Stochastic Kinetics of Nascent RNA

Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding

Phys. Rev. Lett. **117**, 128101 — Published 13 September 2016

DOI: 10.1103/PhysRevLett.117.128101

The Stochastic Kinetics of Nascent RNA

Heng Xu[1,2,3,*], Samuel O. Skinner[1,2,3], Anna Marie Sokac[3] and Ido Golding[1,2,3,*]

[1]Center for Theoretical Biological Physics, Rice University, Houston, Texas, USA

[2]Center for the Physics of Living Cells, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[3]Verna & Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas, USA

* hengx@bcm.edu, golding@bcm.edu

Abstract:

The stochastic kinetics of transcription is typically inferred from the distribution of RNA numbers in individual cells. However, cellular RNA reflects additional processes downstream of transcription, hampering this analysis. In contrast, nascent (actively transcribed) RNA closely reflects the kinetics of transcription. We present a theoretical model for the stochastic kinetics of nascent RNA, which we solve to obtain the probability distribution of nascent RNA per gene. The model allows us to evaluate the kinetic parameters of transcription from single-cell measurements of nascent RNA. The model also predicts surprising discontinuities in the distribution of nascent RNA, a feature which we verify experimentally.

Transcription, the production of RNA from a gene, is a stochastic process consisting of multiple single-molecule events [1,2]. The inference of transcription kinetics is typically addressed as an inverse problem, using the ergodic assumption that population statistics contain the signature of single-cell kinetics. Specifically, the number of RNA molecules from the gene is measured in many individual cells simultaneously using microscopy-based methods [3-5], and the measured RNA copy-number distribution is then compared to the prediction from a stochastic model for transcription kinetics [6-9]. This approach has been successfully used to demonstrate the bursty, non-Poissonian nature of transcription [6-8] and to examine how transcription kinetics are modulated by transcription factors [10-12].

However, mapping cellular RNA number to the underlying kinetics of transcription is hampered by the fact that this number reflects additional processes downstream of transcription, such as RNA degradation and its partitioning during cell division. The stochasticity of both processes may mask that of the transcription process [13,14]. Moreover, cellular RNA represents the combined contributions from multiple copies of the same gene, whose number changes through the cell cycle [14,15] and whose activity may be correlated [15-18].

In contrast to total cellular RNA, nascent RNA—the RNA molecules still actively transcribed at the gene—is not subject to these effects, and therefore bears more closely the signature of the transcription process. Recent progress in fluorescence microscopy has allowed measuring the amount of nascent RNA at individual genes in single cells [8,15,16,19-24]. However, the theoretical modeling of nascent RNA kinetics is only at its infancy [8,16,23,25-27]. We still lack a theoretical framework for mapping the single-cell measurements back to the stochastic kinetics of transcription. The goal of this paper is to develop such a framework.

The model: We model the kinetics of nascent RNA as consisting of four steps (**Fig. 1(a)**): Gene activation, transcription initiation, RNA synthesis (elongation), and release [8,16,23]. The gene fluctuates between two states, active (state 1), where transcription initiation is allowed, and inactive (state 0), where it is forbidden. Transitions between states and the initiation of transcription in the active state are modeled as Poisson processes, with rates k_{01}, k_{10} and k_{INI}, respectively [6,9,28,29]. Following initiation, RNA synthesis proceeds with a constant elongation speed V_{EL} [25,30], to a final length L. The completed RNA molecule remains on the gene for a (deterministic) duration T_{S} before being released [22,31]. See **Supplemental Material** [32] for a detailed discussion of model assumptions and of possible extensions to the model.

The state of the system is defined by two random variables, the gene state n ($n = 0, 1$) and the amount of nascent RNA m ($m \geq 0$). m is obtained by summing over all nascent RNA molecules present at the gene, and is measured in units of a single complete (mature) RNA [8,14,16]. Since nascent RNAs may be incomplete [14,22], m can have non-integer values. Here we generalize m to represent the experimentally measured signal from the nascent RNA. The actual value of m thus depends on the specific experimental observable (**Fig. 1(b)**). For example, in the case of single-molecule fluorescence *in situ* hybridization (smFISH, [3-5]), commonly used for RNA detection, m corresponds to the fluorescent signal emitted by oligonucleotide probes bound to the RNA. In all cases, the signal m at time t is determined by

initiation events happening within a time window $T_{RES} = L/V_{EL} + T_S$ (the residence time of RNA at the gene) prior to t, and the contribution from each nascent RNA molecule depends only on its length at time t. We define the contribution function $G(l)$ to describe the signal from a single RNA of length l [23]. Since l is determined by the difference between the RNA initiation time t_i and the observation time t, we can rewrite G as a function of this time difference, $g(\tau) = G(l(\tau))$, with $\tau = t_i - t$ ($-T_{RES} \le \tau \le 0$) and $l(\tau) = \min\{L, -V_{EL}\tau\}$ [16]. The observed signal is then given by $m(t) = \sum_{t-T_{RES} \le t_i \le t} g(t_i - t)$. The form of $g(\tau)$ reflects the experimental observable. A few examples are depicted in **Fig. 1(c)** and discussed in more detail below. In all cases, $g(\tau)$ is non-increasing, with the delay T_S in RNA release represented as a time period with $g = 1$.

