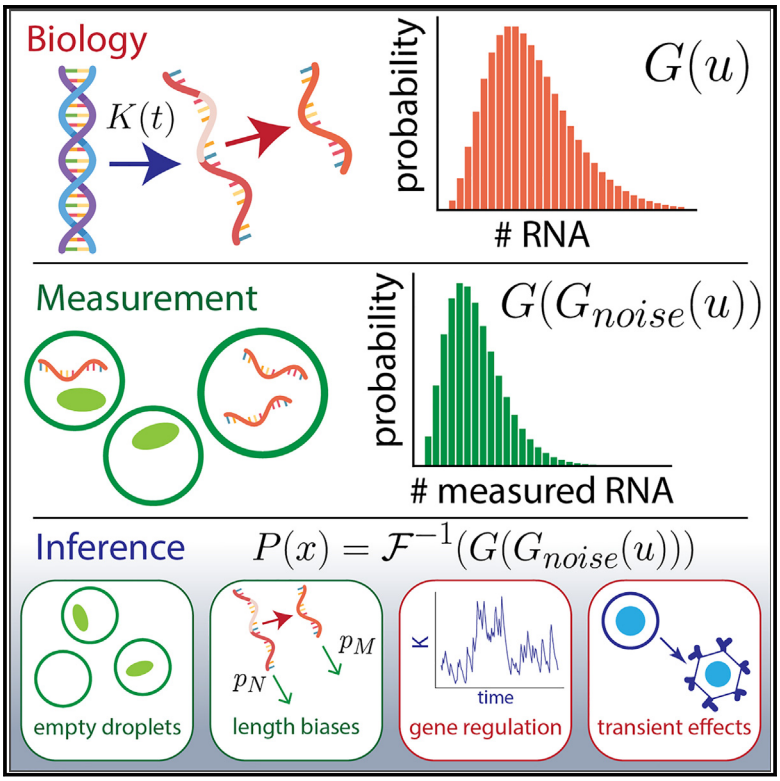
Synthesis

Cell Systems

Studying stochastic systems biology of the cell with single-cell genomics data

Graphical abstract



Highlights

- Fitting single-cell omics data requires quantitatively treating stochasticity
- Generating functions provide a unifying framework for modeling stochasticity
- Transcription, splicing, catalysis, gene switching, and technical artifacts are modeled
- Framework informative about which types of noise are important to model

Authors

Gennady Gorin, John J. Vastola, Lior Pachter

Correspondence

lpachter@caltech.edu

In brief

We construct and apply a mathematical framework that uses generating functions to model biological stochasticity and technical noise for single-cell genomics experiments in a unified fashion. The approach provides biological insights and guidance on which technical artifacts are important to model.

Gorin et al., 2023, Cell Systems *14*, 1–22
October 18, 2023
https://doi.org/10.1016/j.cels.2023.08.004

CellPress

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Cell Systems

CellPress

Synthesis

Studying stochastic systems biology of the cell with single-cell genomics data

Gennady Gorin,[1] John J. Vastola,[2] and Lior Pachter[3,4,5,*]
[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[4]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA
[5]Lead contact
*Correspondence: lpachter@caltech.edu
https://doi.org/10.1016/j.cels.2023.08.004

SUMMARY

Recent experimental developments in genome-wide RNA quantification hold considerable promise for systems biology. However, rigorously probing the biology of living cells requires a unified mathematical framework that accounts for single-molecule biological stochasticity in the context of technical variation associated with genomics assays. We review models for a variety of RNA transcription processes, as well as the encapsulation and library construction steps of microfluidics-based single-cell RNA sequencing, and present a framework to integrate these phenomena by the manipulation of generating functions. Finally, we use simulated scenarios and biological data to illustrate the implications and applications of the approach.

INTRODUCTION

In his classic systems biology textbook,[1] D.J. Wilkinson notes that "Improvements in experimental technology are enabling quantitative real-time imaging of expression at the single-cell level, and improvement in computing technology is allowing modeling and stochastic simulation of such systems at levels of detail previously impossible. The message that keeps being repeated is that the kinetics of biological processes at the intracellular level are stochastic and that cellular function cannot be properly understood without building that stochasticity into *in silico* models." From this perspective, systems biology studies control over randomness and the ways in which living cells exploit variability to grow and function. Counterintuitively, this stochastic *weltanschauung* relies on mental models that are inherently deterministic: differentiation landscapes,[2–6] gene expression manifolds,[7] cellular state graphs,[8,9] gene regulatory networks,[10,11] and kinetic parameters.[12] Analysis of experimental data therefore requires reconciling underlying deterministic structure with biological stochasticity and experimental technical variability, or noise. In particular, distinguishing technical noise from biological stochasticity involves the statistical modeling of experimental readouts, expected noise sources, and the signal-to-noise ratio, and requires consideration of the theoretical and computational tractability of the model.

How can we model these features—latent deterministic structure, biological stochasticity, and technical noise—in a way that balances our models' ability to adequately describe available data with our own ability to adequately understand the mathematical behavior and biological interpretation of our models? Answering this question is particularly challenging in the context of single-cell genomics, where datasets are large and sparse, the signal-to-noise ratio is low, and stochasticity is one of the defining features of the underlying biophysics.[13–15] Here, we explain why many naive approaches to understanding the stochastic systems biology of single cells fall short and describe a theoretical framework that can serve as an alternative. Our framework extends recent work on the mechanistic modeling of single-cell RNA count distributions[16–21] and addresses both how models can be efficiently fit to single-cell data and what features of the underlying biology we can hope to learn.

After introducing the general framework, we illustrate its consequences through a series of vignettes. In each case, we consider modeling particular aspects of biological and technical noise and ask the following questions: (1) what do our models help us learn about the underlying biology and (2) what could go wrong if we ignored these features of our data? We find that certain kinds of noise must be carefully modeled, others are poorly identifiable, whereas others still cannot be identified at all and can be safely ignored.

Systems biology and single-cell genomics

Standard approaches to systems biology

If an experiment has ample controls and provides a readout with a high signal-to-noise ratio in the relevant variables, coarse-grained, moment-based models can be ideal. For example, investigations of cell growth have effectively used least-squares regression to fit scaling relationships between cell volume and molecular abundance that hold on average.[22,23] Analogously, experiments leveraging the integration of multiple fluorescent reporters have successfully decomposed molecular noise sources into intrinsic and extrinsic components,[24] leading to numerous

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

analytical[25–28] and experimental[29–31] extensions that leverage the lower moments of poorly characterized biological drivers to describe or delimit the system variability. These approaches, which have found application to new experimental techniques, have origins in the Onsager and Langevin theories of the early twentieth century,[32] which specify the moment behaviors of near-equilibrium statistical thermodynamic systems using Gaussian terms.

Alongside studying biology on a gene-by-gene basis, considerable effort has been dedicated to the discovery of regulatory networks. This problem is considerably more challenging: the number of candidate network modules rapidly grows with the number and size of motifs of interest, and simple moment-based models risk distorting key qualitative features, such as multistability. From the perspective of statistics, network inference requires specifying or bypassing likelihood functions for joint gene expression, which may combine various noise sources in addition to the "signal" of regulation. Typical ways of addressing this challenge include[33,34]:

(1) The purely descriptive approach, which interprets an expression correlation matrix as a graph but does not provide an easily interpretable way to extract its signal.
(2) Thresholding, which bins the unknown observed distribution to obtain a known, but lower-information distribution, as with binarization used to construct Boolean networks[35] or implement the phixer algorithm.[36]
(3) Distributional assertion, which fits static observations by assuming statistics or observations are Gaussian, as in a variety of popular Bayesian,[34] information-theoretic,[37] and regression-based[38] methods; this assumption may[39] or may not[40] provide accurate results.
(4) The dynamic approach, which fits a time-dependent trajectory to data using assuming Gaussian residuals; this assumption may reflect stochastic differential equation (SDE) dynamics[41] or isotropic observation noise added to a latent process.[42–44]

This overview is far from exhaustive, but it demonstrates a key theme: relatively robust signals, such as the lower moments or the absence/presence of gene expression, can be treated using fairly simple models that rely on highly optimized, well-understood methods and algorithms developed in the context of signal processing and dynamical systems analysis. Which simple model may perform best is not known *a priori* and heavily depends on the task.[33] Ideally, methods are benchmarked on simulated[39,45] or well-characterized "gold standard"[33,46] datasets to glean partial insights about their performance and limitations. In this framework, improving the signal-to-noise ratio requires either designing more precise readouts or sacrificing a portion of the obtained data.

The challenge of single-cell data

Advances in sequencing technologies, most dramatically the rapid commercialization and adoption of single-cell RNA sequencing (scRNA-seq), which can profile millions of cells on a genome-wide scale,[47,48] have been heralded as a promising frontier for systems biology.[49–51] This potential is more striking, yet due to simultaneous advances in multiomics or the measurement of multiple modalities (transient and non-coding RNA species, DNA methylation, chromatin accessibility, and surface protein abundance) in individual cells,[52,53] facilitating "integrated" analysis.[54–56] The "big data" from single-cell sequencing have thus served as a substrate for a plethora of investigations that are, at the first glance, analogous to the research program of systems biology at large: the identification of cell types, their aggregation into trajectories, the discovery of gene modules that consistently differ between cell types or throughout a differentiation trajectory, and the visualization of low-dimensional summaries reflecting some component of the data structure.

To identify these coarse-grained motifs in the structure of single-cell datasets, it is common practice to analyze cell-cell graphs, constructed from measures of expression similarity, to attempt to construct cliques (cell types), shortest paths (trajectories), and neighborhood-preserving low-dimensional embeddings (visualizations). In addition, relatively simple parametric distributions are widely used, with the Gaussian assumption popular for the lower moments (e.g., to compute measures of differential expression) and the lognormal or negative binomial used to describe count distributions.[57,58] Standard scRNA-seq data provide snapshots of processes, rendering dynamical analysis fairly complex, but it is common to fit a "pseudotemporal" curve through the dataset by minimizing a Gaussian error term between this curve and some transformation of the cells' expression levels.[59,60]

Here, however, the underlying assumptions break down. Single-cell data are intrinsically and qualitatively different from readouts of typical systems biology experiments, with drastic implications for analysis. Single-cell data are large and sparse, with a preponderance of technical noise effects, poorly characterized batch- and gene-level biases and low per-cell copy numbers.[13–15] Improving the signal-to-noise ratio by designing more targeted experiments is challenging, as commercial technology is designed to quantify molecules on a genome-wide scale. More problematically, typical distributional assumptions and data transformations risk losing a considerable amount of signal in the low-copy number regime. This challenge informs part of the broader discussion of the relative roles of data analysis and mechanistic hypotheses in genomics,[19,20,61] as analyses that are not constrained by mechanism or theory may contradict existing knowledge.

More specific critiques have considered whether various analyses are appropriate or excessively heavy handed. For example, sparsity has led to *ad hoc* procedures to "correct" the data, which may in turn lead to incorrect conclusions.[62–64] Normalization and log-transformation, which attempt to remove technical biases and prepare the data for dimensionality reduction, rely on assumptions, such as high copy numbers and homogeneity, that are routinely violated in single-cell datasets.[65,66] Dimensionality reduction risks distorting both local and global relationships between data points.[19,67,68] Finally, the use of cell-cell graphs constructed from noisy data reifies relationships that may not reflect those in the original tissue and risks introducing hard-to-diagnose errors into downstream analysis.[19,69] Although these issues span the entire process of analysis, all, at least partially, trace back to uncomfortable compromises in the treatment of uncertainty and variation in a regime unforgiving of approximations.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Stochastic modeling of intracellular network dynamics

Stochasticity is, then, mandatory, and we ignore it at our own risk. Therefore, we advocate for probabilistic alternatives to the "extraction" of signals from scRNA-seq datasets. Since biology is stochastic, the noise *is* the signal. To quantify and characterize the components of deterministic mental models—differentiation landscapes, kinetic parameters, and similar low-dimensional abstractions[70]—in a principled way, we need to combine them with stochastic terms that result from specific hypotheses about the underlying biophysics and chemistry[20] or risk confirmation bias.[19]

The development of stochastic models offers advantages beyond loss function bookkeeping. If multiomic data are available, there is typically a self-consistent way to extend the models accordingly.[71] Although likelihoods induced by stochastic processes are challenging to analyze and implement, they provide appealing statistical properties. When the data are sufficiently informative, full distributions provide better estimates than moments.[40] When they are not, probabilistic approaches are appropriately conservative, as they report, rather than elide, the parameter degeneracies. A thorough mathematical understanding of model behaviors—i.e., precisely which parameters are identifiable and which are degenerate, as well as how much data must be collected—enables the design of informative experiments.[20,72] Finally, the use of mechanistic models, parametrized by rate constants, allows us to draw conclusions about the mechanistic bases and effects of perturbations.[73]

These principles have guided fluorescence-based single-cell transcriptomics for nearly 20 years. To obtain as much information as possible from entire copy-number distributions,[40,74] the field has developed a considerable arsenal of theoretical tools[75,76] and solution strategies.[77–79] It is, then, particularly natural to build scRNA-seq models that *extend* processes consistent with fluorescence imaging: this approach allows us to leverage existing theory, as well as encode the intuition that technology-dependent effects should be independent of biological ones. A particularly popular class of models involves the bursty production of RNA and its Markovian degradation,[73,80] which can be analyzed in the chemical master equation (CME) framework.[81,82] The key theoretical points have already been applied in the context of single-cell sequencing; for example, the Poisson, Poisson-gamma, and Poisson-beta distributions, which are common in sequencing analyses,[58,63,83,84] are three of the limiting distributions induced by this class of models.[20,80,85] However, this possible mechanistic basis is only rarely[84,86–88] invoked in the development of analysis methods.

Outlook

Unfortunately, we cannot simply apply the existing methods from fluorescence transcriptomics; the scale and chemistry of single-cell technologies create additional desiderata. General CME solutions are computationally prohibitive and challenging to scale to thousands of genes,[89] requiring careful study of narrow model classes with tractable solutions.[17,20] In addition, connecting biological models to observations requires explicitly representing the experimental process. The existing models for fluorescence data are sophisticated[79] but cannot be directly applied to sequencing data. Although a variety of models have been proposed for technical noise in single-cell technologies,[13,14,90,91] their chemical foundations, as well as implications for biological parameter identifiability, have been understudied.[21]

In light of this lacuna, we seek to produce a mathematical framework that (1) integrates biological and technical variability in a coherent, modular way; (2) scales to large, multimodal data; (3) can be used to simulate datasets and make testable, quantitative predictions; and (4) affords a thorough mathematical analysis of its components, if not the entire model.

Stochastic modeling of single-cell biology

Constructing a general-purpose framework for the stochastic modeling of single-cell biology necessitates working at a relatively high level of abstraction, since we would in principle like to account for a range of processes with one formalism. In this section, we motivate our abstract formalism using a collection of concrete, biologically relevant examples.

One of the simplest models of transcription is the constitutive model, which assumes that RNA is produced at a constant rate.[20,92] It is defined by the chemical reactions

$$\varnothing \xrightarrow{K} \mathcal{X}, \mathcal{X} \xrightarrow{\gamma} \varnothing, \qquad \text{(Equation 6)}$$

where $\mathcal{X}$ is a single species of RNA, K is the (constant) transcription rate, and γ is the degradation rate. The CME that corresponds to this system is

$$\begin{aligned}\frac{\partial P(x,t)}{\partial t} &= K[P(x-1,t) - P(x,t)] \\ &+ \gamma[(x+1)P(x+1,t) - xP(x,t)],\end{aligned} \qquad \text{(Equation 7)}$$

where $P(x,t)$ is the probability that the system has $x \in \mathbb{N}_0$ RNA at time t. Solving the above master equation allows us to compare its predictions with experimental scRNA-seq data. There are several theoretical approaches for doing this—including using a special ansatz,[85] the Poisson representation,[93] the Doi-Peliti path integral,[17,94–96] and operator techniques[97]—but we would like to highlight a straightforward method that we know works for far more general problems. The idea is to consider a certain transformed version of the probability distribution, which satisfies a PDE instead of a differential-difference equation. This PDE, for a large class of biologically relevant systems, can then be solved using the method of characteristics,[98] which converts the problem of solving a PDE into integrating a system of ordinary differential equations (ODEs). This is mathematically equivalent to using certain path integral methods.[17,20,99]

Define the generating functions (GFs; see Box 1)

$$G(g,t) := \sum_{x=0}^{\infty} g^x P(x,t) \text{ and } \phi(u,t) := \log G(g,t), \qquad \text{(Equation 8)}$$

where g is on the complex unit circle and $u := g - 1$. It is easy to show that G and ϕ satisfy the PDEs

$$\begin{aligned}\frac{\partial G}{\partial t} &= (g-1)\left[KG - \gamma \frac{\partial G}{\partial g}\right], \\ \frac{\partial \phi}{\partial t} &= Ku - \gamma u \frac{\partial \phi}{\partial u}.\end{aligned} \qquad \text{(Equation 9)}$$

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Box 1. Generating function methods for studying stochastic biological systems

Generating functions are ubiquitous tools in stochastic modeling. They are central to the analysis of discrete master equations, as they cast difficult-to-solve infinite-dimensional systems to a finite number of coupled partial differential equations (PDEs), which can be treated using standard analytical or numerical methods. A (one-variable) probability distribution $P(x)$ and its generating function $G(g)$ are related according to the formulas.

$$G(g) = \sum_{x=0}^{\infty} g^x P(x), P(x) = \oint \frac{dg}{2\pi i} \frac{1}{g^{x+1}} G(g) = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} e^{-i\theta x} G(e^{i\theta}). \quad \text{(Equation 1)}$$

In the stochastic modeling of transcription, certain distributions, such as the Poisson and negative binomial, frequently appear. Because G uniquely specifies P, we can often invert G simply by recognizing its form and matching terms. Below are some generating functions of common distributions (Bernoulli, Poisson, geometric, and negative binomial):

$$P(x) = (1-p)\delta_{0x} + p\delta_{1x}, \qquad G(g) = g, \quad \text{(Equation 2)}$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad G(g) = e^{\lambda(g-1)}, \quad \text{(Equation 3)}$$

$$P(x) = \frac{\theta}{1+\theta}\left(\frac{1}{1+\theta}\right)^x, \qquad G(g) = \frac{1}{1-\theta(g-1)}, \quad \text{(Equation 4)}$$

$$P(x) = \frac{\Gamma(\nu+x)}{x!\Gamma(\nu)}\left(\frac{\theta}{1+\theta}\right)^{\nu}\left(\frac{1}{1+\theta}\right)^x, \qquad G(g) = \left(\frac{1}{1-\theta(g-1)}\right)^{\nu}. \quad \text{(Equation 5)}$$

The generating function expressions can often be made more compact by applying the substitution $u := g - 1$.

We can use the method of characteristics to find that

$$\phi(u,t) = \phi^0(U(t)) + K\int_0^t U(s)\, ds, \qquad \frac{dU}{ds} = -\gamma U, \quad \text{(Equation 10)}$$

where the $U(s)$ ODE has initial condition $U(s=0) = u$, and ϕ^0 is the initial (log-) GF of the system. In order to determine $P(x,t)$ from $\phi(u,t) = \log G(g,t)$, we can use an inverse Fourier transform:

$$P(x,t) = \oint \frac{dg}{2\pi i} \frac{1}{g^{x+1}} G(g,t) = \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} e^{-i\theta x} G(e^{i\theta}, t)$$

where we integrate over all g on the complex unit circle. In practice, this step is done numerically using an inverse fast Fourier transform.

The constitutive model, which produces Poisson distributions at steady state, is too simple for single-cell biology.[20] However, fortunately, the technique we have just described can be adapted to predict the behavior of substantially more complex models.

Multiple types of RNA

One possible generalization of the constitutive model is to so-called monomolecular systems,[17,85] which allow phenomena like RNA splicing to be accommodated. An example is the addition of splicing to the constitutive model:

$$\varnothing \xrightarrow{K} \mathcal{X}_N, \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M, \mathcal{X}_M \xrightarrow{\gamma} \varnothing. \quad \text{(Equation 11)}$$

In general, any number of production, conversion, and degradation reactions can be modeled:

$$\varnothing \xrightarrow{K_i} \mathcal{X}_i, \mathcal{X}_i \xrightarrow{c_{ij}} \mathcal{X}_j, \mathcal{X}_i \xrightarrow{c_{i0}} \varnothing. \quad \text{(Equation 12)}$$

Using the same technique we described earlier, the probability $P(\mathbf{x},t)$ that the system is in state $\mathbf{x} \in \mathbb{N}_0^n$ at time t can be shown to be equivalent to the GF

$$\phi(\mathbf{u},t) = \phi^0(\mathbf{U}(t)) + \int_0^t \mathbf{K}^T\mathbf{U}(s)\, ds, \qquad \frac{d\mathbf{U}}{ds} = C\mathbf{U}, \quad \text{(Equation 13)}$$

where $\mathbf{U}(s=0) = \mathbf{u}$, and the C matrix is defined via

$$C_{ij} = c_{ij}\ (i \neq j), C_{ii} = -\sum_{j=0}^{n} c_{ij}, \quad \text{(Equation 14)}$$

and where $c_{ii} := 0$ by convention.

Multiple gene states

Although the monomolecular model is a step forward, it still does not account for nontrivial transcription rate dynamics. One possibility is that there are multiple gene states, as in the telegraph model[76,97,100]:

$$\mathcal{S}_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} \mathcal{S}_{\text{on}}, \mathcal{S}_{\text{on}} \xrightarrow{k_{\text{init}}} \mathcal{S}_{\text{on}} + \mathcal{X}, \mathcal{X} \xrightarrow{\gamma} \varnothing. \quad \text{(Equation 15)}$$

The corresponding three-variable GF is

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$\phi(u, u_{on}, u_{off}, t) = \phi^0(U(t), U_{on}(t), U_{off}(t)), \quad \text{(Equation 16)}$$

$$\frac{dU}{ds} = -\gamma U,$$
$$\frac{dU_{off}}{ds} = -k_{on}(U_{off} - U_{on}),$$
$$\frac{dU_{on}}{ds} = -k_{off}(U_{on} - U_{off}) + k_{init}(U_{on} + 1)U,$$

where $U(0) = u$, $U_{off}(0) = u_{off}$, and $U_{on}(0) = u_{on}$. If we want to marginalize over gene state, which we usually do since it is not observable, we can set $u_{off} = u_{on} = 0$. Notice that the relevant ODEs are now nonlinear (Riccati-type) equations, which make them difficult to solve by hand. In general, considering multiple gene states, or other kinds of added complexity like autocatalytic reactions, yields nonlinear characteristic ODEs. This is no obstacle to numerical integration, however.

Gene regulation

Another possibility we would like to account for is nontrivial gene regulation. In previous work,[20] we considered two models of transcription rate variation: the gamma Ornstein-Uhlenbeck (Γ-OU) model, which assumes variation is due to changes in the mechanical state of DNA, and the Cox-Ingersoll-Ross (CIR) model, which assumes it is due to fluctuations in the concentration of an abundant regulator molecule. Analyzing them can be mathematically challenging, since the discrete stochastic dynamics of RNA production and degradation are coupled to the continuous stochastic process that controls the transcription rate. Fortunately, both models and many generalizations of them can be solved using the method of characteristics. For example, the CIR model (assuming two RNA species) is defined by a SDE[81] and three reactions:

$$\frac{dK}{dt} = a\theta - \kappa K + \sqrt{2\kappa\theta K}\xi(t),$$
$$\varnothing \xrightarrow{K(t)} \mathcal{X}_N, \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M, \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \quad \text{(Equation 17)}$$

and its solution is[20]

$$\phi(u_N, u_M, u_K, t) = \phi^0(U_N(t), U_M(t), U_K(t)) + a\theta \int_0^t U_K(s; u_N, u_M, u_K)\, ds, \quad \text{(Equation 18)}$$

$$\frac{dU_M}{ds} = -\gamma U_M, \qquad U_M(0) = u_M,$$
$$\frac{dU_N}{ds} = \beta (U_M - U_N), \qquad U_N(0) = u_N,$$
$$\frac{dU_K}{ds} = U_N - \kappa U_K + \kappa\theta U_K^2, \qquad U_K(0) = u_K.$$

Thus, it is straightforward to couple dynamics defined on different types of state spaces: categorical (e.g., gene states), continuous (e.g., transcription rates), and discrete (e.g., RNA counts), using the GF approach. In all cases, one obtains a GF solution in terms of a finite set of (possibly nonlinear) ODEs. The total number of ODEs is equal to the total number of degrees of freedom.

One feature of single-cell biology that is challenging to capture using this approach is feedback. For example, proteins expressed by a gene may affect the transcription rate of that gene. Although exact solutions for systems involving feedback are available in certain simple cases,[101–104] particularly when there is only one chemical species, more general results have proven elusive. From the point of view of our approach, including chemical reactions that involve feedback yields GF PDEs that are not first order and that cannot be solved in terms of ODEs via the method of characteristics (as explored in more detail in the supplemental information).

Transient effects

In the context of development or reprogramming, we are especially interested in using single-cell genomics data to study transient processes. In particular, certain cell types or subtypes (like neural progenitor cells) only exist for a certain window of time, and by collecting single-cell data, we are taking a snapshot of many cells, each of which may be in a different part of the process. How does this affect observed RNA counts?

Different cells being observed at different times means we are not interested in $P(\mathbf{x}, t)$, but $P(\mathbf{x}, t)$ averaged over some distribution that indicates how likely we are to sample different times. The shape of the sampling distribution $f(t)$ depends on when cells tend to exit a given state (e.g., by differentiating into a different cell type). Nontrivial sampling distributions are compatible with our GF approach, since we can simply modify the distribution that appears. For a model with one discrete species, we can write the full GF G_{tot} as

$$G_{tot}(g) = \sum_{x=0}^{\infty} g^x \int_0^T P(x, t) f(t) dt = \int_0^T G(g, t) f(t) dt,$$

i.e., we can obtain it by integrating the GF that captures intrinsic noise.

Technical noise

In single-cell genomics experiments, we do not directly observe a given cell's RNA counts, but those numbers filtered through a noisy sequencing process.[21] In microfluidics-based sequencing, noise can come from some combination of droplets not capturing all molecules (especially types of RNA with low-copy numbers), errors in amplification, and reads not being uniquely identifiable. We would like to account for these kinds of technical noise in a way that is both principled and compatible with our GF approach to modeling intrinsic noise.

Consider a simple example, in which the relevant biology is described by the one-species constitutive model (Equation 7), and each RNA molecule is observed independently with probability p. The probability of observing x_{obs} molecules of RNA, given a biological distribution $P(x, t)$, is

$$P(x_{obs}, t) = \sum_{x=0}^{\infty} P(x_{obs}|x) P(x, t) = \sum_{x=0}^{\infty} \binom{x}{x_{obs}} p^{x_{obs}} (1-p)^{x - x_{obs}} P(x, t). \quad \text{(Equation 19)}$$

The corresponding GF G_{tot} is

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$G_{\text{tot}}(g,t) = \sum_{x=0}^{\infty}\sum_{x_{\text{obs}}=0}^{x} g^{x_{\text{obs}}} P(x_{\text{obs}}|x)P(x,t) = \sum_{x=0}^{\infty}[gp+(1-p)]^x P(x,t), \quad \text{(Equation 20)}$$

i.e., the result is the same as without technical noise, except that we have $g \to gp + (1 - p)$. In general, including technical noise requires us to replace the usual g^x factor with $G_{\text{noise}}(g,x)$, the GF associated with the observation model:

$$G_{\text{tot}}(g,t) = \sum_{x=0}^{\infty} G_{\text{noise}}(g,x)\ P(x,t). \quad \text{(Equation 21)}$$

For certain common observation models, like the Bernoulli model just described, or a Poisson noise model, we can say more: since

$$G_{\text{noise}}(g,x) = G^*(g)^x \quad \text{(Equation 22)}$$

for some G^*, including technical noise amounts to replacing g with G^*, so that $G_{\text{tot}} = G(G^*)$ is a composition of GFs. In this paper, we typically assume that all technical noise models satisfy Equation 22 for some G^*.

RESULTS

Theoretical framework for stochastic systems biology

We are ready to present our general framework for stochastic systems biology, which accommodates all of the sources of stochasticity described in the preceding section: intrinsic noise, transient effects, and technical noise. In order to balance the amount of biology our models can capture with the mathematical tractability of those models, we restrict our analysis to a fairly general class of systems that can be solved using the method of characteristics. For such systems, we can obtain likelihoods by integrating characteristic ODEs, using the obtained characteristics to construct the GF, and then doing an inverse (fast) Fourier transform.

