Biophysical Reports

ARTICLE

Length biases in single-cell RNA sequencing of pre-mRNA

Gennady Gorin[1] and Lior Pachter[2,3,*]

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California; [2]Division of Biology and Biological Engineering, Pasadena, California; and [3]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California

ABSTRACT Single-cell RNA sequencing data can be modeled using Markov chains to yield genome-wide insights into transcriptional physics. However, quantitative inference with such data requires careful assessment of noise sources. We find that long pre-mRNA transcripts are over-represented in sequencing data. To explain this trend, we propose a length-based model of capture bias, which may produce false-positive observations. We solve this model and use it to find concordant parameter trends as well as systematic, mechanistically interpretable technical and biological differences in paired data sets.

WHY IT MATTERS Single-cell RNA sequencing is a method to quantify the amount of individual RNA molecules in cells. RNA reflects the extent of gene expression, which ultimately controls cell function. However, the method is imperfect, and some molecules are lost in the process. To understand the biophysics that control gene expression in the living cell, we need to produce and fit models that include both biological and technical sources of variability. Here, we show that unprocessed and mature RNA molecules exhibit counterintuitively different trends in their RNA expression and propose a mechanism of technical variability to account for these differences. This framework allows us to systematically explain differences in expression by specific physical mechanisms.

INTRODUCTION

The development of quantitative single-cell RNA sequencing (scRNA-seq) has made it increasingly tractable to fit single-molecule data to models of the RNA life cycle, thus facilitating a mechanistic view of genome-wide transcriptional regulation. Specifically, protocols with cell barcodes and unique molecular identifiers (UMIs) (1) allow for parameterization of discrete probabilistic models, with contents of cells conceptualized as draws from distributions over the nonnegative integers. When these models represent biophysical phenomena, fitting them provides information about the phenomena or about the overall plausibility of the model.

The standard framework for describing the microscopic biophysics of reactions in living cells is the chemical master equation (CME), which models mRNA counts by Markov chains that traverse a discrete state space (2–4). To fit biophysical parameters (the "inverse" problem of inference), one must solve the CME (the "forward" problem of prediction). This workflow requires computationally facile solutions that can be applied to thousands of genes. In mammalian and bacterial systems, the specific form of the CME is based on a random telegraph model of gene regulation, which describes a single gene locus that randomly switches between active and inactive states (2). A common simplification, supported by genome-wide fluorescence studies (5), treats the active state's duration as vanishingly small: mRNA is produced in geometrically distributed bursts that arrive according to a Poisson process. This model can be extended to describe rather general downstream processes of splicing, degradation (6), and translation. We focus on newly available data with spliced and unspliced mRNA, which can be fit to a tractable bursting model (7), and which has seen recent use in the inference of biological dynamics from static snapshots (8,9).

A remaining barrier to the application of this classical framework for inferring the biophysics underlying scRNA-seq data is modeling of technical artifact. The

Submitted September 15, 2022, and accepted for publication December 22, 2022.

*Correspondence: lpachter@caltech.edu

Editor: Ulrike Endesfelder.

https://doi.org/10.1016/j.bpr.2022.100097

sequencing process is probabilistic, and some molecules may not always be measured. Some studies attempt to "regress out" technical artifacts (10), but these methods are informal and incompatible with a discrete stochastic picture of transcription. Thus, treating both biological and technical stochasticity remains a significant lacuna in single-cell transcriptional models with no satisfactory and rigorous solutions.

We begin by exploring the biophysical interpretability of scRNA-seq data in light of the length bias seen in pre-mRNA expression. In some data sets, average spliced mRNA counts do not seem to show a length dependence (Fig. 1 *a*, *top*), which is consistent with previous studies of UMI-based protocols (11). On the other hand, unspliced mRNA counts strongly correlate with gene length (12) (Fig. 1 *a*, *bottom*). This prompted us to investigate whether the discrepancy has biological origins and raised questions about the consequences of ignoring this bias. We find that comprehensive, integrated stochastic models of biology and experiment are mandatory for interpreting sequencing data sets and appeal to the chemistry of sequencing to propose a class of plausible models.

MATERIALS AND METHODS

A model with no technical noise

To begin, we performed a naive analysis, fitting joint unspliced and spliced count data using a conventional (5,7) stochastic transcriptional model, namely a two-stage birth-death process coupled to a bursting promoter:

$$\varnothing \xrightarrow{k} B \times \mathscr{U} \xrightarrow{\beta} \mathscr{S} \xrightarrow{\gamma} \varnothing, \tag{1}$$

where $\mathscr{U}$ and $\mathscr{S}$ are unspliced and spliced mRNA species; k, β, and γ are the rates of Markovian transcription, splicing, and degradation processes, respectively; and B is a geometrically distributed burst size with mean b. We assumed the system had reached its unique steady state. The generating function solution to this system has been reported by Singh and Bokes (7).

