective' (sometimes called 'productive' ([25],[26])) initiation rate and our estimates (Supplementary Figure S7B). Finally, we found that our estimation accuracy for α (as measured by the ratio of the estimated to true values) had a close negative correlation with the landing-pad occupancy (Supplementary Figure S7C). Together, these findings indicate that new initiation events in our simulations are frequently blocked by RNAPs occupying the region just downstream of the TSS, and that this phenomenon explains much, if not all, of the bias in estimation of α.

We therefore extended our probabilistic model to address this problem and permit a quantitative description of steric hindrance of initiation at steady state. Briefly, we introduced a distinction between an effective initiation rate ω and a potential initiation rate α, letting $\omega = (1 - \phi)\alpha$, where ϕ is the landing-pad occupancy. The model assumes that a fraction ϕ of new initiation events are blocked by RNAPs occupying the landing pad, and the remaining fraction $1 - \phi$ are allowed to proceed (see Materials & Methods for complete details). It allows for more than one RNAP per pause region, because the landing pad will likely be blocked in some cases by RNAPs stacking up behind the pause site. Based on our observations above, we reinterpreted our previous estimator for α instead as an estimator for the effective rate ω (and hence, redefined $\chi = \lambda\frac{\omega}{\zeta}$). This new model allows us to compute ϕ in terms of α, β, k, and an assumed minimum spacing between RNAPs, denoted s_p. An extended EM algorithm allows joint estimation of ϕ, α, and β from the data, with a Beta prior distribution to ensure that ϕ remains in the allowable range, $\phi \in (0, 1)$ (Materials & Methods). In effect, this strategy allows us to correct the assumed model for steric hindrance, yielding estimates not only of the effective initiation rate ω and pause-escape rate β, as in the previous case, but also of the landing-pad occupancy ϕ and, hence, the potential initiation rate α.

When we applied this new model to our simulated data, we found that it generally appeared to work as intended. The estimates of ϕ were broadly consistent with an empirical measure of the landing-pad occupancy across a range of values of α and β (Figure [5]A), Not surprisingly, given the indirect information about ϕ in the data, the variance in these estimates was substantial; nevertheless, their mean values were close to the truth even for large ϕ, although the Beta prior did sometimes produce a slight downward bias in this case. The estimates of the pause-escape rate β remained quite good overall even in the presence of substantial amounts of steric hindrance (Supplementary Figure S8).

The most difficult parameter to recover was the potential initiation rate α, about which the data are only weakly informative via the effective initiation rate ω and the landing-pad occupancy ϕ. Nevertheless, α was estimated well when $\alpha \leq \beta$ and $\phi \leq 0.5$ (Supplementary Figure S9). When $\alpha \gg \beta$ and $\phi \rightarrow 1$, however, the denominator of the estimator $\alpha = \frac{\omega}{1-\phi}$ becomes unstable, and in the most extreme cases (e.g. $\alpha\zeta = 10$, $\beta\zeta = 0.1$), it is no longer possible to estimate α accurately. In these cases, the effective initiation rate is still estimated well, but it is not possible to extrapolate from it to the potential rate, because only a tiny fraction of potential initiation events are allowed to occur.