This class of systems permits gene state interconversion, as well as the production and processing of RNA and proteins, which could be treated as discrete or continuous variables depending on their concentration. We allow zero- and first-order reactions, including state-dependent bursting, interconversion, degradation, and catalysis. However, we disallow higher-order reactions (e.g., binding reactions $A + B \to C$), including feedback-based regulation like protein-promoter binding. Therefore, our analysis focuses on Markovian systems that possess N categorical degrees of freedom, corresponding to gene states; n discrete ones, corresponding to low-copy number molecular species; and m continuous ones, corresponding to transcription rates or high-concentration species. This class of reactions is schematically represented in Figure 1A; crucially, it consists of distinct "upstream" and "downstream" components.

Given all of a model's possible reactions, one can write down a corresponding master equation that keeps track of how microstate probabilities change with time:

$$\frac{dP(s,\mathbf{x},\mathbf{y},t)}{dt} = \psi(s,\mathbf{x},\mathbf{y},t), \quad \text{(Equation 27)}$$

where each microstate consists of s, the categorical dimension; $\mathbf{x} \in \mathbb{N}^n$, the n discrete dimensions; and $\mathbf{y} \in \mathbb{R}^m$, the m continuous dimensions. The generally complicated function ψ aggregates all reaction rates. Master equations like Equation 27 typically consist of an infinite system of coupled ODEs and hence are difficult to solve in general. This is one reason we chose a particular class of systems: to solve Equation 27 using the method of characteristics and hence determine a given model's predictions, all we need to do is solve (a finite number of) ODEs satisfied by the characteristics and GF.

The N-dimensional GF $\mathbf{G} = (G_1, \ldots, G_N)^T$ of the system, which is a function of spectral variables $\mathbf{g}$ and $\mathbf{h}$, is defined by

$$G_s(\mathbf{g},\mathbf{h},t) := \int_{\mathbf{y}}\sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^T\mathbf{y}} P(s,\mathbf{x},\mathbf{y},t)d\mathbf{y}. \quad \text{(Equation 28)}$$

Equation 27 can be converted into a PDE satisfied by $\mathbf{G}$:

$$\frac{\partial \mathbf{G}}{\partial t} = -\mathcal{H}(\mathbf{u},t)\ \mathbf{G} + J[C\mathbf{u} + \text{diag}\ \mathbf{u}\ D\mathbf{u}]$$
$$\mathcal{H}(\mathbf{u},t) = -H(t)^T - \mathcal{A}(\mathbf{u},t)\odot \quad \text{(Equation 29)}$$
$$\mathbf{u} := \begin{bmatrix} \mathbf{g} - 1 \\ \mathbf{h} \end{bmatrix},$$

where $\odot$ is the Hadamard/elementwise matrix product, J is the Jacobian matrix of the GF, and $\mathbf{u}$ combines the discrete and continuous degrees of freedom. The time-dependent matrix H contains the kinetics of state transitions, whereas the operator $\mathcal{A}$ describes the drift and bursty production processes, which may depend on state. Therefore, the operator $\mathcal{H}$ aggregates the upstream components of the system. The matrix C contains interconversion, degradation, and mean reversion-like terms, whereas D contains the catalysis and square-root noise terms. $\mathcal{H}$, C, and D encode a quasi-linear, deterministic, and first-order N-component system of PDEs in $n + m$ spectral variables.

Applying the method of characteristics to solve Equation 29 tells us that the downstream part of the system is fully determined by a set of characteristics $\mathbf{U}$, which are defined by the ODEs

$$\frac{d\mathbf{U}(s)}{ds} = C\mathbf{U}(s) + \text{diag}\ \mathbf{U}(s)\ D\mathbf{U}(s) \quad \text{(Equation 30)}$$

where s is an integration variable, and $\mathbf{U}(s = 0) = \mathbf{u}$. Meanwhile, the GF $\mathbf{G}$ can be determined from

$$\frac{d\mathbf{G}(s)}{ds} = \mathcal{H}(\mathbf{U}, t - s)\ \mathbf{G}, \quad \text{(Equation 31)}$$

which has initial condition $\mathbf{G}^0(\mathbf{U}(t))$, where $\mathbf{G}^0$ is the GF of the initial distribution. The upstream components describe how molecule production occurs and hence depend on $\mathcal{H}$; their influence on the final answer is through the above integral.

The detailed form of Equation 27 is complicated, and the arithmetic exercise of converting it into Equation 29 is tedious. We show how to construct the biological master equation in the section "master equation models of transcription," write it out in full in the section "the full master equation," and discuss at a

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

CellPress

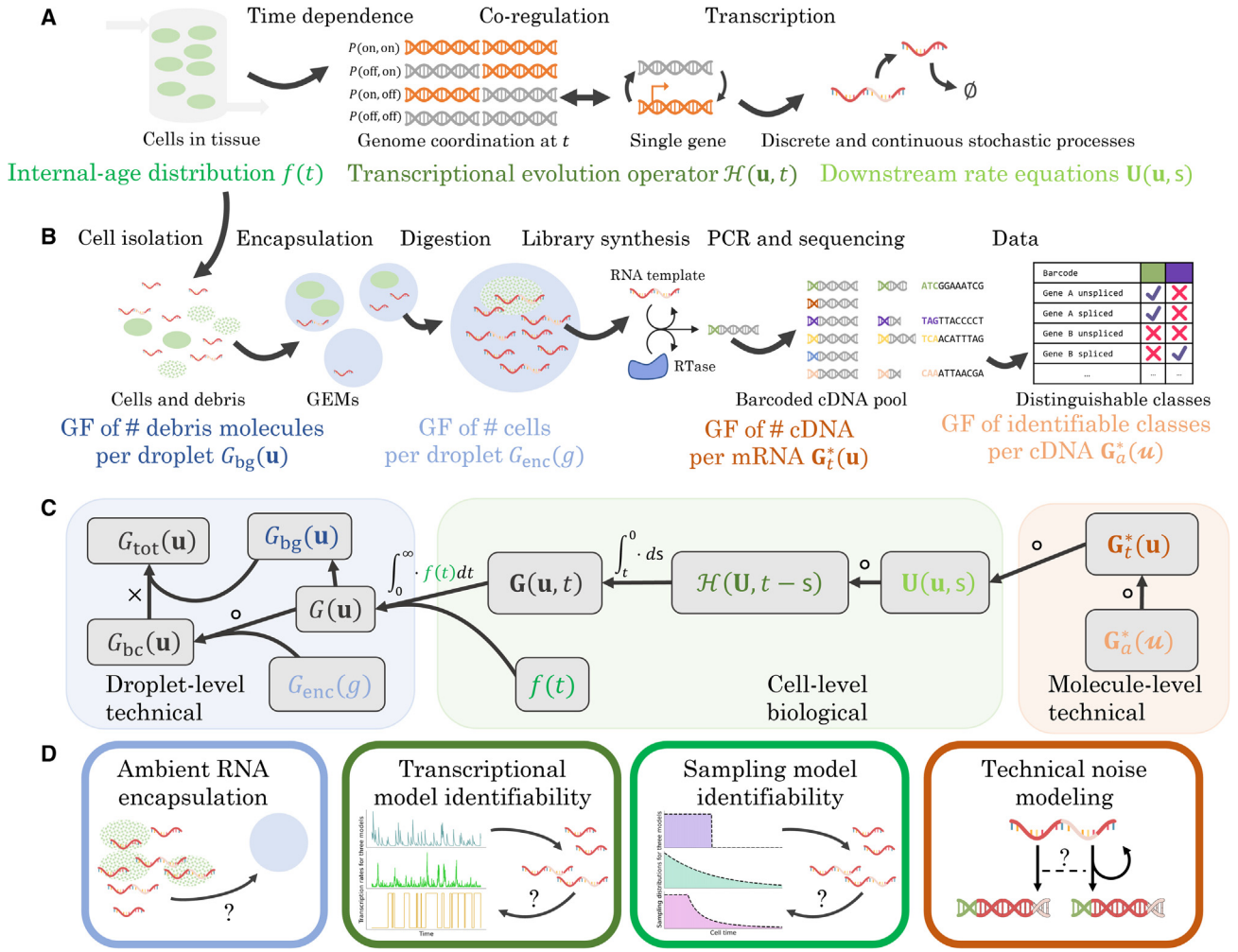


Figure 1. The biophysical and chemical phenomena of interest, as well as the relationships between their generating functions

(A) The biological phenomena of interest: cell influx and efflux into a tissue observed by sequencing; the time-dependent transcriptional regulation of one or more genes; downstream continuous and discrete processes.

(B) The technical phenomena of interest: the encapsulation of cells and cell debris; cDNA library construction; the loss of information in transcript identification (GF, generating function; RTase, reverse transcriptase).

(C) The structure of the full generating function of the system in (A) and (B): to obtain the solution, we variously compose, integrate, and multiply the generating functions of the constituent processes.

(D) The stochastic and statistical properties of four components of the full system: the background debris, the transcriptional regulation, the cell/tissue relationship, and the technical noise mechanism.

high level how to solve it using our GF approach in the section "generating function methods for biological stochasticity." The terms of the full master equation are annotated in Table S1, and the solution process is described in more detail in the supplemental information.

In special cases, the ODEs we obtain can be solved exactly (see, e.g., Box 2). For example, whenever $D = 0$, the downstream ODE system can be solved analytically by eigendecomposition. If, in addition, only a single gene state is present, H vanishes, and the upstream component can be evaluated by numerical integration.[16] Finally, in the simplest case of a linear operator $\mathcal{A}$, we obtain an analytically tractable system equivalent to a deterministic system of reaction rate equations.[17,85]

Although this formulation nominally describes a single gene, it may be exploited to represent multi-gene systems. Conceptually, this strategy entails constructing a model where the transcription of multiple species is controlled by a common regulator. We discuss potential candidate models in the section "coupling multiple genes"; these models instantiate hypotheses to produce $\mathcal{H}$ and $\mathbf{U}$ that represent co-regulation.

To explain the observation of transient processes, such as the simultaneous capture of progenitor and descendant cells from a differentiation process, we take inspiration from previous work in sequencing[86] as well as chemical reactor modeling[105,106] and extend the theoretical framework originally proposed in our recent RNA velocity analysis.[19] In brief, the simplest model that accounts for such desynchronization proposes that cells enter a tissue, receive a signal that triggers changes in transcriptional rates $\mathcal{H}(t)$, and leave at some later point. Sequencing is the observation of cells within the tissue; to find the distribution of

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Box 2. An illustration of the solution procedure

Here, we will illustrate how to solve two simple transcription models using our framework. We assume that RNA is produced with burst event frequency α and degrades at a rate γ. In the constitutive model, each transcription event creates one RNA. In the bursty model, each transcription event creates a random number of RNA, distributed according to a geometric random variable with mean b. Both models have $N = 1$, $n = 1$, and $m = 0$. Since these models are one-dimensional, the C and D matrices are 1×1. For both of them, $C = [-\gamma]$ and $D = [0]$. The ODE for the single characteristic U (with initial condition $U(s = 0) = u$) is

$$\frac{dU(u,s)}{ds} = -\gamma U(u,s) \Rightarrow U(s) = ue^{-\gamma s}. \quad \text{(Equation 23)}$$

For a general burst distribution $p(z)$, the transcriptional evolution operator is $\mathcal{H}(u) = -\alpha(F(1+u) - 1)$, where F is the GF of the number of molecules produced per transcription event. For our two models, we have

$$p(z) = \delta_{1,z}, \qquad F(1+u) = 1+u, \qquad \mathcal{H}(u) = -\alpha u, \quad \text{(Equation 24)}$$

$$p(z) = (1+b)^{-z-1}b^z, \qquad F(1+u) = \frac{1}{1-bu}, \qquad \mathcal{H}(u) = -\alpha\frac{bu}{1-bu}. \quad \text{(Equation 25)}$$

To compute the stationary log-generating functions $\log G$, we evaluate the integrals:

$$\log G = \int_0^\infty \alpha u e^{-\gamma s} ds = \frac{u\alpha}{\gamma} \text{ for the constitutive model and}$$
$$\log G = \int_0^\infty \alpha\left[\frac{1}{1-bue^{-\gamma s}} - 1\right] ds = -\frac{\alpha}{\gamma}\log(1-bu) \text{ for the bursty model.} \quad \text{(Equation 26)}$$

The constitutive model yields a Poisson distribution with mean α/γ (cf. Equation 3), whereas the bursty model yields a negative binomial distribution with shape α/γ and scale b (cf. Equation 5).

RNA counts, we need to condition on the distribution of times since receiving the signal.

As we discuss in the section "transient phenomena," this latter distribution is not arbitrary and reflects the kinetics of cell entry and exit. In the parlance of chemical reaction engineering, the times are drawn from $f(t)$, the internal-age distribution induced by those kinetics.[105,106] This model affords a particularly simple representation of the GF:

$$G = \int_t \sum_s G_s(t) f(t) dt, \quad \text{(Equation 32)}$$

where we marginalize over the gene state, which is typically not observable. Conveniently, this model possesses time symmetry: although the cells within the tissue are all out of equilibrium, the tissue as a whole is at steady state.

We consider the technical noise phenomena shown in Figure 1B, i.e., the encapsulation of cells and background debris into droplets, as well as the stochasticity in complementary DNA (cDNA) library construction and sequencing. Under the assumption of independent encapsulation, the GF of molecule count distributions on a per-droplet level takes the following form:

$$G_{\text{tot}} = G_{\text{enc}}(G)\, G_{\text{bg}}(G), \quad \text{(Equation 33)}$$

where G_{enc} is the GF of the cells per droplet, whereas G_{bg} is the GF of background molecules per droplet, which depends on the entire cell population (section droplet encapsulation noise). Finally, to represent sequencing variability and uncertainty, we evaluate the GF at a set of transformed coordinates:

$$G_{\text{tot,ta}} = G_{\text{tot}}\left(\mathbf{G}_t^*\left(\mathbf{G}_a^*(\boldsymbol{u})\right)\right), \quad \text{(Equation 34)}$$

where $\mathbf{G}_t^*$ reflects the distribution of cDNA produced per molecule of RNA (e.g., Bernoulli, as in Tang et al.[107,108]), whereas $\mathbf{G}_a^*$ reflects the distribution of ambiguous sequenced fragments, which depends on transformed variables $\boldsymbol{u}$ (section "library construction and sequencing noise" and supplemental information).

The full GF of the molecule distribution is given by the composition and integration of the model components, as shown in Figure 1C. To evaluate this GF, it is necessary to specify all components that make up the model. In the analysis below, we take advantage of the modularity of the system definition to investigate four kinds of modeling choices, their statistical implications, and their compatibility with sequencing data. Specifically, we treat the subsystems illustrated in Figure 1D: background noise in single droplets, stochastic transcription rate models, sampling from a transient process, and variation in technical noise.

Empty droplets

One of the first steps in scRNA-seq data analysis is cell quality control, which excludes cell barcodes that appear to originate from empty droplets from further analysis.[57] For computational tractability, this procedure typically relies on "hard" assignment, such that barcodes associated with a total molecule count above some threshold are treated as cells, whereas barcodes below the threshold are treated as empty droplets. Threshold selection is necessary because even "empty" droplets contain ambient RNA. This ambient RNA, which appears to originate

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

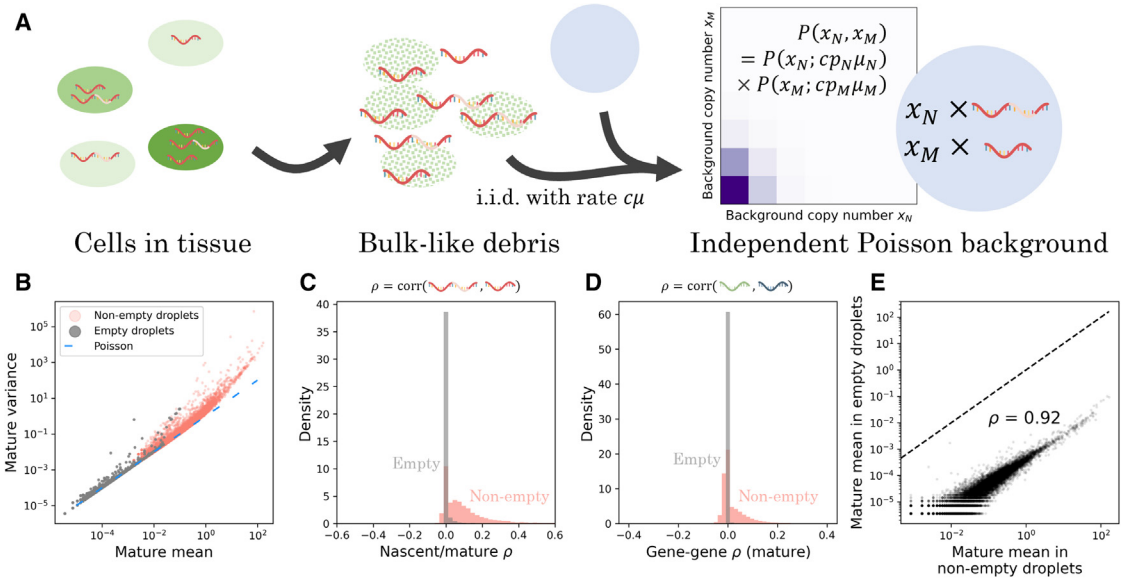


Figure 2. The pseudo-bulk model of background noise is quantitatively consistent with counts from the pbmc_1k_v3 dataset

(A) The simplest explanatory model for background noise invokes the lysis of cells (green), which creates a pool of RNA that reflects the overall transcriptome composition but retains none of the cell-level information. If the loose RNA molecules diffuse into droplets (blue) according to a memoryless and independent arrival process, the resulting background distribution (purple: higher probability mass; white: lower probability mass) observed in empty droplets should be a series of mutually independent Poisson distributions, with the mean controlled by the composition in non-empty droplets.

(B) The mature transcriptome in empty droplets has a mean-variance relationship near identity (gray points, n = 12,298), consistent with Poisson statistics (blue line); the non-empty droplets demonstrate considerable overdispersion (red points, n = 17,393).

(C) The mature and nascent transcripts in empty droplets have sample correlation coefficients ρ near zero, consistent with distributional independence (gray histogram, n = 9,362); the non-empty droplets demonstrate nontrivial statistical relationships (red histogram, n = 14,365).

(D) The mature transcripts of different genes in empty droplets have sample correlation coefficients ρ near zero, consistent with distributional independence (gray histogram, n = 75,614,253); the non-empty droplets demonstrate nontrivial statistical relationships (red histogram, n = 151,249,528).

(E) When both are nonzero, the mature count mean in empty droplets is highly correlated with the mean in the non-empty droplets, consistent with the pseudo-bulk interpretation (black points, n = 12,107; dashed line: identity).

from cells lysed in the preparation process, contaminates empty and cell-containing droplets alike.[57]

The observation of ambient RNA resulting in unwanted molecule counts has led to the development of statistical methods for removing this source of noise, either by estimating and subtracting it[109] or incorporating it into a stochastic model.[110–112] Conceptually, Equation 33 reflects the latter approach: each droplet contains one or more cells, each with biological GF G and background, with a GF G_{bg} that depends on G. To accurately model the background counts, we need to propose and justify a specific functional form for G_{bg}. Thus, under the assumption that empty and cell-containing droplets are similarly susceptible to contamination, the former provides a reasonable estimate of ambient distributions in the latter.[109]

The simplest model holds G_{bg} to be equivalent to a "pseudo-bulk" experiment, with molecules randomly sampled from the lysed cell population. If each cell is equally likely to contribute to the pool of free RNA, and diffusion occurs by a simple independent arrival process, we find that the distribution of background should be Poisson, with the mean for each species proportional to its mean in the original cell population, as in, e.g., Fleming et al.[110] This functional form immediately induces a set of testable predictions: not only are the distributions Poisson, but they are independent Poisson, with no meaningful statistical structure remaining between transcripts of a single gene, as well as between different genes, as illustrated in Figure 2A.

To characterize the accuracy of these predictions, we inspected six datasets (Table S2) pseudoaligned with *kallisto | bustools*[113] and compared the data for barcodes passing *bustools* quality control with data for barcodes that were filtered out. As a shorthand, we call the former "non-empty" and the latter "empty" droplets, keeping in mind that this identification is approximate. We fully describe the analysis procedure in the section "empty droplets," illustrate the results for the human blood dataset pbmc_1k_v3, and display the results for all datasets in the supplemental information.

As shown in Figure 2B, data from non-empty droplets are substantially overdispersed relative to Poisson, whereas data from empty droplets are largely consistent with the Poisson identity mean-variance relationship. However, a small number of relatively high-expression genes are overdispersed. In addition, intra-gene (Figure 2C) and inter-gene (Figure 2D) correlations are typically nontrivial in non-empty droplets, but consistently near zero for empty droplets, supporting distributional independence of the background counts. Finally, the mean expression in empty droplets is highly correlated with mean expression in non-empty droplets, albeit lowered by approximately four orders of magnitude (Figure 2E), supporting the assumption that the original cells are lysed in a uniform fashion.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

To characterize the deviations from the pseudo-bulk model, we identified the genes that demonstrated overdispersion in empty droplets (Table S3). A considerable fraction of these genes were associated with mitochondria or blood cells. For example, of the 21 annotated genes overdispersed in the empty droplets of the mouse neuron dataset neuron_1k_v3, nine were mitochondrial (*mt-Nd1*, *mt-Nd2*, *mt-Co1*, *mt-Co2*, *mt-Atp6*, *mt-Co3*, *mt-Nd3*, *mt-Nd4*, and *mt-Cytb*), three coded for hemoglobin subunits (*Hba-a1*, *Hba-a2*, and *Hbb-bs*), and two coded for blood cell-specific proteins (*Bsg*, *Vwf*).[114,115] On the other hand, of the 10 annotated genes overdispersed in the empty droplets of the desai_dmso dataset generated from cultured mouse embryonic stem cells,[116] six (*mt-Nd1*, *mt-Co2*, *mt-Atp6*, *mt-Co3*, *mt-Nd4*, and *mt-Cytb*) were mitochondrial, and none were blood cell specific[114] (Table S4).

Since overdispersion implies that contamination involves non-independent encapsulation of these molecules, the results suggest that the cell-free debris contains, among other structures, entire mitochondria, or erythrocytes, when they are present in the source tissue. These membrane-bound structures may diffuse into droplets, then lyse and release all of their contents at once. In other words, empty droplets do not merely have disproportionally high mitochondrial content, as has been noted previously[110,117,118]; they have *nontrivially distributed* mitochondrial content, which can hint at the mechanism of its incorporation, and improve interpretation where simple thresholds may be misleading.[118] We hypothesize that cases where the model fails can be leveraged to discover more complicated forms of contamination, such as molecular aggregates.[112]

In addition, we examined the total UMI counts in empty droplets, which should be Poisson (Fano = 1) if each individual gene's distribution is Poisson. For the human blood dataset demonstrated in Figure 2, the empty droplets had fairly significant overdispersion (Fano $\approx$43), which decreased, but did not disappear (Fano $\approx$7.6), once the 53 significantly overdispersed genes were excluded. This result suggests that although the pseudo-bulk model is approximately valid, some residual variance, possibly due to variability in per-droplet capture rates, is present and needs to be modeled to fully describe the stochasticity in single-cell datasets.

Noise-corrupted candidate models of transcriptional variation

A considerable fraction of the variability in single-cell datasets arises from cell-to-cell and time-dependent variation in the transcription rates. These sources of variation control distribution shapes. By carefully analyzing candidate models, we can characterize the prospects for model selection: for example, if different models produce nearly identical distributions, selection is impossible and the choice of model is somewhat arbitrary. More interestingly, such analysis can guide the design of experiments: models may be indistinguishable based on some kinds of data, but not others.[20] This perspective has guided the interest in characterizing noise behaviors[74,119]: distributions provide strictly more information than averages and allow us to distinguish between regulatory mechanisms. Similarly, multivariate distributions provide more information than marginal distributions. Obtaining different data (multiple molecular modalities) is qualitatively more useful than obtaining more data (a larger number of cells) or better data (observations less corrupted by noise).

We illustrate this key point using the simple model system depicted in Figure 3A, which features intrinsic, extrinsic, and technical noise. The continuous stochastic process denoted by K drives the rate of transcription of nascent RNA. We consider three different possibilities for K: the gamma Ornstein-Uhlenbeck process, which models DNA winding and relaxation; the Cox-Ingersoll-Ross process, which models the fluctuations in a high-copy number activator[20]; and the telegraph process, which models variation due to random exposure of the locus to transcriptional initiation.[76,97,100] All three transcription rate models are described by three parameters.[20,100] After a Markovian delay, nascent RNA are converted to mature RNA; after another Markovian delay, the mature RNA is degraded. When the system reaches steady state, it is sequenced; each biological molecule has a probability p of being observed in the final dataset. We seek to use imperfect count data to fit parameters and distinguish models. We fully describe the procedures in the section "noise-corrupted candidate models of transcriptional variation."

Even if we have perfect information about the true averages of the transcriptional strength and the molecular species, the systems can exhibit a wide variety of distribution shapes and statistical behaviors. This variety can be summarized by a two-dimensional parameter space, which was introduced in Figure 2 of Gorin and Vastola et al.[20] The "timescale separation" governs the relative timescales of the transcriptional and molecular processes; if it is high, the transcriptional process is faster than RNA turnover. The "noise intensity" governs the variability in the transcriptional process: if it is high, the process exhibits substantial variability that translates to overdispersion in the RNA distributions. The bottom edge of this parameter space produces Poisson distributions of RNA, the top left corner produces Poisson mixtures of the law of K, and the top right corner yields bursty dynamics that do not typically have simple analytical solutions.[20]

Although these regimes reflect very different transcriptional kinetics, they can produce indistinguishable distributions. The first panel of Figure 3B demonstrates the likelihood landscape of a dataset generated from the gamma Ornstein-Uhlenbeck (Γ-OU) transcriptional model, evaluated using the nascent marginal and $p = 1$. The mixture-like true parameters are indicated by a red point, and the top decile of likelihoods is indicated by hatching. The Γ-OU model's transcription rate has a gamma stationary distribution, which produces approximately Poisson-gamma, or negative binomial, RNA marginals in this regime. However, the bursty regime, indicated by a blue point, also yields a negative-binomial-like marginal,[20] preventing us from identifying the kinetics.