General approach to solving the model. Because m exhibits a finite deterministic memory (over duration T_{RES}), we cannot easily write the master equation for the probability distribution $P(n,m)$. To overcome this problem and solve for the state of the system at time t, we first define the pseudo-observables $\nu(\tau,t) \equiv n(t+\tau)$, which indicates the gene state n at $t+\tau$, and $\mu(\tau,t) \equiv \sum_{t-T_{RES} \le t_i \le t+\tau} g(t_i - t)$, which describes the accumulation of m over the history from $t - T_{RES}$ to $t+\tau$. Here, τ varies from $-T_{RES}$ to 0. Notably, $\mu = 0$ for $\tau = -T_{RES}$ and $\mu = m$ for $\tau = 0$. Next, we write the master equation for the probability distribution $P(\nu,\mu)$ [16]:

$$\frac{d\mathbf{P}(\mu)}{d\tau} = (\mathbf{K} - \mathbf{K}_{INI})\mathbf{P}(\mu) + \mathbf{K}_{INI}\mathbf{P}(\mu - g(\tau)). \tag{1}$$

Here, $\mathbf{K} = \begin{bmatrix} -k_{01} & k_{10} \\ k_{01} & -k_{10} \end{bmatrix}$, $\mathbf{K}_{INI} = \begin{bmatrix} 0 & 0 \\ 0 & k_{INI} \end{bmatrix}$, and $\mathbf{P}(\mu) = \begin{bmatrix} P(0,\mu) \\ P(1,\mu) \end{bmatrix}$. Note that we allow μ to be negative, but **Eq. (1)** guarantees that $\mathbf{P}(\mu < 0) = 0$ as long as the initial conditions satisfy that condition. To obtain the distribution of the true observables (n,m), we solve **Eq. (1)** for the pseudo-observables (ν,μ) and substitute $\tau = 0$ (Alternatively, **Eq. (1)** can be used to derive an equation for $P(n,m)$, see **Supplemental Material** [32]).

We focus on the steady-state behavior of $P(n,m)$. Using the definition of μ and the (easily calculable) steady-state distribution for the gene state n, we obtain the initial condition $\mathbf{P}_{\tau=-T_{RES}}(\mu) = \frac{\delta(\mu)}{k_{01}+k_{10}} \begin{bmatrix} k_{10} \\ k_{01} \end{bmatrix}$. To solve **Eq. (1)**, we transform $P(\nu,\mu)$ to its characteristic function $\Psi(\nu,\omega) \equiv \int_0^\infty e^{i\mu\omega} P(\nu,\mu) d\mu$ [46] to obtain

$$\frac{d\mathbf{\Psi}(\omega)}{d\tau} = (\mathbf{K} + (e^{i\omega g(\tau)} - 1)\mathbf{K}_{INI})\mathbf{\Psi}(\omega), \tag{2}$$

with $\mathbf{\Psi}(\omega)=\begin{bmatrix}\Psi(0,\omega)\\ \Psi(1,\omega)\end{bmatrix}$ and the initial condition $\mathbf{\Psi}_{\tau=-T_{\mathrm{RES}}}(\omega)=\frac{1}{k_{01}+k_{10}}\begin{bmatrix}k_{10}\\ k_{01}\end{bmatrix}$. **Eq. (2)** is analogous to a quantum mechanical spin system with a time-dependent interaction term. Its solution is therefore given by the Dyson series [54]:

$$\mathbf{\Psi}_{\tau=0}(\omega)=\left\{\mathbf{I}+\sum_{N=1}^{\infty}\int_{-T_{\mathrm{RES}}}^{0}d\tau_1\cdots\int_{-T_{\mathrm{RES}}}^{\tau_{N-1}}d\tau_N\prod_{\tau_1\geq\cdots\tau_i\cdots\geq\tau_N}e^{i\omega g(\tau_i)}\mathbf{V}(\tau_i)\right\}e^{(\mathbf{K}-\mathbf{K_{INI}})T_{\mathrm{RES}}}\mathbf{\Psi}_{\tau=-T_{\mathrm{RES}}}, \tag{3}$$

where $\mathbf{V}(\tau)=e^{-(\mathbf{K}-\mathbf{K_{INI}})\tau}\mathbf{K_{INI}}e^{(\mathbf{K}-\mathbf{K_{INI}})\tau}$. Applying the inverse transformation, we obtain the steady-state distribution,

$$\begin{aligned}\mathbf{P}(m)&=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-im\omega}\mathbf{\Psi}_{\tau=0}(\omega)d\omega=\sum_{N=0}^{\infty}\mathbf{P}_N(m)\\&=\left\{\delta(m)+\sum_{N=1}^{\infty}\int_{-T_{\mathrm{RES}}}^{0}d\tau_1\cdots\int_{-T_{\mathrm{RES}}}^{\tau_{N-1}}d\tau_N\delta\left(m-\sum_{i=1}^{N}g(\tau_i)\right)\mathrm{T}\left[\prod_{i=1}^{N}\mathbf{V}(\tau_i)\right]\right\}e^{(\mathbf{K}-\mathbf{K_{INI}})T_{\mathrm{RES}}}\mathbf{\Psi}_{\tau=-T_{\mathrm{RES}}}.\end{aligned} \tag{4}$$

where T is the time-ordering operator. $\mathbf{P}_N(m)=\begin{bmatrix}P(0,m|N)\\ P(1,m|N)\end{bmatrix}$ is the vectorized probability of observing m, given that the number of initiation events in the time interval $-T_{\mathrm{RES}}\leq\tau\leq 0$ was exactly N. In the general case, $\mathbf{P}_N(m)$ depends on the contribution function $g(\tau)$, and therefore solving **Eq. (4)** requires knowing the specific form of $g(\tau)$. Below we describe the solution for a number of experimentally-relevant examples. A closed-form solution may not be always possible, but $P(n,m)$ can be calculated numerically using the finite state projection method [16,23,33,47](**Supplemental Material** [32]). For the purpose of comparing with experimental data, the calculated distribution is typically marginalized over n, i.e. $P(m)=\sum_n P(n,m)$. The moments of $P(m)$ can be directly calculated from **Eq. (2)** (**Supplemental Material** [32]):

$$\begin{aligned}\left\langle m^N\right\rangle&=\mathbf{u}\cdot(-i)^N\frac{d^N}{d\omega^N}\mathbf{\Psi}_{\tau=0}(0)\\&=\mathbf{u}\cdot\sum_{I=1}^{N}\sum_{\underbrace{0=k_0<k_i<k_I=N}_{i=1,\cdots,I-1}}\int_{-T_{\mathrm{RES}}}^{0}d\tau_1\cdots\int_{-T_{\mathrm{RES}}}^{\tau_{I-1}}d\tau_I\mathrm{T}\left[\prod_{i=1}^{I}\binom{k_i}{k_{i-1}}g(\tau_i)^{k_i-k_{i-1}}\mathbf{W}(\tau_i)\right]\mathbf{\Psi}_{\tau=-T_{\mathrm{RES}}}(0),\end{aligned} \tag{5}$$

with $\mathbf{u}=(1,1)$ and $\mathbf{W}(\tau)=e^{-\mathbf{K}\tau}\mathbf{K_{INI}}e^{\mathbf{K}\tau}$. Below we use these moments to explore the shape of $P(m)$ as a function of model parameters.