A technical noise model

In the current section, we motivate, solve, and apply a stochastic model of sequencing that addresses technical artifacts to scRNA-seq data. We use the CME framework to derive the model from a microscopic Markov description of transcription in model definition. Finally, we report the model solution in model solution and fully describe the derivation in section S1.1.

In brief, we build a model that explicitly incorporates the stochastic sequencing steps taking place in fixed media (Fig. 2 *a*). Consistent with previous work on modeling pre-mRNA (8), we assume that the library construction step in the 10x sequencing workflow (1) includes molecules that have been captured at off-target binding sites. We posit that unspliced mRNA are primarily captured at internal poly(A) tracts, whereas spliced mRNA are captured at the poly(A) tail. To quantitatively model this effect, we introduce the concept of UMI "false positives": if a molecule has sufficiently many poly(A) sites, it is likely to be captured and reverse transcribed multiple times. As a first-order approximation, we model this bias as a length-dependent capture rate. Thus, each molecule in a cell gives rise to a Poisson distribution of cDNA. The downstream sequencing and alignment steps are treated as binomial sampling from the cDNA distribution.

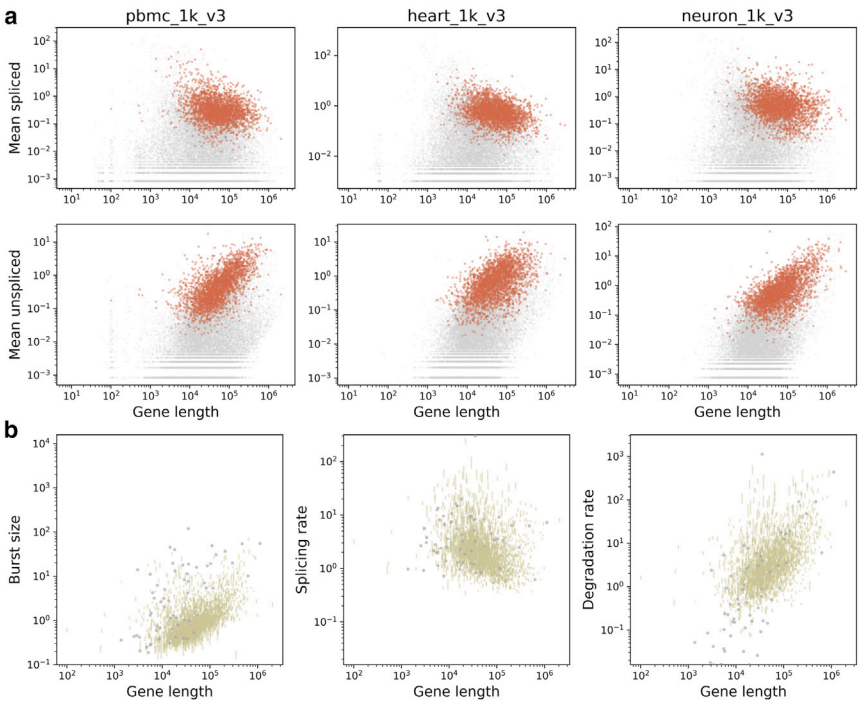

FIGURE 1 Spliced and unspliced single-cell RNA sequencing data demonstrate counterintuitive trends in data moments and model fits. (*a*) Length dependence of average mRNA counts in three data sets (*orange*: high-expression genes; *gray*: discarded low-expression genes; *top row*: spliced RNA; *bottom row*: unspliced RNA). (*b*) Transcriptional parameter estimates without a stochastic model of sequencing demonstrate pervasive length-dependent trends (pbmc_10k_v3; *gold*: lower bounds on 99% confidence intervals; *gray*: fits rejected by statistical testing; splicing and degradation rates are reported in units of burst frequency).