On the other hand, if we evaluate likelihoods using the entire two-species dataset, we obtain the landscape in the second panel of Figure 3B: the symmetry is broken, and the parameters can be localized to the mixture-like regime. The source of this improved performance is evident from examining the distributions, shown in the third and fourth panels of Figure 3B. The nascent marginals are essentially identical; no amount of purely nascent count data can distinguish between them. However, the bivariate distributions show subtle differences, such as higher nascent/mature correlations in the true regime, which can be

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

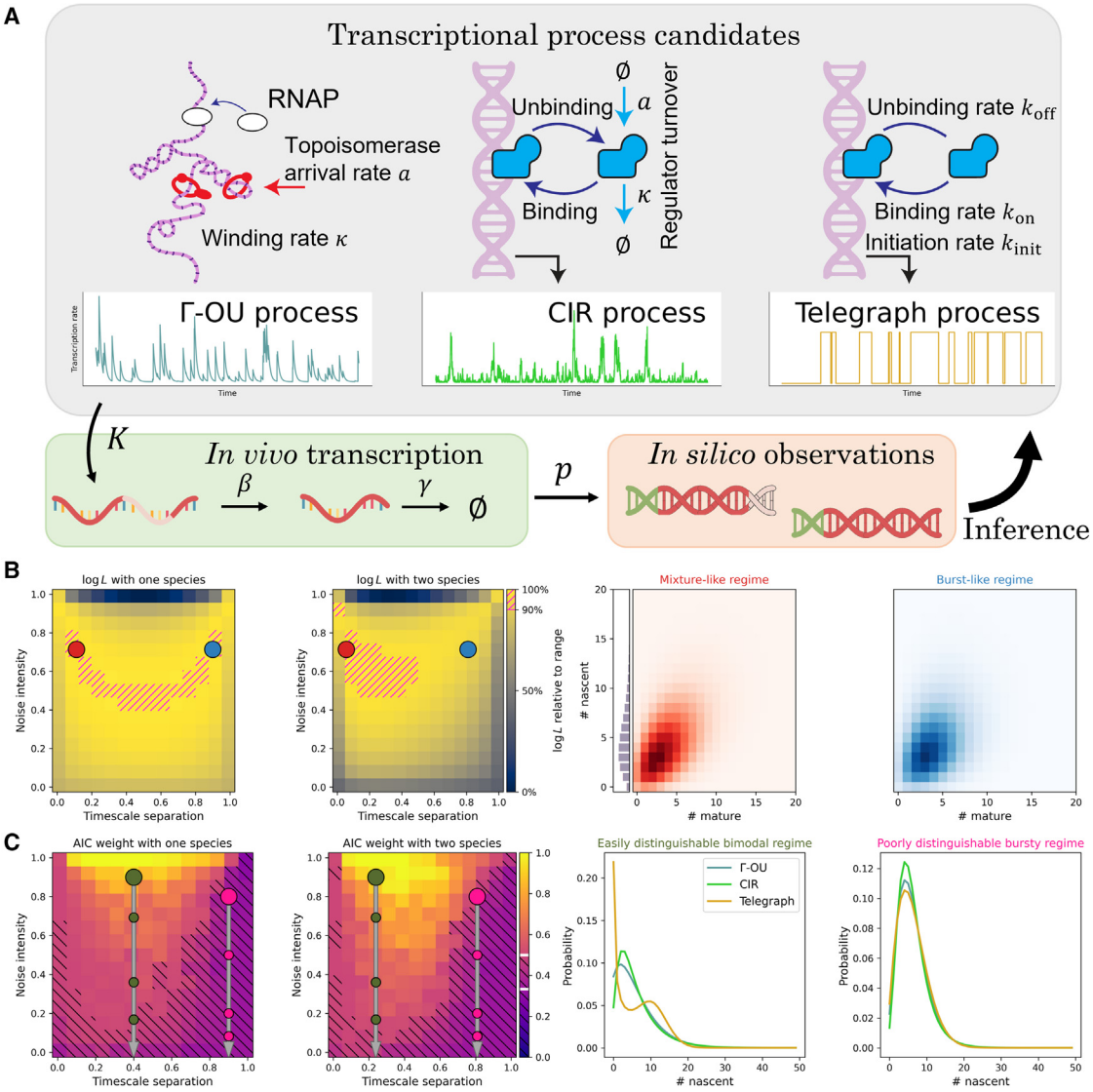


Figure 3. The stochastic analysis of biological and technical phenomena facilitates the identification and inference of transcriptional models

(A) A minimal model that accounts for intrinsic (single-molecule), extrinsic (cell-to-cell), and technical (experimental) variability: one of three time-varying transcriptional processes K generates molecules, which are spliced with rate β, degraded with rate γ, and observed with probability p. Given a set of observations, we can use statistics to narrow down the range of consistent models.

(B) Given a particular model, parameter regimes indistinguishable using a single modality become distinguishable with two. The mixture-like and burst-like regimes both produce negative binomial marginal distributions, but have different correlation structures (left: data likelihoods over the parameter space, computed from 200 simulated cells; Γ-OU ground truth; red point: true parameter set in the mixture-like regime; color: log-likelihood of data, yellow is higher, 90th percentile marked with magenta hatching; blue: an illustrative parameter set in a burst-like parameter regime with a similar nascent marginal but drastically different joint structure. Right: nascent marginal and joint distributions at the points indicated on the left. Nascent distributions nearly overlap).

(C) Given a location in parameter space, models are easier to distinguish using multiple modalities. However, the performance varies widely based on the location in parameter space and the specific candidate models: for example, the telegraph model has a well-distinguishable bimodal limit when the process autocorrelation is slower than RNA dynamics. In addition, all else held equal, dropout noise effectively decreases the noise intensity, lowering identifiability (left: Γ-OU Akaike weights under Γ-OU ground truth, average of n = 50 replicates using 200 simulated cells; color: Akaike weight of correct model, yellow is higher, regions with weight <0.5 marked with black hatching; large circles: illustrative parameter sets; smaller circles: distributions obtained by applying p = 50%, 75%, and 85% dropout to illustrative parameter sets while keeping the averages constant. Right: the three candidate models' nascent marginal distributions at the large points indicated on the left).

used for inference. This approach is analogous to Figure 4B of Gorin et al.,[21] where bivariate data are used to disambiguate differences that would otherwise be indistinguishable due to the degeneracies of steady-state distributions.

In addition, the timescale separation and noise intensity determine the model distinguishability. To quantify this, we use the Akaike weight w_{ϖ}, which transforms log-likelihood differences into model probabilities.[120] For example, if the Akaike weight is

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

near 1/3, the models are indistinguishable; if the correct model's weight is near 1, we can confidently identify the model from the data. The first panel of Figure 3C demonstrates the average Akaike weight landscape of datasets generated from the Γ-OU model, computed using the nascent distribution at the same coordinate. We indicate the region $w_{\varpi} < 1/2$ by hatching. As the Akaike weight may be interpreted as a posterior model probability,[120] this somewhat arbitrary threshold gives even odds for choosing the correct model, on average.

The intermediate regime, indicated by a large olive green point, tends to yield fairly high Akaike weights, consistent with the two-model case explored in Figure 3A of Gorin and Vastola et al.[20] On the other hand, the burst-like regime, indicated by a large pink point, provides considerably less ability to distinguish the models. As expected, the situation improves somewhat when using bivariate data (second panel of Figure 3C): the Akaike weights increase throughout the parameter space, and the bursty regime data move closer to even odds for model selection.

To illustrate the source of the identifiability challenges, we plot the nascent marginals of the models at the two points. In the intermediate regime, the Γ-OU and CIR models yield moderately different distributions, whereas the telegraph model is immediately distinguishable by its bimodality (third panel of Figure 3C). In contrast, in the bursty regime, the distributions are all unimodal and less identifiable (fourth panel of Figure 3C); the Γ-OU and telegraph marginals are particularly similar, as they converge to the same negative binomial limit.[20]

Interestingly, this formulation fully characterizes the effect of certain forms of technical noise. If the transcriptional and observed molecular averages are fixed, but the experiment fails to capture some molecules, the distributions are identical to those obtained by deflating the transcriptional noise intensity. In other words, although technical noise affects the molecules, its theoretical effects are indistinguishable from decreasing the variability of the transcriptional process. As the noise levels increase, the RNA distributions are pushed toward the indistinguishable Poisson limit at the bottom edge of the reduced parameter space. We quantify how rapidly the information degrades by plotting smaller circles on the first and second panels of Figure 3C to indicate the effect of 50%, 75%, and 85% dropout, in that order from top to bottom.

Distributions obtained from a transient process

Due to the interest in understanding developmental processes, the characterization of transient process dynamics is a key problem in single-cell analyses. The use of mechanistic models with multimodal data, which we emphasize here, was originally pioneered in the context of the RNA velocity framework, which attempts to exploit the causal relationship between nascent and mature RNA to fit transient processes.[86] However, the implementations proposed so far use relatively simple noise behaviors,[59,86,121] which do not recapitulate the bursty transcription observed in living cells. As discussed in our recent analysis of RNA velocity methods,[19] this leads us to hold some reservations about the robustness and appropriate interpretation of results obtained by this class of methods.

The inference of transient dynamics from snapshot data is a formidable problem due to a combination of theoretical and practical factors. Most fundamentally, it is not precisely clear what a snapshot is: how does a single measurement simultaneously capture the early and late states in a differentiation process? To develop an explanatory model, we take inspiration from the existing work on cyclostationary processes,[122,123] cell cycle ensemble measurement modeling,[124–126] Markov chain occupation measure theory,[127–129] and chemical reactor engineering.[105,106] In the typical stochastic modeling context, we fit count data using stationary distributions $P(\mathbf{x})$, obtained as the limit $\lim_{t\to\infty} P(\mathbf{x}, t)$ of a transient distribution. By the ergodic theorem,[130–132] this distribution, when it exists, coincides with the occupation measure $\lim_{T\to\infty} \frac{1}{T}\int_0^T P(\mathbf{x}, t)dt$, i.e., observations drawn from a single trajectory over a sufficiently long time horizon, rather than from multiple trajectories at once. Conveniently, the ergodic limit has time symmetry with respect to measurement: the distribution does not depend on the timing of the experiment. In the transient case, we cannot take these limits. However, we can retain time symmetry by proposing that the experiment samples cells at almost surely finite times t since the beginning of the process. Therefore, we conceptualize data as coming from a set of cells indexed by c, such that each cell's time t_c is sampled from $f(t)$, and counts are drawn from some distribution $P(\mathbf{x}, t_c)$, which is not typically available in closed form. This formulation yields Equation 32, which requires specifying the distribution f.

We illustrate some of the challenges and implications using the model system shown at the bottom of Figure 4A. The underlying transient structure involves transitions through three cell types, each characterized by a particular transcriptional burst size. The transient transcription process produces nascent and mature RNA trajectories for each cell; however, we only obtain a single data point per trajectory. Even if we have perfect information about the cell times, it is far from clear that we can accurately reconstruct the transcriptional dynamics from snapshot data (center of Figure 4A).

In addition, we wish to know whether we can identify the mechanism of the snapshot collection. We can imagine cells entering and exiting the observed tissue in multiple ways, which correspond to different choices of $f(t)$. Some natural choices are uniform, which implies that the cells stay in the tissue for a deterministic time[86]; decreasing over time, so cells can exit immediately; or uniform, then decreasing, so cells must stay in the tissue for some duration but are free to leave afterward. These choices can be modeled by Dirac, exponential, and Pareto residence distributions. In the parlance of chemical reactor engineering, these configurations are known as the plug flow reactor, the continuously stirred tank reactor, and the laminar flow reactor, respectively. Their $f(t)$, which are the reactor internal-age distributions, are well-known in the chemical engineering literature[105,106] and shown at the top of Figure 4A. It is not *a priori* obvious that the configurations are mutually distinguishable from count data. If they are not, the choice of $f(t)$ is immaterial for inference.

We generated snapshot data from the Dirac model and fit it under all three models. To efficiently evaluate snapshot distributions, we designed an algorithm that essentially "recycles" t_c for trapezoidal quadrature. The method is fully described in the section "distributions obtained from a transient process." As shown in Figure 4B, despite only having access to a single observation per time point, all models yield results visually close to the

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

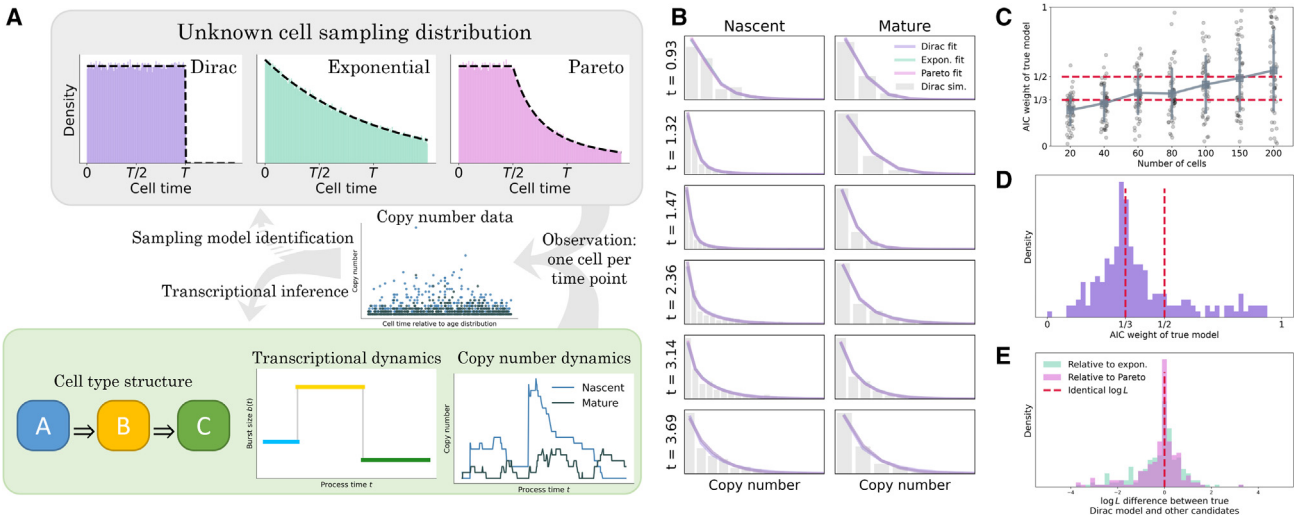


Figure 4. Given ordered and labeled snapshot data obtained from a transient differentiation process, we can typically fit the copy number data, but identifying the mechanism of the snapshot is more challenging

(A) A minimal model that accounts for the observation of transient differentiation processes in scRNA-seq: cells enter a "reactor" and receive a signal to begin transitioning from cell type A through B and to C. The change in cell type is accompanied by a step change in the burst size, which leads to variation in the nascent and mature RNA copy numbers over time. Given information about the cell type abundances and the cells' time along the process, we may fit a dynamic process to snapshot data and attempt to identify the underlying reactor type, which determines the probability of observing a cell at a particular time since the beginning of the process.

(B) In spite of the considerable differences between the reactor architectures, they produce nearly identical molecular count marginals (histogram: data simulated from the Dirac model, 200 cells; colored lines: analytical distributions at the maximum likelihood transcriptional parameter fits for each of the three reactor models. Analytical distributions nearly overlap).

(C) The true reactor model may be identified from molecule count data, but statistical performance is typically poor (points: Akaike weight values for n = 50 independent rounds of simulation and inference under a single set of parameters; blue markers and vertical lines: mean and standard deviation at each number of cells; blue line connects markers to summarize the trends; red lines: the Akaike weight values 1/3, which contains no information for model selection, and 1/2, which gives even odds for the correct model; two-species data generated from the Dirac model; uniform horizontal jitter added).

(D) The reactor models are poorly identifiable across a range of parameters, and rarely produce Akaike weights above 1/2 (histogram: Akaike weight values for n = 200 independent rounds of parameter generation, simulation, and inference under the true Dirac model; red line: the Akaike weight values 1/3 and 1/2; two-species data for 200 cells generated from the Dirac model; parameters were restricted to the low-expression regime $\mu + 4\sigma \leq 25$ for both species).

(E) The challenges in reactor identification arise because all three models produce similar likelihoods (histograms: likelihood differences between candidate models and the true Dirac model for n = 200 independent rounds of parameter generation, simulation, and inference; red line: no likelihood difference; two-species data for 200 cells generated from the Dirac model; parameters were restricted to the low-expression regime $\mu + 4\sigma \leq 25$ for both species).

true marginals. However, despite these superficial similarities, quantitative model identification is possible: for the simulated dataset shown, the true Dirac model achieves an Akaike weight of $w_\varpi \approx 79\%$, whereas the exponential and Pareto both achieve $\approx 10\%$. Decreasing the dataset size substantially degrades the identifiability (Figure 4C). Even at higher sizes, spread is considerable; for example, a 150-cell dataset gives approximately even odds $(w_\varpi > 1/2)$ on average, but individual realizations vary from confidently correct $(w_\varpi \approx 1)$ to confidently wrong $(w_\varpi \approx 0)$.

To understand the robustness of model identifiability, we generated 200 synthetic datasets at random parameter values, constrained to have fairly low expression. We observed poor identifiability, with even or better odds for the correct model in only 20% of the cases (Figure 4D). This performance appears to be attributable to quantitative similarities between all three models' likelihoods. As shown in Figure 4E, given data of this quality, we cannot even narrow the scope down to two models, as neither of the candidate models performs conspicuously worse than the true Dirac configuration. Therefore, it is possible to fit snapshot data approximately equally well using a variety of models; candidates for $f(t)$ are identifiable in principle but challenging to distinguish from any particular dataset. This simulated analysis implies that the details of the reactor configuration may not matter much, providing a basis for omitting this model identification problem for real data.

Variability in library construction

To properly interpret single-cell data, we need to exhibit caution regarding the technical noise behaviors and consider multiple possible candidate models. However, before fitting distributions, we must fully characterize the models and understand which of their parameters are actually identifiable with the data at hand. For example, the two-species models explored in the section "noise-corrupted candidate models of transcriptional variation" produce distributional forms that are closed under the assumption $p_N = p_M = p$, i.e., the magnitude of the observation probability p is impossible to identify from count data alone. Interestingly, when $p_N \neq p_M$ (that is, when nascent and mature RNA may have different observation probabilities), what we can learn about technical noise heavily depends on the form of the biological noise. For example, under slow transcriptional variation (as in the mixture and Poisson

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

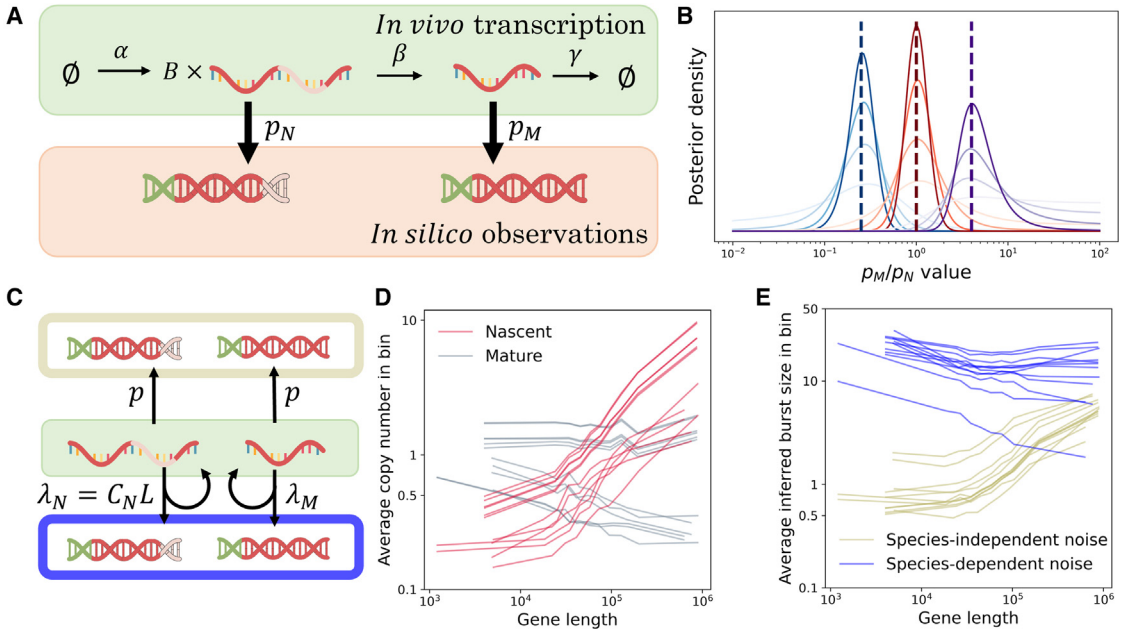


Figure 5. Technical noise models may be identified from count data, either by direct application of statistics or by imposing informal priors about the biological variability

(A) A minimal model that accounts for non-homogeneous noise: transcriptional events occur with frequency α, generating geometrically distributed bursts B with mean size b; the molecules are spliced with rate β and degraded with rate γ. Nascent molecules are observed with probability p_N and mature molecules are observed with probability p_M.

(B) Given information about the nascent distribution and the mature mean, it is possible to use joint distributions to obtain information about the ratio of observation probabilities (curves: average posterior likelihoods, computed from 200 independent synthetic datasets; color: true value of p_M/p_N, blue: $1/4$, red: 1, purple: 4; dashed lines: location of each true value; color intensity: from lightest to darkest, synthetic datasets with 20, 50, 100, and 200 cells).

(C) Two models considered in Gorin et al.[21]: the species-independent bias model for length dependence in averages, which proposes that nascent and mature RNA are sampled with equal probabilities, and the species-dependent bias model, which proposes that the nascent RNA sampling rate scales with length (top, gold: kinetics of species-independent model; bottom, blue: kinetics of species-dependent model; center, green: the source RNA molecules used to template cDNA).

(D) A variety of single-cell datasets produce consistent and counterintuitive length-dependent trends in nascent RNA observations (lines: average per-species gene expression, binned by gene length; red: nascent RNA observations; gray: mature RNA statistics; data for 2,500 genes analyzed in Gorin et al.[21]).

(E) Fits to the species-independent model show a strong positive gene length dependence for inferred burst sizes, whereas fits to the species-dependent model show a modest negative gene length dependence, which is more coherent with orthogonal data (lines: average per-gene burst size inferred by *Monod*,[133] binned by gene length; gold: results for species-independent model; blue: results for species-dependent model; data for genes analyzed in Gorin et al.[21] after goodness-of-fit).

limits of the models explored in the section "noise-corrupted candidate models of transcriptional variation"), the RNA distributions contain no identifiable information whatsoever about the technical noise, regardless of the amount of data. On the other hand, if transcription is bursty, the distributions depend on the ratio of p_N and p_M, but not their absolute values (section "variability in library construction"). This theoretical result calls for further investigation: how much information can we obtain in practice, given finite data?

To understand the prospects for distinguishing parameters, we consider the simple model system shown in Figure 5A, which involves bursty transcription with average burst size b, splicing, degradation, and molecular capture with species-specific probabilities. To characterize how much information about p_M/p_N we can identify from count data, we simulated 200 datasets at the ratio values $1/4, 1$, and 4 and calculated their likelihoods over $(10^{-2}, 10^{2})$. We repeated this analysis using synthetic datasets with 20, 50, 100, and 200 cells and plotted the average of the posterior distributions for each condition. As shown in Figure 5B, color-coded by the ground truth p_M/p_N and intensity-coded by the number of cells, the posteriors are, on average, consistent with the true value. However, even with perfect information about the averages and the nascent RNA distribution, the uncertainty is considerable; at larger dataset sizes, we can typically localize the ratio to an order of magnitude, but not much further.

Given the statistical challenges illustrated by simulations, we speculate that it may be more fruitful to use prior information about biology and physical intuition about sequencing to construct technical noise models. For example, in a recent paper,[21] we fit models that represent two competing hypotheses (Figure 5C). The first has identical, gene-specific observation probabilities p for the nascent and mature species. In this model, the inferred burst size is bp, as these two parameters are not mutually identifiable. The second has a gene-length-dependent technical noise term for the nascent species, which coarsely represents a higher rate of priming for long molecules with abundant intronic poly(A) tracts, and a shared genome-wide term for the

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

mature species, which represents priming at the poly(A) tail. In this model, the inferred burst size is b.

These models attempt to explain the trend summarized in Figure 5D: across a wide range of datasets, nascent RNA averages exhibit a pronounced length dependence[134] not evident in mature RNA.[135] The first model explains the trend by a species-independent bias, as b and p control nascent as well as mature RNA levels. Conversely, the second model explains it by a species-dependent bias. Both models produce fair fits to the data (as demonstrated, e.g., by the low rate of rejection by goodness-of-fit in sections S7.4 and S7.5.2 of Gorin et al.[21]).

However, the trends in the resulting inferred parameters are strikingly different: the species-independent bias model predicts that longer genes have higher bp. Ascribing this trend to the b term—longer genes have higher burst sizes—contradicts burst size trends from fluorescence microscopy.[136] Ascribing it to the p term—longer genes have higher sampling probabilities—is physically unrealistic because mature RNA molecules are depleted of the internal poly(A) tracts necessary for priming.[137] On the other hand, the species-dependent model predicts a modest negative relationship between length and burst size, which is more coherent with orthogonal data.

This technical noise model is a relatively simplistic low-order approximation, since all genes have the same mature molecule capture rate λ_M and length scaling C_N. Nevertheless, it foregrounds a key modeling principle of the investigation: in the absence of prior information, biological parameters need to be fit on a gene-by-gene basis, but technical noise should be constructed using a common genome-wide model that varies in a mechanistic rather than arbitrary way. In sum, mathematics enables us to define and fit systems, but to understand whether the fits are sensible, we need to contextualize and compare them with previous results and physical intuition.

DISCUSSION

The results we have derived provide a blueprint for the holistic modeling of single-cell biology and sequencing experiments. First, we have outlined a generic mathematical framework for treating stochasticity in living cells. By exploiting the GF representation, we reduce discrete, continuous, and mixed reactions to operators in a system of differential equations. These ODEs can be straightforwardly solved via numerical integration to compute model properties, including likelihoods. This approach recapitulates and subsumes a wide range of previous results.[16,17,20,21,75,76,85,100,138,139]

By treating the discrete and continuous degrees of freedom on equal footing, our approach makes certain otherwise challenging problems straightforward to solve, as illustrated in the section "special theoretical cases." By making simplifying assumptions—chiefly, the assumption of independent and identically distributed sampling—we reduce the modeling of technical variation to the composition of GFs. Our framework may be used in its current form or as a substrate for developing more sophisticated models of transcriptional regulation and sequencing that subsume it in turn. This process simply involves instantiating hypotheses, converting them into probabilistic models, and constructing model solutions using a procedure analogous to the one presented in Figure 1C.

We believe that this framework comprises a productive vision for the interpretation of large datasets, but many technological and mathematical challenges remain. For example, the library construction biases are dependent on molecule-specific factors that we do not yet fully understand because their effect is heavily convolved with biological variability. In Figure 5, we considered two extreme cases, where the noise strength/length scaling is either unconstrained or forced to be identical for all genes. We anticipate that careful investigation of technical biases will be necessary to construct models that constrain the technical biases based on RNA chemistry while allowing for gene-to-gene and droplet-to-droplet variabilities.

In the section "library construction and sequencing noise" and supplemental information, we discuss the challenges associated with modeling ambiguous species, motivated by the limitations of short-read sequencing for distinguishing between spliced and unspliced forms of the same RNA gene product.[140] It is worth noting that even the spliced/unspliced binary is a convenient simplification primarily adopted because of data availability[86,113]; we stress that a truly comprehensive treatment requires defining intermediate states,[19] their relationships, and their mutually indistinguishable classes. These computational foundations do not yet exist, although we have attempted a partial solution in recent work[16] and outlined some promising directions in the supplemental information. Therefore, despite our immediate interest in bivariate RNA distributions, our framework is designed to generalize to other modalities as they become practical to quantify. In addition, although we focus on Markovian systems here, non-Markovian processing can be represented by appropriately defining $\mathbf{U}$,[141] which suggests avenues for the treatment of systems with molecular memory.[142,143]

The full GF solutions we have outlined here are typically not computable directly. By construction, the GF needs to be evaluated on a grid; Fourier inversion produces a grid of microstate probabilities, which needs to be quite large to avoid artifacts.[139] If the grid dimension is $\mathfrak{s}_i$ for each discrete species i, the overall state space size is $\mathfrak{s} = N\prod_i \mathfrak{s}_i$. Even in the simplest case, where we only quantify and fit discrete counts, evaluating the probability mass function requires storing and inverting an n-dimensional array, which usually has a size $\mathfrak{s}$ far too large to be practical (e.g., Figure S5B of our prior work on bursty models[16]).

When applicable, the GF approach has numerical advantages over the stochastic simulation algorithm (SSA),[144–146] which approximates distributions by the empirical distributions of trajectories, and finite state projection (FSP),[78] which directly integrates a version of the master equation confined to a finite $\mathfrak{s}$. Specifically, if we only care about a particular species i, we can evaluate its marginal using a grid of size $N\mathfrak{s}_i$ with $\mathfrak{s}\log\mathfrak{s}$ time complexity. In the worst-case scenario, FSP requires a grid of size $\mathfrak{s}$ with $\mathfrak{s}^3$ time complexity, as evaluating a particular marginal requires explicitly evaluating the probabilities for the entire grid, then marginalizing. Similarly, SSA requires explicitly simulating the entire system to obtain the marginals and has the drawback of the usual inverse square-root Monte Carlo convergence.[147,148] In addition, FSP is not compatible with the

GF manipulations used to represent technical noise, SSA is relatively challenging to adapt to time-dependent rates,[149] and neither FSP nor SSA is readily compatible with continuous stochastic processes (although exact[20] and approximate[20,150,151] hybrid schema can be constructed with some work). In the future, the "curse of dimensionality"—the reliance on grid evaluation—may be possible to bypass altogether by training neural networks to predict probability distributions, but this approach is as of yet in its nascence[152–155] and will require considerable further development to apply to general systems.