Solutions for specific contribution functions. **Case 1: $g = 1$**. This corresponds to measuring the number of RNA polymerases (RNAPs) currently transcribing the gene (**Fig. 1(c)**, panel I), or, equivalently, the number of nascent RNA molecules present, irrespective of their lengths [8]. Here and below we assume for simplicity that $T_S = 0$ (i.e. RNA is released from the gene immediately upon completion [16,19]), and (without loss of generality) set $T_{RES} = 1$. Since g in this case does not take fractional values, we replace the characteristic functions with generating functions, $F_\nu(z,\tau) \equiv \sum_{m=0}^{\infty} z^m P_\tau(\nu,m)$ and $F(z,\tau) \equiv F_0(z,\tau) + F_1(z,\tau)$, and transform **Eq. (1)** to obtain

$$\ddot{F} + (k_{01} + k_{10} + (1-z)k_{INI})\dot{F} + (1-z)k_{INI}k_{01}F = 0, \tag{6}$$

with the initial conditions $F(z,-1) = 1$, $\dot{F}(z,-1) = (z-1)\dfrac{k_{INI}k_{01}}{k_{01}+k_{10}}$. Solving **Eq. (6)** and performing the inverse transformation allows us to calculate the marginal probability distribution of m (see **Supplemental Material** [32]),

$$\begin{aligned} P(m) = \frac{e^{-\frac{k_{01}+k_{10}+k_{INI}}{2}}}{m!}\left(\frac{k_{INI}}{2}\right)^m \Bigg\{ \left[\frac{k_{01}+k_{10}+k_{INI}}{2} - \frac{k_{INI}k_{01}}{k_{01}+k_{10}}\right]\sum_{i=0}^{m}\binom{m}{i}M_{1,i} \\ + \sum_{i=0}^{m}\binom{m}{i}M_{0,i} + m\frac{k_{01}-k_{10}}{k_{01}+k_{10}}\sum_{i=0}^{m-1}\binom{m-1}{i}M_{1,i}\Bigg\}, \end{aligned} \tag{7}$$

with $M_{s,i} = \sum_{2l\geq i}\sum_{w=\max(0,i-l)}^{\min(l,i)}\binom{l}{w}\binom{l}{i-w}\frac{(-1)^i i!}{(2l+s)!}\left(\frac{k_{INI}+\kappa_1}{2}\right)^{l-w}\left(\frac{k_{INI}+\kappa_2}{2}\right)^{l-i+w}$, $\kappa_{1,2} = k_{10} - k_{01} \pm 2i\sqrt{k_{10}k_{01}}$.

Eq. (7) provides the exact solution for the distribution of the number of transcribing RNAPs at the gene.

Figure 2(a) depicts $P(m)$, calculated from **Eq. (7)**, for a few parameter values. Stochastic simulations of the model, also shown, agree with the analytical calculation (**Supplemental Material** [32]). For insight into the shape of $P(m)$, we first note that gene-state transitions are typically believed to be slow compared to both the rate of initiation and the time to complete one RNA [8,16,55]. Specifically, in the limit $(k_{01}\,\&\,k_{10}) \ll k_{INI}$ and $(k_{01}\text{or}\,k_{10}) \ll 1$, **Eq. (7)** can be written as the weighed sum of two Poisson distributions, with rates 0 and k_{INI} (**Supplemental Material** [32]). In this limit, $P(m)$ is also identical to the solution for the commonly used two-state model for cellular RNA kinetics [6,8,9,29], if we replace the residence time T_{RES} with the RNA degradation rate k_D. Outside that limiting case, however (as e.g. in [16]), the two distributions can be quite different (**Fig. S1** in the **Supplemental Material** [32]).

To map how the shape of $P(m)$ varies with transcription parameters, we defined the bimodality coefficient, $\beta \equiv 1/(\kappa - \gamma^2)$, where γ is the skewness and κ the kurtosis of $P(m)$ [51]. Calculating β over a broad range of kinetic rates, and using a threshold of $\beta_{th} = 5/9$ (corresponding to a uniform distribution, see **Supplemental Material** [32]), we found that $P(m)$ is bimodal for $k_{01} \sim k_{10} \lesssim 1$ and $k_{INI} \gtrsim 1$, and unimodal outside this region (**Fig. 2(b)**). The unimodal region can be further divided based on the position of the distribution peak, at $m = 0$ or $m > 0$ (**Fig. 2(b)**).

Case 2: $g = -\tau$. This corresponds to measuring the total length of nascent RNA, summed over multiple molecules present at the gene (**Fig. 1(c)**, panel II). Experimentally, this is achieved by using multiple smFISH probes covering the length of the target gene [4]. In contrast to Case 1 above, m is now continuous, and **Eq. (2)** can be transformed to a single equation for $\Psi(1,\omega)$ (**Supplemental Material** [32]):

$$\ddot{\Psi}(1,\omega) + \left[k_{01} + k_{10} + k_{INI}\left(1 - e^{-i\omega\tau}\right)\right]\dot{\Psi}(1,\omega) - \left[k_{INI}\left(k_{01} - i\omega\right)e^{-i\omega\tau} - k_{INI}k_{01}\right]\Psi(1,\omega) = 0, \qquad (8)$$

with the initial conditions $\Psi_{\tau=-1}(1,\omega) = \dfrac{k_{01}}{k_{01}+k_{10}}$, $\dot{\Psi}_{\tau=-1}(1,\omega) = \dfrac{k_{01}k_{INI}}{k_{01}+k_{10}}(e^{i\omega} - 1)$. By solving **Eq. (8)**, we obtain the exact expression for $\Psi(\omega) \equiv \Psi(0,\omega) + \Psi(1,\omega)$ as a combination of confluent hypergeometric functions. Since transforming $\Psi(\omega)$ back to an analytical form of $P(m)$ is challenging, we proceed to calculate $P(m)$ using finite state projection [16,23,33]. The calculated $P(m)$ exhibits the same three characteristic shapes as in Case 1, but the boundaries in parameter space between regions exhibiting different shapes are shifted by up to 2-fold (**Fig. S2** in the **Supplemental Material** [32]). Thus, the difference in contribution functions can lead to different shapes of $P(m)$ for the same transcription parameters (another example of this effect is described below).

