

Joint Distribution of Nuclear and Cytoplasmic mRNA Levels in Stochastic Models of Gene Expression: Analytical Results and Parameter Inference

Yiling Wang,¹ Juraj Szavits-Nossan², Zhixing Cao^{1,3,*} and Ramon Grima^{2,†}

¹State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China

²School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

³Department of Chemical Engineering, Queen's University, Kingston K7L 3N6, Canada



(Received 30 April 2024; accepted 7 July 2025; published 4 August 2025)

Common stochastic models of gene expression predict the gene-specific distribution of total mRNA per cell but lack subcellular resolution. Here, for a broad class of transcription initiation models, we derive an exact steady-state solution for the joint distribution of nuclear and cytoplasmic mRNA, and demonstrate that fitting this solution to spatially resolved mRNA data enhances parameter identifiability. By accounting for extrinsic noise, we use the model to precisely quantify bursty expression across thousands of human genes and link it to their biological functions.

DOI: 10.1103/q5sd-tpms

Gene expression is inherently stochastic [1–4], leading to substantial variability in mRNA and protein levels across cells. Stochastic models help elucidate the biophysical origins of this variability [5–7] and enable gene-specific parameter inference from experimental data [8–11]. Most models are based on the telegraph framework, where a gene switches between inactive and active states, producing mRNA in the latter [12,13]. Traditionally, such models assume total mRNA counts, but recent advances allow simultaneous measurement of nuclear and cytoplasmic mRNA [14–17]. This necessitates refined models that account for subcellular compartmentalization and its potential role in noise regulation [18].

Multispecies models that track nascent, nuclear, and cytoplasmic mRNA have been developed and solved either numerically [15,19] or analytically for special cases [20–27]. Most analytical results provide expressions for moments or marginal distributions; only Ref. [20] derives the full joint distribution of nuclear and cytoplasmic mRNA. However, this model assumes the gene is active for an infinitesimally short time, producing a random burst of transcripts—an assumption that may not hold for eukaryotic genes, which often remain active for minutes [28]. Additionally, its use of exponentially distributed off times contrasts with experimental evidence for some mammalian genes [30].

In this Letter, we present a general stochastic model of gene expression that avoids the limiting assumptions of earlier work and derive the exact joint distribution of mRNA levels at two distinct stages of the mRNA life cycle—interpretable as nascent and mature nuclear, unspliced and spliced, or nuclear and cytoplasmic mRNA. We validate the solution via Monte Carlo simulations and use it to develop an accurate, computationally efficient method for inferring gene-specific kinetic parameters from commonly available data.

General theory—We consider a general stochastic model of nuclear and cytoplasmic mRNA fluctuations in which nuclear mRNA production follows a renewal process [31]; that is, interarrival times between successive production events are independent, identically distributed random variables with density $f(t)$ and finite mean $\alpha = \int_0^\infty tf(t)dt$. This assumption is supported by experimental evidence [32]. After a fixed delay τ , nuclear mRNA molecules (N) are exported to the cytoplasm and become cytoplasmic mRNA (M), which then degrade at rate d . For analytical tractability, we assume τ is deterministic. Following previous work [33], we omit nuclear mRNA degradation, as coding transcripts are typically protected from decay in the nucleus [34]. The system is summarized by the reaction scheme



where $\xrightarrow{f(t)}$ denotes a stochastic delay with distribution $f(t)$, and $\xrightarrow{\tau}$ denotes a fixed delay τ . A representative example of this model is illustrated in Fig. 1.

To compute the steady-state joint distribution of nuclear and cytoplasmic mRNA, we apply queueing theory, as recently reviewed in Ref. [35]. Queueing theory models systems where “customers” arrive, receive service, and

*Contact author: z.cao@queensu.ca

†Contact author: ramon.grima@ed.ac.uk

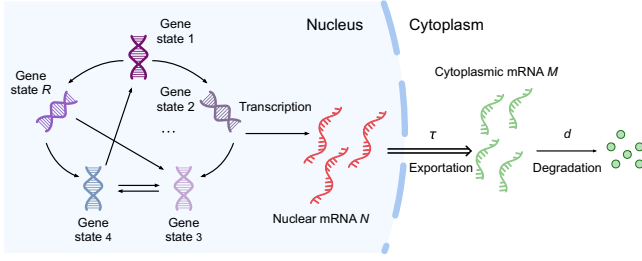


FIG. 1. Illustration of a gene expression model involving initiation (gene switching among R inactive states and an active state), transcription from the active state, and nuclear export after a delay τ , followed by cytoplasmic degradation. This represents a broad class of systems where nuclear mRNA production follows a renewal process, for which we derive the steady-state joint distribution of nuclear and cytoplasmic mRNA counts.

depart [36], and is commonly described using Kendall's notation $A/S/n$, where A is the arrival process, S the service process, and n the number of servers [37]. The

reaction scheme in Eq. (1) is equivalent to two serial queues. The first models nuclear mRNA production and export, corresponding to a $G/D/\infty$ queue: G denotes general (renewal) interarrival times, and D represents deterministic export delay. Its output feeds into a second queue, modeling cytoplasmic mRNA degradation as a $G/M/\infty$ queue with memoryless (exponential) service. Because export is deterministic, both queues share the same arrival distribution. We assume mRNAs are available for processing immediately upon creation, justifying the use of infinitely many servers in both queues.

Let $N(t)$ and $M(t)$ denote the numbers of nuclear and cytoplasmic mRNA at time t , respectively. We seek the joint distribution $P(n, m) = \Pr[N(t) = n, M(t) = m]$ at steady state ($t \rightarrow \infty$). This is obtained via the joint probability generating function (PGF), $G(z_1, z_2) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} P(n, m) z_1^n z_2^m$. For the general model in Eq. (1), using queueing theory we show that the PGF admits an analytical expression,

$$G(z_1, z_2) = \sum_{m=1}^{\infty} \frac{C_{m-1}}{\alpha d m} (z_2 - 1)^m + \mathcal{L}_{\tau}^{-1} \left(\frac{1}{s} + \frac{(z_1 - 1)[1 - \phi(s)]}{\alpha s^2 [1 - z_1 \phi(s)]} + \frac{(z_1 - 1)}{\alpha s [1 - z_1 \phi(s)]} \sum_{r=1}^{\infty} C_r \frac{\mathcal{L}_x \{ \phi(r d, x) \}(s)}{\phi(r d)} (z_2 - 1)^r \right), \quad (2)$$

where $\phi(s) = \int_0^{\infty} e^{-st} f(t) dt$, and $C_0 = 1$, $C_r = \prod_{i=1}^r \phi(i d) / [1 - \phi(i d)]$ for $r \geq 1$. Let $\mathcal{L}_{\tau}[g](s)$ be the Laplace transform of $g(\tau)$; $\mathcal{L}_{\tau}^{-1}$ denotes its inverse operator. The function $\phi(r d, x) = \int_0^{\infty} e^{-rdy} f(x + y) dy$, and $\mathcal{L}_x \{ \cdot \}(s)$ is its Laplace transform over x . A full derivation is given in Supplemental Material (SM) I [38].