Our initial experiments assumed a value of $s_p = 50$ nt for the minimum center-to-center spacing between adjacent RNAPs (following ([25])), but the true value of s_p is not known with any certainty; therefore, we examined the robustness of our model to other plausible values of this parameter. In general, smaller values of s_p will increase the chances that a single paused RNAP will not block the landing pad, but will also allow more RNAPs to stack up behind the pause site. In contrast, when s_p grows larger and approaches k, the distance to the pause-site, at most one RNAP can occupy the pause region at a time, and that RNAP will necessarily block the landing pad (see Materials & Methods for details). We used SimPol to generate data sets for two alternative choices of s_p: a more generous value of $s_p = 70$ nt and the minimum possible value of $s_p = 33$ nt ([24]). The case of $s_p = 33$ nt is probably
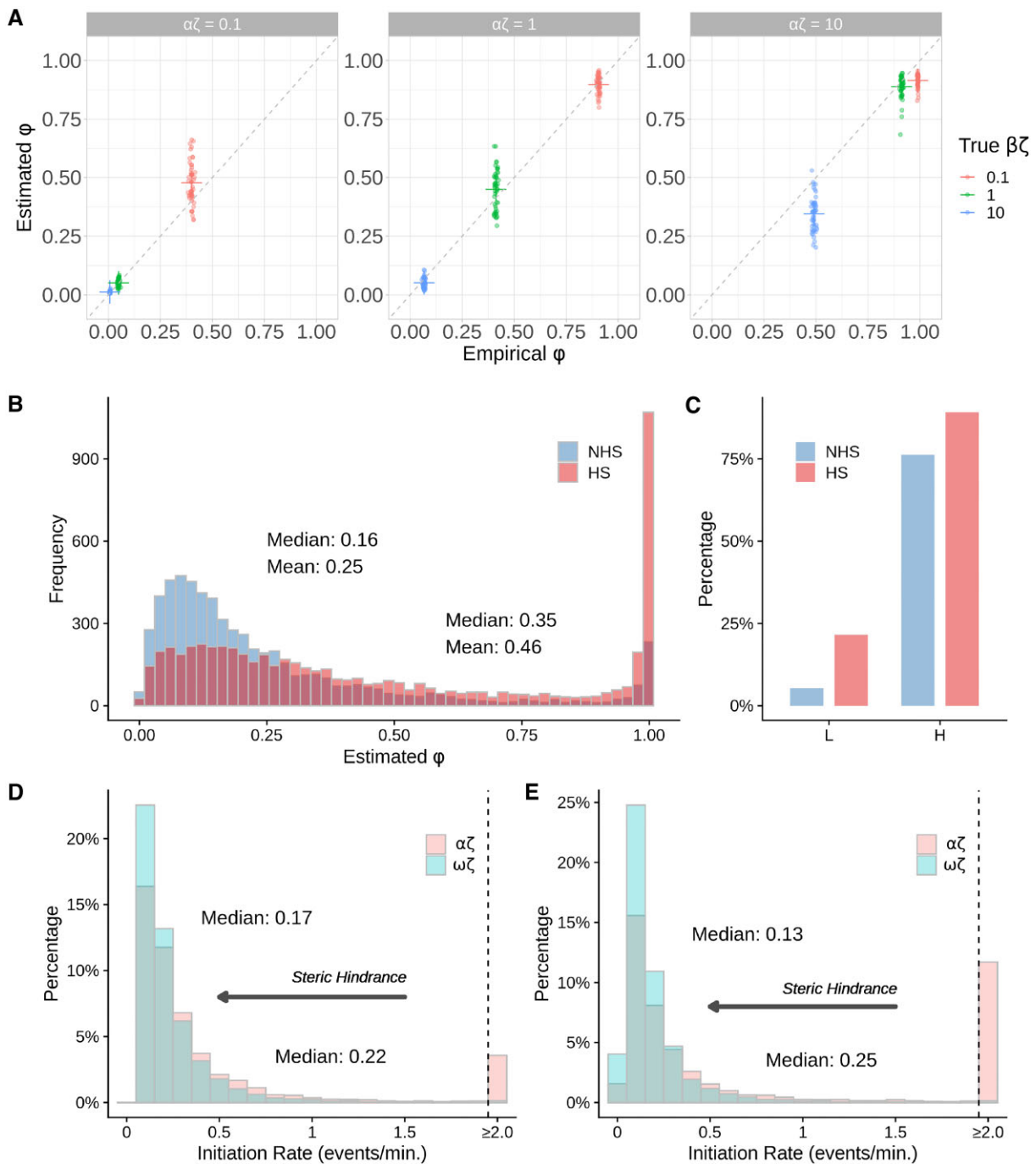
Figure 5. (**A**) Accuracy of estimated landing-pad occupancy φ under the version of the model that allows for steric hindrance in initiation and multiple RNAPs per pause region. Scatter plots show the fraction of simulated cells for which the first 50 nt (the 'landing-pad') are occupied by an RNAP at steady state ('Empirical φ') vs. the fraction predicted to be occupied under the model ('Estimated φ') based on the simulated NRS data, assuming a minimum spacing of $s_p = 50$ nt. Results are shown for simulated true values of $\alpha\zeta \in \{0.1, 1, 10\}$ (left to right) and $\beta\zeta \in \{0.1, 1, 10\}$ (see key), with 50 simulations per parameter combination. Dashed line indicates $y = x$, and colored crosses represent the means of the corresponding points. A value of $\zeta = 2$ kb/min is assumed, so that $\alpha\zeta$ and $\beta\zeta$ are in events per minute. (**B**) Distribution of estimated φ for 6182 robustly expressed genes in K562 cells before (NHS) and after (HS) heat shock under the low (L) calibration (20) (see Materials & Methods for details). (**C**) Percentages of genes having fully occupied landing-pads (φ > 0.95) before (NHS) and after (HS) heat shock, under the low (L) and high (H) calibrations. (**D, E**) Distributions of scaled estimates of the 'effective' (ωζ) and 'potential' ($\alpha\zeta$) rates of transcription initiation, in events per minute per cell, for the same genes. Panel D represents the NHS case and panel E represents the HS case. The *x*-axes are truncated to highlight the bulk of the distributions. Gray arrows indicate effects of steric hindrance.

unrealistic given the physical space required for the other components of the PIC (e.g., (45,46)) but is nevertheless useful as a bound. In our simulations, where k has a mean of 50 and standard deviation of 25, ≥ 2 paused RNAPs are possible at 82%, 54%, and 23% of cells for $s_p = 33$, $s_p = 50$, and $s_p = 70$ nt, respectively; and ≥ 3 paused RNAPs are possible at 28%, 2.4% and 0.02%, respectively.

The estimates of ϕ did not differ dramatically as s_p was varied, but this sensitivity analysis did reveal two types of bias (Supplementary Figure S10). First, when s_p was small and $\alpha \approx \beta$, ϕ tended to be somewhat overestimated. This overestimation in ϕ appears to be a consequence of a slight underestimation of β in the presence of multiple RNAPs per pause region (Supplementary Figure S11), which are ignored by our variable pause-site model. Second, we observed a downward bias when $\alpha\zeta$ and $\beta\zeta$ both took the maximum value of 10, apparently because RNAPs in transition between paused positions, which are ignored by the model, make a nonnegligible contribution to landing-pad occupancy in this highly saturated case. This effect is most evident when $s_p \geq 50$. Overall, however, the ϕ and β estimates were reasonably accurate across all parameter values, indicating that the model is fairly robust to assumptions about RNAP spacing.