Nevertheless, SSA and FSP are substantially more general than the approach we outline here. The simulation- and matrix-based methods only require a list of reactions, whereas the GF methods also require those reactions to produce readily solvable PDEs. We have omitted phenomena that would be trivial to treat using FSP and SSA, such as regulation involving feedback. (In principle, one can always construct "synthetic likelihoods" for inference by fitting a function approximator to the results of stochastic simulations, even for highly nonlinear and chaotic systems.[156–158]) To our knowledge, these phenomena, which are mathematically analogous to adding multi-molecular interaction terms, cannot be directly treated with the method of characteristics. Instead, mathematically precise treatment of them requires perturbative methods[77] or fairly complicated special function manipulations,[101–104] which do not easily generalize. We illustrate the challenges in the supplemental information, using the example of downstream species catalyzing gene state transitions.

On the other hand, there are a number of ways to treat systems involving feedback approximately. Approaches like the linear-mapping approximation[159] permit the derivation of approximate but accurate GFs for such systems, which can then be used in standard inference pipelines. Alternatively, using only the results presented here, the net effect of feedback can be captured in the time-dependence of certain parameters (e.g., burst sizes) if dynamics are sufficiently chaotic, or if the time scale of feedback is slow compared with other system time scales.

We have, until now, stressed applications to "snapshot" single-cell data from dissociated tissues; however, our framework may be extended to spatial single-cell data; for instance, we can define transcriptional parameters that depend on the cell's coordinates in the tissue. In this case, the typical systems biology goals translate to fitting a time- and space-dependent function that governs these parameters. However, the GF formulation relies on the assumption of cells being stochastically independent; it is far from clear that this should hold for densely sampled spatial data, and more sophisticated alternatives, such as agent-based models, may be needed.[160,161]

Despite these challenges, the framework is already quantitatively useful. To fully "explain" a dataset, we need to fit gene-specific transcriptional mechanisms, genome-wide technical noise and co-expression parameters, and cell type structure while controlling for potential misspecification. However, at this time, it may be more fruitful to focus on narrower questions, using assumptions, orthogonal data, or simulated benchmarking to justify omitting some parts of the problem.[19] We have applied this "bottom-up" approach to single-cell data, considering, in turn, the estimation of transcriptional kinetics and technical noise,[21,133] the identification of transcriptional models,[20] the analysis of co-regulation patterns,[16] and the determination of nuclear transport kinetics.[141] Conversely, it may be valuable to apply a "top-down" approach, augmenting an existing method with biophysically meaningful noise, as we have proposed in the context of transient processes[19] and neural network dimensionality reduction.[71]

We anticipate that making meaningful progress on the stochastic modeling project championed by Wilkinson will require extended "real contact"[162] between systems biology, genomics, and mathematics. The general framework we propose, which unifies a variety of previous work, represents one step toward this synthesis. The role of mathematics here is key; as Wilkinson noted, the stochastic systems biology of single cells cannot be "properly understood" without stochastic mathematical models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Master equation models of transcription
  - The full master equation
  - Generating function methods for biological stochasticity
  - Coupling multiple genes
  - Transient phenomena
  - Droplet encapsulation noise
  - Library construction and sequencing noise
  - Example systems

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.08.004.

ACKNOWLEDGMENTS

G.G. and L.P. were partially funded by NIH 5UM1HG012077-02 and NIH U19MH114830. J.J.V. was partially funded by NIH 1U19NS118246-01. The RNA, DNA, and cDNA illustrations were derived from the DNA Twemoji by Twitter, Inc., used under the CC-BY 4.0 license. The authors thank Dr. A. Sina Booeshaghi, Maria Carilli, Tara Chari, Taleen Dilanyan, Dr. Kristján Eldjárn Hjörleifsson, Meichen Fang, Catherine Felce, and Delaney Sullivan for fruitful discussions of co-regulation, contamination, transient behaviors, catalysis, fragmentation, genomic alignment, and a variety of other phenomena and processes. Part of this work was performed during G.G.'s Data Sciences Co-op with Celsius Therapeutics, Inc.

AUTHOR CONTRIBUTIONS

G.G. performed all computational experiments. G.G. and J.J.V. developed the theoretical framework. All authors conceptualized the work and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 29, 2023
Revised: August 16, 2023
Accepted: August 25, 2023
Published: September 25, 2023

REFERENCES

1. Wilkinson, D.J. (2018). Stochastic Modelling for Systems Biology (Chapman and Hall/CRC).
2. Waddington, C.H. (1957). The Strategy of the Genes (Routledge).
3. Huang, S. (2009). Reprogramming cell fates: reconciling rarity with robustness. BioEssays *31*, 546–560. https://doi.org/10.1002/bies.200800189.
4. Huang, S. (2012). The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? BioEssays *34*, 149–157. https://doi.org/10.1002/bies.201100031.
5. Rand, D.A., Raju, A., Sáez, M., Corson, F., and Siggia, E.D. (2021). Geometry of gene regulatory dynamics. Proc. Natl. Acad. Sci. USA *118*, e2109729118. https://doi.org/10.1073/pnas.2109729118.
6. Coomer, M.A., Ham, L., and Stumpf, M.P.H. (2022). Noise distorts the epigenetic landscape and shapes cell-fate decisions. Cell Syst. *13*, 83–102.e6. https://linkinghub.elsevier.com/retrieve/pii/S2405471221003392.
7. Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1663-x.
8. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982. http://www.nature.com/articles/nmeth.4402.
9. Zhou, J., and Troyanskaya, O.G. (2021). An analytical framework for interpretable and generalizable single-cell data analysis. Nat. Methods *18*, 1317–1321. https://www.nature.com/articles/s41592-021-01286-1.
10. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.
11. Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc. Natl. Acad. Sci. USA *102*, 4936–4942.
12. Érdi, P., and Lente, G. (2014). Stochastic Chemical Kinetics: Theory and (Mostly) Systems Biological Applications. Springer Complexity (Springer).
13. Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. *6*, 8687. http://www.nature.com/articles/ncomms9687.
14. Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640. http://www.nature.com/articles/nmeth.2930.
15. Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics *19*, 562–578. https://academic.oup.com/biostatistics/article/19/4/562/4599254.
16. Gorin, G., and Pachter, L. (2022). Modeling bursty transcription and splicing with the chemical master equation. Biophys. J. *121*, 1056–1069. https://www.cell.com/biophysj/fulltext/S0006-3495(22)00104-7.
17. Vastola, J.J. (2021). Solving the chemical master equation for monomolecular reaction systems and beyond: a Doi-Peliti path integral view. J. Math. Biol. *83*, 48. https://doi.org/10.1007/s00285-021-01670-7.
18. Vastola, J.J. (2021). In search of a coherent theoretical framework for stochastic gene regulation. Ph.D. thesis. Vanderbilt. https://ir.vanderbilt.edu/handle/1803/16646.
19. Gorin, G., Fang, M., Chari, T., and Pachter, L. (2022). RNA velocity unraveled. PLOS Comp. Biol. *18*, e1010492. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010492.
20. Gorin, G., Vastola, J.J., Fang, M., and Pachter, L. (2022). Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. Nat. Commun. *13*, 7620. https://www.nature.com/articles/s41467-022-34857-7.
21. Gorin, G., and Pachter, L. (2023). Length biases in single-cell RNA sequencing of pre-mRNA. Biophys. Rep. (N Y) *3*, 100097. https://linkinghub.elsevier.com/retrieve/pii/S2667074722000544.
22. Belliveau, N.M., Chure, G., Hueschen, C.L., Garcia, H.G., Kondev, J., Fisher, D.S., Theriot, J.A., and Phillips, R. (2021). Fundamental limits on the rate of bacterial growth and their influence on proteomic composition. Cell Syst. *12*, 924–944.e2. https://linkinghub.elsevier.com/retrieve/pii/S240547122100209X.
23. Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352. https://linkinghub.elsevier.com/retrieve/pii/S1097276515001707.
24. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.
25. Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA *99*, 12795–12800. http://www.pnas.org/cgi/doi/10.1073/pnas.162041399.
26. Hilfinger, A., and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc. Natl. Acad. Sci. USA *108*, 12167–12172. http://www.pnas.org/cgi/doi/10.1073/pnas.1018832108.
27. Fu, A.Q., and Pachter, L. (2016). Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. Stat. Appl. Genet. Mol. Biol. *15*, 447–471. https://www.degruyter.com/doi/10.1515/sagmb-2016-0002.
28. Hilfinger, A., Norman, T.M., and Paulsson, J. (2016). Exploiting natural fluctuations to identify kinetic mechanisms in sparsely characterized systems. Cell Syst. *2*, 251–259. https://linkinghub.elsevier.com/retrieve/pii/S2405471216301107.
29. Finkenstädt, B., Woodcock, D.J., Komorowski, M., Harper, C.V., Davis, J.R.E., White, M.R.H., and Rand, D.A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. Ann. Appl. Stat. *7*, 1960–1982. https://projecteuclid.org/euclid.aoas/1387823306.
30. Baudrimont, A., Jaquet, V., Wallerich, S., Voegeli, S., and Becskei, A. (2019). Contribution of RNA degradation to intrinsic and extrinsic noise in gene expression. Cell Rep. *26*, 3752–3761.e5. https://linkinghub.elsevier.com/retrieve/pii/S2211124719303080.
31. Hausser, J., Mayo, A., Keren, L., and Alon, U. (2019). Central dogma rates and the trade-off between precision and economy in gene expression. Nat. Commun. *10*, 68. https://www.nature.com/articles/s41467-018-07391-8.
32. Keizer, J. (1987). Statistical Thermodynamics of Nonequilibrium Processes (Springer).
33. Saint-Antoine, M.M., and Singh, A. (2020). Network inference in systems biology: recent developments, challenges, and applications. Curr. Opin. Biotechnol. *63*, 89–98. https://linkinghub.elsevier.com/retrieve/pii/S0958166919301399.
34. Xing, L., Guo, M., Liu, X., Wang, C., Wang, L., and Zhang, Y. (2017). An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. BMC Genomics *18*, 844. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4228-y.

35. Shmulevich, I., and Dougherty, E.R. (2010). Probabilistic Boolean Networks: the Modeling and Control of Gene Regulatory Networks (Society for Industrial and Applied Mathematics).

36. Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435. http://www.nature.com/articles/nature22794.

37. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R.D., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7*, S7. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S1-S7.

38. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, e12776. https://dx.plos.org/10.1371/journal.pone.0012776.

39. Silk, D., Kirk, P.D.W., Barnes, C.P., Toni, T., and Stumpf, M.P.H. (2014). Model selection in systems biology depends on experimental design. PLoS Comp. Biol. *10*, e1003650. https://dx.plos.org/10.1371/journal.pcbi.1003650.

40. Munsky, B., Li, G., Fox, Z.R., Shepherd, D.P., and Neuert, G. (2018). Distribution shapes govern the discovery of predictive models for gene regulation. Proc. Natl. Acad. Sci. USA *115*, 7533–7538.

41. Huynh-Thu, V.A., and Sanguinetti, G. (2015). Combining tree-based and dynamical systems for the inference of gene regulatory networks. Bioinformatics *31*, 1614–1622. https://academic.oup.com/bioinformatics/article/31/10/1614/176842.

42. Bansal, M., Della Gatta, G.D., and di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics *22*, 815–822. https://academic.oup.com/bioinformatics/article/22/7/815/202299.

43. Henriques, D., Rocha, M., Saez-Rodriguez, J., and Banga, J.R. (2015). Reverse engineering of logic-based differential equation models using a mixed-integer dynamic optimization approach. Bioinformatics *31*, 2999–3007. https://academic.oup.com/bioinformatics/article/31/18/2999/241026.

44. Stumpf, P.S., Smith, R.C.G., Lenz, M., Schuppert, A., Müller, F.J., Babtie, A., Chan, T.E., Stumpf, M.P.H., Please, C.P., Howison, S.D., et al. (2017). Stem cell differentiation as a non-markov stochastic process. Cell Syst. *5*, 268–282.e7. https://linkinghub.elsevier.com/retrieve/pii/S2405471217303423.

45. Cannoodt, R., Saelens, W., Deconinck, L., and Saeys, Y. (2021). Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat. Commun. *12*, 3942. http://www.nature.com/articles/s41467-021-24152-2.

46. Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., DREAM5 Consortium, Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804. http://www.nature.com/articles/nmeth.2016.

47. Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc. *13*, 599–604. http://www.nature.com/articles/nprot.2017.149.

48. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. http://www.nature.com/articles/ncomms14049.

49. Stumpf, M.P.H. (2021). Inferring better gene regulation networks from single-cell data. Curr. Opin. Syst. Biol. *27*, 100342. https://linkinghub.elsevier.com/retrieve/pii/S2452310021000275.

50. Wang, L., Zhang, Q., Qin, Q., Trasanidis, N., Vinyard, M., Chen, H., and Pinello, L. (2021). Current progress and potential opportunities to infer single-cell developmental trajectory and cell fate. Curr. Opin. Syst. Biol. *26*, 1–11. https://linkinghub.elsevier.com/retrieve/pii/S2452310021000093.

51. Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Mol. Syst. Biol. *14*, e8046.

52. Packer, J., and Trapnell, C. (2018). Single-cell multi-omics: an engine for new quantitative models of gene regulation. Trends Genet. *34*, 653–665. https://linkinghub.elsevier.com/retrieve/pii/S0168952518301082.

53. Stein-O'Brien, G.L., Ainsile, M.C., and Fertig, E.J. (2021). Forecasting cellular states: from descriptive to predictive biology via single-cell multiomics. Curr. Opin. Syst. Biol. *26*, 24–32. https://linkinghub.elsevier.com/retrieve/pii/S245231002100010X.

54. Gligorijević, V., and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. J. R. Soc. Interface *12*, 20150571. https://royalsocietypublishing.org/doi/10.1098/rsif.2015.0571.

55. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833.

56. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods *18*, 272–282. http://www.nature.com/articles/s41592-020-01050-x.

57. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. http://msb.embopress.org/lookup/doi/10.15252/msb.20188746.

58. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. http://www.nature.com/articles/s41592-018-0229-2.

59. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. http://www.nature.com/articles/s41587-020-0591-3.

60. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4772-0.

61. Huang, S. (2018). The tension between big data and theory in the "omics" era of biomedical research. Perspect. Biol. Med. *61*, 472–488. https://muse.jhu.edu/article/713156.

62. Jiang, R., Sun, T., Song, D., and Li, J.J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. *23*, 31. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02601-5.

63. Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. Nat. Biotechnol. *38*, 147–150. https://www.nature.com/articles/s41587-019-0379-5.

64. Andrews, T.S., and Hemberg, M. (2018). False signals induced by single-cell imputation. F1000Res *7*, 1740. https://f1000research.com/articles/7-1740/v2.

65. Booeshaghi, A.S., Hallgrímsdóttir, I.B., Gálvez-Merchán, A., and Pachter, L. (2022). Depth normalization for single-cell genomics count data. https://doi.org/10.1101/2022.05.06.490859.

66. Booeshaghi, A.S., and Pachter, L. (2021). Normalization of single-cell RNA-seq counts by log(x + 1) or log(1 + x). Bioinformatics *37*, 2223–2224. https://academic.oup.com/bioinformatics/article/37/15/2223/6155989.

67. Cooley, S.M., Hamilton, T., Ray, J.C.J., and Deeds, E.J. (2020). A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. Preprint. bioRxiv, 689851. https://www.biorxiv.org/content/10.1101/689851v4.

68. Chari, T., Banerjee, J., and Pachter, L. (2021). The specious art of single-cell genomics. https://doi.org/10.1101/2021.08.25.457696.

69. Zheng, S.C., Stein-O'Brien, G., Boukas, L., Goff, L.A., and Hansen, K.D. (2022). Pumping the brakes on RNA velocity – understanding and

interpreting RNA velocity estimates. https://doi.org/10.1101/2022.06.19.494717.

70. François, P. (2023). New wave theory. Development *150*, dev201647. https://journals.biologists.com/dev/article/150/4/dev201647/287679/New-wave-theory.
71. Carilli, M.T., Gorin, G., Choi, Y., Chari, T., and Pachter, L. (2023). Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data. https://doi.org/10.1101/2023.01.13.523995.
72. Fox, Z.R., and Munsky, B. (2019). The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments. PLoS Comp. Biol. *15*, e1006365. https://dx.plos.org/10.1371/journal.pcbi.1006365.
73. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol. *4*, e309. https://dx.plos.org/10.1371/journal.pbio.0040309.
74. Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. Science *336*, 183–187.
75. Shahrezaei, V., and Swain, P.S. (2008). Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. USA *105*, 17256–17261. http://www.pnas.org/cgi/doi/10.1073/pnas.0803850105.
76. Iyer-Biswas, S., Hayot, F., and Jayaprakash, C. (2009). Stochasticity of gene products from transcriptional pulsing. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *79*, 031911. https://link.aps.org/doi/10.1103/PhysRevE.79.031911.
77. Veerman, F., Marr, C., and Popović, N. (2018). Time-dependent propagators for stochastic models of gene expression: an analytical method. J. Math. Biol. *77*, 261–312. http://link.springer.com/10.1007/s00285-017-1196-4.
78. Munsky, B., and Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. J. Chem. Phys. *124*, 044104.
79. Xu, H., Skinner, S.O., Sokac, A.M., and Golding, I. (2016). Stochastic kinetics of nascent RNA. Phys. Rev. Lett. *117*, 128101. https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.117.128101.
80. Stinchcombe, A.R., Peskin, C.S., and Tranchina, D. (2012). Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *85*, 061919. https://link.aps.org/doi/10.1103/PhysRevE.85.061919.
81. Gardiner, C. (2004). Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences, Third Edition (Springer).
82. Gillespie, D.T. (1992). A rigorous derivation of the chemical master equation. Phys. A *188*, 404–425. https://linkinghub.elsevier.com/retrieve/pii/037843719290283V.
83. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1.
84. Vu, T.N., Wills, Q.F., Kalari, K.R., Niu, N., Wang, L., Rantalainen, M., and Pawitan, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics *32*, 2128–2135. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw202.
85. Jahnke, T., and Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. J. Math. Biol. *54*, 1–26. http://link.springer.com/10.1007/s00285-006-0034-x.
86. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. http://www.nature.com/articles/s41586-018-0414-6.
87. Kim, J.K., and Marioni, J.C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. *14*, R7. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-1-r7.
88. Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics *17*, 110. http://www.biomedcentral.com/1471-2105/17/110.
89. Vo, H.D., Fox, Z., Baetica, A., and Munsky, B. (2019). Bayesian estimation for stochastic gene expression using multifidelity models. J. Phys. Chem. B *123*, 2217–2234. https://pubs.acs.org/doi/10.1021/acs.jpcb.8b10946.
90. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095. http://www.nature.com/articles/nmeth.2645.
91. Bacher, R., Chu, L.-F., Argus, C., Bolin, J.M., Knight, P., Thomson, J.A., Stewart, R., and Kendziorski, C. (2021). Enhancing biological signals and detection rates in single-cell RNA-seq experiments with cDNA library equalization. Nucleic Acids Res. *50*, e12.
92. Thattai, M., and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. Proc. Natl. Acad. Sci. USA *98*, 8614–8619. http://www.pnas.org/cgi/doi/10.1073/pnas.151588598.
93. Gardiner, C.W., and Chaturvedi, S. (1977). The poisson representation. I. A new technique for chemical master equations. J. Stat. Phys. *17*, 429–468. http://link.springer.com/10.1007/BF01014349.
94. Doi, M. (1976). Stochastic theory of diffusion-controlled reaction. J. Phys. A: Math. Gen. *9*, 1479–1495. https://doi.org/10.1088/0305-4470/9/9/009.
95. Doi, M. (1976). Second quantization representation for classical many-particle system. J. Phys. A: Math. Gen. *9*, 1465–1477. https://doi.org/10.1088/0305-4470/9/9/008.
96. Peliti, L. (1985). Path integral approach to birth-death processes on a lattice. J. Phys. France. *46*, 1469–1483. https://doi.org/10.1051/jphys:019850046090146900.
97. Vastola, J.J., Gorin, G., Pachter, L., and Holmes, W.R. (2021). Analytic solution of chemical master equations involving gene switching. I: Representation theory and diagrammatic approach to exact solution. https://doi.org/10.48550/arXiv.2103.10992.
98. Ebert, M.R., and Reissig, M. (2018). Methods for Partial Differential Equations (Springer International Publishing).
99. Vastola, J.J., and Holmes, W.R. (2020). Chemical Langevin equation: A path-integral view of Gillespie's derivation. Phys. Rev. E *101*, 032417. https://link.aps.org/doi/10.1103/PhysRevE.101.032417.
100. Peccoud, J., and Ycart, B. (1995). Markovian modeling of gene product synthesis. Theor. Popul. Biol. *48*, 222–234.
101. Grima, R., Schmidt, D.R., and Newman, T.J. (2012). Steady-state fluctuations of a genetic feedback loop: an exact solution. J. Chem. Phys. *137*, 035104. http://aip.scitation.org/doi/10.1063/1.4736721.
102. Huang, L., Yuan, Z., Liu, P., and Zhou, T. (2014). Feedback-induced counterintuitive correlations of gene expression noise with bursting kinetics. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *90*, 052702. https://link.aps.org/doi/10.1103/PhysRevE.90.052702.
103. Kumar, N., Platini, T., and Kulkarni, R.V. (2014). Exact distributions for stochastic gene expression models with bursting and feedback. Phys. Rev. Lett. *113*, 268105. https://link.aps.org/doi/10.1103/PhysRevLett.113.268105.
104. Liu, P., Yuan, Z., Huang, L., and Zhou, T. (2015). Feedback-induced variations of distribution in a representative gene model. Int. J. Bifurcation Chaos *25*, 1540008. https://www.worldscientific.com/doi/abs/10.1142/S0218127415400088.
105. Fogler, H.S. (2006). Elements of chemical reaction engineering. In Prentice Hall PTR International Series in the Physical and Chemical Engineering Sciences, Fourth Edition (Prentice Hall PTR).
106. Roberts, G.W. (2008). Chemical Reactions and Chemical Reactors (John Wiley & Sons).

107. Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., and Shahrezaei, V. (2020). bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. Bioinformatics *36*, 1174–1181. https://academic.oup.com/bioinformatics/article/36/4/1174/5581401.

108. Tang, W., Jørgensen, A.C.S., Marguerat, S., Thomas, P., and Shahrezaei, V. (2023). Modelling capture efficiency of single cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. https://doi.org/10.1101/2023.03.06.531327.

109. Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience *9*, giaa151. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa151/6049831.

110. Fleming, S.J., Chaffin, M.D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J.C., Philippakis, A.A., Ellinor, P.T., and Babadi, M. (2019). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. https://doi.org/10.1101/791699.

111. Sheng, C., Lopes, R., Li, G., Schuierer, S., Waldt, A., Cuttat, R., Dimitrieva, S., Kauffmann, A., Durand, E., Galli, G.G., et al. (2022). Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics. https://doi.org/10.1101/2022.01.14.476312.

112. Yin, Y., Yajima, M., and Campbell, J.D. (2023). Characterization and decontamination of background noise in droplet-based single-cell protein expression data with DecontPro. https://doi.org/10.1101/2023.01.27.525964.

113. Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H.J., da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat. Biotechnol. *39*, 813–818. http://www.nature.com/articles/s41587-021-00870-2.

114. National Library of Medicine (2004). Gene. https://www.ncbi.nlm.nih.gov/gene/.

115. Lutsch, G., Vetter, R., Offhauss, U., Wieske, M., Gröne, H.J., Klemenz, R., Schimke, I., Stahl, J., and Benndorf, R. (1997). Abundance and location of the small heat shock proteins HSP25 and aB-crystallin in rat and human heart. Circulation *96*, 3466–3476. https://www.ahajournals.org/doi/abs/10.1161/01.CIR.96.10.3466.

116. Desai, R.V., Chen, X., Martin, B., Chaturvedi, S., Hwang, D.W., Li, W., Yu, C., Ding, S., Thomson, M., Singer, R.H., et al. (2021). A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. Science *373*, eabc6506. https://www.sciencemag.org/lookup/doi/10.1126/science.abc6506.

117. Heiser, C.N., Wang, V.M., Chen, B., Hughey, J.J., and Lau, K.S. (2021). Automated quality control and cell identification of droplet-based single-cell data using dropkick. Genome Res. *31*, 1742–1752. http://genome.cshlp.org/lookup/doi/10.1101/gr.271908.120.

118. Hippen, A.A., Falco, M.M., Weber, L.M., Erkan, E.P., Zhang, K., Doherty, J.A., Vähärautio, A., Greene, C.S., and Hicks, S.C. (2021). miQC: an adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. PLoS Comp. Biol. *17*, e1009290. https://dx.plos.org/10.1371/journal.pcbi.1009290.

119. Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. Mol. Syst. Biol. *5*, 318. https://www.embopress.org/doi/full/10.1038/msb.2009.75.

120. Burnham, K.P., and Anderson, D.R. (2002). Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach, Second Edition (Springer).

121. Qin, Q., Bingham, E., Manno, G.L., Langenau, D.M., and Pinello, L. (2022). Pyro-Velocity: probabilistic RNA Velocity inference from single-cell data. https://doi.org/10.1101/2022.09.12.507691v2.

122. Dattani, J. (2015). Exact solutions of master equations for the analysis of gene transcription models. PhD Dissertation (Imperial College Press).

123. Dattani, J., and Barahona, M. (2017). Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. J. R. Soc. Interface *14*, 20160833. https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0833.

124. Thomas, P. (2017). Making sense of snapshot data: ergodic principle for clonal cell populations. J. R. Soc. Interface *14*, 20170467. https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0467.

125. Perez-Carrasco, R., Beentjes, C., and Grima, R. (2020). Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. J. R. Soc. Interface *17*, 20200360. https://royalsocietypublishing.org/doi/10.1098/rsif.2020.0360.

126. Beentjes, C.H.L., Perez-Carrasco, R., and Grima, R. (2020). Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. Phys. Rev. E *101*, 032403. https://link.aps.org/doi/10.1103/PhysRevE.101.032403.

127. Pitman, J.W. (1977). Occupation measures for markov chains. Adv. Appl. Probab. *9*, 69–86. https://www.jstor.org/stable/1425817.

128. Yang, Y., Nurbekyan, L., Negrini, E., Martin, R., and Pasha, M. (2023). Optimal transport for parameter identification of chaotic dynamics via invariant measures. https://doi.org/10.48550/arXiv.2104.15138.

129. Kuntz, J., Thomas, P., Stan, G.-B., and Barahona, M. (2019). The exit time finite state projection scheme: bounding exit distributions and occupation measures of continuous-time markov chains. SIAM J. Sci. Comput. *41*, A748–A769. https://epubs.siam.org/doi/10.1137/18M1168261.

130. Birkhoff, G.D. (1931). Proof of the ergodic theorem. Proc. Natl. Acad. Sci. USA *17*, 656–660. https://www.pnas.org/doi/abs/10.1073/pnas.17.2.656.

131. Neumann, J.V. (1932). Proof of the quasi-ergodic hypothesis. Proc. Natl. Acad. Sci. *18*, 70–82. https://www.pnas.org/doi/abs/10.1073/pnas.18.1.70.

132. Moore, C.C. (2015). Ergodic theorem, ergodic theory, and statistical mechanics. Proc. Natl. Acad. Sci. USA *112*, 1907–1911. https://www.pnas.org/doi/abs/10.1073/pnas.1421798112.

133. Gorin, G., and Pachter, L. (2023). Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using *Monod*.

134. Gupta, A., Shamsi, F., Altemose, N., Dorlhiac, G.F., Cypess, A.M., White, A.P., Yosef, N., Patti, M.E., Tseng, Y.H., and Streets, A. (2022). Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. Genome Res. *32*, 242–257. http://genome.cshlp.org/lookup/doi/10.1101/gr.275509.121.

135. Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. F1000Res *6*, 595. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5428526/.

136. Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature *565*, 251–254. http://www.nature.com/articles/s41586-018-0836-1.

137. Patrick, R., Humphreys, D.T., Janbandhu, V., Oshlack, A., Ho, J.W.K., Harvey, R.P., and Lo, K.K. (2020). Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol. *21*, 167. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02071-7.

138. Singh, A., and Bokes, P. (2012). Consequences of mRNA transport on stochastic variability in protein levels. Biophys. J. *103*, 1087–1096. https://linkinghub.elsevier.com/retrieve/pii/S0006349512007904.

139. Bokes, P., King, J.R., Wood, A.T.A., and Loose, M. (2012). Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. J. Math. Biol. *64*, 829–854. http://link.springer.com/10.1007/s00285-011-0433-5.