Inferring transcription kinetics from single-cell measurements of nascent RNA. To demonstrate how the model can be used to interpret experimental data, we first examined the transcription of the *hunchback* (*hb*) gene in embryos of the fruit fly, *Drosophila melanogaster* ([16] and **Supplemental Material** [32]). Early in development, *hb* is regulated by the transcription factor Bicoid (Bcd), whose concentration forms a gradient along the embryo [56] (**Fig. 3(a)**). We measured the amount of nascent RNA at individual copies of the *hb* gene [16], and examined the distribution of nascent RNA over all cell nuclei within a given region of the embryo (corresponding to a given Bcd concentration) (**Fig. 3(a)**). Next, we solved **Eq. (1)** using $g(\tau)$ that corresponds to the set of smFISH probes used in the experiment [16], and used maximum likelihood estimation to fit the model to the experimental data. The model was able to capture the change in $P(m)$ shape along the embryo (**Fig. 3(a)**). We found that the regulatory effect of Bcd is to increase k_{01} (>50 fold along a single embryo) while k_{10} and k_{INI} remain almost unchanged (**Fig. 3(a)**). Thus, the model allowed us to identify what aspect of *hb* kinetics is modulated during gene regulation [16].

In the second example, we labeled the two halves of the same gene using two different smFISH probe sets carrying two different fluorescent dyes (**Fig. 3(b)** and **Supplemental Material** [32]). In the experiment, the two probe sets yielded very different signal distributions $P(m)$ (both normalized to the signal from a single full-length RNA). In particular, the signal from the first half of the gene was spread ~2 fold wider on the m axis than that from the second half (**Fig. 3(b)**). Since both probe sets label the same gene, the two data sets should be describable using the same kinetic parameters, the only difference being the form of $g(\tau)$, which we calculated directly from the probe positions on the gene (**Fig. 3(b)**). In agreement with this hypothesis, we were able to fit the two experimental distributions (as well as the joint distribution) using a single set of transcription parameters (**Fig. 3(b)** and **Supplemental Material** [32]).

Discontinuities in P(m). As noted above, a distinctive feature of nascent RNA, in contrast to mature cellular RNA, is that it can be approximated as continuous [4,5,16]. When examining the behavior of our model in the case $g=-\tau$ (i.e. measuring the total amount of nascent RNA at the gene), we found that, for multiple parameter choices, $P(m)$ appears discontinuous at integer values of m (insets of **Fig. S2(a)** in the **Supplemental Material** [32]). This discontinuity was consistent with the appearance of terms of order $1/\omega$ in the characteristic function $\Psi(\omega)$ [57]. The source of the discontinuity can be understood by noting that, in **Eq. (4)**, $P(m)$ is written as the sum of $P_N(m)$, the probabilities of observing m given that the number of initiation events in the time interval $-T_{\mathrm{RES}} \le \tau \le 0$ is N (equivalently, the number of RNAPs present at the gene is N). Since, for a given N, m cannot exceed N, the result may be a discontinuity of $P(m)$ or its derivatives at integer values. Specifically, since $P_0(m) \propto \delta(m)$, $P(m)$ has an infinite discontinuity at $m=0$. $P_1(m)$ is nonzero only for $m \le 1$, hence $P(m)$ has a jump discontinuity at $m=1$. For higher values of N, it can be shown that the $(N-1)^{\mathrm{th}}$ derivative of $P(m)$ has a jump discontinuity at $m=N$ (**Supplemental Material** [32]). For each point of discontinuity, the magnitude of the jump is

$$\Delta P_N = \left.\frac{d^{N-1}P(m)}{dm^{N-1}}\right|_{m=N^-} - \left.\frac{d^{N-1}P(m)}{dm^{N-1}}\right|_{m=N^+} = \frac{(-1)^{N-1}}{N!}\mathbf{u}\cdot e^{(\mathbf{K}-\mathbf{K}_{\mathrm{INI}})}\mathbf{K}_{\mathrm{INI}}^{N}\boldsymbol{\Psi}_{\tau=-1}. \qquad (9)$$

We explore this feature in **Fig. 4**. For the parameters used ($k_{01}=k_{10}=0.1$, $k_{\mathrm{INI}}=50$), zooming in to the low range of m reveals a sharp drop of $P(m)$ at $m=1$ (**Fig. 4(a)**). At higher integer m's, the drop becomes smaller and is shifted to the left (**Fig. 4(a)**). The drop reflects the discontinuity of $P(m)$ (or its derivatives) at integer m's. Each drop is preceded by an increase of $P(m)$, resulting in a peak at $m \to N^-$ (**Fig. 4(b)**). This peak, in turn, is due to the fact that, when $k_{\mathrm{INI}} \gg N$ and gene transitions are slow ($k_{01}, k_{10} \ll 1$), the two most probable ways of observing exactly N initiation events are for the gene to be active only at the beginning ($\tau \to -T_{\mathrm{RES}}^{+}$) or the end ($\tau \to 0^-$) of the time window, resulting in maxima of $P_N(m)$ at $m \to N^-$ and $m \to 0^+$, respectively (**Fig. 4(b)**).