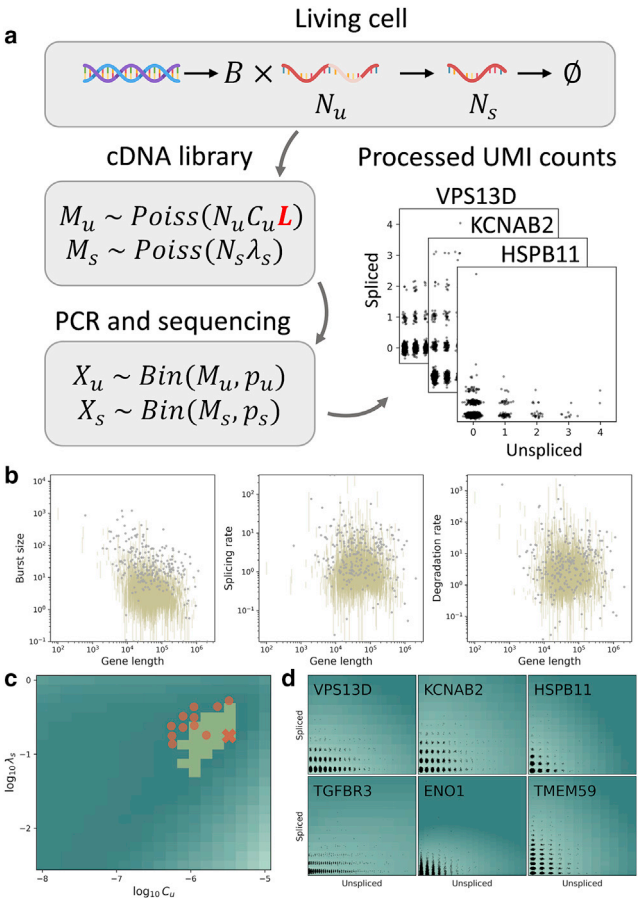

FIGURE 2 A length-biased technical noise model produces more physically interpretable results. (*a*) The integrated stochastic model of transcription and sequencing, with length dependence of the library construction step indicated in red. (*b*) Inferred transcriptional parameters do not appear to have strong length dependence (pbmc_10k_v3; *gold*: lower bounds on 99% confidence intervals; *gray*: fits rejected by statistical testing; splicing and degradation rates are reported in units of burst frequency). (*c*) The sampling parameter likelihood landscape shows a single optimum (*dark teal*: lower, *light teal*: higher total Kullback-Leibler divergence between fit and data from pbmc_10k_v3; *highlighted yellow region*: 5% quantile region for the displayed landscape; *orange cross*: optimal sampling parameter fit for the displayed landscape; *orange points*: optimal sampling parameter fits for other analyzed v3 data sets; C_u: coefficient for length-dependent unspliced capture rate; λ_s: spliced capture rate). (*d*) The parameter fitting procedure successfully recapitulates empirical copy-number distributions (*dark teal*: lower, *light teal*: higher log probability mass; *black points*: raw data UMI counts).

Model definition

The biological processes are defined in Eq. 1. This live-cell stage yields the unobservable distribution $P(N_u = n_u, N_s = n_s) := P(n_u, n_s)$, where N_z is the random variable describing true physiological counts of species z and n_z is the molecule count. This distribution has the probability-generating function (PGF) $G(g_u, g_s)$.

After equilibration, cDNA library construction begins, and all physiological processes halt due to cell fixation (1). Due to the possibility of multiple priming, each molecule of mRNA produces $Poisson(D_z)$ molecules of cDNA. D_u is presumed to be length dependent and governed by internal priming, whereas D_s is presumed to be length independent and governed by poly(A) tail priming.

Finally, amplification and sequencing take place. Unlike the library construction, these are strictly depleting processes: we suppose they cannot generate new UMIs, but they can lead to loss of UMIs. We assume the PCR amplification and product fragmentation are not substantially biased from gene to gene; further, the downstream fragments do not retain length information. Nevertheless, the overall identifiability of unspliced and mature mRNA may be different. Therefore, we suppose that each *in vitro* cDNA UMI gives rise to $Bernoulli(p_z)$ amplified, sequenced, and corrected in silico UMIs. The corresponding overall joint PGF takes the following form:

$$\begin{aligned} G_{X_u,X_s}(g_u, g_s) &= G(G_{1,u}(G_{2,u}(g_u)), G_{1,s}(G_{2,s}(g_s))) \\ &:= H(g_u, g_s), \end{aligned} \tag{2}$$

where $G_{i,z}$ is the PGF for sampling step i and species z. The parameters D_z and p_z are not independently identifiable, leading us to define net sampling rates $\lambda_z := D_z p_z$.

We use a first-order model of length dependence $\lambda_u = C_u L$: the rate of capture of any particular molecule scales directly with its length, acting as a proxy for the number of poly(A) tracts in the molecule. Even short poly(A) sequences can be captured by the oligo(dT) primers used in sequencing (13), and the number of poly(A) sequences in a given gene is strongly correlated with length (Fig. S2). We do not directly consider the number of tracts, as the determination of appropriate length thresholds or weights is a distinct thermodynamics challenge. The spliced mRNA parameter λ_s is kept constant, modeling capture at the poly(A) tail. For convenience, the model random variables and parameters are summarized in Tables S1 and S2.