140. Eldjárn Hjörleifsson, K., Sullivan, D.K., Holley, G., Melsted, P., and Pachter, L. (2022). Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. https://doi.org/10.1101/2022.12.02.518832.

141. Gorin, G., Yoshida, S., and Pachter, L. (2022). Transient and delay chemical master equations. https://doi.org/10.1101/2022.10.17.512599.

142. Fu, X., Patel, H.P., Coppola, S., Xu, L., Cao, Z., Lenstra, T.L., and Grima, R. (2022). Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions. eLife *11*, e82493. https://elifesciences.org/articles/82493.

143. Jiang, Q., Fu, X., Yan, S., Li, R., Du, W., Cao, Z., Qian, F., and Grima, R. (2021). Neural network aided approximation and parameter inference of non-Markovian models of gene expression. Nat. Commun. *12*, 2618. http://www.nature.com/articles/s41467-021-22919-1.

144. Gillespie, D.T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Phys. *22*, 403–434. https://linkinghub.elsevier.com/retrieve/pii/0021999176900413.

145. Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. *81*, 2340–2361. https://pubs.acs.org/doi/abs/10.1021/j100540a008.

146. Gillespie, D.T. (2007). Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. *58*, 35–55.

147. Geyer, C.J. (1992). Practical Markov chain Monte Carlo. Stat. Sci. *7*, 473–483. http://www.jstor.org/stable/2246094.

148. Mauch, S., and Stalzer, M. (2010). An efficient method for computing steady state solutions with Gillespie's direct method. J. Chem. Phys. *133*, 144108. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2973983/.

149. Prados, A., Brey, J.J., and Sánchez-Rey, B. (1997). A dynamical Monte Carlo algorithm for master equations with time-dependent transition rates. J. Stat. Phys. *89*, 709–734. http://link.springer.com/10.1007/BF02765541.

150. Shahrezaei, V., Ollivier, J.F., and Swain, P.S. (2008). Colored extrinsic fluctuations and stochastic gene expression. Mol. Syst. Biol. *4*, 196. https://onlinelibrary.wiley.com/doi/10.1038/msb.2008.31.

151. Voliotis, M., Thomas, P., Grima, R., and Bowsher, C.G. (2016). Stochastic simulation of biomolecular networks in dynamic environments. PLoS Comp. Biol. *12*, e1004923. https://dx.plos.org/10.1371/journal.pcbi.1004923.

152. Wang, S., and Bianco, S. (2021). AI-assisted biology: predict the conditional probability distributions from noisy measurements. https://doi.org/10.1101/2021.10.07.463577.

153. Wang, S., Fan, K., Luo, N., Cao, Y., Wu, F., Zhang, C., Heller, K.A., and You, L. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. Nat. Commun. *10*, 4354. http://www.nature.com/articles/s41467-019-12342-y.

154. Gorin, G., Carilli, M., Chari, T., and Pachter, L. (2022). Spectral neural approximations for models of transcriptional dynamics. https://doi.org/10.1101/2022.06.16.496448.

155. Sukys, A., Öcal, K., and Grima, R. (2022). Approximating solutions of the Chemical master equation using neural networks. iScience *25*, 105010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9474291/.

156. Wood, S.N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. Nature *466*, 1102–1104. https://doi.org/10.1038/nature09319.

157. Drovandi, C.C., Pettitt, A.N., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. Stat. Sci. *30*, 72–95. https://doi.org/10.1214/14-STS498.

158. Öcal, K., Gutmann, M.U., Sanguinetti, G., and Grima, R. (2022). Inference and uncertainty quantification of stochastic gene expression via synthetic models. J. R. Soc. Interface *19*, 20220153. https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2022.0153.

159. Cao, Z., and Grima, R. (2018). Linear mapping approximation of gene regulatory networks with stochastic dynamics. Nat. Commun. *9*, 3305. https://doi.org/10.1038/s41467-018-05822-0.

160. Thorne, B.C., Bailey, A.M., and Peirce, S.M. (2007). Combining experiments with multi-cell agent-based modeling to study biological tissue patterning. Brief. Bioinform. *8*, 245–257. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbm024.

161. Thomas, P., and Shahrezaei, V. (2021). Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations. J. R. Soc. Interface *18*, 20210274. https://royalsocietypublishing.org/doi/10.1098/rsif.2021.0274.

162. Kac, M., Rota, G.-C., and Schwartz, J.T. (2009). Discrete Thoughts: Essays on Mathematics, Science and Philosophy (Springer Science & Business Media).

163. 10x Genomics. (2018). 1k PBMCs from a Healthy Donor (v3 chemistry). https://www.10xgenomics.com/resources/datasets/1-k-pbm-cs-from-a-healthy-donor-v-3-chemistry-3-standard-3-0-0.

164. 10x Genomics. (2018). 1k Heart Cells from an E18 mouse (v3 chemistry). https://www.10xgenomics.com/resources/datasets/1-k-heart-cells-from-an-e-18-mouse-v-3-chemistry-3-standard-3-0-0.

165. 10x Genomics. (2018). 1k Brain Cells from an E18 Mouse (v3 chemistry). https://www.10xgenomics.com/resources/datasets/1-k-brain-cells-from-an-e-18-mouse-v-3-chemistry-3-standard-3-0-0.

166. 10x Genomics. (2018). 1k PBMCs from a Healthy Donor (v2 chemistry). https://www.10xgenomics.com/resources/datasets/1-k-pbm-cs-from-a-healthy-donor-v-2-chemistry-3-standard-3-0-0.

167. 10x Genomics. (2018). 5k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Nuclei. https://www.10xgenomics.com/resources/datasets/5-k-mouse-e-18-combined-cortex-hippocampus-and-subventricular-zone-nuclei-3-1-standard-6-0-0.

168. Cariboni, J., and Schoutens, W. (2009). Jumps in intensity models: investigating the performance of Ornstein-Uhlenbeck processes in credit risk modeling. Metrika *69*, 173–198. http://link.springer.com/10.1007/s00184-008-0213-4.

169. Risken, H. (1996). The Fokker-Planck equation: methods of solution and applications. In Springer series in synergetics, Second Edition (Springer-Verlag).

170. Montroll, E.W. (1972). On coupled rate equations with quadratic nonlinearities. Proc. Natl. Acad. Sci. USA *69*, 2532–2536. https://www.jstor.org/stable/61810.

171. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. USA. *115*, E2467–E2476. http://www.pnas.org/lookup/doi/10.1073/pnas.1714723115.

172. Sanders, S., Joshi, K., Levin, P., and Iyer-Biswas, S. (2022). Single cells tell their own story: an updated framework for understanding stochastic variations in cell cycle progression in bacteria. https://doi.org/10.1101/2022.03.15.484524.

173. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868. http://www.nature.com/articles/nmeth.4380.

174. 10x Genomics (2021). Interpreting intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data. Technical Note CG000376, 10x Genomics. https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-intronic-and-antisense-reads-in-10-x-genomics-single-cell-gene-expression-data.

175. Cox, J.C., Ingersoll, J.E., and Ross, S.A. (1985). A theory of the term structure of interest rates. Econometrica *53*, 385. https://www.jstor.org/stable/1911242?origin=crossref.

176. Fredriksson, T. (2017). Fokker Planck for the Cox-Ingersoll-Ross Model. Ph.D. thesis (Uppsala Universitet).

177. Sabino, P., and Cufaro Petroni, N.C. (2021). Gamma-related Ornstein–Uhlenbeck processes and their simulation. J. Stat. Comput. Simul. *91*, 1108–1133. https://www.tandfonline.com/doi/full/10.1080/00949655.2020.1842408.

178. Melsted, P., Ntranos, V., and Pachter, L. (2019). The barcode, UMI, set format and BUStools. Bioinformatics *35*, 4472–4473.

179. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170. https://www.nature.com/articles/s41592-021-01346-6.

180. Skinner, S.O., Xu, H., Nagarkar-Jaiswal, S., Freire, P.R., Zwaka, T.P., and Golding, I. (2016). Single-cell analysis of transcription kinetics across the cell cycle. eLife *5*, e12175. https://elifesciences.org/articles/12175.

181. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://www.nature.com/articles/s41586-020-2649-2.

182. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. http://www.nature.com/articles/s41592-019-0686-2.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
H. sapiens peripheral blood 10x v3 scRNA-seq data	10x Genomics[163]	pbmc_1k_v3
M. musculus heart 10x v3 scRNA-Seq Data	10x Genomics[164]	heart_1k_v3
M. musculus neuron 10x v3 scRNA-seq data	10x Genomics[165]	neuron_1k_v3
M. musculus cultured embryonic stem cells treated with DMSO 10x v2 scRNA-seq data	Desai et al.[116]	desai_dmso
H. sapiens peripheral blood 10x v2 scRNA-seq data (technical replicate of pbmc_1k_v3)	10x Genomics[166]	pbmc_1k_v2
M. musculus neuron 10x v3 snRNA-seq data	10x Genomics[167]	brain_nuc_5k_v3
Supplementary Data for GP_2021_3	Gorin and Pachter[21]	Zenodo: 7388133
Software and Algorithms		
Python	python.org	3.9.1; RRID: SCR_008394
NumPy	numpy.org	1.22.1; RRID: SCR_008633
SciPy	scipy.org	1.7.3; RRID: SCR_008058
pandas	pandas.pydata.org	1.2.4; RRID: SCR_018214
kallisto | bustools	Melsted et al.[113]	0.26.0; RRID: SCR_018213
Monod	Gorin and Pachter	2.5.0
Other		
Count matrices for all datasets	This manuscript	Zenodo: 8132976
Custom analysis notebooks	This manuscript	GitHub: https://github.com/pachterlab/GVP_2023 (version of record deposited at Zenodo: 8132976)

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Lior Pachter (lpachter@caltech.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Pseudoaligned count matrices in the *mtx* format have been deposited at Zenodo: 8132976. The data, *Monod* fits, and analysis scripts used to generate Figures 5D and 5E, originating from Gorin et al.,[21] were previously deposited at Zenodo: 7388133.
- All original code has been deposited at https://github.com/pachterlab/GVP_2023 and Zenodo: 8132976, and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Master equation models of transcription

We are interested in continuous-time stochastic processes that combine categorical, nonnegative discrete, and (usually nonnegative) continuous degrees of freedom. To solve these systems, we begin by separately defining their allowed transitions and converting them to master equation forms.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

The categorical variable, denoted by $s \in \{1,\ldots,N\}$, represents the instantaneous state of a multi-state gene. By assuming that the state interconversions are Markovian and independent of all other components of the system, we can define H_{ij}, the rates of transitioning from state i to state j:

$$\mathcal{S}_i \overset{H_{ij}}{\rightarrow} \mathcal{S}_j. \tag{Equation 35}$$

These rates can be summarized in the state transition matrix $H \in \mathbb{R}_{\geq 0}^{N\times N}$, such that $H_{ii} = -\sum_{j\neq i} H_{ij}$ and $\sum_j H_{ij} = 0$ to enforce the conservation of probability. This set of transitions can be represented by a master equation involving finitely many ODEs, which tracks the probabilities of each state s at a time t:

$$\begin{aligned} \frac{\partial P(s,t)}{\partial t} &= \sum_{i=1}^{N} H_{is} P(i,t), \text{ or more compactly} \\ \frac{\partial \mathbf{P}(t)}{\partial t} &= H^T \mathbf{P}. \end{aligned} \tag{Equation 36}$$

As this system is expressed in terms of a differential equation for an arbitrary time t, the relation holds for time-dependent H. For simplicity, we assume that H is deterministic.

The nonnegative discrete variables, denoted by $\mathbf{x} \in \mathbb{N}_0^n$, represent molecular copy numbers. We assume that n molecular species participate in four classes of transitions, and can summarize their effect by considering their reaction schema and effect on x_i, the number of molecules of species i:

$$\begin{aligned} \mathcal{X}_i &\overset{c_{ij}}{\rightarrow} \mathcal{X}_j, \\ \mathcal{X}_i &\overset{c_{i0}}{\rightarrow} \varnothing, \\ \mathcal{X}_i &\overset{q_{ij}}{\rightarrow} \mathcal{X}_i + \mathcal{X}_j, \\ \varnothing &\overset{\alpha_\omega}{\rightarrow} B_{i_1}\mathcal{X}_{i_1} + \cdots + B_{i_{\ell_\omega}}\mathcal{X}_{i_{\ell_\omega}}. \end{aligned} \tag{Equation 37}$$

First, species i can be converted to species j with rate $c_{ij}x_i$. Second, species i can spontaneously degrade with rate $c_{i0}x_i$. These classes of monomolecular transitions, which either maintain or reduce the total number of molecules in the system, can be summarized in the matrix $C^{dd} \in \mathbb{R}^{n\times n}$, such that $C^{dd}_{ij} = c_{ji}$ and $C^{dd}_{ii} = -c_{i0} - \sum_{j\neq i} c_{ij}$; C^{dd} is the matrix governing the associated reaction rate equations.[17,85] Third, species i participate in autocatalysis at the rate q_{ii}, or catalysis of species j at the rate q_{ij}. These reactions can be summarized by the matrix $Q^d \in \mathbb{R}_{\geq 0}^{n\times n}$, such that $Q^d_{ij} = q_{ji}$. Finally, molecules can be produced. In the general case, a burst of production simultaneously creates molecules of ℓ_ω discrete species $\{i_1,\ldots,i_{\ell_\omega}\}$. We assume bursts are described by a Poisson arrival process, with burst frequency α^d_ω and the nontrivial ℓ_ω-variate joint distribution $p^d_\omega(\mathbf{z})$ of non-negative burst sizes $\{B_{i_1},\ldots,B_{\ell_\omega}\}$.[16] This formulation includes the trivial case of Poisson point process production of species i, for which $\ell_\omega = 1$ and $p^d_\omega(\mathbf{z}) = \delta_{ij}$, the degenerate distribution located at unity for species i and zero for all other species.

This mass action model, which tracks molecule counts, can be represented by an equivalent discrete CME, which tracks the probability of each microstate $\mathbf{x}$:

$$\begin{aligned} \frac{\partial P(\mathbf{x},t)}{\partial t} &= \sum_{i=1}^{n} c_{i0}[(x_i+1)P(x_i+1,t) - x_i P(\mathbf{x},t)] \\ &+ \sum_{ij=1}^{n} c_{ij}[(x_i+1)P(x_i+1,x_j-1,t) - x_i P(\mathbf{x},t)] \\ &+ \sum_{i=1}^{n} Q^d_{ii}[(x_i-1)P(x_i-1,t) - x_i P(\mathbf{x},t)] \\ &+ \sum_{ij=1}^{n} Q^d_{ji}[x_i P(x_j-1,t) - x_i P(\mathbf{x},t)] \\ &+ \sum_{\omega} \alpha^d_\omega \left[\sum_{\mathbf{z}} p^d_\omega(\mathbf{z}) P(\mathbf{x}-\mathbf{z},t) - P(\mathbf{x},t)\right]. \end{aligned} \tag{Equation 38}$$

For simplicity of notation, species that do not occur in a reaction are elided from the master equation, as in previous work on modeling bursty transcription.[16] As above, this equation holds even if the rates are time-dependent. For the purposes of this report, we assume only α_ω and p_ω can vary over time.

The nonnegative continuous variables, denoted by $\mathbf{y} \in \mathbb{R}^m_{\geq 0}$, represent concentrations or coarsely-modeled noise sources. We assume that these variables are governed by Ornstein-Uhlenbeck-type SDEs:

$$d\mathbf{y}_t = C^{cc}\mathbf{y}_t dt + Q^c(\mathbf{y}_t) d\mathbf{W}_t + \sum_{\omega} d\mathbf{L}_\omega(t), \tag{Equation 39}$$

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

where $\mathbf{y}_t$ is a realization of the process, $\mathbf{W}_t$ is an w-dimensional Brownian motion, and $\mathbf{L}_\omega$ is a subordinator. The matrix $C^{cc} \in \mathbb{R}^{m\times m}$ sets the mean-reversion terms, whereas the operator $Q^c(\mathbf{y}_t) : \mathbb{R}^m_{\geq 0} \rightarrow \mathbb{R}^{m\times w}_{\geq 0}$ sets the level of noise. We assume that each $\mathbf{L}_\omega$ only includes drift or compound Poisson terms. The drift terms have the form $\alpha^c_i \delta_{ij} t$. To slightly lighten the notation, we can aggregate all drift terms under $\omega = 1, \ldots, m$ as $\{\alpha^c_1 dt, \ldots, \alpha^c_m dt\}$; some of these entries may be zero. The compound Poisson terms have the form $\sum_{k=0}^{N_\omega(t)} (\mathbf{B}_\omega)_k$,[168] such that $N_\omega(t)$ is a Poisson random variable with mean $\alpha^c_\omega t$ and $(\mathbf{B}_\omega)_k$ is a set of independent and identically distributed realizations of the random variable $\mathbf{B}_\omega$. This random variable has a nontrivial ℓ_ω-variate joint density $p^c_\omega(\mathbf{z})$ on $\mathbb{R}^m_{\geq 0}$, with the remaining $m - \ell_\omega$ dimensions concentrated at zero. We note that this formulation entails a slight abuse of notation, as ω is used to index over discrete burst processes as well as continuous drift and jump components.

For simplicity, we assume the noise term takes the form of an uncoupled square-root diffusion, such that $w = \mathbf{m}$ and $Q^c(\mathbf{y}_t) = \mathrm{diag}(\boldsymbol{\sigma} \odot \sqrt{\mathbf{y}_t})$. The symbol $\odot$ denotes the elementwise/Hadamard product of two vectors, the square root should be interpreted as elementwise, and all elements of the constant volatility vector $\boldsymbol{\sigma}$ are non-negative. Although this choice of Q^c is somewhat restrictive, it produces a particularly simple diffusion tensor Σ:

$$\Sigma(\mathbf{y}) : = \frac{1}{2} Q^c(\mathbf{y}) Q^c(\mathbf{y})^T = \frac{1}{2} \mathrm{diag}(\boldsymbol{\sigma}^2 \odot \mathbf{y}), \quad \text{(Equation 40)}$$

where the square $\boldsymbol{\sigma}^2$ should be interpreted as elementwise. This formulation can be reframed as a Fokker-Planck equation,[169] which tracks the probability density of each microstate $\mathbf{y}$:

$$\frac{\partial P}{\partial t} = -\sum_{i,j=1}^{m} C^{cc}_{ji} \frac{\partial}{\partial y_j}[y_i P] + \frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 \frac{\partial^2}{\partial y_i^2}[y_i P] - \sum_{i=1}^{m} \alpha^c_i \frac{\partial P}{\partial y_i} + \sum_{\omega > m} \alpha^c_\omega \left[\int_{\mathbf{z}} p^c_\omega(\mathbf{z}) P(\mathbf{y} - \mathbf{z}, t) d\mathbf{z} - P(\mathbf{y}, t) \right]. \quad \text{(Equation 41)}$$

As above, we assume that only the components of $\mathbf{L}_\omega$ can vary in time.

In addition to these discrete- and continuous-only terms, we need to account for these components' interactions. For example, we may want to represent the production of a discrete species controlled by a continuous variable, e.g., a time-varying transcription rate[20]:

$$\varnothing \xrightarrow{y_i c_{ij}} \mathcal{X}_j. \quad \text{(Equation 42)}$$

This reaction has the rate $y_i c_{ij}$. This class of reactions can be summarized in the matrix $C^{cd} \in \mathbb{R}^{m\times n}_{\geq 0}$, such that $C^{cd}_{ij} = c_{ji}$. In other words, this class of reactions contributes the following terms to the overall master equation:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} C^{cd}_{ji} [y_i P(x_j - 1, \mathbf{y}, t) - y_i P(\mathbf{x}, \mathbf{y}, t)]. \quad \text{(Equation 43)}$$

Finally, we may want to represent the production of a continuous species from a discrete one, e.g., the rapid translation of high-abundance protein from low-abundance RNA.[139] This class of reactions simply adds a term proportional to $C^{dc}\mathbf{x}\ dt$ to the expression for $\mathbf{y}_t$. The matrix $C^{dc} \in \mathbb{R}^{n\times m}_{\geq 0}$ contains the relevant rates, such that C^{dc}_{ij} is the rate of producing the continuous species i from discrete species j. Therefore, we append a set of drift-like terms to the Fokker-Planck equation:

$$-\sum_{i=1}^{n} \sum_{j=1}^{m} C^{dc}_{ji} x_i \frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial y_j}. \quad \text{(Equation 44)}$$

To construct the full master equation, we need to define a system of N coupled equations. To do so, we essentially add Equations 36, 38, 41, 43, and 44, replacing all instances of P with $\mathbf{P}(s, \mathbf{x}, \mathbf{y}, t)$. However, to account for differences in transcription between gene states, we allow the ω-associated terms to vary with s. The full master equation is reported below in Equation 45.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

The full master equation

The full master equation for $P(s, \mathbf{x}, \mathbf{y}, t)$ is:

$$\begin{aligned}
\frac{\partial P}{\partial t} &= \sum_{i=1}^{N} H_{is}(t) P(i, \mathbf{x}, \mathbf{y}, t) \\
&+ \sum_{i=1}^{n} c_{i0}[(x_i + 1)P(s, x_i + 1, \mathbf{y}, t) - x_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&+ \sum_{ij=1}^{n} c_{ij}[(x_i + 1)P(s, x_i + 1, x_j - 1, \mathbf{y}, t) - x_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&+ \sum_{i=1}^{n} Q_{ii}^{d}[(x_i - 1)P(s, x_i - 1, \mathbf{y}, t) - x_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&+ \sum_{ij=1}^{n} Q_{ji}^{d}[x_i P(s, x_j - 1, \mathbf{y}, t) - x_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&+ \sum_{\omega} \alpha_{s,\omega}^{d}(t) \left[\sum_{\mathbf{z}} p_{s,\omega}^{d}(\mathbf{z}, t) P(s, \mathbf{x} - \mathbf{z}, \mathbf{y}, t) - P(s, \mathbf{x}, \mathbf{y}, t) \right] \\
&- \sum_{ij=1}^{m} C_{ji}^{cc} \frac{\partial}{\partial y_j}[y_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&+ \frac{1}{2} \sum_{i=1}^{m} \sigma_i^2 \frac{\partial^2}{\partial y_i^2}[y_i P(s, \mathbf{x}, \mathbf{y}, t)] \\
&- \sum_{i=1}^{m} \alpha_{s,i}^{c}(t) \frac{\partial P(s, \mathbf{x}, \mathbf{y}, t)}{\partial y_i} \\
&+ \sum_{\omega > m} \alpha_{s,\omega}^{c}(t) \left[\int_{\mathbf{z}} p_{\omega}^{c}(\mathbf{z}) P(\mathbf{y} - \mathbf{z}, t) d\mathbf{z} - P(\mathbf{y}) \right] \\
&+ \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ji}^{cd}[y_i P(x_j - 1, \mathbf{y}, t) - y_i P(\mathbf{x}, \mathbf{y}, t)] \\
&- \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ji}^{dc} x_i \frac{\partial P(\mathbf{x}, \mathbf{y}, t)}{\partial y_j}.
\end{aligned} \tag{Equation 45}$$

We annotate the terms in Table S1.

Generating function methods for biological stochasticity

The full master equation is fairly cumbersome and challenging to analyze directly. Therefore, analysis has to proceed by spectral methods. We use the generating function (GF), a length-N vector function $\mathbf{G}$, such that each component is

$$\begin{aligned}
G_s(\mathbf{g}, \mathbf{h}, t) &= \int_0^{\infty} \cdots \int_0^{\infty} \sum_{x_1=0}^{\infty} \cdots \sum_{x_n=0}^{\infty} \left(\prod_{i=1}^{n} g_i^{x_i} \right) \left(\prod_{i=1}^{m} e^{h_i y_i} \right) \\
&P(s, \mathbf{x}, \mathbf{y}, t) dy_m \cdots dy_1 \\
&:= \int_{\mathbf{y}} \sum_{\mathbf{x}} \mathbf{g}^{\mathbf{x}} e^{\mathbf{h}^T \mathbf{y}} P(s, \mathbf{x}, \mathbf{y}, t) d\mathbf{y},
\end{aligned}$$

where the lowest line is the definition expressed in useful shorthand notation. Formally, the generating function is the combination of a probability-generating function (PGF) in the discrete variables and moment-generating function (MGF) in the continuous variables. The arguments $\mathbf{g} \in \mathbb{C}^n$ and $\mathbf{h} \in \mathbb{C}^m$ are spectral variables. By computing the generating function of both sides of Equation 45, we find (see supplemental information) that the master equation is equivalent to a much more compact system of PDEs:

$$\frac{\partial \mathbf{G}}{\partial t} = H^T \mathbf{G} + \mathbf{G} \odot \mathcal{A}(\mathbf{u}) + J[C\mathbf{u} + \text{diag } \mathbf{u}\ D\mathbf{u}]. \tag{Equation 46}$$

This formulation relies on defining the unified variables $\mathbf{u}$:

$$\mathbf{u} := \begin{bmatrix} \mathbf{g} - 1 \\ \mathbf{h} \end{bmatrix} \text{ and } J_{si} = \frac{\partial G_s}{\partial u_i}, \tag{Equation 47}$$

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

as well as unified matrices:

$$C := \begin{bmatrix} \left(C^{dd}\right)^T + \left(Q^d\right)^T & \left(C^{dc}\right)^T \\ \left(C^{cd}\right)^T & (C^{cc})^T \end{bmatrix}$$
$$D := \begin{bmatrix} \left(Q^d\right)^T & \left(C^{dc}\right)^T \\ 0 & \frac{1}{2}\text{diag } \boldsymbol{\sigma}^2 \end{bmatrix} \quad \text{(Equation 48)}$$
$$:= \begin{bmatrix} \left(Q^d\right)^T & \left(C^{dc}\right)^T \\ 0 & (Q^c)^T \end{bmatrix}.$$

Each entry of the length-N matrix function $\mathcal{A}$ consists of the burst and drift terms:

$$\begin{aligned} \mathcal{A}_s &= \left(\boldsymbol{\alpha}^d\right)_s^T (\mathbf{F}_s(\mathbf{u}+1) - 1) + \left(\boldsymbol{\alpha}^c\right)_s^T (\mathbf{M}_s(\mathbf{u}) - 1) \\ &= \boldsymbol{\alpha}_s^T (\mathcal{M}_s(\mathbf{u}) - 1). \end{aligned} \quad \text{(Equation 49)}$$

The vector $\boldsymbol{\alpha}_s^d$ contains the frequencies of all discrete burst processes for state s. The first m entries of $\boldsymbol{\alpha}_s^c$ contain the continuous species' drifts in state s. The remaining entries contain the corresponding rates of continuous burst processes. $\boldsymbol{\alpha}_s$ aggregates these quantities. The vector function $\mathbf{F}_s$ contains the joint PGF of the discrete burst processes, and only depends on the first n variables. The vector function $\mathbf{M}_s$ contains the drift terms, as well as the joint MGF of the continuous burst processes, and only depends on the last m variables. The parameters of the $\mathcal{A}_s$ operator may vary in time.

To obtain the generating function at t, we apply the method of characteristics. First, we calculate the characteristics parametrized by the scalar variable s:

$$\begin{aligned} T(\mathrm{s}) &= t - \mathrm{s} \\ \frac{d\mathbf{U}(\mathrm{s})}{d\mathrm{s}} &= C\mathbf{U}(\mathrm{s}) + \text{diag } \mathbf{U}(\mathrm{s})\ D\mathbf{U}(\mathrm{s}) \end{aligned} \quad \text{(Equation 50)}$$

where $\mathbf{U}(\mathrm{s} = 0) = \mathbf{u}$. This is the "downstream" ODE, which governs abundances in isolation from production and regulation.

Therefore, $\mathbf{G}$ is governed by the following system of ODEs:

$$\begin{aligned} \frac{d\mathbf{G}(\mathbf{U}(\mathrm{s}), T(\mathrm{s}))}{d\mathrm{s}} &= -H(T(\mathrm{s}))^T \mathbf{G} - \mathbf{G} \odot \mathcal{A}(\mathbf{U}(\mathrm{s}), T(\mathrm{s})) \\ &:= \mathcal{H}(\mathbf{U}, T)\ \mathbf{G}. \end{aligned} \quad \text{(Equation 51)}$$

To obtain $\mathbf{G}$ at t, we integrate this matrix system from $\mathrm{s} = t$ to $\mathrm{s} = 0$. We use $\mathbf{G}^0(\mathbf{U}(t))$ as the initial condition, where $\mathbf{G}^0$ is the generating function of the initial distribution. This is the "upstream" ODE, which governs the full generating function.