To ask whether these features of $P(m)$ can be detected experimentally, we first defined the discontinuity factor $r \equiv \Delta P_1 / P(m=1^+)$ to characterize the magnitude of the jump in the distribution of nascent RNA. Calculating r over a wide range of kinetic rates indicated that it would be high (>0.1) for $k_{01} \lesssim 10^1$ (**Fig. 4(c)**). This range covers the estimated parameters in multiple biological systems [16,23,58], including our measurements in *Drosophila* (**Fig. 3** above). To then try and detect this feature in our experimental data, we focused on the small m (<6.5) range, where the peaks in $P(m)$ are expected to be the highest (**Fig. 4(a)**). To improve data sampling, we defined the variable $m_0 = m - [m]$ (where $[\cdot]$ denotes the nearest integer) such that all m values are mapped into the range $[-0.5, 0.5)$. Using this procedure, we detected a peak to the left of $m_0 = 0$, as predicted by the model (**Fig. 4(d)**). Allowing for the finite binding probability of smFISH probes [16,53], we were able to successfully reproduce the shape of the folded probability distribution (**Fig. 4(d)**, see **Supplemental Material** [32]). Thus, the experimental data supports the theoretical prediction of discontinuity in the distribution of nascent RNA. The periodic discontinuities can be used to identify the signal intensity corresponding to a single RNA, thus improving the precision of RNA counting using smFISH [3,5,14,16].

Conclusion. We presented a theoretical framework for connecting the stochastic kinetics of transcription with the resulting probability distribution of nascent RNA at the gene. By changing the form of the contribution function $g(\tau)$, the model can be used to describe different experimental observables. The model allowed us to interpret experimental data, extract the kinetic parameters of gene activity, and identify how the kinetics vary under the regulatory influence of a transcription factor. The model also predicted a hitherto unobserved feature of discontinuities and periodic peaks in nascent RNA distribution, which we were able to validate experimentally. To further improve the estimation of transcription parameters, the model for nascent RNA can be combined with one for the total cellular RNA [15] and compared to experimental measurements of both species simultaneously [3,6,8,15] (**Supplemental Material** [32]). Beyond the steady-state distribution discussed here, solving for the time-dependent behavior of the model (**Supplemental Material** [32]) can allow a direct comparison with live-cell measurements of nascent RNA [19,21,22].

Acknowledgements:
We are grateful to the following people for generous advice: H. Garcia, D. Larson, H. Levine, A. Sanchez and N. Wingreen. Work in the Golding lab is supported by grants from NIH (R01 GM082837), NSF (PHY 1147498, PHY 1430124 and PHY 1427654), The Welch Foundation (Q-1759) and The John S. Dunn Foundation (Collaborative Research Award). H.X. is supported by the Burroughs Wellcome Fund Career Award at the Scientific Interface. A.M.S. is supported by a grant from the NIH (R01 GM115111). We gratefully acknowledge the computing resources provided by the CIBR Center of Baylor College of Medicine.

References:

[1] A. Sanchez and I. Golding, Science **342**, 1188 (2013).

[2] A. Coulon, C. C. Chow, R. H. Singer, and D. R. Larson, Nat. Rev. Genet. (2013).

[3] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, Science **280**, 585 (1998).

[4] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, Nat. Methods **5**, 877 (2008).

[5] S. O. Skinner, L. A. Sepulveda, H. Xu, and I. Golding, Nat. Protoc. **8**, 1100 (2013).

[6] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, PLoS Biol. **4**, e309 (2006).

[7] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).

[8] D. Zenklusen, D. R. Larson, and R. H. Singer, Nat. Struct. Mol. Biol. **15**, 1263 (2008).

[9] V. Shahrezaei and P. S. Swain, Proc. Natl. Acad. Sci. U. S. A. **105**, 17256 (2008).

[10] L. H. So, A. Ghosh, C. Zong, L. A. Sepulveda, R. Segev, and I. Golding, Nat. Genet. **43**, 554 (2011).

[11] D. L. Jones, R. C. Brewster, and R. Phillips, Science **346**, 1533 (2014).

[12] L. A. Sepulveda, H. Xu, J. Zhang, M. Wang, and I. Golding, Science **351**, 1218 (2016).

[13] D. Huh and J. Paulsson, Proc. Natl. Acad. Sci. U. S. A. **108**, 15004 (2011).

[14] S. C. Little, M. Tikhonov, and T. Gregor, Cell **154**, 789 (2013).

[15] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, eLife **5** (2016).

[16] H. Xu, L. A. Sepulveda, L. Figard, A. M. Sokac, and I. Golding, Nat. Methods **12**, 739 (2015).

[17] M. J. Levesque, P. Ginart, Y. Wei, and A. Raj, Nat. Methods (2013).

[18] C. H. Hansen and A. van Oudenaarden, Nat. Methods **10**, 869 (2013).

[19] H. G. Garcia, M. Tikhonov, A. Lin, and T. Gregor, Curr. Biol. **23**, 2140 (2013).

[20] A. Coulon, M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow, and D. R. Larson, eLife **3** (2014).

[21] T. Lucas, T. Ferraro, B. Roelens, J. De Las Heras Chanes, A. M. Walczak, M. Coppey, and N. Dostatni, Curr. Biol. **23**, 2135 (2013).

[22] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer, Science **332**, 475 (2011).

[23] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq, Cell. Rep. (2014).

[24] K. Bahar Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg, and S. Itzkovitz, Mol. Cell **58**, 147 (2015).

[25] S. Choubey, J. Kondev, and A. Sanchez, PLoS Comp. Biol. **11**, e1004345 (2015).

[26] S. Klumpp and T. Hwa, Proc. Natl. Acad. Sci. U. S. A. **105**, 18159 (2008).

[27] M. Voliotis, N. Cohen, C. Molina-Paris, and T. B. Liverpool, Biophys. J. **94**, 334 (2008).

[28] J. Peccoud and B. Ycart, Theor. Popul. Biol. **48**, 222 (1995).

[29] B. Munsky, G. Neuert, and A. van Oudenaarden, Science **336**, 183 (2012).

[30] L. Rosenfeld, E. Kepten, S. Yunger, Y. Shav-Tal, and Y. Garini, Phys. Rev. E **92**, 032715 (2015).

[31] D. L. Bentley, Nat. Rev. Genet. **15**, 163 (2014).

[32] See Supplemental Material at xxxx, which includes Refs. [33-53].