Models—The joint distribution $P(n, m)$ can be computed once the interarrival time distribution $f(t)$ between successive nuclear mRNA productions is specified. We consider three common models with a two-compartment extension [Figs. 2(a)–2(c), insets], where transcription is governed by a Markov process [52] in which the gene switches among 1, 2, or 3 discrete states—referred to as the 1-, 2-, and 3-state models. In the 1-state model, expression is constitutive, with the gene always in a transcriptionally active state. The 2-state model generalizes the telegraph model [12], allowing stochastic switching between inactive and permissive states. The 3-state model extends the refractory model [30], introducing a two-step activation to capture “off” intervals that peak at nonzero durations in some mammalian genes.

Using the general result from queueing theory [Eq. (2)], we derive the analytical expression for the PGF of the 1-state model,

$$G(z_1, z_2) = \exp[\rho \tau (z_1 - 1) + \rho / d (z_2 - 1)], \quad (3)$$

showing that both nuclear and cytoplasmic mRNA follow Poisson distributions. The PGF for the 2-state model is

$$G = \frac{\nu_+ e^{-\nu_- \tau} - \nu_- e^{-\nu_+ \tau}}{d \theta} M \left(\frac{\sigma_{\text{on}}}{d}, \frac{\sigma_{\text{on}} + \sigma_{\text{off}}}{d}, \frac{\rho u_2}{d} \right) + \rho u_1 \frac{e^{-\nu_- \tau} - e^{-\nu_+ \tau}}{d \theta} G_{M1}. \quad (4)$$

Here $G_{M1} = \alpha M(1 + \sigma_{\text{on}}/d, 1 + \sigma_{\text{off}}/d + \sigma_{\text{on}}/d, x_2)$ with $\alpha = \sigma_{\text{on}}/(\sigma_{\text{off}} + \sigma_{\text{on}})$, $x_2 = \rho u_2/d$ and $M(a, b, x)$ being the Kummer function. Additionally, $\nu_{\pm} = d(r \pm \theta - 1)/2$, $r = 1 - x_1 + (\sigma_{\text{off}} + \sigma_{\text{on}})/d$, $\theta = [(x_1 - \sigma_{\text{off}}/d - \sigma_{\text{on}}/d)^2 + 4\sigma_{\text{on}}x_1/d]^{(1/2)}$, $x_1 = \rho u_1/d$ and $u_{1,2} = z_{1,2} - 1$. Derivations of Eqs. (3) and (4) are provided in SM II A, B, [38] with an alternative derivation of Eq. (4) based on the delay master equation [39,40] presented in SM III and SM Fig. S1 [38].

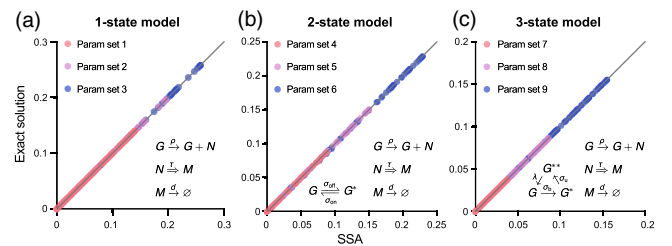


FIG. 2. (a)–(c) P-P plots comparing the joint distributions from the exact solutions of the 1-, 2-, and 3-state models (insets), derived via queueing theory, with those obtained from the delay SSA. The exact solutions are given by Eqs. (3)–(5) and (S33). The nine corresponding kinetic parameter sets are listed in SM Table S1 [38]. Gray lines: reference lines $y = x$.

The marginal generating function for the 2-state model, $G(z_1, z_2 = 1)$, describing nuclear mRNA dynamics, has been previously derived [41,55]; in those works, the variable N refers to bound RNA polymerase rather than mature nuclear mRNA. The PGF for the 3-state model, derived in SM II C, involves a sum of four ${}_2F_2$ hypergeometric functions. The joint distribution is given by

$$P(n, m) = \frac{1}{n!m!} \frac{\partial^n}{\partial u_1^n} \frac{\partial^m}{\partial u_2^m} G \Big|_{u_1=-1, u_2=-1}, \quad (5)$$

which can be efficiently evaluated using symbolic computation. Importantly, unlike existing approaches [20], our solution for the 2-state model [Eqs. (4) and (5)] does not assume that the gene state G is short-lived or that transcription occurs in instantaneous bursts.

In Figs. 2(a)–2(c), we use probability-probability (P-P) plots to assess the accuracy of the analytical solution of the three models given by Eqs. (3)–(5) and (S33) [38], by comparing it to the joint distribution $\hat{P}(n, m)$ obtained from 10^6 independent runs of the delay stochastic simulation algorithm (delay SSA, using the rejection method in Ref. [56]). The plot displays all pairs of conditional cumulative probabilities $\sum_{i=1}^n F(i, m)$ for $F \in \{P, \hat{P}\}$ and all $n, m \in [0, N_{\max}]$, where $N_{\max}$ is the maximum mRNA count observed in the simulations. The close alignment of all points along the line $y = x$ confirms the accuracy of the analytical solution.

In SM IV, V and SM Figs. S2, S3 [38], we compare the non-Markovian models to their Markovian counterparts, in which export times are exponentially distributed, i.e., $N \xrightarrow{1/\tau} M$. We show that when nuclear export is fast compared to mRNA degradation, the non-Markovian and Markovian models become indistinguishable.

To align the theory with eukaryotic cell data, we incorporate extrinsic noise arising from transcription volume coupling [57,58] by setting the transcription rate ρ proportional to the normalized cell volume β (cell volume divided by its mean). This leads to the modified reaction $G \xrightarrow{\rho\beta} G + N$ in the 1-, 2-, and 3-state models, hereafter referred to as *extrinsic-noise-corrected (ENC) models* [see End Matter (EM) Fig. 5(b), left] [59]. We further assume two independent gene copies per cell [62], simplifying more detailed models of growing and dividing cells [63].