Model solution

Following previous work (7), the steady-state PGF for the joint distribution of unspliced and spliced mRNA is $G(g_u, g_s) = e^{\varphi(v_u, v_s)}$, where

$$\begin{aligned} v_z &:= g_z - 1 \\ \varphi(v_u, v_s) &= k \int_0^\infty \frac{bU}{1 - bU} ds \\ f &:= \frac{\beta}{\beta - \gamma} \\ U &= v_s f e^{-\gamma s} + [v_u - v_s f] e^{-\beta s}. \end{aligned} \tag{3}$$

The PGF of a distribution under two steps of independent sampling is given in Eq. 2. Using the model assumptions outlined above, the overall PGF takes the following form:

$$H(g_u, g_s) = G\left(e^{\lambda_u (g_u - 1)}, e^{\lambda_s (g_s - 1)}\right). \tag{4}$$

The corresponding joint probability distribution $P(x_u, x_s)$ is easily computed by evaluating g_u and g_s around the complex unit circle and performing an inverse Fourier transform (7,14).

The moments of the model can be calculated by differentiating the PGF at $g_u = g_s = 1$. We report the lower moments of the noise-free model and the full model in Table I. The full derivations are provided in section S1.2. For convenience, the definitions of the summary statistics are given in Table S3.

TABLE 1Comparison of models' lower moments

Moment	Noise-free model	Technical noise model
μ_u	$\frac{kb}{\beta}$	$\frac{\lambda_u kb}{\beta}$
μ_s	$\frac{kb}{\gamma}$	$\frac{\lambda_s kb}{\gamma}$
$\sigma_u^2 - \mu_u$	$\mu_u b$	$\mu_u \lambda_u (1+b)$
$\sigma_s^2 - \mu_s$	$\mu_s \frac{b\beta}{\beta+\gamma}$	$\mu_s \lambda_s \left(1+\frac{b\beta}{\beta+\gamma}\right)$
$\mathrm{Cov}(u,s)$	$\frac{kb^2}{\beta+\gamma}$	$\frac{\lambda_u \lambda_s kb^2}{\beta+\gamma}$

Data processing and inference

We downloaded the human and mouse genomes from the Ensembl (15) database, computed gene lengths, and partitioned each gene's sequence into a set of contiguous poly(A) sequences. These sequences were used to compute cumulative histograms of the number of poly(A) tracts.

The scRNA-seq processing procedure is summarized in Fig. S1 and fully described in section S4.2. We downloaded scRNA-seq reads and processed them with the *kallisto|bustools* workflow (16), thereby obtaining spliced and unspliced count matrices. The analyzed data sets are motivated in section S4.1 and summarized in Table S4; nine were generated by 10x Genomics, and seven were generated by the Allen Institute for Brain Science (17,18). Genes without length annotations were discarded. As shown in section S7.3, all data sets demonstrated the previously encountered (Fig. 1 *a*) expression bias. For each inference batch, we selected the top genes according to the number of data sets in which they passed an expression filter. We used 2,500 genes for whole-data set analyses, 3,500 for cell-type difference analysis in blood cells, and 5,000 for cell-type difference analysis in neurons.

We estimated the parameters by scanning over a grid of sampling parameters, computing the conditional maximum likelihood estimates (MLEs) of all gene-specific parameters by gradient descent, and identifying the $\{C_u, \lambda_s\}$ MLE. In some cases, the fits were unreliable due to the sparsity of the data, suboptimal gradient descent fits, or model misspecification. To control for these sources of error, we discarded fits that were too close to the search domain bounds. Further, we performed a chi-squared test and discarded all genes with $p < 0.01$ and Hellinger distance > 0.05 as a measure of goodness of fit with an effect size component. We estimated a lower bound on 99% confidence intervals for MLEs through the Fisher information matrix; as we omit uncertainty in $\{C_u, \lambda_s\}$, these intervals necessarily underestimate the error. We detail the procedure in section S4.3. The analysis was performed using the *Monod* 0.2.5.0 Python package (19).

RESULTS

A model with no length bias produces implausible parameter estimates

At first glance, the rates we obtained by fitting the noise-free model to bivariate copy-number distributions seemed reasonable (Fig. 1 *b*; section S7.4). Two other noise models without a sequencing length bias produced qualitatively identical results (section S2). However, comparison with previous transcriptome-wide analyses suggested that the results were biophysically implausible.