A quantitative characterization of steric hindrance in K562 cells

Using this model we re-analyzed the untreated K562 sample from Vihervaara et al. (20), focusing on 6,182 robustly expressed genes (the top 80% by $\hat{\chi}$). When fitting the model to real data, however, a problem of scale arises: the effective initiation rate ω can only be estimated up to a scale factor, which is determined by the (unknown) ratio of the read depth to the elongation rate, λ/ζ. At the same time, the pause-escape rate β—whose estimator is a ratio of two summary statistics (equations 7 and 11)—has a fixed scale; thus, the estimator for the landing-pad occupancy, $\hat{\phi} \approx \omega/\beta$ (equations 20 and 25), depends on the choice of scale for ω.

To address this problem, we calibrated the scale of ω using published estimates of the initiation rate. Because there is considerable uncertainty about this quantity in the literature, we selected two different calibration points, one on the low end of the reported range for median productive initiation rates in eukaryotes, at 0.2 events per minute (54,66), and another on the higher end, at ~ 1.0 events/min (median of estimates reported by (26)). We used a set of 'housekeeping' genes (58) for calibration, reasoning that these genes would minimize sensitivity to differences among species, cell types, and conditions. We scaled our estimates of ω in the control samples such that the median value of $\omega\zeta$ within housekeeping genes in our set was either equal to 0.2 (the *low* (L) calibration) or equal to the median value in matched housekeeping genes from ref. (26), which was 1.5 events/min (the *high* (H) calibration; see Materials & Methods for details).

In the case of the L calibration, we found that the distribution of ϕ estimates across genes spanned a fairly broad range, with a median of 0.16 and a mean of 0.25 (Figure 5B), suggesting that, on average, about a quarter of landing pads are occupied at steady state, but considerable fractions of genes have substantially lower or higher occupancies. About 5% of genes were predicted to have their landing-pads fully occupied ($\phi \geq 0.95$; Figure 5C). The scaled estimates of the effective initiation rate $\omega\zeta$ were roughly exponentially distributed, with a median of 0.17 events per minute per cell (Figure 5D). By combining our estimates of ϕ and $\omega\zeta$, we were additionally able to obtain per-gene estimates of the potential initiation rate $\alpha\zeta$. These estimates were modestly inflated, with a median of 0.22 and $\sim 3\%$ of values ≥ 2.0 events per minute (Figure 5D). Thus, steric hindrance appears to result in a reduction in the effective initiation rate in these cells, but the effect is fairly subtle in this condition.

When we assumed the H calibration, the ϕ distribution shifted accordingly to the right. Indeed, in this case 76.3% of genes were predicted to have fully occupied landing pads ($\phi \geq 0.95$; Figure 5C), and as a result, the majority of $\alpha\zeta$ values became poorly defined, suggesting that this initiation-rate calibration is perhaps too aggressive. Nevertheless, by considering a broad range of potential initiation-rate calibrations, we estimated that the landing-pad occupancy is generally fairly substantial, with the median value of ϕ ranging from 16% to full occupancy. Thus, our model predicts that steric hindrance has a substantial impact on initiation rates at steady state, even in the untreated condition.

For comparison, we repeated the analysis with the treated sample, following heat shock (HS). Consistent with observations of increased pausing after HS (20,26), we found a dramatic shift toward larger ϕ estimates in this sample (Figure 5B). With the L calibration, the fraction of genes with fully occupied landing pads ($\phi \geq 0.95$) increased from 5.4% to 21.5% (Figure 5C). Accordingly, steric hindrance was predicted to have a substantially stronger impact on the effective initiation rate after HS, decreasing from a median of 0.25 events/min to 0.13 events/min (Figure 5E). We also analyzed untreated and treated samples from the K562/celastrol study (19) using 5,964 genes with initiation rates calibrated using similar L and H strategies (Materials & Methods). The results were qualitatively similar to those from the heat-shock analysis, but the absolute ϕ estimates were somewhat lower, shifting from a median of 0.13 in the untreated case to 0.30 after treatment (L calibration; Supplementary Figure S12A). Application of celastrol resulted in a striking increase in fully occupied landing-pads, from 2.7% to 14.8% of genes under the L calibration, but as in the heat shock analysis, the H calibration indicated nearly complete landing-pad occupancy before treatment and a limited shift after treatment (Supplementary Figure S12B). Prediction of $\alpha\zeta$ indicated a modest reduction in initiation rates from steric hindrance before treatment (Supplementary Figure S12C) and a more pronounced one after treatment (Supplementary Figure S12D). Overall, we found that steric hindrance has a clear impact on productive initiation, and that impact is particularly striking during responses to cellular stress.

Characterization of heat-induced and heat-repressed genes

The previous analysis grouped all genes together, potentially obscuring important differences among up- and down-regulated genes. Therefore, we re-examined the K562 heat-shock data from Vihervaara *et al.* (20), this time focusing on groups of genes designated in that study as heat-induced (503 genes) or heat-repressed (4399 genes). We first compared estimates of the effective initiation rate ($\omega\zeta$) and the pause-escape rate ($\beta\zeta$) for these two sets of genes, separately considering the untreated, or non-heat shock (NHS), and heat-shock (HS) conditions. As expected, the model did

find that, under HS, the effective initiation rates of designated heat-repressed genes significantly decreased, whereas those of designated heat-induced genes significantly increased (Supplementary Figure S13A). In addition, we found that heat-repressed genes showed a significant decrease in the pause-escape rate under HS, suggesting that promoter-proximal pausing contributes to down-regulation of these genes (Supplementary Figure S13B) as concluded by the authors (20) (see also (21)). By contrast, the heat-induced genes showed relatively little change in the pause-escape rate under HS. While increased rates of pause-escape are known to be important in the HS response (67), initiation rates have also been reported to increase at up-regulated genes (20,22). Our observations suggest that the increased initiation rates are more durable, and dominate once a new equilibrium is reached. We carried out a parallel analysis of the untreated and treated samples from the K562/celastrol study (19) with similar results (Supplementary Figure S14).