PGF-based parameter inference—To connect theory with experimental data, we develop a PGF-based parameter inference method. In principle, the exact joint distribution of molecule numbers [Eqs. (3)–(5) and (S33) [38]] allows transcriptional parameters to be estimated via maximum likelihood using single-cell measurements of nuclear and cytoplasmic mRNA. However, this approach is computationally intensive, requiring symbolic evaluation of high-order derivatives across the parameter space. To address this, we propose an alternative method based solely on the generating function, which involves (i) constructing an

empirical generating function from the data and (ii) identifying the parameter set that minimizes its distance to the analytical PGF. Given nuclear-cytoplasmic mRNA counts (n_i, m_i) from a population of n_c cells, the empirical generating function is

$$\bar{G}(z_1, z_2) = \sum_{i=1}^{n_c} \frac{1}{n_c} z_1^{n_i} z_2^{m_i}. \quad (6)$$

The transcriptional parameters are inferred by minimizing the loss function

$$J(\phi) = \int_0^1 \int_0^1 |\mathcal{G}(z_1, z_2, \phi) - \bar{G}(z_1, z_2)|^2 dz_1 dz_2, \quad (7)$$

where $\mathcal{G}(z_1, z_2, \phi)$ is the analytical generating function of a user-defined gene expression model, and $\hat{\phi} = \arg \min_{\phi} J(\phi)$ is the inferred parameter vector.

To assess the performance of the proposed inference method, we apply it to synthetic joint count datasets generated via delay SSA simulations of the 2-state ENC model with two independent gene copies [64]. The model's PGF solution is given by

$$\mathcal{G}(z_1, z_2, \phi) = \int_0^\infty G^2(z_1, z_2, \phi, \beta) \hat{p}(\beta) d\beta, \quad (8)$$

where $\phi = \{\bar{\rho}, \sigma_{\text{on}}, \sigma_{\text{off}}, \tau\}$. The distribution $\hat{p}(\beta)$ is estimated from simulated cell volume β_i using kernel density estimation (KDE) [67]. The inference procedure, accounting for extrinsic noise, minimizes the loss in Eq. (7) with $\mathcal{G}(z_1, z_2, \phi)$ from Eq. (8) and $G(z_1, z_2, \phi, \beta)$ from Eq. (4), where the exponent 2 on G reflects the two independent gene copies. See EM and EM Fig. 5(a) for details on the inference implementation.

In Fig. 3(a), the blue bars show the relative errors of the PGF-based inference method across the 15 simulated datasets. These errors are small and comparable to those obtained using conventional maximum likelihood estimation via the joint distribution computed numerically with the finite state projection (FSP) method (purple bars; SM VI). However, the PGF method is significantly faster at achieving speedups of several thousandfold over FSP [Fig. 3(a), inset], making it particularly advantageous for fitting stochastic models to single-cell transcriptomic data across many genes, especially at subcellular resolution [68]. Additionally, the PGF method generally outperforms an alternative approach based on matching the first and second moments of mRNA fluctuations [Fig. 3(b), SM VII [38]].

Next, we compare two inference approaches: fitting the 2-state ENC model to nuclear-cytoplasmic counts and fitting the telegraph model to total mRNA counts [69]. We apply the same simulation protocol to three kinetic regimes: $\sigma_{\text{off}} = \sigma_{\text{on}} = \{0.05, 5, 500\}$ with all other parameters given in [70].

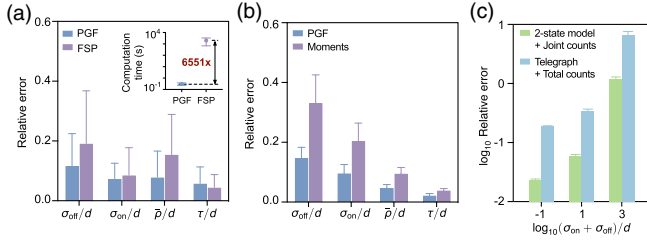


FIG. 3. (a) Comparison of the PGF-based and maximum likelihood-based inference methods using FSP on simulated data shows similar accuracy but higher computational efficiency (0.30s for PGF vs 1991.52s for FSP), with the relative error averaged over 15 datasets. (b) The PGF-based method outperforms the method of moments in accuracy. Parameters for (a), (b) are summarized in SM Tables S2 and S3 [38]. (c) Fitting nuclear-cytoplasmic mRNA counts with the two-mRNA species, 2-state ENC model [inset of Fig. 2(b)] improves parameter identifiability over the single-mRNA species model (standard telegraph model [12]) for total counts. Bars: the standard error of the mean (SEM).

For each regime, we generate 10^2 replicate datasets, recording both joint and total mRNA counts. As shown in Fig. 3(c), the average relative error increases with the switching rate $(\sigma_{on} + \sigma_{off})/d$ in both models. This aligns with [71,72], which shows that in the fast switching limit, the 2-state model converges to the 1-state model, making reliable parameter inference difficult—an issue that also affects the telegraph model. Notably, fitting joint counts yields significantly lower errors (p value $< 10^{-3}$ in all cases), demonstrating improved identifiability.

Model selection and application to experimental data— We apply the developed theory and inference method to identify gene activity—constitutive (1-state model) vs bursty (2-state model) from 2 types of fluorescent in site hybridization (FISH) data [see [73], SM Fig. S4, Table S4 [38]], corrected for extrinsic noise. To this end, we develop a tenfold cross-validation model selection pipeline (EM Fig. 5). Models are ranked by increasing complexity (i.e., number of parameters), and the dataset is split into 10 equal parts. For each fold, parameters are inferred from nine subsets using the PGF-based inference method, and the remaining subset is used to compute a validation score $J(\hat{\phi}_j)$ via Eq. (7), yielding a score vector $\mathcal{J}_{model_i} = [J(\hat{\phi}_1), \dots, J(\hat{\phi}_{10})]^T$ for each model i . The model with the lowest mean score is indexed by $\mathcal{I}$, with score vector $\mathcal{J}_{best}$, mean $\bar{\mathcal{J}}_{best}$, and standard deviation σ_{best} . To balance accuracy and complexity, we compare this best model with all simpler models ($i < \mathcal{I}$) using a modified one-standard-error rule [74],

$$\bar{\mathcal{J}}_{model_i} \leq \bar{\mathcal{J}}_{best} + \sigma_{best} \sqrt{1 - \text{cor}(\mathcal{J}_{best}, \mathcal{J}_{model_i})}, \quad (9)$$

selecting the model with the smallest i that satisfies Eq. (9) as the final best model [75]. See EM and Fig. 5(b) for implementation details. The proposed model selection