In the general case, evaluating this system requires two applications of quadrature: first, solving the $n+m$-dimensional downstream system to obtain the values of characteristics $\mathbf{U}$ at a set of grid points over $[0, t]$; then, solving the N-dimensional upstream system to obtain the value of the generating function.

Some special cases afford simpler solutions. If $D \neq 0$, the downstream ODE takes a Riccati-like form and generally resists exact analysis.[17,170] However, if $D = 0$ and C is diagonalizable, the system takes the tractable linear form

$$\begin{aligned} &\frac{d\mathbf{U}(\mathrm{s})}{d\mathrm{s}} = C\mathbf{U}(\mathrm{s}) := V^{-1}\Lambda V\mathbf{U}(\mathrm{s}), \text{ with the solution} \\ &\mathbf{U}(\mathrm{s}) = V^{-1}e^{\Lambda \mathrm{s}}V\mathbf{u} \end{aligned} \quad \text{(Equation 52)}$$

whenever all eigenvalues of C are distinct. When they are not, the ODE can be solved in a similar way using generalized eigenvectors. Practically, this means that only one application of quadrature is required.

If, in addition, $N = 1$, the upstream ODE reduces to a single integral:

$$\begin{aligned} \phi(t) &= \int_t^0 \frac{d\phi(\mathbf{U}(\mathrm{s}), T(\mathrm{s}))}{d\mathrm{s}} d\mathrm{s} \\ &= \phi^0(\mathbf{U}(t)) + \int_0^t \mathcal{A}(\mathbf{U}(\mathrm{s}), T(\mathrm{s}))d\mathrm{s}, \end{aligned} \quad \text{(Equation 53)}$$

where $\phi := \log G$, $\phi^0 = \log G^0$, and the generating function G is no longer boldfaced because only a single gene state exists.

If $\mathcal{A}$ is a linear operator $a_1u_1 + \cdots + a_{n+m}u_{n+m}$, the system is in the drift-only regime; no bursting occurs. In this case, the system reduces to

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$\phi(t) = \phi^0(\boldsymbol{U}(t)) + \sum_{i=1}^{n+m} \int_0^t a_i(t - \mathrm{s}) U_i(\mathrm{s}) d\mathrm{s}, \quad \text{(Equation 54)}$$

where U_i are the components of $\mathbf{U}$. As each U_i is, in turn, a weighted sum of u_i, the second term of the log-generating function is given by a sum of fairly simple convolutions that scale as $\int_0^t a_i(t - \mathrm{s}) e^{\lambda_j \mathrm{s}} d\mathrm{s}$.

Finally, in the simplest case, if all eigenvalues λ_i of C are negative, the transient part of Equation 54 vanishes as $t \to \infty$ and the stationary log-generating function is a linear combination of u_i. This implies that the distribution converges to a product of independent Poisson distributions.[17,85]

Coupling multiple genes

The results solve master equations with abstracted production and processing reactions. To connect them to systems phenomena, such as the co-regulation of multiple genes, we need to specify how upstream interactions lead to co-expression. As the simplest illustrative model system, we can consider the co-regulation of two genes, indexed by j, with $U_j = u_j e^{-\gamma_j \mathrm{s}}$. We outline several relatively simple classes of candidate models which induce expression coupling.

In the simplest case, $\mathcal{H}(\mathbf{u}, t) = \sum_j \mathcal{H}_j(u_j, t)$. In other words, the genes' dynamics are fully separable, and produce solutions in the form $G(\mathbf{u}, t) = \prod_j G_j(u_j, t)$. This formulation produces independent distributions at each t, but the *trajectories* may possess nontrivial statistical relationships. For example, if both genes start at $x_1 = x_2 = 0$, their trajectories will be correlated over a finite timespan $[0, T]$, with the correlation decaying as $T \to \infty$.

In the next simplest case, co-regulation is the consequence of parameter differences in subpopulations. For example, the full cell population may consist of cell types indexed by κ. If we suppose each cell type has the abundance π_κ and transcriptional parameters Θ_κ, we obtain

$$G(\mathbf{u}, t) = \sum_\kappa \pi_\kappa G(\mathbf{u}, t; \Theta_\kappa) = \sum_\kappa \pi_\kappa \prod_j G_j(u_j, t; \Theta_{j,\kappa}); \quad \text{(Equation 55)}$$

i.e., the generating function decomposes into a product of independent generating functions *conditional on* a particular cell type, but not globally. In other words, even if transcriptional processes are independent, cell type structure can produce nontrivial relationships between genes.

Alternatively, we can propose a model of co-regulation by the categorical variables. For example, two neighboring genes may prefer to have the same or opposite accessibility, depending on the polymeric properties of DNA. Assuming, for the purposes of illustration, that the system is symmetric, we obtain the following $N = 4$ form:

$$H = \begin{bmatrix} -2k_{\mathrm{on}} & k_{\mathrm{on}} & k_{\mathrm{on}} & 0 \\ \varepsilon^{-1} k_{\mathrm{off}} & -\varepsilon^{-1}(k_{\mathrm{on}} + k_{\mathrm{off}}) & 0 & \varepsilon^{-1} k_{\mathrm{on}} \\ \varepsilon^{-1} k_{\mathrm{off}} & 0 & -\varepsilon^{-1}(k_{\mathrm{on}} + k_{\mathrm{off}}) & \varepsilon^{-1} k_{\mathrm{on}} \\ 0 & k_{\mathrm{off}} & k_{\mathrm{off}} & -2k_{\mathrm{off}} \end{bmatrix}$$

$$\mathcal{A} = \begin{bmatrix} 0 \\ k_{\mathrm{init}} u_1 \\ k_{\mathrm{init}} u_2 \\ k_{\mathrm{init}}(u_1 + u_2) \end{bmatrix}. \quad \text{(Equation 56)}$$

This form encodes the co-regulation of two genes, such that $s \in \{\text{both off}, \text{gene 1 on}, \text{gene 2 on}, \text{both on}\}$. If $\varepsilon \ll 1$, the intermediate states are unstable and the genes tend to be either both on or both off. If $\varepsilon \gg 1$, the intermediate states are particularly stable, and only one of the genes tends to be on at a time. If $\varepsilon = 1$, we recover the independent case.

We can define a similar model for co-regulation by a continuous variable y_1. For example, there may be a latent regulator, such as the concentration of an activator, that controls multiple loci: if it is high, both have a high transcription rate; otherwise, both are inactive.[20] This amounts to appending the following reactions to the master equation:

$$C^{cd}_{j1} y_1 [P(x_j - 1) - P(x_j)], \quad \text{(Equation 57)}$$

where the C^{cd} matrix encodes the relationship between the concentration and the transcription rate. Therefore, the genes become mutually correlated through the trajectory of y_1, although the extent of correlation depends on the dynamics.

If the categorical or continuous driving process is bursty, we can approximate it by a co-bursting module. For example, in the limit of $\varepsilon \to 0$, the dynamics of the system in Equation 56 converge to the $N = 2$ formulation

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$H = \begin{bmatrix} -k^*_{\text{on}} & k^*_{\text{on}} \\ k^*_{\text{off}} & -k^*_{\text{off}} \end{bmatrix} \text{ and } \mathcal{A} = \begin{bmatrix} 0 \\ k_{\text{init}}(u_1 + u_2) \end{bmatrix}, \text{ where}$$
$$k^*_{\text{on}} = \frac{2k^2_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \text{ and } k^*_{\text{off}} = \frac{2k^2_{\text{off}}}{k_{\text{on}}+k_{\text{off}}}. \quad \text{(Equation 58)}$$

If, in addition, $k^*_{\text{off}}, k_{\text{init}} \to \infty$, we obtain the $N = 1$ module characterized by

$$\mathcal{A} = k^*_{\text{on}}\left[\frac{1}{1 - b(u_1+u_2)} - 1\right], \quad \text{(Equation 59)}$$

where $b := k_{\text{init}}/k^*_{\text{off}}$.[16] This is the bursty limit of Equation 56. Interestingly, that mechanism also possesses a slow mixture limit. If $\varepsilon \to \infty$ while $k_{\text{on}}, k_{\text{off}} \to 0$, we obtain a special case of Equation 55, with $\pi_\kappa = 1/2$ and mutually exclusive expression in the "cell types," or long-lived gene states.

Even when we restrict our analysis to simple feed-forward regulation, this outline of motifs is nowhere near exhaustive. Nevertheless, the "mixture" and "bursty" limits are particularly natural starting points, as their distributions are straightforward to construct. In other words, we speculate that the careful analysis of co-expression models can distinguish relationships due to "slow" variation between cell types and "fast" variation due to coupled transcriptional events.

Transient phenomena

This result yields a fairly simple numerical recipe for the determination of probabilities at a particular time t. Typically, analysis proceeds by assuming H and $\mathcal{A}$ are time-independent and letting $t \to \infty$, i.e., considering the stationary limit of the process. However, this may not be strictly justifiable: much of single-cell analysis involves the determination of trajectories from intrinsically transient data representing differentiation pathways[171] If the transient process occurs on a timescale comparable to RNA turnover, using a stationary model may not be appropriate.[16]

To rigorously fit transient data, we need to posit just *how* a snapshot of cells may capture multiple cell states, such that some states are the progenitors of others. The solution is not yet clear, and multiple reasonable explanations exist; for example, we may suppose that the differentiation process "lags" in certain cells (in the vein of the models of variability proposed in Stumpf et al.[44] for development, and in Sanders et al.[172] and Perez-Carrasco et al.[125] for the cell cycle). In other words, all cells are captured at a time t since the beginning of a process, but H and $\mathcal{A}$ have different time-dependence for different cells. Although such an explanatory model can be instantiated, it may be too challenging to fit. Further, it does not appear to be compatible with processes that operate continuously; the choice of t becomes somewhat challenging to motivate.

We propose that the simplest model for observations relies on minimal synchronization between the biology and the experimental process. To mathematically formalize it, we take inspiration from the theory of reactor modeling in chemical engineering[105] and extend preliminary work from our recent RNA velocity methods analysis.[19] A cell enters a medium; this entrance triggers a chemical signal that begins a transient process. The dynamics of this transient process are only dependent on time since receiving the signal, and identical between cells. After a delay, the cells exit the medium. In this framework, sequencing is the uniform random sampling of cells present within this medium. Although this formulation is admittedly simplistic—it excludes the cell cycle and stochastic driving—it allows us to take the first steps with a systematic study of using snapshot data to fit transient stochastic processes. This toy model is numerically tractable, which is useful for its simulation and characterization, and possesses a stationary state that is independent of the time at which the experiment is performed, which is useful for biological admissibility and realism.

Therefore, to marginalize over t, we need to augment the model with an additional property: the relationship between time along a transient process and the probability of capturing a cell. In the parlance of reactor engineering, this relationship is given by the internal-age distribution f. The simulations of transient processes in La Manno et al.[86] and Bergen et al.[59] implicitly adopt this model and assume a particular functional form of f. We might suppose cells enter the observation window at $t = 0$ and leave it at $t = T$, with a Dirac residence time distribution $\delta(t - T)$ and uniform sampling throughout this window. The resulting age distribution is uniform, with $f = T^{-1}$, and formally corresponds to the ideal plug flow reactor (PFR) architecture.[105] As $T \to \infty$, we obtain the $t \to \infty$ ergodic limit, if such a limit exists. On the other hand, if $f \to \delta(t - T)$, we recover the instantaneous distribution at time T; this limit formally corresponds to the batch reactor (BR).

To obtain the generating function for the cells inside a tissue, we represent the tissue as a reactor, specify its influx and efflux properties, and solve for the internal-age distribution f. This internal-age distribution yields the occupation measure of the process times, as discussed in our RNA velocity review,[19] and induces the following reactor-wide generating function:

$$G = \int_t G(t)f(t)dt, \text{ where}$$
$$G(t) = \sum_s G_s(t). \quad \text{(Equation 60)}$$

We have marginalized over the instantaneous gene state s because this variable is typically not observable.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Droplet encapsulation noise

The generating function G describes the biological variability due to molecular processes, transcriptional driving, and the capture of cells from a reaction medium. However, single-cell RNA sequencing data do not quantify cells—they quantify *barcodes*. Cells are randomly encapsulated into droplets with barcoded beads; to avoid the formation of "doublets," with two cells per droplet, the microfluidic protocols typically have a fairly low encapsulation rate. If we assume that a droplet may have either zero or one cells, we obtain the following generating function for the distribution of RNA on a per-barcode level:

$$G_{\text{enc}} = p_1 G + p_0 = pG + (1 - p) = G_{\text{bc}}(G), \quad \text{(Equation 61)}$$

where G_{bc} is the PGF of the Bernoulli distribution, with $p_1 = p$ the probability of capturing a single cell and $p_0 = 1 - p$ that of capturing none. Analogously, if we assume that doublets can occur, and the encapsulation of cells is independent and identically distributed (i.i.d.), we find

$$\begin{aligned} G_{\text{enc}} &= p_2 G^2 + p_1 G + p_0 = p^2 G^2 + 2p(1 - p)G + (1 - p)^2 \\ &= [pG + (1 - p)]^2 = G_{\text{bc}}(G), \end{aligned} \quad \text{(Equation 62)}$$

where G_{bc} is now the PGF of the binomial distribution. It is straightforward to extend this to the unconstrained case, with per-cell encapsulation *rate* λ, and obtain the analogous expression

$$\begin{aligned} G_{\text{enc}} &= p_0 + p_1 G + p_2 G^2 + p_3 G^3 + \ldots \\ &= e^{\lambda(G-1)} = G_{\text{bc}}(G), \end{aligned} \quad \text{(Equation 63)}$$

where G_{bc} is the PGF of the Poisson distribution.

However, even empty droplets typically contain some "background" molecules. Removing the empty droplets by filtering for cells with relatively high expression, as well as correcting for the background, is a standard part of sequencing workflows.[57,109–112] To model the joint distribution of biological and background RNA, we need to instantiate a mechanistic hypothesis about its source. The simplest hypothesis consists of two parts. First, we impose the *pseudobulk* interpretation of background: we assume that a fraction of the cells loaded in the library construction step are lysed, and produce a pool of loose molecules. Next, we assume that these molecules are free to be encapsulated into the droplets in an i.i.d. fashion. This implies the Poisson functional form for the distribution of debris entering each droplet:

$$G_{\text{bg}} = \exp\left(c \sum_i \mu_i u_i\right), \quad \text{(Equation 64)}$$

where c is some shared constant that reflects the pool size and the rate of diffusion, whereas $\mu_i = \frac{\partial G}{\partial u_i}\Big|_{u_i = 0}$ is the expectation of species i over the entire cell population. This simplest model assumes that all cells are equally likely to lyse and release their contents; if this assumption is violated, μ_i needs to be obtained by computing an expectation with respect to a measure biased toward the less stable cells. Finally, the full per-droplet distribution of molecules is

$$G_{\text{tot}} = G_{\text{bc}} G_{\text{bg}}, \quad \text{(Equation 65)}$$

i.e., each droplet contains contributions from the encapsulated cells, as well as the background. With some abuse of notation, we occasionally use the expression $G_{\text{bc}}(G)G_{\text{bg}}(G)$, where the first argument denotes composition, whereas the second denotes functional dependence.

Library construction and sequencing noise

We cannot observe the biological molecule content of each droplet: we are restricted to analyzing counts of complementary DNA (cDNA). In a typical dual-index 3′ microfluidic workflow (e.g., the commercialized 10x chemistry[48]), these cDNA are quantified by the following sequence of reactions. First, a synthetic primer captures a poly(A) stretch in RNA, which may be an endogenous molecule or a synthetic tag.[173] The primer contains a poly(dT) oligonucleotide, a sequencing primer, a cell barcode, and a unique molecular identifier (UMI). Next, reverse transcriptase (RTase) attaches to the RNA-primer complex and synthesizes the complementary strand. When the first strand is complete, a template-switching oligonucleotide (TSO) attaches to the end, allowing RT to synthesize the second strand of cDNA. After library construction, the droplet emulsion is broken, producing a pool of long cDNA; polymerase chain reaction (PCR) is used to amplify this pool. The long cDNA molecules are enzymatically fragmented, and another sequencing primer is attached at the end of the molecule that formerly contained the TSO. Finally, another round of PCR amplifies the pool and appends sample indices and Illumina adaptors to both sides of the molecule. The pool of cDNA is loaded onto a sequencing machine and sequenced from both sides, producing two reads. One read contains the barcode and UMI bases, whereas the other contains partial information about the 3′ end of the molecule, beginning at the fragmentation site. This sequence of reactions represents the ideal-case scenario, and the products may well include artifacts due to off-target reactions.[174]

To understand the effect of technical variability on the per-barcode distributions, we need to summarize this workflow in a mechanistic model. First, we assume that the library preparation reactions occur in an i.i.d. fashion relative to each RNA molecule in the droplet, allowing us to construct a separate description of technical noise for each discrete molecular species indexed by i. At this

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

stage, we omit the modeling of continuous species. As we quantify the number of UMIs, we can considerably simplify the description by splitting the workflow into the initial cDNA synthesis and all downstream steps. For the cDNA synthesis, we may choose one of two models:

$$\begin{array}{l}\mathcal{X}_i \rightarrow \mathcal{X}_i + \mathcal{T}_i \text{ or} \\ \mathcal{X}_i \rightarrow \mathcal{T}_i.\end{array} \quad \text{(Equation 66)}$$

In the first model, the formation of a UMI-tagged cDNA $\mathcal{T}_i$ is non-sequestering, and the template RNA $\mathcal{X}_i$ can participate in further cDNA synthesis. In other words, a single RNA molecule can produce more than one cDNA with distinct UMIs. In the second model, the cDNA synthesis is sequestering, and each RNA can template at most one cDNA with a particular UMI. For the downstream steps, if we assume the PCR and sequencing steps produce results that are reasonably faithful to their templates, we are essentially restricted to a single model:

$$\mathcal{T}_i \rightarrow \varnothing. \quad \text{(Equation 67)}$$

In other words, the sequence of steps after the formation of cDNA $\mathcal{T}_i$ may lose some UMIs, but it cannot create them. Aggregating these steps, we find the shifted per-molecule generating function for technical noise:

$$\begin{aligned} G_{ti}^* &= G_{ti} - 1 = e^{\lambda_i(g_i-1)} - 1 = e^{\lambda_i u_i} - 1 \ \text{(non-sequestering)} \\ &= p_i g_i + (1-p_i) - 1 = p_i u_i \ \text{(sequestering)}, \end{aligned} \quad \text{(Equation 68)}$$

where $\lambda_i = \lambda_{i,c} p_{i,p}$ and $p_i = p_{i,c} p_{i,p}$. $\lambda_{i,c}$ is the overall Poisson rate of the catalytic production of cDNA $\mathcal{T}_i$ with distinct UMIs, $p_{i,c}$ is the probability of producing a single cDNA $\mathcal{T}_i$ in a non-catalytic fashion, and $p_{i,p}$ is the probability of retaining a molecule of $\mathcal{T}_i$ through the PCR steps. It is straightforward to use a Taylor expansion to observe that the limit $\lambda_{i,c} \ll 1$ yields the Bernoulli form: if non-sequestering sequencing is relatively slow or inefficient, the probability of obtaining multiple cDNA from a single RNA is low, and the mathematically simpler Bernoulli noise form approximately holds.[16,133]

Using the properties of PGFs,[21] we find that the overall generating function is given by a simple composition, substituting G_{ti} for g_i:

$$G_{\text{tot,t}} = G_{\text{tot}}\left(\mathbf{G}_t^*\right), \quad \text{(Equation 69)}$$

where we use the $G_{\text{tot}}(\mathbf{u})$ parametrization, and each entry of $\mathbf{G}_t^*$ contains the shifted generating function G_{ti}^* for a particular species i.

Finally, the reads associated with each cDNA $\mathcal{T}$ are not always uniquely identifiable: for example, the sequence content is typically sufficient to identify the gene, but if a read only covers an exonic portion of the gene, it is impossible to distinguish whether or not the original molecule has been spliced.[140] To correctly represent this ambiguity, we need to transform the arguments of the generating function from a length-n vector to a length n-vector, such that n is the total number of mutually distinguishable classes of molecules. The simplest form of this transformation is a linear categorical partition:

$$\mathbf{g} = \mathcal{P}^a \boldsymbol{g}, \quad \text{(Equation 70)}$$

where $\mathcal{P}^a$ is an $n \times n$ ambiguity matrix with $\mathcal{P}^a_{i,j}$ giving the probability of molecule i being identifiable in the equivalence class i. We assume that each molecule can be assigned to at least one class, implying $\sum_i \mathcal{P}^a_{i,j} = 1$. In principle, only the constraint $\sum_i \mathcal{P}^a_{i,j} \leq 1$ is mandatory, but the loss of molecules can be equivalently reframed as a technical noise component in $\mathbf{G}_t^*$.

We discuss the general case of this model component in supplemental information section notes on ambiguity. In summary, the entries of $\mathcal{P}^a$ are challenging to identify, but it may be possible to exploit genomic information, polymer physics, and orthogonal long-read sequencing data to construct it from first principles. This formulation admits several special cases. For example, if we cannot distinguish any distinct species at all and can only quantify the total RNA content, $n = 1$ and $\mathcal{P}^a_{i,j} = 1$ for each i. Then we obtain

$$\begin{array}{l}(\mathbf{g})_i = g \text{ for all } i \text{ and} \\ G(g) = G\left(\begin{bmatrix} g \\ \vdots \\ g \end{bmatrix}\right).\end{array} \quad \text{(Equation 71)}$$

On the other hand, if all species are perfectly identifiable, we obtain $n = n$ and $\mathcal{P}^a = I_n$, the n-dimensional identity matrix. If, say, we have $n = 2$ but $n = 3$, as in the case of nascent, mature, and ambiguous molecules described in La Manno et al.[86] and Eldjárn Hjörleifsson et al.,[140] we obtain

$$G(\mathbf{g}) = G\left(\begin{bmatrix} \mathcal{P}^a_{1,1} g_1 + \mathcal{P}^a_{1,3} g_3 \\ \mathcal{P}^a_{2,2} g_2 + \mathcal{P}^a_{2,3} g_3 \end{bmatrix}\right), \quad \text{(Equation 72)}$$

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

where g_1 and g_2 correspond to two unambiguously identifiable species, whereas g_3 corresponds to ambiguous cDNA which may have come from either. In the general case, we find

$$\begin{aligned}\mathbf{u} &= \mathcal{P}^a\boldsymbol{g} - 1\\ &= \mathcal{P}^a(\boldsymbol{u}+1) - 1\\ &= \mathcal{P}^a\boldsymbol{u}\\ &= \mathbf{G}_a(\boldsymbol{u}) - 1 := \mathbf{G}_a^*(\boldsymbol{u}),\end{aligned} \tag{Equation 73}$$

where each entry of the vector $\mathbf{G}_a$ contains the generating function of the relevant categorical distribution that governs how species i is parsed as one of the n identifiable species:

$$(\mathbf{G}_a(\boldsymbol{u}))_i = \sum_i \mathcal{P}^a_{ij} g_i. \tag{Equation 74}$$

Therefore, the overall GF takes the following form:

$$G_{\text{tot,ta}} = G_{\text{tot,t}}\left(\mathbf{G}_a^*(\boldsymbol{u})\right). \tag{Equation 75}$$

Example systems

The equation above provides a generic, modular framework for characterizing variability in sequencing experiments. To fit it to data, we need to specify a particular set of models for each step of the process. To do so, we should first strive to understand which modular components are realistic based on relatively simple summaries of data. Further, the process of evaluating and fitting these models is fairly involved, and often requires substantial up-front work to design scalable solvers. Therefore, it is useful to understand their qualitative behaviors relevant to statistical inquiry. In the current section, we characterize some analytically tractable systems, as well as their identifiability properties, such as our ability to distinguish between different models and parameter regimes. To illustrate these points, we apply the models to real and simulated data and speculate about their implications and physical relevance.

Special theoretical cases

We revisit section generating function methods for biological stochasticity to emphasize the implications and advantages of unifying the discrete and continuous degrees of freedom of the biological model in a common framework. The similarity of the discrete and continuous generating function terms is not accidental, and follows directly from the Poisson representation.[93] Occasionally, we can exploit this representation to bypass calculations for discrete processes by referring to results from the study of continuous processes, and vice versa. This approach consists of writing down the generating function PDE for a discrete process, identifying a continuous process governed by the same PDE, obtaining its solution from the stochastic process literature, and asserting that the discrete process distribution is given by compounding a Poisson distribution with the continuous law.

For instance, we may consider the case of a system with constitutive transcription at rate α, autocatalysis at rate q, and degradation at rate γ ($N = 1, n = 1, m = 0$):

$$\begin{aligned}\varnothing \xrightarrow{\alpha} \mathcal{X} \xrightarrow{\gamma} \varnothing,\\ \mathcal{X} \xrightarrow{q} 2\mathcal{X}.\end{aligned} \tag{Equation 76}$$

We can represent these reactions by the matrices $C = -\gamma + q$ and $D = q$, as well as the operator $\mathcal{A}(u) = \alpha u$. This system was introduced, but not treated, in Jahnke and Huisinga,[85] and, to our knowledge, first solved with master equation and generating function calculations by Vastola.[17] However, we can also solve it merely by matching terms, without any new calculations. We provide the full details of the parameter-matching process in supplemental information section "Poisson representation isomorphisms." The derivation consists of noticing that the functional form of C, D, and $\mathcal{A}$ can also arise from an $N = 1, n = 0, m = 1$ system with drift α, square-root noise $\sigma = \sqrt{2q}$, and mean-reversion at the rate $\gamma - q$. This is the Cox–Ingersoll–Ross (CIR) process, a popular mathematical finance model of interest rates.[175,176] Its stationary distribution is gamma with shape α/q and scale $\frac{q}{\gamma - q}$. This immediately implies the distribution of the discrete process is negative binomial with the same shape and scale. This matches the result obtained by directly solving the master equation.[18] We find, then, that autocatalysis with constitutive transcription yields a stationary distribution equivalent to bursty transcription with no autocatalysis.

Obtaining this result, we may ask how the distribution changes if the molecules are produced in geometric bursts B with mean size b:

$$\varnothing \xrightarrow{\alpha} B \times \mathcal{X} \xrightarrow{\gamma} \varnothing\, \mathcal{X} \xrightarrow{q} 2\mathcal{X}. \tag{Equation 77}$$

By changing the drift operator to a jump operator, we obtain a PDE with $\mathcal{A}(u) = \alpha\left[\frac{1}{1-bu} - 1\right]$. In other words, the continuous version of this process is a combination of CIR and gamma Ornstein–Uhlenbeck (Γ-OU) processes,[20] with the mean-reversion terms of both, the square-root noise of the former, and the exponentially-distributed jumps of the latter.

Define the parameter combinations

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$c := \gamma - q$$
$$\nu := \frac{\alpha b}{bc - q}. \quad \text{(Equation 78)}$$

By direct integration, we find the characteristic and the stationary distribution

$$U(\mathrm{s}) = \frac{cue^{-c\mathrm{s}}}{c+qu(e^{-c\mathrm{s}} - 1)}$$
$$G = \exp\left[\alpha \int_0^\infty \frac{bU(\mathrm{s})}{1 - bU(\mathrm{s})} d\mathrm{s}\right] = \left(\frac{1 - qc^{-1}u}{1 - bu}\right)^\nu. \quad \text{(Equation 79)}$$

Curiously, this distribution exactly matches the *transient* MGF of the Γ-OU process, as well as the equivalent transient PGF of the bursty transcription process with no autocatalysis[16]:

$$G = \left(\frac{1 - bue^{-\kappa\tau}}{1 - bu}\right)^\nu; \quad \text{(Equation 80)}$$

we may take advantage of the fact that qc^{-1} can be equivalently expressed as $be^{-\kappa\tau} < b$ for some positive k and t, because $bc - q > 0$ to have a steady state (i.e., positive ν). In the continuous setting, this process is known[177] to have a law consisting of a mixture of gamma distributions with scale $be^{-\kappa\tau}$ and shape k; in turn, k is drawn from a negative binomial distribution with shape ν and scale $(1+e^{-\kappa\tau})^{-1}$. This immediately implies that the distribution of the corresponding discrete process is a negative binomial-negative binomial mixture with equivalent parameters, which may be confirmed by the considerably more involved direct derivation in supplemental information section "Poisson representation isomorphisms." Although this distribution cannot be expressed in closed form, its construction makes the simulation of the bursty transient and stationary autocatalytic processes trivial, and suggests that simple finite approximations (i.e., up to a modest k) may be developed.