Even among up- or down-regulated genes, changes in gene expression can in principle be driven more at the transcription-initiation stage or at the pause-escape stage. Our steady-state model cannot fully reconstruct the dynamics of a change in transcriptional output, but we hypothesized that it would, in at least some cases, provide an indication of which stage had been dominant. To address this question, we searched for examples of genes that fell in each of four distinct classes: (A) down-regulated primarily at pause-escape; (B) down-regulated primarily at initiation; (C) up-regulated primarily at pause-escape; or (D) up-regulated primarily at initiation. The normalized PRO-seq read counts, together with model-based estimates of the effective initiation rate ($\omega\zeta$) and the pause-escape rate ($\beta\zeta$) for two examples of each class are shown in Supplementary Figure S15A–D. In each case, the PRO-seq signal in the gene body indicates clear up-regulation or down-regulation, but the behavior at the pause peak relative to the gene body indicates differences in the primary driver of the change in expression. For example, the gene *EMC2*, which is down-regulated nearly four-fold in response to heat shock (Supplementary Figure S15A), shows a marked increase in the PRO-seq signal at the pause peak relative to the gene body, leading to a 15-fold decrease in its estimated pause-escape rate, suggesting that pause-escape is the driver of down-regulation. By contrast, the gene *AGL*, which is down-regulated about two-fold, shows almost no change in the estimated pause-escape rate (Supplementary Figure S15B), suggesting that down-regulation is driven by initiation. These examples demonstrate that our steady-state model can identify cases where changes in expression are likely to have been driven by initiation or pause-escape. Future extensions to nonequilibrium models and time-course data will improve this type of analysis.

Discussion

The widespread occurrence of promoter-proximal pausing has been one of the major surprises of the past ~15 years in the study of gene regulation. Most studies of this phenomenon have focused on its molecular mechanisms and its direct impact on rates of productive elongation (17). In addition, however, there have been indications that such pausing, when sufficiently pronounced, also imposes indirect limits on rates of transcription initiation (24–27). In this article, we have shown through analysis of simulated and real data that many aspects of this complex interplay between initiation and pause-escape can be captured by relatively simple probabilistic models.

Our models not only describe raw read counts as a function of variable rates of initiation and pause-release, enabling estimation of these rates from NRS data, but they also allow for variable pause sites across cells, for reductions in initiation rates via steric hindrance, and for the effects of multiple RNAPs stacking up behind the pause site. We have shown by simulation that both variable pause sites and steric hindrance can have major impacts on the estimated rate parameters. Our analysis of real data indicates not only that pause sites tend to be highly variable, but that their degree of variability is correlated with particular sequence motifs as well as with stress responses. Similarly, steric hindrance of transcription initiation appears to occur at many genes and is intensified in stress responses. Both of these phenomena appear to play major roles in shaping the patterns of aligned reads from NRS data, particularly near the 5′ ends of transcription units. We have focused in our initial work on applications to PRO-seq data, where read counts can be assumed to reflect the density of engaged, transcriptionally competent RNAPs. Additional work will be required to see if our methods can be adapted for use with protocols that more broadly identify chromatin-associated RNAs or Pol II molecules, such as mNET-seq or caRNA-seq (reviewed in (35)).

Our analysis shows that, even at steady-state, it is possible to obtain accurate estimates of relative initiation and pause-escape rates. If, in addition, separate estimates are available of the average elongation rate per gene, then the initiation and pause-escape rates can be expressed in absolute terms as numbers of events per cell per unit time. In turn, these estimated rates can be used to obtain various downstream quantities of interest, such as pausing half-lives (Supplementary Figure S3), landing-pad occupanies (Figure 5B, C, Supplementary Figure S12A, B), or, through combination with other data sources, half-lives of RNA molecules or related quantities (68). Even when estimated elongation rates are not available, it may be adequate for some applications to estimate 'ballpark' rates based on the average rate of elongation across genes, which appears to be fairly stable at ~2 kb/min in mammalian cells (18,25,26,59). We have adopted this approach here in approximating absolute initiation rates, landing-pad occupancies (Figure 5; Supplementary Figure S12), and pausing half-lives (Supplementary Figure S3) from publicly available NRS data. We should emphasize, however, that our per-gene estimates are only approximate and may be biased by differences across genes in elongation rate.