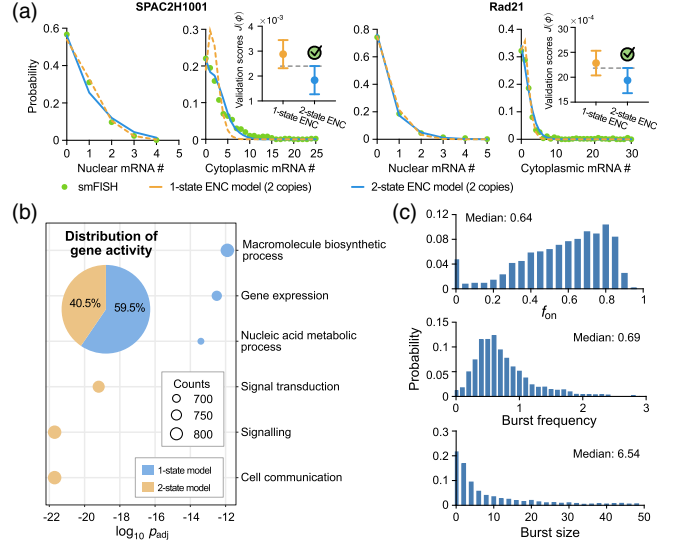


FIG. 4. (a) Application of the PGF-based model selection pipeline to smFISH data for the fission yeast genes SPAC2H1001 and Rad21 reveals bursty expression in both. Marginal distributions of nuclear and cytoplasmic mRNA, based on parameters from the 2-state ENC model, show better fits than the 1-state model, supported by validation scores and the model selection threshold (gray dashed lines, Eq. (9) inset). Inferred parameters (tenfold) are summarized in SM Table S5 [38]. Bars: standard deviation. (b) The pipeline applied to MERFISH data identifies the fraction of genes best fit by the 1- and 2-state models, with gene ontology analysis linking each group to relevant biological processes. p_{adj} : Benjamini-adjusted p value indicating the expected false discovery rate among the identified terms. (c) PGF-based inference applied to 2,184 mammalian genes best fit by the 2-state ENC model and passing a moment-based fit quality test (SM Fig. S7). Genome-wide distributions of inferred parameters include active-state fraction time f_{on} , burst frequency (normalized by degradation rate), and burst size.

pipeline performs comparably to the widely used Bayesian information criterion (BIC) (SM Fig. S5 [38]) and demonstrates improved model identification accuracy when using joint counts over total counts (SM Fig. S6 [38]).

We apply the pipeline to single-cell measurements of cell size and nuclear and cytoplasmic mRNA levels for the fission yeast genes SPAC2H1001 and Rad21 [17] (see EM for details). mRNA levels are quantified using single-molecule FISH (smFISH) in mononucleated cells, each containing two gene copies. For both genes, model selection using Eq. (9) [Fig. 4(a), insets] supports bursty expression, with marginal distributions derived from the inferred 2-state ENC model providing a better fit to the smFISH data than the 1-state model (Fig. 4(a)).

We next examine genome-wide gene activity corrected for extrinsic noise, using a multiplexed error-robust FISH (MERFISH) dataset of single-cell nuclear and cytoplasmic mRNA counts for 6724 mammalian genes across 1368 human osteosarcoma cells [14] (see EM for data processing details). Since cell volume measurements are unavailable,

we approximate volume using the total transcript count per cell, normalized by the population mean, to compute β_i —a reasonable proxy supported by [76]. While extrinsic noise from volume variability is accounted for, technical noise is not, as it is minimal in this dataset due to high transcript detection efficiency (80%) [14]. Using the PGF-based model selection pipeline, we identified 4002 genes best fit by the 1-state ENC model and 2722 genes by the 2-state model. The proportions are shown in the pie chart in Fig. 4(b), inset. The 60% of genes favoring the 1-state model aligns with previous studies [77–79]. Gene ontology analysis with DAVID [80] further revealed functional differences between the two clusters (see EM for details). As shown in Fig. 4(b), genes best fit by the 1-state model are enriched for macromolecule production, while those better fit by the 2-state model are associated with signaling functions. Figure 4(c) presents the inferred distributions of burst size, burst frequency (normalized by degradation rate), and active-state fraction $f_{\text{on}} = \sigma_{\text{on}}/(\sigma_{\text{on}} + \sigma_{\text{off}})$ for genes best fit by the 2-state ENC model and passing a fit-quality threshold (average relative error in first- and second-order moments < 35%; see SM Fig. S7 [38]). The median values of f_{on} , burst size, and normalized burst frequency are 0.64, 6.83, and 0.69, respectively.

In SM Table S6 [38], we repeat the model selection and parameter inference procedure under alternative assumptions and data types. Specifically, we vary the assumed gene copy number in the 2-state model (4 instead of 2), use total mRNA counts per cell instead of joint nuclear-cytoplasmic data, and consider only intrinsic noise (no correction for extrinsic noise). We find that neglecting extrinsic noise leads to a significant overestimation of genes best fit by the 2-state model (90% vs 40% in the corrected case). Additionally, using total counts instead of joint data results in systematic underestimation of f_{on} and overestimation of burst frequency. These results emphasize the importance of subcellular-resolution transcriptomic data for accurate parameter inference.

In summary, we applied queueing theory to derive the exact steady-state joint distribution of nuclear and cytoplasmic mRNA in gene expression models with general initiation mechanisms and an arbitrary number of gene states. We also developed a PGF-based parameter inference and model selection method, fitting the exact PGF to the empirical PGF derived from transcriptomic data. Applied to a large experimental dataset, this approach uncovered a clear link between gene expression type and biological function.

Acknowledgments—Z. C. acknowledges support from NSFC Grant (62073137), Shanghai Action Plan for Technological Innovation Grants No. 22ZR1415300, No. 22511104000, and No. 23S41900500, the Natural Science and Engineering Research Council of Canada’s (NSERC’s) Discovery Grant (RGPIN-2024-06015) and

Shanghai Center of Biomedicine Development. R. G. and J. S.-N. acknowledge support from a Leverhulme Trust grant (RPG-2020-327).

Data availability—All codes and data have been deposited at [81].