We found that the inferred burst size increased with transcript length, in stark contrast with the previously observed modest inverse relationship (20). The degradation rate, normalized to burst frequency, displayed a similar positive trend. Previous studies found little to no gene length effect on burst frequency (20) and no effect on the rate of mRNA degradation (21). The latter is primarily controlled by open reading frame features rather than the length of the source gene. The decreasing splicing rates are more challenging to analyze: the splicing timescales given in the literature vary over several orders of magnitude depending on system and technology (22). However, length-based effects should be minimal, as cotranscriptional splicing is ubiquitous in mammalian cells (23,24) and widely varying intron sizes have little impact on splicing time (25).

Aside from empirical data, there are theoretical reasons to question these results: for example, the splicing rate is likely governed by spliceosome kinetics at individual introns, which is a local, rather than gene-wide, effect. Similarly, the cytoplasmic coding isoform is degraded, and its length is only weakly related to that of the parent transcript. In summary, the observed UMI counts of spliced transcripts cannot be plausibly treated the same way as those of unspliced transcripts. Such a simplification is incompatible with empirical evidence and currently accepted models.

The technical noise model produces consistent and physically interpretable results

Fitting the length-dependent Poisson technical noise model yielded transcriptional parameters (Fig. 2 *b*; section S7.5.2) without systematic length dependence. Therefore, we suggest that this integrated description of transcription and sequencing provides a more realistic and physically interpretable picture than available by considering the two sources of stochasticity separately.

All optima discovered by the coordinate scan procedure for the 10x v3 data sets lie within the square $\log_{10} C_u = -5.87 \pm 0.40$ and $\log_{10} \lambda_s = -0.56 \pm 0.32$. The Kullback-Leibler divergence (KLD) landscapes suggest that the data sets have unique optima and that the model is appropriate (Fig. 2 *c*; section S7.6). Furthermore, empirical joint mRNA count histograms were consistent with the fits (Fig. 2 *d*).

The inferred parameter distributions were consistently well fit (26) to a log normal-inverse Gaussian law (Fig. 3 *a*; section S7.7), although the mechanistic import of this finding is unclear. We performed a set of technical replicates, fitting distinct libraries generated from the same organism, and biological replicates, fitting libraries from multiple organisms. The