With these caveats, our estimates of the half-lives of paused RNAPs—with mean values of 2–3 min. in untreated K562 cells and ~6 min. after heat shock or celastrol treatment (Supplementary Figure S3)—agreed reasonably well with previous experimental estimates. For example, Jonkers et al. measured a mean half-life of 6.9 min. at ~3200 genes in mouse ES cells (59); Shao et al. found most half-lives were between 5 and 20 min at 2300 genes in *Drosophila* Kc167 cells (27); and Gressel *et al.* estimated a median pause duration of 1.4 min. (corresponding to a half-life of 1.0 min.) for 2135 genes in human Raji B cells (25). (The same group subsequently estimated a similar rate for 6355 protein-coding genes in K562 cells (26).) In addition, Henriques et al. reported half-lives of promoter Pol II complexes from *Drosophila* S2 cells of ~2–15 min at selected loci (69). However, differences in species,

cell types, treatments and sets of genes make these estimates difficult to compare precisely. Notably, the largest estimates (27,59) were obtained after treatment with triptolide, which could potentially lead to some inflation in the estimates (31). Despite these differences, our method seems to be in general agreement overall with previous findings that paused Pol II is typically stable for minutes, and sometimes for tens of minutes, although we detect fewer extremely long half-lives than some previous studies. It is likely that our estimator for β reaches saturation when pause peaks become unusually elevated, leading to reduced sensitivity for the extreme tail of the distribution of pausing half-lives. At the same time, our estimator may be more sensitive to short half-lives than some experimental methods.

Despite such limitations, the use of steady-state rather than time-course data also has some important advantages. This approach allows us to analyze all expressed genes, not just a subset at which expression can be induced. In addition, it requires no chemical treatment to block initiation or pause escape, and therefore avoids potential off-target effects. At the same time, it is worth noting that our framework can be extended to the nonequilibrium setting and the use of time-course data (see bioRxiv: https://doi.org/10.1101/2021.01.12.426408). In that setting, it could be used to infer gene-specific elongation rates together with the other quantities. This version of the model, however, is considerably more complicated mathematically, more difficult to fit to data, and harder to interpret. Additional work will be required before it can be applied to real data.

Our estimator for the pause-escape rate β is closely related to the quantity known as the 'pausing index' (I_P), which is frequently used to measure the prominence of promoter-proximal pausing (14,44). In the case where the pause site is constant across cells, we have shown that I_P is precisely the inverse of the maximum likelihood estimator for the rate parameter β. When the pause site varies across cells, however, it becomes less straightforward to interpret I_P in physical terms, and a naive interpretation may lead to a strongly biased characterization of the pause-escape rate (e.g. Figure 4D). The details of how I_P is calculated—e.g. by averaging the read counts in the pause region or using their maximum value (Figure 3C and Supplementary Figure S1)– -also become important in this case. We have shown that our model has a natural extension to variable pause sites, which permits estimation of both the rate parameter β and the distribution of pause sites per gene by expectation maximization. In this setting, β no longer has a simple relationship to I_P. Interestingly, however, it is possible to find an approximate closed-form relationship in the case where I_P is calculated by averaging, namely, $\frac{1}{\beta} \approx L(I_P - 1)$, where $L = k_{\max} - k_{\min} + 1$ is the length of the pause region (see Materials & Methods, equation 17). We anticipate that this relationship may help to standardize definitions of I_P and clarify its physical meaning.

Our model also has implications for how to interpret the average read depth in the gene body ($\hat{\chi}$ in our notation), which is a natural measure of the transcriptional output at each gene. The model makes clear—as other investigators have previously argued (25,26)—that $\hat{\chi}$ is a measure of the rate of 'effective' or 'productive' transcription initiation at steady state, rather than of the 'potential' rate. That is, in our notation, χ is proportional to ω rather than α. Moreover, in the presence of steric hindrance, $\omega \approx \phi\beta$ (equation 19; see also equation 25),

which means that ω (and hence χ) is effectively limited by the pause-escape rate β. Notably, this equation provides a simple and interpretable characterization of the 'pause-initiation limit' described by Gressel *et al.* (25,26). It says that the effective initiation rate ω can never be greater than the pause-escape rate β, with equality when the landing-pad is fully occupied (φ = 1). The physical meaning of the 'potential' initiation rate α is somewhat less clear but, at least when φ is not too close to one, our model does allow extrapolation to a larger rate at which initiation events would hypothetically occur in the absence of steric hindrance. These values can be contrasted with the effective initiation rate in assessing the importance of steric hindrance.

We should emphasize that our model does depend on some crucial simplifying assumptions. Perhaps the most restrictive of these is the assumption that any initiating RNAP will eventually make its way along the entire DNA template, with no possibility of premature termination (PT). Similarly, we assume that the elongation process will not be blocked or substantially reduced by exogenous factors, such as UV-induced DNA damage. It seems likely that PT does occur to a degree, although its prevalence remains controversial. Several investigators have argued that the dynamics of RNAP occupancy near the promoter is dominated by pausing, rather than PT, because promoter-associated RNAP complexes are quite stable and predominantly localize to the pause region (27,59,69) (reviewed in (17)). However, others have argued, based on imaging or footprinting techniques, that PT may play a much more prominent role (31,70). Even if PT does occur at appreciable rates, its effect on our estimates will depend on precisely where and when it occurs. For example, if PT tends to occur immediately after initiation, it may have little effect on our estimates of the effective initiation or pause-escape rates. Indeed, if it occurs within a few nucleotides of the TSS, it may be largely invisible to NRS assays. If, on the other hand, PT tends to occur farther into the pause region, it will tend to influence estimates of the pause-escape rate—causing them to be too low (because some apparently paused RNAPs are actually lost to PT)—but not those of the effective initiation rate. We cannot see any way of fully incorporating PT in a probabilistic model based on NRS data or any other readout of RNAP density—even one that applies to the nonequilibrium setting—without relying on external estimates of the position-dependent rate at which PT occurs; otherwise it will necessarily be confounded with other rates. Until such estimates become available, model-based estimates of the pause-escape rate, in particular, must be considered with this caveat in mind.