- [1] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Stochastic gene expression in a single cell, *Science* **297**, 1183 (2002).
- [2] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins, Noise in eukaryotic gene expression, *Nature (London)* **422**, 633 (2003).
- [3] J. Rodriguez, G. Ren, C. R. Day, K. Zhao, C. C. Chow, and D. R. Larson, Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity, *Cell* **176**, 213 (2019).
- [4] A. Sanchez and I. Golding, Genetic determinants and cellular constraints in noisy gene expression, *Science* **342**, 1188 (2013).
- [5] N. Kumar, T. Platini, and R. V. Kulkarni, Exact distributions for stochastic gene expression models with bursting and feedback, *Phys. Rev. Lett.* **113**, 268105 (2014).
- [6] N. Friedman, L. Cai, and X. S. Xie, Linking stochastic dynamics to population distribution: An analytical framework of gene expression, *Phys. Rev. Lett.* **97**, 168302 (2006).
- [7] L. Ham, R. D. Brackston, and M. P. Stumpf, Extrinsic noise and heavy-tailed laws in gene expression, *Phys. Rev. Lett.* **124**, 108101 (2020).
- [8] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koepl, Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings, *Nat. Methods* **11**, 197 (2014).
- [9] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, Genomic encoding of transcriptional burst kinetics, *Nature (London)* **565**, 251 (2019).
- [10] X. Fu, H. P. Patel, S. Coppola, L. Xu, Z. Cao, T. L. Lenstra, and R. Grima, Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions, *eLife* **11**, e82493 (2022).
- [11] Z. Kilic, M. Schweiger, C. Moyer, D. Shepherd, and S. Pressé, Gene expression model inference from snapshot RNA data using Bayesian non-parametrics, *Nat. Comput. Sci.* **3**, 174 (2023).
- [12] J. Peccoud and B. Ycart, Markovian modeling of gene-product synthesis, *Theor. Popul. Biol.* **48**, 222 (1995).
- [13] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, Stochastic mRNA synthesis in mammalian cells, *PLoS Biol.* **4**, e309 (2006).
- [14] C. Xia, J. Fan, G. Emanuel, J. Hao, and X. Zhuang, Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 19490 (2019).
- [15] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, Distribution shapes govern the discovery of predictive models for gene regulation, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 7533 (2018).

- [16] G. Li and G. Neuert, Multiplex RNA single molecule FISH of inducible mRNAs in single yeast cells, *Sci. Data* **6**, 94 (2019).
- [17] D. E. Weidemann, J. Holehouse, A. Singh, R. Grima, and S. Hauf, The minimal intrinsic stochasticity of constitutively expressed eukaryotic genes is sub-Poissonian, *Sci. Adv.* **9**, eadh5138 (2023).
- [18] T. Stoeger, N. Battich, and L. Pelkmans, Passive noise filtering by cellular compartmentalization, *Cell* **164**, 1151 (2016).
- [19] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, Single-cell analysis of transcription kinetics across the cell cycle, *eLife* **5**, e12175 (2016).
- [20] A. Singh and P. Bokes, Consequences of mRNA transport on stochastic variability in protein levels, *Biophys. J.* **103**, 1087 (2012).
- [21] M. Sturrock, S. Li, and V. Shahrezaei, The influence of nuclear compartmentalisation on stochastic dynamics of self-repressing gene expression, *J. Theor. Biol.* **424**, 55 (2017).
- [22] G. Gorin and L. Pachter, Modeling bursty transcription and splicing with the chemical master equation, *Biophys. J.* **121**, 1056 (2022).
- [23] G. Gorin, S. Yoshida, and L. Pachter, Assessing Markovian and delay models for single-nucleus RNA sequencing, *Bull. Math. Biol.* **85**, 114 (2023).
- [24] T. Filatova, N. Popovic, and R. Grima, Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination, *Bull. Math. Biol.* **83**, 1 (2021).
- [25] T. Filatova, N. Popović, and R. Grima, Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: Analytical distributions, *Math. Biosci.* **347**, 108828 (2022).
- [26] J. Szavits-Nossan and R. Grima, Mean-field theory accurately captures the variation of copy number distributions across the mRNA life cycle, *Phys. Rev. E* **105**, 014410 (2022).
- [27] C. Shi, X. Yang, J. Zhang, and T. Zhou, Stochastic modeling of the mRNA life process: A generalized master equation, *Biophys. J.* **122**, 4023 (2023).
- [28] Time spent in the active state can be estimated from data in Table 1 of Ref. [29] since the mean active time is the inverse of σ_b .
- [29] Z. Cao and R. Grima, Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4682 (2020).
- [30] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Mammalian genes are transcribed with widely different bursting kinetics, *Science* **332**, 472 (2011).
- [31] D. Cox, *Renewal Theory* (Methuen, London, 1967).
- [32] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer, Real-time observation of transcription initiation and elongation on an endogenous yeast gene, *Science* **332**, 475 (2011).
- [33] K. B. Halpern, I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz, Nuclear retention of mRNA in mammalian tissues, *Cell Rep.* **13**, 2653 (2015).
- [34] S. Bresson and D. Tollervy, Surveillance-ready transcription: Nuclear RNA decay as a default fate, *Open Biol.* **8**, 170270 (2018).
- [35] J. Szavits-Nossan and R. Grima, Solving stochastic gene-expression models using queueing theory: A tutorial review, *Biophys. J.* **123**, 1034 (2024).
- [36] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, 4th ed. (Wiley-Interscience, Hoboken, New Jersey, 2008).
- [37] D. G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov Chain, *Ann. Math. Stat.* **24**, 338 (1953).
- [38] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/q5sd-tpms>, which includes Refs. [12,29,31,33,39–51], for deriving PGF solutions and developing PGF-based inference method.
- [39] A. Leier and T. T. Marquez-Lago, Delay chemical master equation: Direct and closed-form solutions, *Proc. R. Soc. A* **471**, 20150049 (2015).
- [40] Q. Jiang, X. Fu, S. Yan, R. Li, W. Du, Z. Cao, F. Qian, and R. Grima, Neural network aided approximation and parameter inference of non-Markovian models of gene expression, *Nat. Commun.* **12**, 2618 (2021).
- [41] J. Szavits-Nossan and R. Grima, Steady-state distributions of nascent RNA for general initiation mechanisms, *Phys. Rev. Res.* **5**, 013064 (2023).
- [42] L. Takács, On a coincidence problem concerning telephone traffic, *Acta Math. Acad. Sci. Hung.* **9**, 45 (1958).
- [43] W. R. Inc., Mathematica, Version 14.1 (2024), Champaign, IL.
- [44] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Vol. 1 (Elsevier, New York, 1992).
- [45] W. R. Inc., HypergeometricPFQ (2024), Champaign, IL.
- [46] M. M. Hansen, R. V. Desai, M. L. Simpson, and L. S. Weinberger, Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability, *Cell Syst.* **7**, 384 (2018).
- [47] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriiti, P. Lönnerberg, A. Furlan *et al.*, RNA velocity of single cells, *Nature (London)* **560**, 494 (2018).
- [48] T. Jahnke and W. Huisinga, Solving the chemical master equation for monomolecular reaction systems analytically, *J. Math. Biol.* **54**, 1 (2007).
- [49] B. Munsky and M. Khammash, The finite state projection algorithm for the solution of the chemical master equation, *J. Chem. Phys.* **124**, 044104 (2006).
- [50] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Stochasticity of gene products from transcriptional pulsing, *Phys. Rev. E* **79**, 031911 (2009).
- [51] NIST digital library of mathematical functions, <https://dlmf.nist.gov/>, Release 1.1.10 of 2023-06-15.
- [52] This corresponds to a Markovian arrival process (MAP) in queueing theory [53]. Since not all MAPs are renewal processes [35], we focus on renewal MAPs where inter-arrival times follow a phase-type distribution [54].
- [53] M. F. Neuts, A versatile Markovian point process, *J. Appl. Probab.* **16**, 764 (1979).
- [54] C. Commault and J.-P. Chemla, An invariant of representations of phase-type distributions and some applications, *J. Appl. Probab.* **33**, 368 (1996).
- [55] H. Xu, S. O. Skinner, A. M. Sokac, and I. Golding, Stochastic kinetics of nascent RNA, *Phys. Rev. Lett.* **117**, 128101 (2016).