[33] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, Science **339**, 584 (2013).
[34] K. Adelman, A. La Porta, T. J. Santangelo, J. T. Lis, J. W. Roberts, and M. D. Wang, Proc. Natl. Acad. Sci. U. S. A. **99**, 13538 (2002).
[35] N. Korzheva, A. Mustaev, M. Kozlov, A. Malhotra, V. Nikiforov, A. Goldfarb, and S. A. Darst, Science **289**, 619 (2000).
[36] C. P. Selby, R. Drapkin, D. Reinberg, and A. Sancar, Nucleic Acids Res. **25**, 787 (1997).
[37] E. A. Galburt, S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante, Nature **446**, 820 (2007).
[38] E. Roldan, A. Lisica, D. Sanchez-Taltavull, and S. W. Grill, Phys. Rev. E **93**, 062411 (2016).
[39] A. Lisica, C. Engel, M. Jahnel, E. Roldan, E. A. Galburt, P. Cramer, and S. W. Grill, Proc. Natl. Acad. Sci. U. S. A. **113**, 2946 (2016).
[40] S. Chong, C. Chen, H. Ge, and X. S. Xie, Cell **158**, 314 (2014).
[41] C. A. Brackley, J. Johnson, A. Bentivoglio, S. Corless, N. Gilbert, G. Gonnella, and D. Marenduzzo, Phys. Rev. Lett. **117**, 018101 (2016).
[42] R. M. Martin, J. Rino, C. Carvalho, T. Kirchhausen, and M. Carmo-Fonseca, Cell. Rep. **4**, 1144 (2013).
[43] N. P. Hoyle and D. Ish-Horowicz, Proc. Natl. Acad. Sci. U. S. A. **110**, E4316 (2013).
[44] J. Singh and R. A. Padgett, Nat. Struct. Mol. Biol. **16**, 1128 (2009).
[45] L. Bai, A. Shundrovsky, and M. D. Wang, J. Mol. Biol. **344**, 335 (2004).
[46] N. G. v. Kampen, *Stochastic processes in physics and chemistry* (Elsevier, Amsterdam ; Boston, 2007).
[47] B. Munsky and M. Khammash, J. Chem. Phys. **124**, 044104 (2006).
[48] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[49] F. W. J. Olver and National Institute of Standards and Technology (U.S.), *NIST handbook of mathematical functions* (Cambridge University Press : NIST, Cambridge ; New York, 2010).
[50] P. Dennery and A. Krzywicki, *Mathematics for physicists* (Dover Publications, Mineola, N.Y., 1996).
[51] T. R. Knapp, J. Mod. App. Stat. Meth. **6**, 3 (2007).
[52] L. R. Shenton and K. O. Bowman, J. Amer. Statist. Assoc. **72**, 206 (1977).
[53] E. Lubeck and L. Cai, Nat. Methods **9**, 743 (2012).
[54] C. J. Joachain, *Quantum collision theory* (North-Holland Pub. Co. , Amsterdam, 1975).
[55] S. Yunger, L. Rosenfeld, Y. Garini, and Y. Shav-Tal, Nat. Methods **7**, 631 (2010).
[56] G. Struhl, K. Struhl, and P. M. Macdonald, Cell **57**, 1259 (1989).
[57] R. N. Bracewell, *The Fourier transform and its applications* (McGraw-Hill, New York, 1986).
[58] D. R. Larson, C. Fritzsch, L. Sun, X. Meng, D. S. Lawrence, and R. H. Singer, eLife **2**, e00750 (2013).

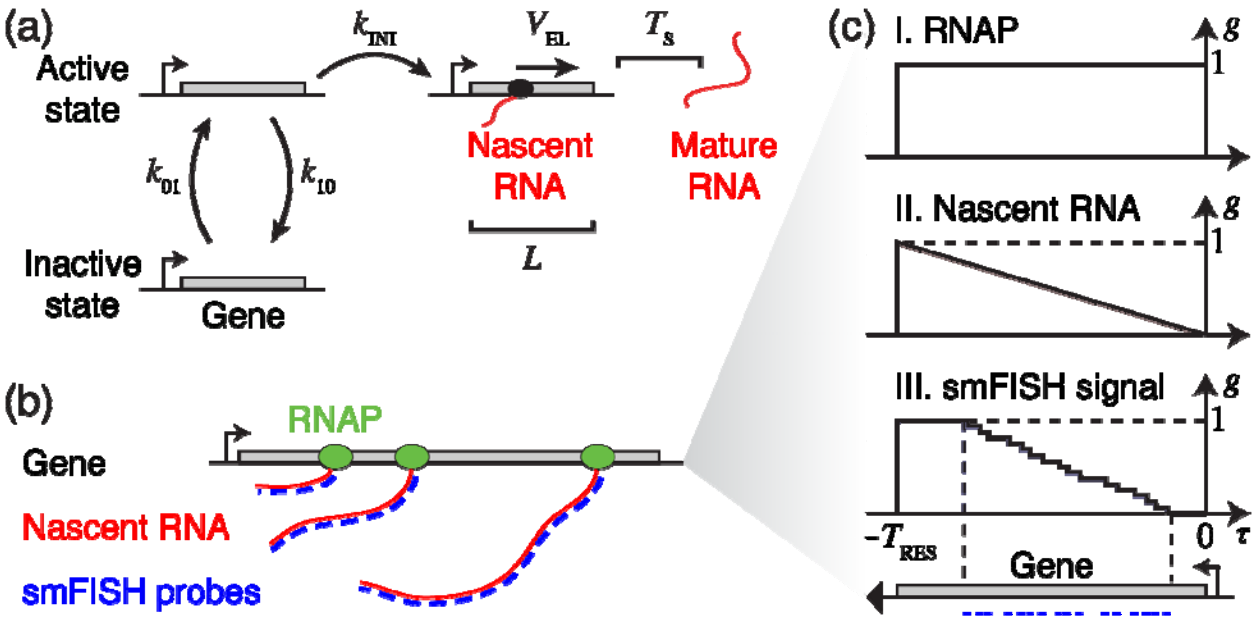

FIG. 1. A stochastic model of nascent RNA kinetics.