A second process we ignore here is 'bursting' of transcription initiation. There is now substantial evidence that rates of transcription initiation do exhibit considerable temporal variation in actively transcribed genes, with apparent oscillations between active and inactive states (28–30,32,71,72). This pattern is strictly inconsistent with our assumption of a time-homogeneous Markov process. It is worth bearing in mind, however, that our model effectively averages over the processes that are occuring in a large population of cells. Unless those cells are synchronized in their bursting behavior, it would seem that averaging over 'on' and 'off' states in a bursting model, as our Markov model will effectively do, may be adequate for our purposes—although it is true that bursting could contribute to additional overdispersion in our read-count data.

Third, while we are able to model interactions among up to three RNAPs in the pause region, we ignore interactions in the gene body. Our simulations suggest that, at typical rates of initiation and pause-escape, collisions between RNAPs in the gene body should be fairly rare and have a limited impact on RNAP dynamics. However, they may become important under some conditions, when rates of transcription are unusually high. Our strategy of focusing on collisions among RNAPs in the pause region only and ignoring them in the gene body is what allows us to avoid working with the more general but difficult-to-analyze 'totally asymmetric simple exclusion process' (TASEP) (37,38) (see also (43)). Essentially, we simplify the modeling problem by exploiting the fact that, at least when promoter-proximal pausing is prominent, most relevant collisions occur in the restricted region immediately downstream of the TSS.

It is now widely accepted that promoter-proximal pausing is rate-limiting at many genes and this step is frequently regulated (17). Our model provides a quantitative demonstration of precisely how steric hindrance can modulate this important regulatory step. The model confirms that, in the presence of rate-limiting pause-escape (small β), promoting initiation (increasing α) will have little or no impact on transcriptional output, because the pause-initiation limit will quickly be reached. In this setting, a more rapid and effective transcriptional response will come from releasing RNAPs from the paused state (increasing β). At the same time, the model also reveals that the dynamic equilibrium between initiation and pause-escape will tend to limit the *duration* of the transcriptional response to pause-release. The reason is that the geometry of the pause region allows only a few RNAPs to be kept 'in the chamber,' ready to be released into productive elongation. As soon as these $\sim$1–3 RNAPs are released in each cell, initiation will again become rate-limiting. Thus, the effects of changes to initiation and pause-release cannot be fully separated; rather, the two processes interact with one another in a kind of dance, each limiting the other under certain circumstances. For this reason, each process is the natural target for regulation of gene expression in a different regime—for example, initiation when a sustained transcriptional response is required, and pause-escape when a rapid and/or synchronous but transient response is needed. Rates of initiation and pause-escape are perhaps best thought of as two 'knobs' for control of a single interrelated dynamical system near the TSS of each gene.

Data availability

PRO-seq fastq files for the heat shock (20) and celastrol (19) datasets were downloaded from GEO with accession numbers GSE89230 and GSE96869, respectively. BigWig files for CoPRO-cap data were downloaded with accession number GSE116472 (16). The bigWig file for TAF1 Chip-seq was downloaded from ENCODE with accession number ENCFF101GBL (73). All data are for human K562 cells, and replicates were combined for analysis.

Code availability

The SimPol simulator is available at https://github.com/CshlSiepelLab/SimPol. Our software for parameter estimation from NRS data is available at https://github.com/CshlSiepelLab/UniMod.

Supplementary data

Supplementary Data are available at NAR Online.

Conflict of interest statement

None declared.

References

1. Ptashne,M. and Gann,A. (1997) Transcriptional activation by recruitment. *Nature*, **386**, 569–577.
2. Rougvie,A.E. and Lis,J.T. (1990) Postinitiation transcriptional control in *Drosophila melanogaster. Mol. Cell Biol.*, **10**, 6041–6045.
3. Strobl,L.J. and Eick,D. (1992) Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO J.*, **11**, 3307–3314.
4. Krumm,A., Meulia,T., Brunvand,M. and Groudine,M. (1992) The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev.*, **6**, 2201–2213.
5. Rasmussen,E.B. and Lis,J.T. (1993) In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc. Natl. Acad. Sci. USA*, **90**, 7923–7927.
6. Plet,A., Eick,D. and Blanchard,J.M. (1995) Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene*, **10**, 319–328.
7. Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
8. Guenther,M.G., Levine,S.S., Boyer,L.A., Jaenisch,R. and Young,R.A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.
9. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
10. Muse,G.W., Gilchrist,D.A., Nechaev,S., Shah,R., Parker,J.S., Grissom,S.F., Zeitlinger,J. and Adelman,K. (2007) RNA polymerase is poised for activation across the genome. *Nat. Genet.*, **39**, 1507–1511.
11. Zeitlinger,J., Stark,A., Kellis,M., Hong,J.W., Nechaev,S., Adelman,K., Levine,M. and Young,R.A. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat. Genet.*, **39**, 1512–1516.

12. Lee,C., Li,X., Hechmer,A., Eisen,M., Biggin,M.D., Venters,B.J., Jiang,C., Li,J., Pugh,B.F. and Gilmour,D.S. (2008) NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Mol. Cell Biol.*, **28**, 3290–3300.