The continuous formulation is a way to exploit existing quantitative results, but does not typically make problems easier. For example, we may be interested in solving an RNA/protein system with transcription, catalytic translation (at rate q), and the degradation of both species (at respective rates γ_R and γ_P). Without specifying the transcriptional dynamics, we find that the downstream ODEs have a nontrivial D matrix, i.e.,

$$C = \left(C^{dd}\right)^T = \begin{bmatrix} -\gamma_R & q \\ 0 & -\gamma_P \end{bmatrix} \text{ and } D = \left(Q^d\right)^T = \begin{bmatrix} 0 & q \\ 0 & 0 \end{bmatrix}. \quad \text{(Equation 81)}$$

Although these matrices *can* be exploited to obtain both characteristics, the solution depends on special functions and is thus challenging to manipulate.[77] Instead, we may ask whether we can simplify the problem by eliding all stochasticity in the protein species and assuming it may be described by a continuous process. Defining the variables for this system, we find:

$$C = \begin{bmatrix} \left(C^{dd}\right)^T & \left(C^{dc}\right)^T \\ 0 & \left(C^{cc}\right)^T \end{bmatrix} = \begin{bmatrix} -\gamma_R & q \\ 0 & -\gamma_P \end{bmatrix}$$
$$D = \begin{bmatrix} 0 & \left(C^{dc}\right)^T \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & q \\ 0 & 0 \end{bmatrix}, \quad \text{(Equation 82)}$$

i.e., in spite of this supposed simplification, the problem is precisely as challenging as it was before. This provides an immediate and intuitive explanation for a range of results, such as the observation that the stationary distribution of proteins under constitutive transcription has a complicated solution in terms of Kummer's hypergeometric function even if one uses a leading-order approximation (cf. Equations 34 and 50 of Bokes[139]).

Empty droplets

Model definition. In Equation 64, we propose the simplest nontrivial model for the background distribution of RNA molecules in each droplet: the RNA content for each species i is described by a set of independent Poisson distributions whose mean is proportional to the mean in the entire cell population. Per Equation 65, the distribution of background is convolved with the endogenous RNA distribution of cell-containing droplets, making it challenging to distinguish technical and biological contributions. However, we *can* make predictions about the empty droplets, which have $G_{\mathrm{bc}} = 1$, and compare these predictions to real datasets.

First, we define a baseline $n = 2$ model of biology, such that

$$\varnothing \xrightarrow{K} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \quad \text{(Equation 83)}$$

where K is a generic, but non-constant (bursty, multistate, or SDE-controlled) transcription process, $\mathcal{X}_N$ is a *n*ascent transcript, $\mathcal{X}_M$ is a *m*ature transcript, and β and γ are Markovian splicing and degradation rates, respectively. As the case of constant K yields a Poisson distribution of $\mathcal{X}_N$ and $\mathcal{X}_M$, the case of variable K induces an overdispersed distribution of RNA in droplets with one or more cells. Further, it implies that certain correlations are nonzero. For a given gene j, the correlation between counts of $\mathcal{X}_N$ and $\mathcal{X}_M$ should

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

be nonzero, as the latter is, conceptually, the moving average of the former. Further, the correlation between the counts of a given species for different genes should be nonzero, as it reflects cell type heterogeneity and gene co-regulation[16] (see section coupling multiple genes).

This model describes the biology in living cells; to connect it to UMI measurements, we assume that $\mathbf{G}_t^*$ is an approximately linear map, i.e., library construction is either sequestering or non-sequestering and slow. Further, we assume $\mathbf{G}_a^*$ is a linear map, as in Equation 74. Therefore, for each species i, we have a per-cell biological distribution with mean μ_i. In a droplet with a single cell, the mean becomes $\mu_i p_i(1+c) \approx \mu_i p_i$, such that p_i is the overall probability of capturing, retaining, sequencing, and identifying each molecule (section library construction and sequencing noise). In a droplet with no cells, the mean is $c\mu_i p_i$. We assume the number of doublets is negligible.

Under the foregoing assumptions, we predict that the empty-droplet marginal per-gene UMI distribution is Poisson with mean $c\mu_i p_i$. This mean is proportional to the mean in non-empty droplets with a small coefficient of proportionality c. Further, we should observe zero correlations on an intra-gene basis, between counts of $\mathcal{X}_{j,N}$ and $\mathcal{X}_{j,M}$, and on an inter-gene basis, e.g., between counts of $\mathcal{X}_{j_1,M}$ and $\mathcal{X}_{j_2,M}$. However, it is not *a priori* clear that this model should even approximately describe real data, even in the case of empty droplets. For example, these data may exhibit considerable "read depth" variability,[65,83] or, in our framework, inter-droplet variation in the probability p_i, which would induce overdispersion or genome-wide correlations between molecule counts. By inspecting the distributional properties of empty droplet data, we can attempt to qualitatively motivate or raise doubts regarding the Poisson model.

Data processing. To build references and pseudoalign datasets, we used *kallisto* | *bustools* 0.26.0. We downloaded pre-built *H. sapiens* and *M. musculus* genomes from https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest (10x Genomics, GRCh38 and mm10, 2020-A versions). Next, we used the kb ref function with the—lamanno option to build references. We obtained the raw FASTQ files for the six datasets reported in Table S2. Then, we used the kb count function with the—lamanno option, as well as the appropriate (10x v2 or v3) whitelist option -x to quantify the datasets, outputting unspliced and spliced RNA matrices. The unspliced counts correspond to molecular barcodes containing introns, whereas the spliced counts correspond to molecular barcodes not containing introns.[140,178] For the reasons outlined in Section S6 of Carilli et al.,[71] we identify unspliced counts with "nascent" RNA species and spliced counts with "mature" RNA species, and elide any ambiguity.

Data analysis. We split the datasets into two categories. The "non-empty" droplets were retained after the *bustools* filter; the "empty" category contains barcodes that were discarded by the filter. Although this split is fairly coarse, as the filtering choices are heuristic, it is coherent with typical processing workflows and allows us to inspect the broad trends of distributional properties.

To investigate the overdispersion, or lack thereof, we separately computed the mean and variance of nascent and mature UMI counts for each gene in each set of cells. We plotted these quantities on a log-log scale, omitting the data points where one or both of these quantities were zero. Under the pseudobulk model, we expect the non-empty droplets to exhibit overdispersion and the empty droplets to be near identity, as the model encodes Poisson statistics for the latter.

To investigate the intra-gene correlation structure, we computed the Pearson correlation coefficient ρ between nascent and mature UMI counts for each gene in each set of cells. We plotted the histograms of these values, as well as their relationship to the mature UMI mean, omitting the data points where ρ was undefined. To investigate the inter-gene correlation structure, we computed the Pearson correlation coefficient between the nascent UMI counts for each pair of genes in each set of cells, and repeated the analysis for mature count data. We plotted the histograms of these values, omitting the data points where ρ was undefined. As the number of gene pairs is fairly large, we first excluded all genes that were not expressed in the dataset. We expect both measures of correlation to be substantial for non-empty droplets and near zero for the empty droplets, as the model encodes statistical independence between marginal distributions for the latter.

To investigate the relationship between the empty and non-empty droplet averages, we plotted the mean mature UMI count for each gene in empty droplets against the mean mature UMI count in cell-containing droplets. As we plotted these quantities on a log-log scale, we omitted the data points where one or both of these quantities were zero. We repeated the analysis for mature RNA data. We expect these averages to be highly correlated, as the pseudobulk model proposes that the background RNA are sampled from a pool representative of the cell population.

Next, we computed and reported the Pearson correlation coefficient between the (well-defined) log-means. To characterize and explain deviations from Poisson behavior, we selected all genes with overdispersion in the mature RNA count distributions in empty droplets ($\sigma_M^2 > 2 \times \mu_M$) and reported their identities. Finally, to quantify the variation not included in the model, we computed the mean and variance of total mature UMI counts in empty droplets, with and without the overdispersed genes. As the sum of independent Poisson distributions is Poisson, we expect the total per-cell UMI count distributions to have a variance approximately equal to the mean.

Noise-corrupted candidate models of transcriptional variation

Model definition. We would like to characterize the mutual distinguishability of superficially similar transcriptional models. In particular, we are interested in the benefits of multimodal data collection and the effects of technical noise.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

As above, we begin by defining a baseline $n = 2$ model of biology, such that

$$\varnothing \xrightarrow{K} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{Equation 84}$$

where K represents one of three candidate transcriptional models. The discrete dynamics are summarized by

$$\begin{aligned} C^{dd} &= \begin{bmatrix} -\beta & 0 \\ \beta & -\gamma \end{bmatrix} \\ U_M &= u_M e^{-\gamma s} \\ U_N &= u_N e^{-\beta s} + \frac{u_M \beta}{\beta - \gamma}\left(e^{-\gamma s} - e^{-\beta s}\right). \end{aligned} \tag{Equation 85}$$

The first transcriptional model is the Γ-OU process, with $N = 1$ and $m = 1$:

$$dy_t = -\kappa y_t dt + dZ_t, \tag{Equation 86}$$

where Z_t is a subordinator with arrival rate a and exponentially distributed jumps with mean size θ. This system is characterized by

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} u_N \\ u_M \\ u_K \end{bmatrix}, C^{cc} = -\kappa, C^{cd} = \begin{bmatrix} 1 & 0 \end{bmatrix} \\ \mathcal{A}(\mathbf{u}) &= a\left[\frac{1}{1 - \theta u_K} - 1\right], \end{aligned} \tag{Equation 87}$$

with all other matrices and operators set to zero.

The second is the CIR process, with $N = 1$ and $m = 1$:

$$dy_t = (a\theta - \kappa y_t)dt + \sqrt{2\kappa\theta y_t} dW_t. \tag{Equation 88}$$

This system is characterized by

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} u_N \\ u_M \\ u_K \end{bmatrix}, C^{cc} = -\kappa, C^{cd} = \begin{bmatrix} 1 & 0 \end{bmatrix}, Q^c = \kappa\theta \\ \mathcal{A}(\mathbf{u}) &= a\theta u_K, \end{aligned} \tag{Equation 89}$$

with all other matrices and operators set to zero.

We previously proposed the Γ-OU and CIR processes as potential explanatory models for gamma-distributed stochastic variability in transcription rates, solved them, and investigated the implications of their kinetics on the model properties and distinguishability.[20] The stationary distribution of the Γ-OU and CIR processes is gamma, with shape a/κ and scale θ, i.e., mean $a\theta/\kappa$ and variance $a\theta^2/\kappa$. In addition, their (appropriately normalized) autocorrelation function is $e^{-\kappa t}$.

Finally, the third is the telegraph process,[100] with $N = 2$ and $m = 0$. This system is characterized by

$$\mathbf{u} = \begin{bmatrix} u_N \\ u_M \end{bmatrix}, H = \begin{bmatrix} -k_{\text{on}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{off}} \end{bmatrix}, \text{and } \mathcal{A}(\mathbf{u}) = \begin{bmatrix} 0 \\ k_{\text{init}} u_N \end{bmatrix}. \tag{Equation 90}$$

The stationary distribution of this process is Bernoulli scaled by k_{init}, with mean $\frac{k_{\text{on}} k_{\text{init}}}{k_{\text{on}} + k_{\text{off}}}$ and variance $\frac{k_{\text{on}} k_{\text{off}} k_{\text{init}}^2}{(k_{\text{on}} + k_{\text{off}})^2}$. Its autocorrelation function is $e^{-(k_{\text{on}} + k_{\text{off}})t}$.[81]

For all three models, assuming a Bernoulli observation model (i.e., that each molecule has an independent probability p of being observed) is equivalent to a parameter redefinition. For the Γ-OU and CIR models, this redefinition is that $\theta \to p\theta$; for the telegraph model, we have analogously that $k_{\text{init}} \to p k_{\text{init}}$.

Let us see why this is true. Recall from section stochastic modeling of single-cell biology that the Bernoulli technical noise model amounts to a redefinition $u_N \to p u_N$, $u_M \to p u_M$. For the Γ-OU model, the steady-state (log-) GF is

$$\phi_{ss}(u_N, u_M) = a \int_0^\infty \frac{\theta U_K(s; u_N, u_M)}{1 - \theta U_K(s; u_N, u_M)} ds, \tag{Equation 91}$$

where $U_K(s; u_N, u_M)$ is the exponential sum solution of

$$\frac{dU_K}{ds} = U_N - \kappa U_K, \quad U_K(0) = 0, \tag{Equation 92}$$

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

and where the characteristics U_N and U_M are as in Equation 85. Because the U_K ODE is linear, U_K depends linearly on u_N and u_M (and hence on p). But ϕ_{ss} only depends on U_K through the combination θU_K, so the problem with technical noise is equivalent to redefining $\theta \to p\theta$.

For the CIR model, the steady-state (log-) GF is

$$\phi_{ss}(u_N, u_M) = a\theta \int_0^\infty U_K(\mathrm{s}; u_N, u_M)\ d\mathrm{s}, \tag{Equation 93}$$

where

$$\frac{dU_K}{d\mathrm{s}} = U_N - \kappa U_K + \kappa\theta\ U_K^2, \quad U_K(0) = 0.$$

The technical noise causes $U_N \to pU_N$. Divide both sides by p, so that the p factor is moved elsewhere; we can see that

$$\phi_{ss}(u_N, u_M) = ap\theta \int_0^t \frac{U_K(\mathrm{s}; u_N, u_M)}{p}\ d\mathrm{s} \tag{Equation 94}$$

$$\frac{d(U_K/p)}{d\mathrm{s}} = U_N - \kappa(U_K/p) + \kappa p\theta\ (U_K/p)^2 \quad U_K(0) = 0$$

is equivalent, i.e., that again the technical noise problem is equivalent to a non-technical-noise problem with $\theta \to p\theta$.

For the telegraph model, the steady-state (log-) GF is

$$\phi_{ss}(u_N, u_M) = \phi^0(U_N(\infty), U_M(\infty), U_{\mathrm{on}}(\infty), U_{\mathrm{off}}(\infty)) \tag{Equation 95}$$

$$\frac{dU_{\mathrm{off}}}{d\mathrm{s}} = -k_{\mathrm{on}}(U_{\mathrm{off}} - U_{\mathrm{on}})$$
$$\frac{dU_{\mathrm{on}}}{d\mathrm{s}} = -k_{\mathrm{off}}(U_{\mathrm{on}} - U_{\mathrm{off}}) + k_{\mathrm{init}}(U_{\mathrm{on}} + 1)U_N,$$

where $U_{\mathrm{off}}(0) = U_{\mathrm{on}}(0) = 0$. Since $U_N(\infty) = U_M(\infty) = 0$, the values of $U_N(\mathrm{s})$ only affect ϕ_{ss} through the combination $k_{\mathrm{init}}U_N$ that appears in the U_{on} ODE; this means we can just redefine $k_{\mathrm{init}} \to pk_{\mathrm{init}}$ as promised to get a completely equivalent problem.

Model analysis. Formally, these models have five parameters each: three for the upstream transcriptional dynamics and two for the downstream molecular conversion. However, their qualitative behaviors at steady state can be effectively summarized by fixing μ_K, β, and γ, and varying two key parameters, the timescale separation and the noise intensity. From a statistical point of view, μ_K/β and μ_K/γ are easily and robustly identifiable from the mean molecular counts; from an experimental point of view, β and γ can, in principle, be fit by orthogonal experiments.[86] At steady state, the value of μ_K is a somewhat arbitrary scaling factor.

For the two-species SDE driver models, the qualitative parameters take the following form:

$$\begin{aligned}\text{timescale separation}: &= \mathrm{x} = \frac{\kappa}{\kappa+\beta+\gamma}\\ \text{noise intensity}: &= \mathrm{y} = \frac{\theta}{a+\theta}.\end{aligned} \tag{Equation 96}$$

These parameters both range in (0,1). When the timescale separation approaches zero, the transcriptional variation is much slower than the turnover, and the distribution of RNA is given by a simple Poisson mixture of the law of K. When the noise intensity approaches zero, the law of K degenerates and the distribution of RNA becomes Poisson. Most interestingly, when the timescale separation and the noise intensity are both high, the system exhibits bursty transcription.[20]

Equation 96 is defined with reference to the process parameters of the Γ-OU and CIR drivers.[20] It remains to define κ, θ, and a in terms of k_{on}, k_{off}, and k_{init} for the telegraph process. The correct identification is:

$$\begin{aligned}\kappa &= k_{\mathrm{on}} + k_{\mathrm{off}} \text{ is the autocorrelation timescale,}\\ a &= \frac{k_{\mathrm{on}}\kappa}{k_{\mathrm{off}}} \text{ is the process scaling, and}\\ \theta &= \frac{k_{\mathrm{off}}k_{\mathrm{init}}}{\kappa} \text{ is the gain.}\end{aligned} \tag{Equation 97}$$

These identifications are not arbitrary, as they endow the system with lower moments that match the SDE formulation: autocorrelation function $e^{-\kappa t}$, mean $a\theta/\kappa$, and variance $\theta\mu_K$. In addition, the system has the correct geometric burst limit $(k_{\text{init}}, k_{\text{off}} \to \infty)$ with burst size $\theta/\kappa \to k_{\text{init}}/k_{\text{off}}$ and burst frequency $a \to k_{\text{on}}$[73]; this limit matches the Γ-OU one.[20]

Given any combination of $\{\mathrm{x}, \mathrm{y}, \mu_K, \beta, \gamma\}$, we can identify the transcriptional parameters:

$$\kappa = \frac{(\beta+\gamma)\mathrm{x}}{1-\mathrm{x}}$$
$$\frac{a}{\theta} = \frac{k_{\text{on}}\kappa}{k_{\text{off}}}\frac{\kappa}{k_{\text{off}}k_{\text{init}}} = \frac{k_{\text{on}}\kappa^2}{k_{\text{off}}^2 k_{\text{init}}}$$
$$\mathrm{y} = \frac{1}{1+a/\theta} \text{ or } \frac{a}{\theta} = \frac{1}{\mathrm{y}} - 1 \qquad \text{(Equation 98)}$$
$$\frac{\mu_K}{\kappa}\left(\frac{1}{\mathrm{y}} - 1\right) = \frac{k_{\text{init}}k_{\text{on}}}{\kappa^2}\frac{k_{\text{on}}\kappa^2}{k_{\text{off}}^2 k_{\text{init}}} = \frac{k_{\text{on}}^2}{k_{\text{off}}^2} = \left(\frac{k_{\text{on}}}{k_{\text{off}}}\right)^2 := c, \text{giving}$$
$$k_{\text{on}} = \frac{\sqrt{c}\kappa}{\sqrt{c}+1}, k_{\text{off}} = \frac{\kappa}{\sqrt{c}+1}, \text{and } k_{\text{init}} = \frac{\mu_K\kappa}{k_{\text{on}}}.$$

This allows us to define a particular set of $\{\mu_K, \beta, \gamma\}$, vary x and y over the constrained domain $(0,1) \times (0,1)$, and compare the model properties for each (x, y) tuple. If we are interested in a one-species model, we simply replace each instance of $\beta + \gamma$ with β. Since the construction in Equation 98 is bijective, if we fairly densely sample the square, we can be confident that the results fully encompass the range of behaviors under a particular set of averages.

Simulated data analysis. To evaluate PMFs, we used trapezoidal quadrature for the Γ-OU generating function integral, the Runge-Kutta method for the CIR characteristic U_k and trapezoidal quadrature for the generating function integral, and the Runge-Kutta method for the telegraph model's coupled differential equations.[18,20] We marginalized over the continuous and categorical dimensions. We evaluated all PMFs on $x_N, x_M \in [0, \ldots, 49] \times [0, \ldots, 50]$. To generate synthetic data, we sampled with replacement from the 2,550 microstates in the domain, using $P(x_N, x_M)$ as sampling probabilities.

To investigate parameter identifiability, we generated 200 realizations from the Γ-OU model under $\kappa = 0.1, a = 0.4, \theta = 1, \beta = 0.8$, and $\gamma = 0.9$. These parameters lie in the "mixture-like" regime, where the transcriptional process is slower than the RNA turnover process. Next, we constructed a uniformly spaced 14×15 grid of x and y, constructed at the true values of μ_K, β, and γ and bounded by [0.01,0.99]. In statistical terms, this model formulation is the best-case scenario where no noise exists and uncertainty in the fixed parameters is negligible.

To investigate the statistical properties of one-species data, we evaluated the log-likelihood $\log L$ of the nascent marginal of the data at each of the 210 x, y coordinates (with the true value being $\mathrm{x} = 1/9$ and $\mathrm{y} = 5/7$). Next, we plotted $\log L$ as a heatmap over x,y. The coordinates with high $\log L$ are not readily distinguishable, i.e., these parameters produce very similar distributions to the data. We highlighted the coordinates in the 90th percentile of $\log L$—the least distinguishable region—using hatching. To illustrate a case where the one-species data are relatively uninformative, we considered a point with $\mathrm{x} = 9/10$ and $\mathrm{y} = 5/7$, which lies in the qualitatively different "burst-like" regime $(\kappa = 7.2)$ but closely resembles the "mixture-like" data at steady state.

To investigate the statistical properties of two-species data, we repeated the analysis above, computing the joint likelihood rather than the marginal likelihood. In the two-species model, the true "mixture-like" parameter set has $\mathrm{x} = 1/18$ and the illustrative "burst-like" parameter set has $\mathrm{x} \approx 0.81$; the other parameters do not change. To demonstrate the source of failure to distinguish between these parameter regimes, we plotted the PMFs in both. We used a transparent bar plot for the nascent PMFs and a heatmap for the joint PMFs, with darker colors representing a higher probability mass.

To investigate the mutual identifiability of models, we computed their Akaike weights over the x, y landscape. The Akaike weight of model ϖ is defined as follows:

$$w_\varpi = \frac{e^{-\frac{1}{2}\Delta_\varpi}}{\sum_k e^{-\frac{1}{2}\Delta_k}}, \text{where}$$
$$\Delta_k = \mathrm{AIC}_k - \mathrm{AIC}_{\min},$$
$$\mathrm{AIC}_{\min} = \min_k \mathrm{AIC}_k, \text{and} \qquad \text{(Equation 99)}$$
$$\mathrm{AIC}_k := -2\log L_k(\widehat{\Theta}_k) + 2\varsigma_k.$$

Thus, AIC_k is the Akaike information criterion (AIC) for model k. The AIC depends on the model log likelihood $\log L_k$ at the maximum likelihood estimate $\widehat{\Theta}_k$, as well as number of model parameters ς_k.[120] Therefore, the Akaike weight essentially transforms and combines the models' relative likelihoods to provide a measure of their agreement with the data.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Although this measure has its caveats and limitations—for example, it cannot account for uncertainty in the model-specific parameters Θ_k—it is a fairly conventional criterion for model selection. Most usefully to our investigation, it admits a simple interpretation: if the Akaike weight of the true model $w_\varpi \approx 1/3$, there is essentially no basis for choosing a particular model, since their distributions are not practically distinguishable. If $w_\varpi > 1/2$, we have a basis for model discrimination: the odds for the correct model are even. In the three-model case, this may reflect both, or only one, of the competing hypotheses being substantially worse at describing the data, so more careful examination of the w_k values is necessary to judge the models.

To investigate model identifiability, we constructed a uniformly spaced 14×15 grid of x and y, bounded by [0.01,0.99]. At each grid point, we generated 200 realizations from the Γ-OU model under $\mu_K = 5$, $\beta = 0.8$, and $\gamma = 0.9$. Next, we computed the log L_k of each model using the nascent marginal and the full data, and used the relative likelihoods to compute the Akaike weights of the Γ-OU model under these two scenarios. Finally, to reduce the impact of stochastic sampling variability, we repeated the process 50 times and computed their average. In other words, we generated 50 independent datasets at each of the 210 grid points, evaluated likelihoods of all models, computed the univariate and bivariate Γ-OU Akaike weight of each, then aggregated the 50 trials at each grid point to obtain two "average-case" performance measures. In statistical terms, this model formulation represents the best-case scenario where the parameters are perfectly known, and the problem solely consists of distinguishing between the models, as in the Γ-OU/CIR case considered in Figure 3 of Gorin and Vastola et al.[20]

To visualize the behavior of the Akaike weights under these assumptions, we plotted its value as a heatmap over x, y. We highlighted the coordinates with $w_\varpi < 1/2$—the poorly distinguishable region—using hatching. To illustrate a case where the one-species data are relatively uninformative, we compared a point with one-species coordinates $\mathrm{x,y} = (0.4, 0.9)$, which lies in the "mixture-like" regime, to one with $\mathrm{x,y} = (0.9, 0.8)$, which lies in the "burst-like" regime. We visualized these points on the x, y axes using large, color-coded circles. From Gorin and Vastola et al.[20] and the properties of low-x processes outlined in the definition of x, we expect the former regime to be highly distinguishable, particularly since the telegraph process converges to a bimodal Bernoulli mixture for $\kappa \to 0$. On the other hand, we expect the latter regime to be somewhat less distinguishable; in this limit, the Γ-OU and telegraph models both converge to the bursty model discussed in Singh and Bokes.[138] We repeated this analysis for two-species Akaike weights, transforming the coordinates appropriately (i.e., $\mathrm{x} \approx 0.24$ for the mixture-like regime and $\mathrm{x} \approx 0.81$ for the burst-like regime).

To demonstrate the basis of statistical distinguishability properties, we plotted the PMFs of the three models in the two parameter regimes. To simultaneously display them, we plotted marginal distributions of the nascent species as line charts, color-coded by the model identity.

To investigate the effect of drop-out technical noise, we did not perform dedicated simulations; instead, we exploited the result, derived above, that the functional form of the solutions is closed under downsampling. In other words, all distributional properties of a system with gain θ and the technical noise parameter p are identical to those of a system with gain $p\theta$ and no technical noise. These properties include the model distinguishability. To illustrate this result, we represented Bernoulli technical noise by arrows in the negative y direction, with small circles located on an arrow corresponding to 50%, 75%, and 85% dropout. To compute the y value under dropout, we use:

$$\mathrm{y}^* = \frac{p\theta}{p^{-1}a + p\theta}, \text{since } \mu_K = \frac{a\theta}{\kappa} = \frac{p^{-1}ap\theta}{\kappa} = \text{const.} \qquad \text{(Equation 100)}$$

The arrows begin at 0% dropout, corresponding to the illustrative base cases (large circles) described above. This demonstrates that increasing the drop-out rate while holding the averages constant leads to the molecular distributions' degeneration to the Poisson limit. If we do not hold the averages constant, we simply obtain the decreased $\mathrm{y}^* = \frac{p\theta}{a+p\theta}$ on the (less identifiable) x, y landscape with mean transcription rate $p\mu_K$.

Distributions obtained from a transient process

Model definition. As motivated in our RNA velocity review,[19] understanding transient developmental processes that occur on a timescale comparable to RNA lifetimes requires fitting transient probabilistic models. Even under the considerable simplifications made in section transient phenomena, fully treating transient transcriptional phenomena requires identifying the *a priori* unknown (1) internal-age distribution $f(t)$ as well as (2) process parameters for $G(t)$. As the time since process start t can be conceptualized as a cell-specific latent variable, this problem can be treated by an expectation–maximization (EM) algorithm, which may proceed by probabilistically constraining the unknown (3) cell-specific times t_c.

Since parameter inference is mandatory for the expectation step of the EM algorithm, we begin by characterizing the upper limit on its performance. In particular, previous attempts to treat the problem have assumed simple Gaussian or Poisson error terms,[59,86,121] or applied graph methods.[179] These approaches do not recapitulate[19] the discrete stochasticity and bursting observed in transient biophysical processes.[125,180] However, the transient distributions of bursty processes are not available in closed form, and require new algorithms. Therefore, we treat the simplest nontrivial formulation, which combines points (1) and (2), while omitting (3): if we have perfect information about the cells' relative times, can we satisfactorily fit a bursty transcriptional model and use the results as a basis for distinguishing between internal-age distributions?