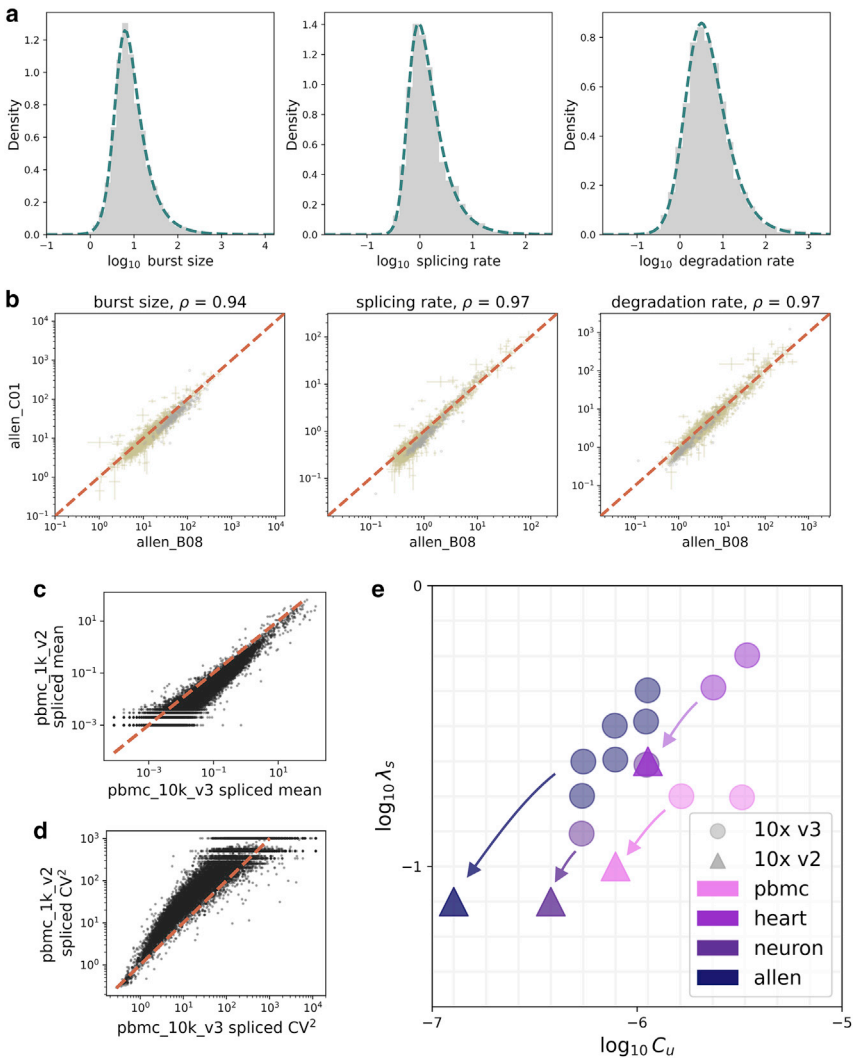

FIGURE 3 The technical noise model fits can be interpreted to analyze experimental effects. (*a*) Inferred transcriptional parameter distributions (pbmc_10k_v3; *gray*: histogram of biological parameters retained after statistical testing; *teal dashed line*: best fit to normal-inverse Gaussian distribution; splicing and degradation rates are reported in units of burst frequency). (*b*) Parameter estimates from biological replicates show largely concordant inferred parameter values (conventions as in Fig. 2 *b*). (*c*) 10x v2 and v3 single-cell RNA sequencing (scRNA-seq) replicates generated from a single sample demonstrate discordant RNA count distributions: the v2 data sets have lower mean values (*orange dashed line*: identity; *black*: genes). (*d*) The v2 data sets have higher CV^2 values (*orange dashed line*: identity; *black*: genes). (*e*) The v2 data sets' distributional differences can be tentatively explained by a combination of identical biological parameters and lower technical noise parameters (C_u: coefficient for length-dependent unspliced capture rate; λ_s: spliced capture rate; colors: data set categories; intersections of grid lines indicate the sampling parameter sets evaluated in the inference process).

results (Fig. 3 *b*; section S7.8) were consistent, with higher correlations among the technical replicates.

The technical noise model provides a framework for studying experimental effects

The obtained estimates for the technical noise parameters demonstrated limited identifiability. The data sets appeared to possess information sufficient to localize the technical noise to a coarse one-order-of-magnitude domain, but no further. When comparing multiple data sets de novo, it is challenging to attribute biases in parameter values: for example, under the current model, an apparent decrease in total RNA content may be caused by transcriptome-wide downregulation of transcription, upregulation of turnover, or decline in the sampling rates.

We can investigate the technical effects more systematically by treating replicates generated by different sequencing technologies and adopting stronger priors. We found that count data generated by the higher-efficiency v3 chemistry consistently yielded higher mean and lower noise (CV^2) levels than those generated by the older v2 chemistry (Figs. 3, *c* and *d*, and S32). We hypothesized that these differences should be appropriately attributed to technical effects, as the source tissues were similar or identical. A naive noise-free fit produced pronounced and nonphysical biases in parameter values (Fig. S33).

Imposing the belief that the underlying biological parameters should be the identical between all technical replicates and treating the results for large v3 samples as a putative ground truth, we identified the set of sampling parameters for the v2 data sets that produced the best agreement to these biological parameter values (Fig. S34; section S4.4). The resulting inferred sampling parameter optima are shown in Fig. 3 *e*: as expected, v2 data sets have lower sampling

parameter values. These values are somewhat challenging to identify without enforcing the consistency criterion between transcriptional parameters: as shown in section S7.6, the v2 KLD landscapes are more susceptible to noise than the v3 KLD landscapes, preventing de novo inference. Although the current comparison is mostly relative, the framework provides a quantitative explanatory mechanism for the technical effect of sequencing chemistry.

Inferred biophysical parameters provide insights into the mechanistic basis of differential expression

Just as technical noise parameters provide a mechanistic route to analyzing the effect of sequencing chemistry, the biological parameters provide a principled mechanistic route to identifying genes that are differentially regulated under varying conditions. Instead of the standard descriptive approach that tests differences in average expression (10), our model can test differences in parameter values. This conceptualization provides multiple advantages. Firstly, it increases statistical power due to reliance on model-specific results rather than nonparametric limiting theorems. For example, a gene may be expressed at nearly identical average levels in two cell types but have very different distributions (12); such an effect is easier to detect using full parametric distribution fits. Secondly, our approach yields greater interpretability, as all parameters explicitly model biophysical processes. For example, a difference in average expression may be directly attributed to the modulation of specific reaction rates, as discussed in previous work using fluorescence-based measurements (5,27) and very recently applied to scRNA-seq data (19,28).