13. Nechaev,S., Fargo,D.C., dos Santos,G., Liu,L., Gao,Y. and Adelman,K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, **327**, 335–338.

14. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, **13**, 720–731.

15. 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*, **339**, 950–953.

16. Tome,J.M., Tippens,N.D. and Lis,J.T. (2018) Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.*, **50**, 1533–1541.

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

18. 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**, 212–222.

19. Dukler,N., Booth,G.T., Huang,Y.F., Tippens,N., Waters,C.T., Danko,C.G., Lis,J.T. and Siepel,A. (2017) Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. *Genome Res.*, **27**, 1816–1829.

20. Vihervaara,A., Mahat,D.B., Guertin,M.J., Chu,T., Danko,C.G., Lis,J.T. and Sistonen,L. (2017) Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nat. Commun.*, **8**, 255.

21. Vihervaara,A., Mahat,D.B., Himanen,S.V., Blom,M. A.H., Lis,J.T. and Sistonen,L. (2021) Stress-induced transcriptional memory accelerates promoter-proximal pause release and decelerates termination over mitotic divisions. *Mol. Cell*, **81**, 1715–1731.

22. Mahat,D.B., Salamanca,H.H., Duarte,F.M., Danko,C.G. and Lis,J.T. (2016) Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Mol. Cell*, **62**, 63–78.

23. Tornaletti,S., Reines,D. and Hanawalt,P.C. (1999) Structural characterization of RNA polymerase II complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template DNA. *J. Biol. Chem.*, **274**, 24124–24130.

24. Ehrensberger,A.H., Kelly,G.P. and Svejstrup,J.Q. (2013) Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell*, **154**, 713–715.

25. Gressel,S., Schwalb,B., Decker,T.M., Qin,W., Leonhardt,H., Eick,D. and Cramer,P. (2017) CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife*, **6**, e29736.

26. Gressel,S., Schwalb,B. and Cramer,P. (2019) The pause-initiation limit restricts transcription activation in human cells. *Nat. Commun.*, **10**, 3603.

27. Shao,W. and Zeitlinger,J. (2017) Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.*, **49**, 1045–1051.

28. Raj,A., Peskin,C.S., Tranchina,D., Vargas,D.Y. and Tyagi,S. (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, **4**, e309.

29. Zenklusen,D., Larson,D.R. and Singer,R.H. (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.*, **15**, 1263–1271.

30. Suter,D.M., Molina,N., Gatfield,D., Schneider,K., Schibler,U. and Naef,F. (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–474.

31. Steurer,B., Janssens,R.C., Geverts,B., Geijer,M.E., Wienholz,F., Theil,A.F., Chang,J., Dealy,S., Pothof,J., van Cappellen,W.A., *et al.* (2018) Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E4368–E4376.

32. Ali,M.Z., Choubey,S., Das,D. and Brewster,R.C. (2020) Probing mechanisms of transcription elongation through cell-to-cell variability of RNA Polymerase. *Biophys. J.*, **118**, 1769–1781.

33. Mahat,D.B., Kwak,H., Booth,G.T., Jonkers,I.H., Danko,C.G., Patel,R.K., Waters,C.T., Munson,K., Core,L.J. and Lis,J.T. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.*, **11**, 1455–1476.

34. Chu,T., Rice,E.J., Booth,G.T., Salamanca,H.H., Wang,Z., Core,L.J., Longo,S.L., Corona,R.J., Chin,L.S., Lis,J.T., *et al.* (2018) Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat. Genet.*, **50**, 1553–1564.

35. Wissink,E.M., Vihervaara,A., Tippens,N.D. and Lis,J.T. (2019) Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.*, **20**, 705–723.

36. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

37. MacDonald,C.T., Gibbs,J.H. and Pipkin,A.C. (1968) Kinetics of biopolymerization on nucleic acid templates. *Biopolymers: Orig. Res. Biomol.*, **6**, 1–25.

38. Zia,R., Dong,J. and Schmittmann,B. (2011) Modeling translation in protein synthesis with TASEP: A tutorial and recent developments. *J. Stat. Phys.*, **144**, 405–428.

39. Reuveni,S., Meilijson,I., Kupiec,M., Ruppin,E. and Tuller,T. (2011) Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput. Biol.*, **7**, e1002127.

40. Edri,S., Gazit,E., Cohen,E. and Tuller,T. (2014) The RNA polymerase flow model of gene transcription. *IEEE T. Biomed. Circuits Syst.*, **8**, 54–64.

41. Azofeifa,J.G. and Dowell,R.D. (2017) A generative model for the behavior of RNA polymerase. *Bioinformatics*, **33**, 227–234.

42. Erdmann-Pham,D.D., Dao Duc,K. and Song,Y.S. (2020) The key parameters that govern translation efficiency. *Cell Syst.*, **10**, 183–192.

43. Fischer,J., Song,Y.S., Yosef,N., di Iulio,J., Churchman,L.S. and Choder,M. (2020) The yeast exoribonuclease Xrn1 and associated factors modulate RNA polymerase II processivity in 5′ and 3′ gene regions. *J. Biol. Chem.*, **295**, 11435–11454.

44. Chen,F.X., Smith,E.R. and Shilatifard,A. (2018) Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **19**, 464–478.

45. Chen,X., Qi,Y., Wu,Z., Wang,X., Li,J., Zhao,D., Hou,H., Li,Y., Yu,Z., Liu,W., *et al.* (2021) Structural insights into preinitiation complex assembly on core promoters. *Science*, **372**, eaba8490.