- [56] X. Cai, Exact stochastic simulation of coupled chemical reactions with delays, *J. Chem. Phys.* **126**, 124108 (2007).
- [57] C. Jia and R. Grima, Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model, *Iscience* **26**, 105746 (2023).
- [58] P. Thomas and V. Shahrezaei, Coordination of gene expression noise with cell size: Analytical results for agent-based models of growing cell populations, *J. R. Soc. Interface* **18**, 20210274 (2021).
- [59] In the 2-state model, this implies that burst size (ρ/σ_{off}) scales with cell volume, consistent with experimental findings in eukaryotic cells [60,61].
- [60] O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj, Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms, *Mol. Cell* **58**, 339 (2015).
- [61] S. Berry and L. Pelkmans, Mechanisms of cellular mRNA transcript homeostasis, *Trends Cell Biol.* **32**, 655 (2022).
- [62] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells, *Science* **343**, 193 (2014).
- [63] Our modeling approach does not explicitly capture fluctuations in mRNA counts due to the cell-cycle phase dependence of kinetic parameters, cell-cycle duration variability, doubling of gene copy number during replication, and cell division, which some other published models account for [29,57]. A major reason for excluding such biological detail is that these models are time dependent, which makes the exact solution of the joint distribution of nuclear and cytoplasmic mRNA a difficult technical problem. Another reason is that these complex models have a considerably large number of parameters which makes it very challenging to reliably infer them using mRNA count data from a few thousand cells of one type (typical sample size).
- [64] Each cell follows the 2-state ENC model with transcription rate proportional to the normalized cell volume $\beta \sim \text{Gamma}(7.5, 0.11)$. A gamma distribution or a similar unimodal distribution (such as a lognormal distribution) is a good approximation to the cell volume distribution when the cell-cycle duration is highly noisy; see Fig. 2A in Ref. [65]. Other papers have used such distributions of parameters to model extrinsic noise on the rates of the telegraph model [7,66]. Delay SSA simulations [56] generated 15 batches of joint count data, each comprising 2000 cells with distinct kinetic parameters summarized in SM Table S2 [38].
- [65] C. Jia, A. Singh, and R. Grima, Cell size distribution of lineage data: Analytic results and parameter inference, *Iscience* **24**, 102220 (2021).
- [66] R. Grima and P.-M. Esmeijaud, Quantifying and correcting bias in transcriptional parameter inference from single-cell data, *Biophys. J.* **123**, 4 (2024).
- [67] Y.-C. Chen, A tutorial on kernel density estimation and recent advances, *Biostat. Epidemiol.* **1**, 161 (2017).
- [68] M. Zhang, X. Pan, W. Jung, A. R. Halpern, S. W. Eichhorn, Z. Lei, L. Cohen, K. A. Smith, B. Tasic, Z. Yao *et al.*, Molecularly defined and spatially resolved cell atlas of the whole mouse brain, *Nature (London)* **624**, 343 (2023).
- [69] The 2-state ENC model was fit using Eqs. (6)–(8) to infer the parameters $\{\bar{\rho}, \sigma_{\text{on}}, \sigma_{\text{off}}, \tau\}$. For the telegraph model [12], transcription was assumed volume-dependent with four reactions: $G \xrightarrow{\bar{\rho}} G + M$, $G \xrightleftharpoons[\sigma_{\text{on}}]{\sigma_{\text{off}}} G^*$, and $M \xrightarrow{d} \emptyset$, where M denotes total mRNA. The empirical PGF for total counts was defined as $\bar{G}(z) = \sum_i z^{n_i+m_i}/n_c$, and the loss function adapted to $J(\phi) = \int_0^1 |\mathcal{G}(z, \phi) - \bar{G}(z)|^2 dz$, with $\mathcal{G}(z, \phi)$ computed by integrating the square of $G = M(\sigma_{\text{on}}/d, (\sigma_{\text{on}} + \sigma_{\text{off}})/d, \bar{\rho}\beta(z-1)/d)$ over $\hat{p}(\beta)$. The modified telegraph model requires inference of three parameters $\{\bar{\rho}, \sigma_{\text{on}}, \sigma_{\text{off}}\}$. The relative errors are averaged over four parameters for the 2-state ENC model and three for the telegraph model, with the mean and standard error computed across 100 replicate datasets.
- [70] The other fixed parameters are $\bar{\rho} = 17$, $\tau = 0.5$, $d = 1$, and $n_c = 10^4$.
- [71] P. Thomas, N. Popović, and R. Grima, Phenotypic switching in gene regulatory networks, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6994 (2014).
- [72] Y. Wang, Z. Shu, Z. Cao, and R. Grima, From noise to models to numbers: Evaluating negative binomial models and parameter estimations in single-cell RNA-SEQ, bioRxiv, 10.1101/2025.05.05.652189 (2025).
- [73] We excluded the 3-state model, as simulations (SM Fig. S4, Table S4 [38]) showed that, with sample sizes of a few thousand cells, it is difficult to distinguish it from the 2-state model.
- [74] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, Cross validation for model selection: A review with examples from ecology, *Ecol. Monogr.* **93**, e1557 (2023).
- [75] Here $\text{cor}(x, y)$ denotes the correlation coefficient between score vectors x and y .
- [76] A. Riba, A. Oravec, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes, and N. Molina, Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning, *Nat. Commun.* **13**, 2865 (2022).
- [77] R. Foreman and R. Wollman, Mammalian gene expression variability is explained by underlying cell state, *Mol. Syst. Biol.* **16**, e9146 (2020).
- [78] R. Ietswaart, S. Rosa, Z. Wu, C. Dean, and M. Howard, Cell-size-dependent transcription of FLC and its antisense long non-coding RNA coolair explain cell-to-cell expression variation, *Cell Syst.* **4**, 622 (2017).
- [79] N. Battich, T. Stoeger, and L. Pelkmans, Control of transcript variability in single mammalian cells, *Cell* **163**, 1596 (2015).
- [80] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, David: Database for annotation, visualization, and integrated discovery, *Genome Biol.* **4**, P3 (2003).
- [81] <https://github.com/edwardcao3026/pgf-inf>
- [82] M. J. Levesque and A. Raj, Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation, *Nat. Methods* **10**, 246 (2013).
- [83] Y. Jiang, N. R. Zhang, and M. Li, Scale: Modeling allele-specific gene expression by single-cell RNA sequencing, *Genome Biol.* **18**, 74 (2017).