(a) Model schematic. (b) Different experimental observables that can be described by the model: The number of RNA polymerases (RNAPs) on the gene (green), the amount of nascent RNA (red), and the signal from single-molecule fluorescence *in situ* hybridization (smFISH) probes (blue). (c) The contribution function corresponding to the three observables in panel **b**. In all cases, $T_S = 0$.

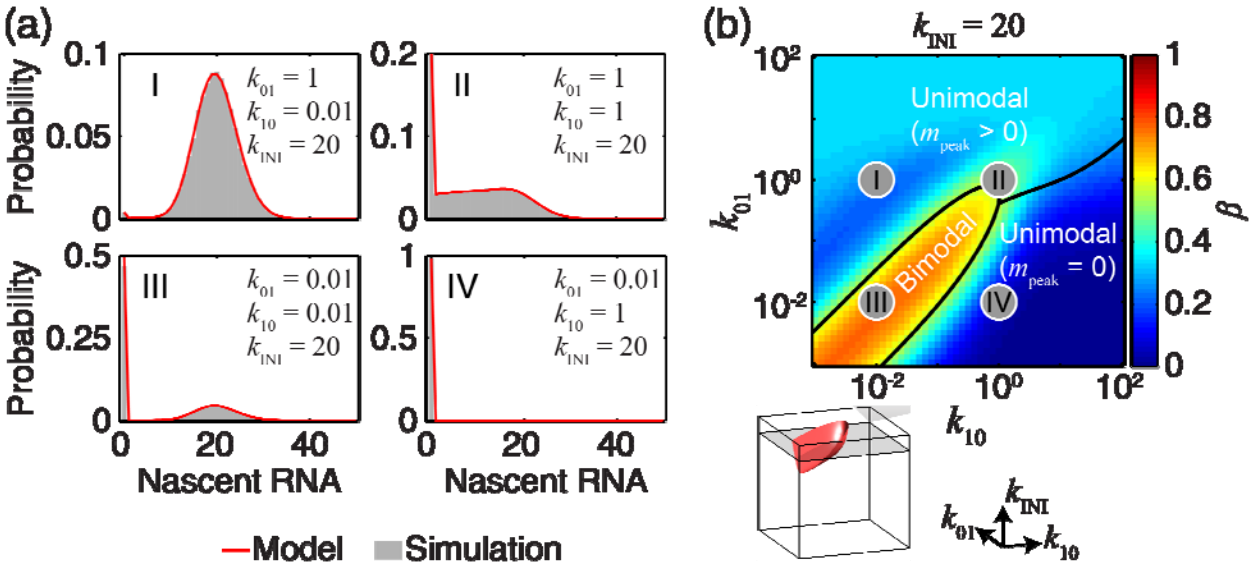

FIG. 2. The probability distribution for the number of RNAPs at the gene.

(a) The exact solution for $P(m)$ (binned to integer values, red) for a few parameter values. Also shown are the results of stochastic simulations (gray). (b) The bimodality coefficient β as a function of k_{01}, k_{10} and k_{INI} was calculated and thresholded ($\beta_{th} = 5/9$, bottom, red surface) to classify $P(m)$ as either bimodal or unimodal. The unimodal distributions were further classified based on the peak position. Parameter values corresponding to panel **a** are marked as gray circles.

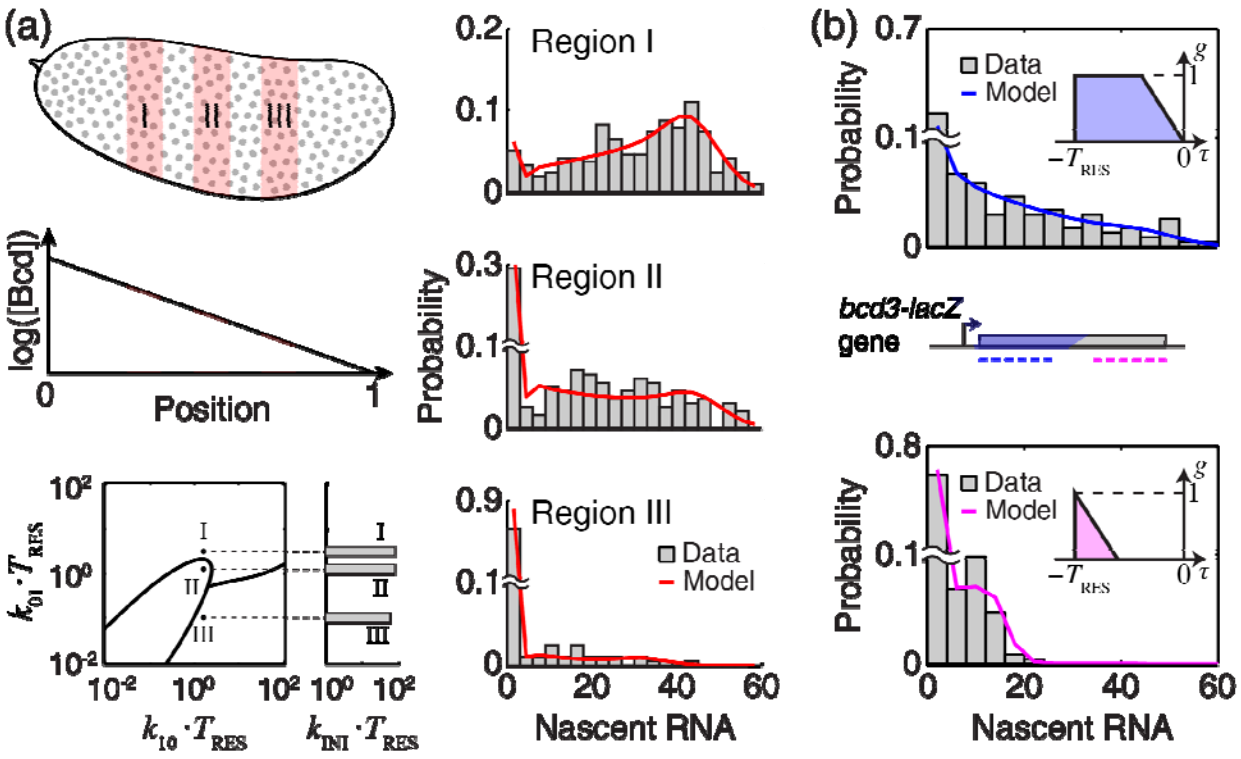

FIG. 3. Estimating transcription kinetics from experimental data.