46. Fant,C.B., Levandowski,C.B., Gupta,K., Maas,Z.L., Moir,J., Rubin,J.D., Sawyer,A., Esbin,M.N., Rimel,J.K., Luyties,O., *et al.* (2020) TFIID enables RNA polymerase II promoter-proximal pausing. *Mol. Cell*, **78**, 785–793.

47. Chu,T., Wang,Z., Chou,S.P. and Danko,C.G. (2019) Discovering transcriptional regulatory elements from run-on and sequencing data using the web-based dREG gateway. *Curr. Protoc. Bioinform.*, **66**, e70.

48. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S., *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.

49. Zhao,Y., Dukler,N., Barshad,G., Toneyan,S., Danko,C.G. and Siepel,A. (2021) Deconvolution of expression for nascent RNA sequencing data (DENR) highlights pre-RNA isoform diversity in human cells. *Bioinformatics*, **37**, 4727–4736.

50. Bailey,T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840.

51. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

52. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A., *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.

53. Akalin,A., Franke,V., Vlahoviček,K., Mason,C.E. and Schübeler,D. (2015) Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, **31**, 1127–1129.

54. Larson,D.R., Zenklusen,D., Wu,B., Chao,J.A. and Singer,R.H. (2011) Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, **332**, 475–478.

55. Claypool,J.A., French,S.L., Johzuka,K., Eliason,K., Vu,L., Dodd,J.A., Beyer,A.L. and Nomura,M. (2004) Tor pathway regulates Rrn3p-dependent recruitment of yeast RNA polymerase I to the promoter but does not participate in alteration of the number of active genes. *Mol. Biol. Cell*, **15**, 946–956.

56. Schneider,D.A., French,S.L., Osheim,Y.N., Bailey,A.O., Vu,L., Dodd,J., Yates,J.R., Beyer,A.L. and Nomura,M. (2006) RNA polymerase II elongation factors Spt4p and Spt5p play roles in transcription elongation by RNA polymerase I and rRNA processing. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 12707–12712.

57. Tongaonkar,P., French,S.L., Oakes,M.L., Vu,L., Schneider,D.A., Beyer,A.L. and Nomura,M. (2005) Histones are required for transcription of yeast rRNA genes by RNA polymerase I. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10129–10134.

58. Hounkpe,B.W., Chenou,F., de Lima,F. and De Paula,E.V. (2021) HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.*, **49**, D947–D955.

59. 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.

60. Louder,R.K., He,Y., López-Blanco,J.R., Fang,J., Chacón,P. and Nogales,E. (2016) Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, **531**, 604–609.

61. Patel,A.B., Louder,R.K., Greber,B.J., Grünberg,S., Luo,J., Fang,J., Liu,Y., Ranish,J., Hahn,S. and Nogales,E. (2018) Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science*, **362**, eaau8872.

62. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

63. Watts,J.A., Burdick,J., Daigneault,J., Zhu,Z., Grunseich,C., Bruzel,A. and Cheung,V.G. (2019) Cis elements that mediate RNA polymerase II pausing regulate human gene expression. *Am. J. Hum. Genet.*, **105**, 677–688.

64. Chou,S.P., Alexander,A.K., Rice,E.J., Choate,L.A. and Danko,C.G. (2022) Genetic dissection of the RNA polymerase II transcription cycle. *Elife*, **11**, e78458.

65. Aoi,Y., Smith,E.R., Shah,A.P., Rendleman,E.J., Marshall,S.A., Woodfin,A.R., Chen,F.X., Shiekhattar,R. and Shilatifard,A. (2020) NELF regulates a promoter-proximal step distinct from RNA Pol II pause-release. *Mol. Cell*, **78**, 261–274.

66. 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**, 796–806.

67. Vihervaara,A., Duarte,F.M. and Lis,J.T. (2018) Molecular mechanisms driving transcriptional stress responses. *Nat. Rev. Genet.*, **19**, 385–397.

68. Blumberg,A., Zhao,Y., Huang,Y.F., Dukler,N., Rice,E.J., Chivu,A.G., Krumholz,K., Danko,C.G. and Siepel,A. (2021) Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC Biol.*, **19**, 30.

69. Henriques,T., Gilchrist,D.A., Nechaev,S., Bern,M., Muse,G.W., Burkholder,A., Fargo,D.C. and Adelman,K. (2013) Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell*, **52**, 517–528.

70. Krebs,A.R., Imanci,D., Hoerner,L., Gaidatzis,D., Burger,L. and Schubeler,D. (2017) Genome-wide single-molecule footprinting reveals high RNA Polymerase II turnover at paused promoters. *Mol. Cell*, **67**, 411–422.

71. Fukaya,T., Lim,B. and Levine,M. (2016) Enhancer control of transcriptional bursting. *Cell*, **166**, 358–368.

72. Bartman,C.R., Hamagami,N., Keller,C.A., Giardine,B., Hardison,R.C., Blobel,G.A. and Raj,A. (2019) Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell*, **73**, 519–532.

73. Gertz,J., Savic,D., Varley,K.E., Partridge,E.C., Safi,A., Jain,P., Cooper,G.M., Reddy,T.E., Crawford,G.E. and Myers,R.M. (2013) Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell*, **52**, 25–36.

74. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.