End Matter

PGF-based parameter inference—Consider a sample of n_c cells under steady-state conditions. For each cell i , we measure the nuclear mRNA count n_i , cytoplasmic mRNA count m_i , and cell volume V_i . The empirical joint distribution across the population is defined as

$$\hat{P}(n, m) = \frac{1}{n_c} \sum_{i=1}^{n_c} \mathbf{1}_{n=n_i, m=m_i},$$

where $\mathbf{1}_{n=n_i, m=m_i} = 1$ if $n = n_i$ and $m = m_i$, and 0 otherwise. The corresponding empirical PGF is $\bar{G}(z_1, z_2) = \sum_{n, m} \hat{P}(n, m) z_1^n z_2^m$, which leads to Eq. (6).

To infer the kinetic parameters ϕ , we minimize the loss function defined by the L^2 norm in Eq. (7). The focus below is on the efficient implementation of this objective.

To account for extrinsic noise—primarily arising from variability in cell volume [57,58]—we adopt ENC models, in which the transcription rate ρ is proportional to the normalized cell volume β . This normalization is given by

$$\beta_i = \frac{V_i}{\sum_i V_i / n_c}. \quad (\text{A1})$$

When direct measurements of V_i are unavailable, the total transcript count per cell (i.e., the sum of $n_i + m_i$ across the genome) can be used as a proxy. The true distribution of β , denoted $p(\beta)$, is typically unknown and is therefore

approximated using kernel density estimation (KDE) [67] based on the observed values β_i . The resulting estimate is denoted by $\hat{p}(\beta)$.

By correcting for extrinsic noise and assuming two independent gene copies per cell [63], the PGF of the ENC models across the cell population can be computed using Eq. (8). This independence assumption is supported by genome-wide experimental evidence [62,82,83]. To improve computational efficiency of evaluating Eq. (8), we apply Gauss quadrature with a change of integration variable. This transforms Eq. (8) into

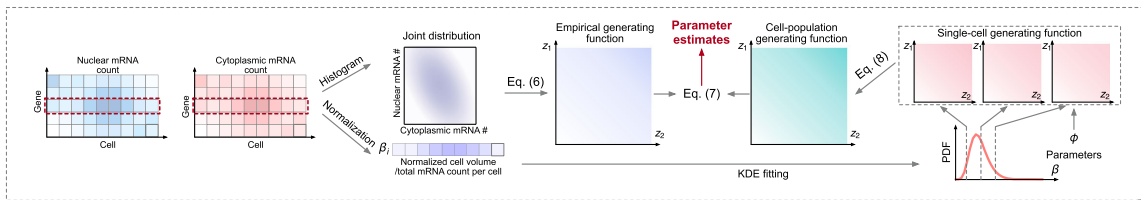
$$\mathcal{G}(z_1, z_2, \phi) = \frac{\beta_{\max} - \beta_{\min}}{2} \int_{-1}^1 G^2(z_1, z_2, \phi, \beta_x) \hat{p}(\beta_x) dx, \quad (\text{A2})$$

where

$$\beta_x = \frac{\beta_{\max} - \beta_{\min}}{2} x + \frac{\beta_{\max} + \beta_{\min}}{2}.$$

Here, $\beta_{\max} = \max_i \beta_i$ and $\beta_{\min} = \min_i \beta_i$, with x as the transformed integration variable. The numerical approximation of Eq. (A2) using N_β -point Gauss-Legendre quadrature is given by

(a) PGF-based parameter inference



(b) PGF-based model selection

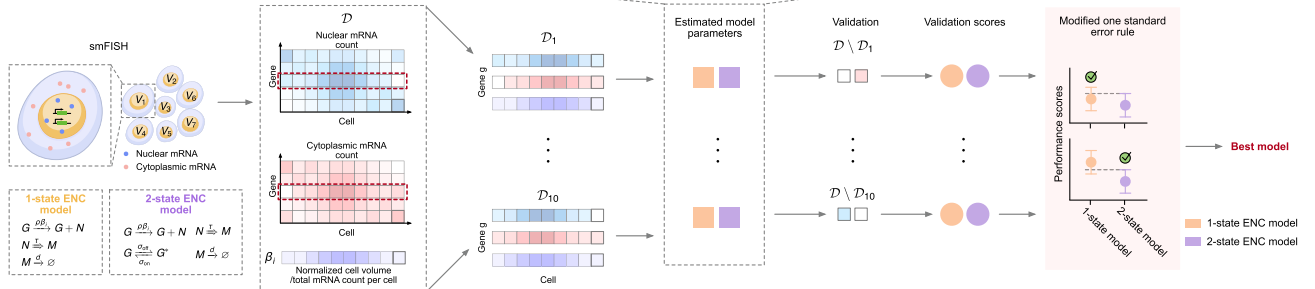


FIG. 5. Schematic of the PGF-based analysis pipeline for single-cell nuclear–cytoplasmic mRNA count data, comprising (a) PGF-based parameter inference (Algorithm 1) and (b) model selection (Algorithm 2).

$$\mathcal{G}(z_1, z_2, \phi) \approx \frac{\beta_{\max} - \beta_{\min}}{2} \sum_{j=1}^{N_\beta} w_j G^2(z_1, z_2, \phi, \beta_{x_j}) \hat{p}(\beta_{x_j}), \quad (\text{A3})$$

where x_j and w_j are the j th quadrature node and weight, respectively, computed via the `gausslegendre` function in Julia.