(a) Regulation of the *hb* gene by Bcd. Top left, Bcd forms a concentration gradient along the anterior-posterior axis of the *Drosophila* embryo. Grey circles indicate individual cell nuclei. Three representative regions of the embryo are highlighted in pink, corresponding to high (I), medium (II) and low (III) Bcd concentrations. Right, the measured distribution of nascent *hb* RNA at each region (smFISH data from a single embryo, >200 data points per histogram, bin width = 3), and the corresponding theoretical fit (red). Bottom left, the estimated transcription parameters (dots), superimposed on the modality phase plane of $P(m)$ calculated as in **Fig. 2(b)**. (b) The effect of smFISH probe positions. Two different sets of probes were designed against the *bcd3-lacZ* reporter gene, targeting the first half (blue) and second half (magenta) of the gene. The two sets yielded different distributions of nascent RNA (top and bottom, >250 data points from a single embryo, at 0.2-0.3 embryo length, bin width = 4). Using the contribution functions calculated from the probe positions on the gene (insets) yielded a good fit between the model and experimental data.

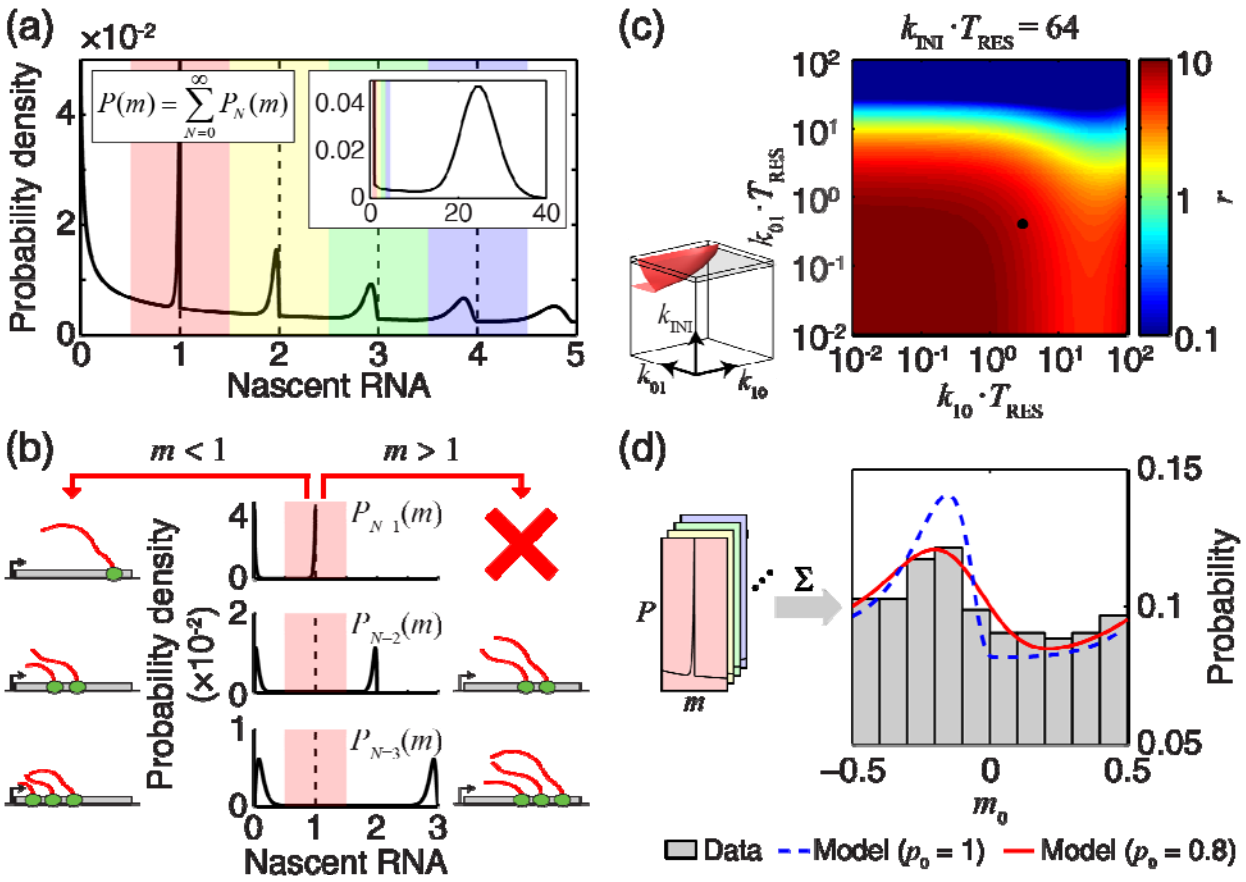

FIG. 4. Discontinuities in nascent RNA distribution at integer *m* values.

(a) The calculated distribution of nascent RNA at small values of m, for $k_{01} = k_{10} = 0.1$, $k_{\mathrm{INI}} = 50$. A larger range of m is shown in the inset. The range of m was divided into windows covering −0.5 to 0.5 around each integer (colored shading). (b) The origin of discontinuity at $m = 1$. The total probability of observing m is a marginalization over different numbers of RNAPs on the gene (plotted for $N = 1, 2, 3$). (c) The discontinuity factor r as a function of k_{01}, k_{10} and k_{INI} was calculated and thresholded ($r_{\mathrm{th}} = 0.1$, left, red surface). Black dot indicates the experimental data analyzed in panel **d**. (d) The experimental signature of $P(m)$ discontinuity. Nascent RNA from *bcd3-lacZ* was measured using smFISH (at 0.1-0.3 embryo length, 23 embryos). The distribution of m_0, the deviation of m from the nearest integer, was calculated (gray, $3.5 \le m < 6.5$, ~500 data points, bin width = 0.1) and compared to model predictions with (red) and without (dashed blue) incorporating the effect of finite probe binding probability p_0.