Thus far, we have tacitly assumed that cells can be described as independent and identically distributed draws from a single stationary probability distribution. This approach is consistent with previous work (29,30), as well as a foundational premise of transcriptomic analyses: cell-type differences and transient phenomena are driven by a small set of marker genes (10,18,31–33), whereas the rest of the transcriptome is roughly static. Therefore, we have consciously omitted intrasample heterogeneity by discarding genes that do not match the model or have particularly high or low expression.

To demonstrate the potential applications of the mechanistic approach to discovery, we separately fit the cell types present in human blood and mouse brain data sets, based on previous clustering results. Disaggregated cell-type grid fits produced technical noise parameter estimates consistent with the full data sets (Fig. S35). For simplicity, we assumed that the technical noise parameters in each cell type were identical to those of the full data set. We found that the marker gene axiom appeared to be satisfied: the matched data sets parameter values were located near the identity line, with a small number of conspicuously off-diagonal genes that included known marker genes (Fig. 4 *a*; section S7.10.2).

The parameters demonstrated patterns of comodulation. In particular, the striking high correlation between differences in β/k and in γ/k suggests that this modulation pattern should be properly interpreted as reflecting modulation of the burst frequency k (Fig. 4 *b*; section S7.10.3). Using the change in β/k as a coarse proxy for the change in k, we can attribute marker gene modulation to a specific transcriptional mechanism: for example, the differences between T and B cells are typically associated with modulation of the burst frequency (Fig. 4 *c*), as previously proposed as a primary driver for cell-type differences (20,28). However, this mechanism is far from universal in our data sets, and we generally see a combination of burst size and burst frequency modulation in cell-type differences (section S7.10.4).

We used multiple biologically independent replicates, combined with a standard t-test, to identify patterns of parameter modulation between glutamatergic and GABAergic cell types (section S4.5.2). The results are shown in Fig. 4 *d*. Most interestingly, we observed several genes that consistently exhibited transcriptional parameter modulation but exhibited approximately constant mean spliced expression between cell types (average $\log_2$ fold change < 1) and would not be identifiable by standard statistical procedures. We identified burst size modulation for the genes *Rnf152*, *Fam174a*, *Nin*, *Rgmb*, *Dpysl3*, *Bach2*, *Igf1r*, *Stx4a*, and *Scg3*. We identified burst frequency modulation (putatively assigned due to changes in either splicing or degradation rate) for the genes *Fam174a*, *A330023F24Rik*, *Socs2*, *Ankrd40*, *Slc39a11*, *Mblac2*, *Itga4*, *Cxxc4*, *Ankrd6*, *Ccdc136*, *Crtc3*, *Egln1*, *Il34*, and *Mid2*. We visualize their distributions in a single neuronal data set in section S7.10.5: the distribution shapes demonstrate visually distinguishable differences and do not appear to suffer from significant failure to fit the data.

The identified genes largely, but not exclusively, relate to neuronal structure and development. *Socs2*, *Igf1r*, *Itga4*, and *Dpysl3* are involved in differentiation and neurite outgrowth (34–37). *Bach2* and *Cxxc4* induce feedback in neuronal development, apparently to maintain differentiated status in neurons (38,39). *Mid2* and *Nin* are associated with neural development regulation through microtubule organization (40,41). *Egln1* is linked to neuronal apoptosis (42). *Fam174a* is involved in lipid metabolism and membrane