To further improve the computational efficiency of evaluating the loss function in Eq. (7), we approximate the integral using Gauss quadrature:

$$J(\phi) \approx \frac{1}{4} \sum_{j=1}^{N_y} \sum_{k=1}^{N_y} w_j w_k \left| \mathcal{G}\left(\frac{y_{1,j} + 1}{2}, \frac{y_{2,k} + 1}{2}, \phi\right) - \bar{G}\left(\frac{y_{1,j} + 1}{2}, \frac{y_{2,k} + 1}{2}\right) \right|^2, \quad (\text{A4})$$

where $y_{1,j}, y_{2,k} \in [-1, 1]$ are the j th and k th quadrature nodes of order N_y , and w_j, w_k are the corresponding weights obtained via the `gausslegendre` function in Julia. Finally, we minimize the loss function $J(\phi)$ using the Nelder–Mead algorithm, implemented via the `OPTIM.JL` package in Julia. This algorithm is chosen for its robustness to initial conditions and its efficiency in avoiding the costly computation of gradients $\partial_\phi J(\phi)$. The complete numerical procedure is summarized in Algorithm 1. Unless otherwise specified, optimization uses a gradient tolerance of `g_tol` = 10^{-20} and a maximum of `iterations` = 1000, as set in the default `optimize` command.

Algorithm 1. PGF-based parameter inference accounting for extrinsic noise and multiple gene copies.

Input: Number of cells n_c , nuclear and cytoplasmic mRNA counts (n_i, m_i) and cell volumes V_i for each cell i

Output: Inferred kinetic parameters ϕ

- 1: Compute normalized cell volumes β_i using Eq. (A1)
- 2: Estimate $\hat{p}(\beta)$ via KDE from $\{\beta_i\}$
- 3: Generate Gauss–Legendre quadrature points and weights (x_j, w_j) for β using `gausslegendre`
- 4: Generate quadrature points and weights $(y_{1,j}, y_{2,k}, w_j, w_k)$ for z_1 and z_2
- 5: Compute the empirical PGF $\bar{G}(z_1, z_2)$ using Eq. (6)
- 6: Initialize parameter vector ϕ
- 7: **while** Stopping criterion not met **do**
- 8: Evaluate the model PGF $\mathcal{G}(z_1, z_2, \phi)$ in Eq. (8) using the efficient approximation in Eq. (A3), where $G(z_1, z_2; \phi, \beta_{x_j})$ is computed from Eq. (3), Eq. (4), or Eq. (S33)
- 9: Compute the loss $J(\phi)$ using Eq. (A4), a numerically efficient approximation to Eq. (7)
- 10: Update ϕ using Nelder–Mead optimization
- 11: **end while**
- 12: **return** Final estimate of kinetic parameters ϕ

Algorithm 2. PGF-based model selection pipeline via ten fold cross validation.

Input: Number of cells n_c ; partitioned dataset $\{\mathcal{D}_1, \dots, \mathcal{D}_{10}\}$; candidate models $\{\text{model}_1, \dots, \text{model}_n\}$ ordered by increasing complexity

Output: Selected best-fitting model

- 1: **for** each candidate model model_i **do**
- 2: **for** each fold $j = 1$ to 10 **do**
- 3: Use Algorithm 1 to infer parameters $\hat{\phi}_j$ from training data $\{\mathcal{D}_k\}_{k \neq j}$
- 4: Evaluate validation score $J(\hat{\phi}_j)$ on $\mathcal{D}_j$
- 5: **end for**
- 6: Store validation scores $\mathcal{J}_{\text{model}_i} = [J(\hat{\phi}_1), \dots, J(\hat{\phi}_{10})]^\top$
- 7: Compute mean $\bar{\mathcal{J}}_{\text{model}_i}$ and standard deviation σ_{model_i}
- 8: **end for**
- 9: Identify model with minimal mean score: index $\mathcal{I}$ and score vector $\mathcal{J}_{\text{best}} := \mathcal{J}_{\text{model}_{\mathcal{I}}}$
- 10: **for** each model model_i with $i \leq \mathcal{I}$ **do**
- 11: Compute correlation $\text{cor}(\mathcal{J}_{\text{best}}, \mathcal{J}_{\text{model}_i})$
- 12: Compute the threshold score $\text{thresh}_{\text{model}_i} = \bar{\mathcal{J}}_{\text{best}} + \sigma_{\text{best}} \sqrt{1 - \text{cor}(\cdot)}$ as given by the right-hand side of Eq. (9)
- 13: **if** $\bar{\mathcal{J}}_{\text{model}_i} \leq \text{thresh}_{\text{model}_i}$ **then**
- 14: Set $\ell := i$ and break
- 15: **end if**
- 16: **end for**
- 17: **return** Best-fitting model model_ℓ

PGF-based model selection pipeline—The identification of gene activity type is formulated as a model selection problem. Since PGF-based inference does not rely on conventional maximum likelihood estimation, standard information criteria (e.g., Akaike or Bayesian information criteria) are not directly applicable. Nevertheless, the proposed pipeline performs comparably to these criteria (SM Fig. S5 [38]). To overcome this limitation, we adopt a tenfold cross-validation strategy described in the main text. The full procedure is summarized in Algorithm 2.

Processing of smFISH and MERFISH data—For the *smFISH* dataset, we obtained the file `sciadv.adh5138_data_s2.xlsx` from the SM of Ref. [17], which reports nuclear and cytoplasmic mRNA counts for 503 and 537 mononucleated cells for the genes SPAC2H1001 and Rad21, respectively, along with corresponding cell volume measurements. These genes were selected because their mRNA Fano factors (variance-to-mean ratios) are greater than or equal to 1, a characteristic consistent with the properties of both the 1- and 2-state models analyzed in this study. For the *MERFISH* dataset, we used total and nuclear mRNA counts from the files `pnas.1912459116.sd12.csv` and `pnas.1912459116.sd14.csv` in Ref. [14], which include measurements for 10050 genes across $n_c = 1368$ human osteosarcoma (U-2 OS) cells. Genes with total mRNA counts less than 1 were excluded,

resulting in 6724 retained genes. For each gene, we constructed nuclear-cytoplasmic mRNA count pairs (n_i, m_i) , where i denotes the cell index. To mitigate the influence of outliers, we removed cells whose counts for a given gene fell outside the range $(\text{mean} \pm 5 \text{ std})$, where the mean and standard deviation were computed across all cells for that gene. Consequently, the number of usable cells varied by gene.

Ontology analysis—Genes classified by their best-fitting model (1- or 2-state ENC models) were subjected to gene ontology (GO) analysis using the DAVID platform [80], with a background of 6724 filtered genes. GO terms were extracted from the GOTERM_BP_FAT category. Terms with a Benjamini-adjusted p value (p_{adj}) below 0.01 and at least 15 enriched genes were retained and ranked by Gene Ratio.