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

We define a baseline $N = 1, n = 2, m = 0$ model of biology with no technical noise, with the reaction schema

$$\varnothing \xrightarrow{\alpha} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{Equation 101}$$

representing bursty transcription with stochastic burst sizes B drawn from a geometric distribution with time-dependent mean $b(t)$:

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} u_N \\ u_M \end{bmatrix}, C^{dd} = \begin{bmatrix} -\beta & 0 \\ \beta & -\gamma \end{bmatrix} \\ \mathbf{U} &= \begin{bmatrix} U_N \\ U_M \end{bmatrix} = \begin{bmatrix} u_N e^{-\beta s} + \frac{u_M \beta}{\beta - \gamma}\left(e^{-\gamma s} - e^{-\beta s}\right) \\ u_M e^{-\gamma s} \end{bmatrix} \\ \mathcal{A}(\mathbf{u}) &= \alpha\left[\frac{1}{1 - b(t)u_N} - 1\right], \end{aligned} \tag{Equation 102}$$

with all other operators set to zero. To specify $b(t)$, we define a three-stage model of cell type transitions, such that

$$b(t) = \begin{cases} b_1 & t < \tau_1, \\ b_2 & t \in (\tau_1, \tau_2), \\ b_3 & t > \tau_2, \end{cases} \tag{Equation 103}$$

i.e., a transition is accompanied by a rapid change in burst size at a deterministic time after starting the process.

Next, we propose candidate internal-age distributions. Drawing on the chemical engineering literature,[105,106] we outline one-parameter reactor models, such that $t = 0$ corresponds to the cell entering the reactor; after some residence time t, which is dependent on reactor architecture and drawn from the distribution f_{res}, the cell exits. The internal-age distribution is given by

$$f(t) = \frac{1}{T}\int_t^\infty f_{\text{res}}(t)dt. \tag{Equation 104}$$

The plug flow reactor (PFR) is the model implicit in previous studies.[59,86] Formally, it represents each cell entering a reactor, then exiting after some deterministic time T. Its residence-time distribution is Dirac or degenerate, with $f_{\text{res}}(t) = \delta(t - T)$, so

$$f(t) = \frac{1}{T}\int_t^\infty f_{\text{res}}(t)dt = \frac{1}{T}\int_t^\infty \delta(t - T)dt = \frac{\mathbb{I}(t < T)}{T}, \tag{Equation 105}$$

the expected uniform distribution. This distribution has the CDF and inverse CDF

$$F(t) = \frac{t}{T}\mathbb{I}(t < T) \text{ and } F^{-1}(p) = pT. \tag{Equation 106}$$

The continuously stirred tank reactor (CSTR) represents a cell entering a homogeneous reactor, then exiting after a random time, in a memoryless fashion. Therefore, the residence-time distribution $f_{\text{res}}(t) = \frac{1}{T}e^{-t/T}$ is memoryless or exponential, yielding

$$f(t) = \frac{1}{T}\int_t^\infty f_{\text{res}}(t)dt = \frac{1}{T^2}\int_t^\infty e^{-t/T}dt = \frac{1}{T}e^{-t/T}; \tag{Equation 107}$$

i.e., memorylessness implies that the properties inside the reactor—including the age distribution—are identical to the properties of the efflux stream. We obtain the CDF and inverse CDF

$$F(t) = 1 - e^{-t/T} \text{ and } F^{-1}(p) = -T\ln(1 - p). \tag{Equation 108}$$

The laminar-flow reactor (LFR) is a configuration between these two extremes: it represents a cell entering a reactor, remaining in it for some time deterministic time, then being able to exit after a power-law delay. Its residence-time distribution $f_{\text{res}}(t) = \frac{T^2}{2t^3}\mathbb{I}(t > T/2)$ is Pareto, yielding

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$\begin{aligned} f(t) &= \frac{1}{T}\int_t^\infty f_{\text{res}}(t)dt = \frac{T}{2}\int_t^\infty \frac{1}{t^3}\mathbb{I}(t > T/2)dt \\ &= \frac{T}{2}\int_{\max t,T/2}^\infty \frac{1}{t^3}dt \\ &= \begin{cases} \frac{T}{2}\int_t^\infty t^{-3}dt = \frac{T}{4t^2} & t > \frac{T}{2} \\ \frac{T}{2}\int_{T/2}^\infty t^{-3}dt = \frac{1}{T} & t < \frac{T}{2}. \end{cases} \end{aligned} \tag{Equation 109}$$

The PDF can be integrated to yield the CDF and inverse CDF

$$F(t) = \begin{cases} \frac{t}{T} & t < \frac{T}{2} \\ 1 - \frac{T}{4t} & t > \frac{T}{2} \end{cases} \text{ and } F^{-1}(p) = \begin{cases} pT & p < \frac{1}{2} \\ \frac{T}{4(1-p)} & p > \frac{1}{2}. \end{cases} \tag{Equation 110}$$

We are interested in the CDFs and inverse CDFs of the internal-age distributions because "perfect information about the cells' relative times" properly requires specifying $\{F_\varpi(t_c)\}$ and $\{F_\varpi(\tau_i)\}$ under the true model ϖ rather than the raw $\{t_c\}$ and $\{\tau_i\}$ values. Otherwise, the model selection problem becomes somewhat trivial; for example, if we know the mean residence time is T and we know one of $t_c > T$, we can immediately eliminate the PFR configuration without performing any calculations.

A synthetic dataset consists of observations $x_{N,c}, x_{M,c}$ for each cell c, generated from the true model ϖ at the true time point t_c. The log-likelihood of parameters $\Theta_k = \{b_1, b_2, b_3, \alpha, \beta, \gamma\}_k$ for model k takes the form

$$\log L_{k,c}(\Theta_k | x_{N,c}, x_{M,c}) = \log P(x_{N,c}, x_{M,c}, t_{c,k} | \Theta_k; \{\tau_i\}_k), \tag{Equation 111}$$

where $t_{c,k} := F_k^{-1}(F_\varpi(t_c))$ and $\{\tau_i\}_k := \{F_k^{-1}(F_\varpi(\tau_k))\}$ are the transformed times. This yields the full log-likelihood under the assumption of independence

$$\begin{aligned} \log L_k(\Theta_k) &= \sum_c \log L_{k,c}(\Theta_k | x_{N,c}, x_{M,c}, \{\tau_i\}_k) \\ &= \sum_c \log P(x_{N,c}, x_{M,c}, t_{c,k} | \Theta_k, \{\tau_i\}_k). \end{aligned} \tag{Equation 112}$$

The problem of identifying the maximum likelihood parameter set consists of optimizing Equation 112 with respect to Θ_k. The problem of reactor identification consists of using the resulting reactor-specific maximum likelihood value $\log L_k(\widehat{\Theta}_k)$ with Equation 99 to obtain the Akaike weights of each reactor configuration.

Simulated data analysis. To generate the illustrations in Figure 4A, we directly simulated cells entering and exiting each reactor configuration. First, we sampled arrival times from a uniform distribution on [0,100]. Next, we sampled residence times by inverse transform sampling from the inverse CDF corresponding to each f_{res}, using the mean residence time $T = 2$. We arbitrarily selected the observation time 75 and selected all cells which had arrived but not exited at this time. We computed the cell age by subtracting the arrival time from the current time. We repeated this procedure 10^7 times for each reactor to obtain the internal-age distribution. Next, we computed the histogram of the distribution on [0,10], using 200 bins. To account for the fact that this histogram only contains part of the CSTR and LFR densities, we rescaled the bins by the internal-age distribution's CDF value at $t = 10$. Finally, we plotted the rescaled histogram as a bar plot, and the analytical f as a line plot for comparison.

To understand the actionable differences between reactors, we simulated data from a single reactor model, then fit all three models to the obtained counts. First, we sampled 200 true reaction times $\{t_c\}$ under the PFR model with $T = 5$ and sorted them. To generate synthetic data, we used Gillespie's stochastic simulation algorithm[141,145] with a time-dependent burst size, storing the state of the system at $\{t_c\}$. We generated 200 realizations, using only one realization per time point to fit the models. To simulate, we used the parameters $\Theta_\varpi = \{b_1, b_2, b_3, \alpha, \beta, \gamma\}_\varpi = \{2, 5, 1, 0.8, 1.2, 3.14\}$. We set $\{\tau_1, \tau_2\}$ to $\{1, 3\}$. We started the system in a bivariate Poisson initial distribution with $\lambda_N^0 = \frac{\alpha b_1}{\beta}$ nascent and $\lambda_M^0 = \frac{\alpha b_1}{\gamma}$ mature molecules on average. Although this initial condition is somewhat arbitrary, as it is out of equilibrium, it is readily tractable and yields a constant mean over the first stage of the process.

The instantaneous probability $P(x_{N,c}, x_{M,c}, t_{c,k} | \Theta_k, \{\tau_i\}_k)$ is not available in closed form, and needs to be obtained by inverting the generating function for each $t_{c,k}$[16,20,138]:

$$\begin{aligned} G(\mathbf{u}, t_{c,k}) &= \exp\left(\lambda_N^0 U_N(\mathbf{u}, t_{c,k}) + \lambda_M^0 U_M(\mathbf{u}, t_{c,k})\right. \\ &\left. + \alpha \int_0^{t_{c,k}} \left[\frac{1}{1 - b(t_{c,k} - s)U_N(\mathbf{u}, s)} - 1\right] ds\right), \end{aligned} \tag{Equation 113}$$

where we elide the dependence of b on the model-specific $\{\tau_i\}_k$. For a given value of $\mathbf{u}$, it is straightforward to propagate the initial condition. However, it is impractical to compute the integral separately for each c. We can bypass this bottleneck by reusing quadrature points. Conceptually, we define the quadrature matrices

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

$$T_Q = \begin{bmatrix} b(t_0 - t_0) & 0 & \cdots & 0 & 0 \\ b(t_1 - t_0) & b(t_1 - t_1) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b(t_{\eta-1} - t_0) & b(t_{\eta-1} - t_1) & \cdots & b(t_{\eta-1} - t_{\eta-1}) & 0 \\ b(t_\eta - t_0) & b(t_\eta - t_1) & \cdots & b(t_\eta - t_{\eta-1}) & b(t_\eta - t_\eta) \end{bmatrix}$$

$$D_Q = \mathrm{diag} \begin{bmatrix} U_N(\mathbf{u}, t_{0,k}) \\ U_N(\mathbf{u}, t_{1,k}) \\ \vdots \\ U_N(\mathbf{u}, t_{\eta-1,k}) \\ U_N(\mathbf{u}, t_{\eta,k}) \end{bmatrix} \quad \text{(Equation 114)}$$

in the general case with η cells. We appended the starting grid point $t_{0,k} := 0$ to properly integrate from zero. We use the notation T_Q because this matrix is Toeplitz in the narrow, but numerically relevant,[19] case of a uniformly spaced grid approximating sampling from a PFR. To lighten the notation, we drop the subscript k from the time points in the definition of T_Q. D_Q is diagonal and does not need to be constructed explicitly; to obtain the product $T_Q D_Q$, we broadcast T_Q with the vector used in the definition of D_Q. Then, we computed $M_Q = (1 - T_Q D_Q)^{\odot(-1)} - 1$, where $\odot(-1)$ is to be interpreted as the elementwise/Hadamard inverse of the matrix. Finally, we approximated the integral by applying the *NumPy* quadrature algorithm trapz along the rows of M_Q, using $\{t_{c,k}\}$ as the integration grid.[181] The GF evaluation grid size was set to $[0, \ldots, \max x_N + 4] \times [0, \ldots, \max x_M + 4]$, where $\max x_i$ is the highest RNA count observed for species i over the entire simulation, in all cells.

Next, we used the *SciPy* algorithm optimize.minimize[182] to minimize the negative log-likelihood of the data under all three models, and obtain a satisfactory set of parameters. Specifically, we varied the 6-dimensional vector $\log_{10} \Theta$, with each log-parameter's bounds set to $(-1.5, 1.5)$. We optimized with the L-BFGS-B solver for a maximum of 20 steps. Since we are primarily interested in the models' relative performance at their maximum likelihood estimates (MLEs), rather than the process of obtaining these estimates, we initialized each search at the parameters used to generate the data.

Next, we sought to illustrate the fit performance and the differences between the models' distributions. We plotted the marginals of the simulated data at each time point t_c as bar plots, now using the counts from all 200 cells to demonstrate the full transient distribution. Next, we plotted the marginal PMFs of the three models at the corresponding time points $t_{c,k}$ as color-coded line charts. We expect the true reactor configuration (PFR) to closely agree with the distribution shapes; however, we have no *a priori* information regarding how well other reactor architectures can recapitulate the same data. To quantify the prospects for model selection, we inserted the optimal log-likelihoods into Equation 99 and calculated the Akaike weights of the model candidates.

To characterize the identifiability properties, we reproduced the simulation and analysis process using the same parameters, but varying the dataset size, with $\eta = \{20, 40, 60, 80, 100, 150, 200\}$. For each η, we generated 50 synthetic datasets, fit them, and computed the Akaike weights of the models. We plotted all w_ϖ as a function of the number of cells, adding uniform jitter to facilitate inspection. To visualize the trends in model identifiability, we plotted the mean and standard deviations of all w_ϖ for a given η, connecting them with a line to guide the eye. We do not *a priori* know whether the reactor configurations are meaningfully distinguishable, but if they are, we expect them to become more so with more data.

Next, we sought to characterize the prospects for distinguishing reactor models for a broader range of transcriptional parameters. We used rejection sampling to draw Θ_ϖ. First, we drew $\log_{10} b_i$ from a normal distribution with mean 0.8 and standard deviation 1, and all other log-parameters from a normal distribution with mean 0 and standard deviation 1. The parameters were clipped to stay in the domain $[10^{-1.4}, 10^{1.4}]$ to avoid "trivial" regimes with excessive timescale separation relative to the reactor residence time. Next, we found the highest b_i, computed the nascent and mature mean and standard deviation corresponding to this set of $b_i, \alpha, \beta, \gamma$,[138] and kept the proposed Θ_ϖ if $\mu_N + 4\sigma_N$ and $\mu_M + 4\sigma_M$ were both lower than 25. Otherwise, we regenerated Θ_ϖ. This is an *ad hoc* way to limit the state space size for PMF evaluation: although we do not know what the maximum observed counts will be until we simulate the system, $\mu + 4\sigma$ is typically provides a reasonable estimate.[97] Rejecting parameters in this fashion approximately limited the state space size to 25 × 25. In this way, we simulated, fit, and computed the Akaike weights for 200 parameter sets. All used the PFR ground truth model, $\{\tau_1, \tau_2\} = \{1, 3\}$, and $T = 5$ as above.

To summarize the model identifiability over this domain of synthetic parameters, we plotted the distribution of AIC weights w_ϖ. Finally, to characterize the relationships between the models, we plotted the distributions of log-likelihood differences $\log L_k(\widehat{\Theta}_k) - \log L_\varpi(\widehat{\Theta}_\varpi)$, where k corresponds to the CSTR and LFR models, as transparent histograms color-coded by k. If such a histogram is skewed toward negative values, the model k produces consistently worse fits than the true PFR model. In the other hand, if it is centered at zero, then model k is typically easily confused with the true model. We restricted this visualization to $(-5, 5)$ to compensate for potential failure to converge, which produces inflated likelihood differences. This visualization provides a basis for explaining the distribution of w_ϖ.

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

Variability in library construction

Model definition. In section noise-corrupted candidate models of transcriptional variation, we considered the parameter and model identifiability for a two-stage model of RNA processing, and found that several interesting distributions are closed under downsampling, so long as the downsampling is Bernoulli with equal parameters for both species. However, this assumption may be too restrictive in practice: for example, nascent RNA may be more or less likely to be captured than mature RNA, depending on the poly(A) content of their introns. In the current section, we investigate the behavior of models with differences in capture probabilities or rates.

The identifiability properties are highly model-dependent. For example, if we consider the Γ-OU or CIR models, with $N = 1$, $n = 2$, $m = 1$, such that

$$\varnothing \xrightarrow{K} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing, \tag{Equation 115}$$

where the autocorrelation of K is $\kappa \ll \beta, \gamma$, the stationary distribution of K is gamma with shape $\nu = a/\kappa$ and scale θ. We find the stationary RNA generating function is bivariate negative binomial, with

$$G = \left(\frac{1}{1 - \frac{\theta u_N}{\beta} - \frac{\theta u_M}{\gamma}}\right)^{\nu}, \tag{Equation 116}$$

which is outlined in the Section S2.5.2 of Gorin and Vastola et al.[20] Under sampling, the distribution stays bivariate negative binomial, with GF

$$G = \left(\frac{1}{1 - \frac{\theta p_N u_N}{\beta} - \frac{\theta p_M u_M}{\gamma}}\right)^{\nu}. \tag{Equation 117}$$

In other words, even if we have perfect information about this distribution's three parameters ν, $\theta p_N/\beta$, and $\theta p_M/\gamma$, we cannot conclude anything about the magnitudes of p_N and p_M, as they are degenerate with θ, β, and γ. If K is telegraph (i.e., $N = 2$, $n = 2$, $m = 0$), we obtain a finite Poisson mixture:

$$G = \frac{k_{\text{off}}}{\kappa} + \frac{k_{\text{on}}}{\kappa} \exp\left(\frac{k_{\text{init}} p_N u_N}{\beta} + \frac{k_{\text{init}} p_M u_M}{\gamma}\right), \tag{Equation 118}$$

which exhibits the same degeneracy with respect to k_{init}, β, and γ. Entirely analogously, if the system is in the Poisson limit $(y \approx 0)$ with average transcriptional strength μ_K, we find that sampling yields

$$G = \exp\left(\frac{\mu_K p_N u_N}{\beta} + \frac{\mu_K p_M u_M}{\gamma}\right), \tag{Equation 119}$$

which is non-identifiable.

Interestingly, the bursty regime *is* partially identifiable. We begin by defining a baseline $N = 1, n = 2, m = 0$ model of biology with technical noise but no ambiguity, such that

$$\varnothing \xrightarrow{\alpha} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing \tag{Equation 120}$$

representing bursty transcription with stochastic burst sizes B drawn from a geometric distribution with constant mean b. Further, we assume that a molecule $\mathcal{X}_i$ is retained with probability p_i, yielding:

$$\begin{aligned} G_t^*(\mathbf{u}) &= \begin{bmatrix} p_N u_N \\ p_M u_M \end{bmatrix}, C^{dd} = \begin{bmatrix} -\beta & 0 \\ \beta & -\gamma \end{bmatrix} \\ \mathbf{U}\left(G_t^*(\mathbf{u}), s\right) &= \begin{bmatrix} p_N u_N e^{-\beta s} + \frac{p_M u_M \beta}{\beta - \gamma}\left(e^{-\gamma s} - e^{-\beta s}\right) \\ p_M u_M e^{-\gamma s} \end{bmatrix} \\ \mathcal{A}(\mathbf{u}) &= \alpha\left[\frac{1}{1 - b u_N} - 1\right]. \end{aligned} \tag{Equation 121}$$

In other words, the stationary generating function is given by

$$\exp\left(\alpha \int_0^{\infty} \mathcal{A}\left(\mathbf{U}\left(G_t^*(\mathbf{u}, s)\right)\right) ds\right). \tag{Equation 122}$$

In principle, this quantity can be integrated, inverted, and optimized with respect to the parameters. However, to be thorough, we need to reformulate the optimization problem in the most compact form available, which involves identifying the distribution's

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

degeneracies. Although this system formally has six parameters $b,\alpha,\beta,\gamma,p_N,p_M$, at steady state only four are identifiable. This is made clear by examining the integrand:

$$\begin{aligned} bU_N &= bp_N u_N e^{-\beta s} + b\frac{p_M u_M \beta}{\beta - \gamma}\left(e^{-\gamma s} - e^{-\beta s}\right) \\ &= bp_N\left[u_N e^{-\beta s} + \frac{p_M}{p_N}\frac{u_M \beta}{\beta - \gamma}\left(e^{-\gamma s} - e^{-\beta s}\right)\right], \end{aligned} \quad \text{(Equation 123)}$$

i.e., the characteristic is invariant so long as bp_N and p_M/p_N are constant. By plugging in zero for u_N or u_M, we observe that the characteristics take the functional form of the characteristics of the noise-free system, implying different values of p_N and p_M may give indistinguishable distributions. Therefore, identifying the relationship between p_N and p_M requires bivariate data. To quantitatively characterize *how* identifiable p_N and p_M are, we need to use simulations.

However, challenges particular to single-cell technologies arise when attempting to apply this model to large datasets with many genes. Although the Bernoulli model is a useful approximation, considering the sequencing process suggests that the non-sequestering technical noise model is more realistic: there is no chemical barrier to an RNA molecule being captured multiple times. In this formulation, each gene's technical noise is parametrized by the species' overall capture rates λ_N and λ_M, which produce the Bernoulli limit when both of these parameters are small.

Furthermore, it appears implausible that $\lambda_{j,N}$ and $\lambda_{j,M}$, where j indexes over genes, vary arbitrarily. In a previous report,[21] we have found that the model $\lambda_{j,N} = C_N L_j$ and $\lambda_{j,M} = \lambda_M$ performs satisfactorily. In this model, the nascent species are identified with unspliced molecules, which are considerably longer than spliced molecules and contain a large number of internal poly(A) priming sites. To a first-order approximation, we may propose that nascent species are captured at a rate proportional to the gene length L_j, where the constant of proportionality C_N is a dataset-wide technical noise parameter. Analogously, we identify the mature species with fully spliced, poly(A)-tailed molecules, and make the zeroth-order approximation that poly(A) tails are chemically identical. The capture rate λ_M is, then, also dataset-wide. Although this model is relatively simplistic, it foregrounds a key challenge. Even if we assume different genes' transcriptional processes are independent, we cannot fit their distributions independently, as we need to account for coupling through the technical noise parameters.

Data analysis To illustrate the identifiability of p_M/p_N under the Bernoulli noise model, we considered the likelihood landscape for the simplest one-parameter formulation. We fixed the parameters $\alpha = 1, bp_N = 4.9, \mu_N = \frac{\alpha bp_N}{\beta} = 7$, and $\mu_M = \frac{\alpha bp_M}{\gamma} = 10$; in other words, the nascent RNA distribution is negative binomial with shape $\frac{\alpha}{\beta} \approx 1.43$ and scale bp_N. We simulated data at $p_M/\ p_N \in \{1/4, 1, 4\}$, with $\eta = \{20, 50, 100, 200\}$ simulated cells. For each of the true p_M/p_N and η values, we generated 200 datasets by sampling from the PMF on $[0,\ldots,99] \times [0,\ldots,99]$. To evaluate the PMF for $p_M > p_N$, we set p_M to unity with no loss of generality. To evaluate it for $p_N < p_M$, we set p_N to unity. This yields $b = \frac{(bp_N)}{p_N}$ and $\gamma = \frac{\alpha bp_M}{\mu_M}$. Next, we computed the likelihood of the data under $\log_{10} p_M/\ p_N \in [-2, 2]$, keeping α, $b\ p_N$, μ_N, and μ_M constant, using the evaluation grid size $[0,\ldots,\max x_N + 3] \times [0,\ldots,\max x_M + 3]$, where $\max x_i$ is the maximum observed for each species in the simulation. We used 200 $\log_{10} p_M/p_N$ grid points, evenly spaced throughout the domain. Next, we computed the posteriors over the grid by dividing each likelihood vector by its sum. Finally, we plotted the average posterior distribution using line charts, with the color indicating the true value of p_M/p_N and the intensity indicating the number of cells, with more saturated colors corresponding to more simulated cells. For ease of comparison, we plotted the true values using dashed lines. From a statistical perspective, this analysis summarizes the parameter identifiability conditional on perfect information about the nascent marginal and the species averages. As we do not *a priori* know whether the differences in the PGF are actionable, the analysis illustrates the sample sizes required to fit the parameter to a particular degree of precision.

We previously motivated and fit the Poisson model of technical noise.[21,133] In Gorin et al.,[21] we inspected a variety of datasets, and observed a pronounced length bias in the nascent RNA count data, which did not appear in mature RNA counts (Section S7.3 of Gorin et al.[21]). This bias may be explained by three naïve models of biology.

The first model posits that the nascent RNA molecules are in the process of being transcribed; higher amounts of nascent RNA for longer genes simply reflect longer elongation delays. Although this explanation is superficially plausible, it is not borne out by the data. The model predicts a geometric-Poisson distribution of nascent RNA and zero correlation between nascent and mature counts.[141,143] Real data, on the other hand, have distinctly negative binomial-like marginals (as evident in, e.g., the third column of Figure 4B of our recent work on delay CMEs,[141] which shows consistently inferior fits under the delay model), and nontrivial nascent/mature correlations (as in the red histogram in Figure 2B).

The second model posits that the differences in expression reflect real differences in the underlying biological parameters, and technical noise may be neglected. However, fitting this model produces pervasive length biases in the parameter values (section S7.4 of Gorin et al.[21]), which are inconsistent with trends observed in orthogonal data. This is the model we explored in Gorin et al.[21]

The third model posits that technical noise *does* occur, but takes the species-independent form p_M/p_N. This formulation is mathematically identical to the second model, but proposes that an *apparent* length bias in the burst size is *actually* a length bias in bp. This model partially bypasses the objection raised for the second model by proposing that p is gene length-dependent, identical for nascent and mature species, and higher for longer genes. However, this model is implausible on physical grounds, as mature transcripts do not have the intronic poly(A) content necessary to produce this length dependence. This is indirectly implied by the consistently low fraction of exonic reads in sequencing datasets, in contrast to introns and the 3′ untranslated region.[137]

Please cite this article in press as: Gorin et al., Studying stochastic systems biology of the cell with single-cell genomics data, Cell Systems (2023), https://doi.org/10.1016/j.cels.2023.08.004

These biases can be largely eliminated by proposing a length-dependent sampling rate for nascent RNA counts, suggesting that this technical noise model is more coherent with known biology. To illustrate this process, we summarize the key results from Gorin et al.[21]

We obtained the raw data for the twelve 10x v3 datasets reported in Table S4 of Gorin et al.[21] The raw data consisted of nascent and mature count matrices for 2,500 genes per dataset. The counts were generated by running the *kallisto|bustools* 0.26.0 kb count command on the raw FASTQs with the—lamanno option, using an intronic/exonic index built from the GRCh38 and mm10 reference genomes, as described in Section 6.8.2. The datasets were filtered to remove low-expression droplets, first using the default *bustools* filter, then using the manually selected knee plot thresholds shown in Table S5 of Gorin et al.[21] Next, they were filtered for the top 2,500 moderate- to high-expression genes using the procedure in section S4.3.1 of Gorin et al.[21] To visualize the broad trends in count averages, we obtained the gene lengths L_j, then binned the values of $\log_{10} L_j$ into ten bins, with the edges given by the deciles $d_0, d_1, \ldots, d_{10}$. Next, we computed the average $\log_{10}$ mean of nascent and mature expression levels for genes falling into each bin. Finally, we plotted these mean levels at each bin center $d_k + \frac{1}{2}(d_{k+1} - d_k)$, connecting the values with a line to guide the eye. We repeated this analysis for all twelve datasets, distinguishing the nascent and mature statistics by color.

Next, we obtained the fit results for these datasets. The fits were performed using *Monod* 0.2.5.0 Python package[133] as described in Gorin et al.[21] Fitting the model with no technical noise entailed gradient optimization over the per-gene joint distributions to fit b_j, β_j, and γ_j. Although the model did not explicitly include technical noise, the theoretical discussion above implies that the results can be interpreted as those from a $p = p_N = p_M$ model, with the inferred "burst size" corresponding to $b_j\, p_j$ for gene j. Fitting the model with technical noise entailed scanning over a grid of C_N and λ_M, obtaining per-gene maximum likelihood estimates of b_j, β_j, and γ_j *conditional on* the technical parameter values at the grid point, then identifying the grid point which produced the lowest sum of Kullback-Leibler divergences over all genes. In both cases, the genes underwent a round of goodness-of-fit filtering to remove fits that did not accurately recapitulate the data, as in Section S4.3.5 of Gorin et al.[21] Next, we computed the average inferred $\log_{10}$ burst size for the genes falling into each length bin. As with the means, we plotted the average burst sizes at each bin center. connecting the values with a line to guide the eye. We repeated this analysis for all twelve datasets, distinguishing the results fit with and without a technical noise component by color.