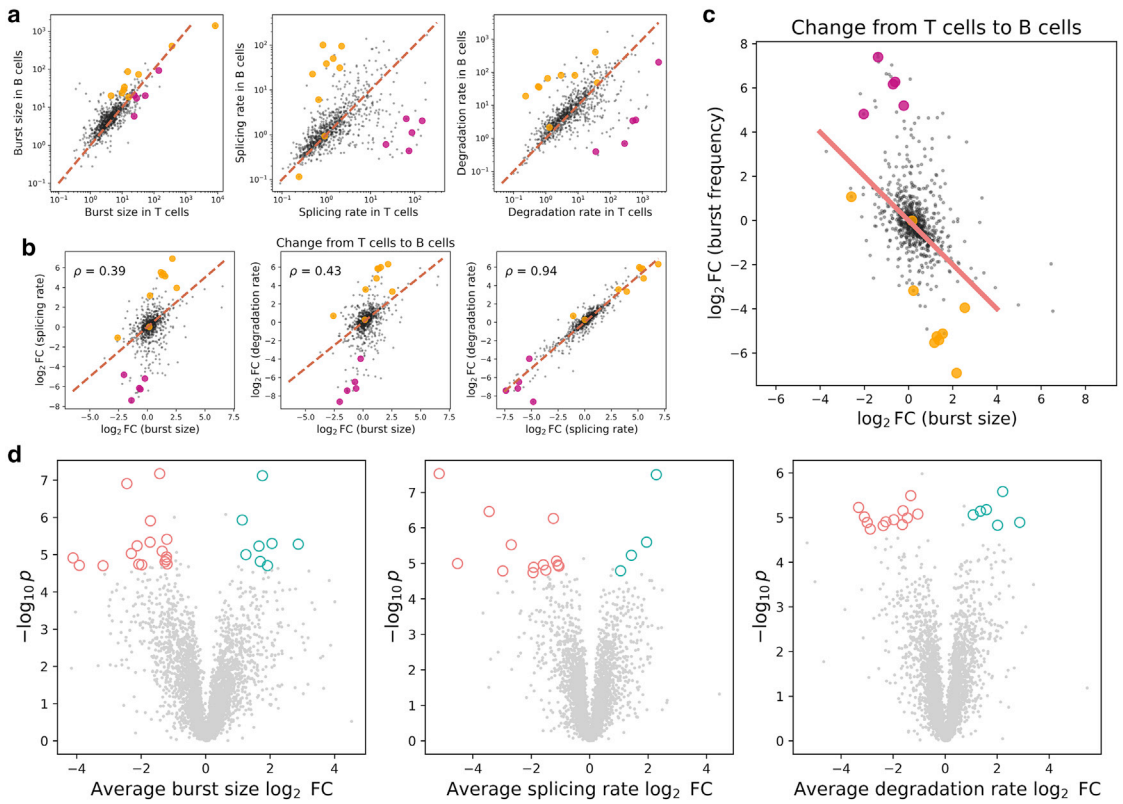

FIGURE 4 The inferred biological parameters provide insight into the biophysical basis of gene expression modulation. (*a*) Cell types in the pbmc_10k_v3 blood cell data set show largely concordant inferred parameter values, with the conspicuous exception of marker genes (*orange dashed line*: identity; *black*: genes retained after statistical testing; *orange*: T cell marker genes; *violet*: B cell marker genes; splicing and degradation rates are reported in units of burst frequency). (*b*) Cell types show strong covariation in splicing and degradation rate differences, suggesting potential burst frequency modulation (conventions as in *a*). (*c*) Cell-type differences can be attributed to combination of mechanisms; marker gene differences between B and T cells appear to be most readily explained by burst frequency modulation (*red line*: parameter combinations that yield identical average expression levels; *black*: genes retained after statistical testing; *orange*: T cell marker genes; *violet*: B cell marker genes; burst frequency modulation is estimated by splicing rate modulation). (*d*) Differential expression analysis identifies genes that exhibit consistent intercell-type parameter modulation in Allen neuron populations (*gray*: parameters for genes not identified as differentially expressed by the *t*-test and a fold change criterion; *light red*: parameters identified as higher in the glutamatergic cell type; *light teal*: parameters identified as higher in the GABAergic cell type).

structure (43). *Rnf152* and *Rgmb* are broadly implicated in neural development (44–46). *Scg3* appears to have a functional role in secretory granule biogenesis (47).

Some identified genes have less clear mechanistic connections to brain structure and function. *Ankrd40* is uncharacterized and is not known to have neural functions (48), but the similar gene product *Ankrd6* has an obscure neurodevelopmental role (49). *Stx4a* is localized on synaptic membranes (50). *Slc39a11* is a zinc transporter involved in brain function (51). *Mblac2* codes for an obscure protein that may have enzymatic activity (52). *Ccdc136* appears to have a DNA-regulatory role (53), but may be involved in neural speech pathology (54). The role of *Crtc3* in the rodent brain appears to be restricted to stress response (55). *Il34* is a microglial marker; microglia have immune and regulatory functions in the brain (56). *A330023F24Rik* is uncharacterized.

Although these distinctions are statistically identifiable, the import and basis of cell-type differences in distribution rather than average expression is, as of yet, obscure. The mechanism may involve expression compensation previously explored using theoretical tools (4) and recently observed under DNA repair stress (57).

DISCUSSION

We have introduced and implemented a stochastic model of intrinsic transcriptional noise that accounts for sequencing artifacts or technical noise. This model addresses an apparent overrepresentation of long unspliced mRNA in a variety of scRNA-seq data sets, and we posit that this bias is unlikely to arise biologically: fitting a simple model of mRNA production, splicing, and degradation produces parameter trends that render the fits suspect. Instead, we propose a model

motivated by the chemistry of the sequencing process: each mRNA can be captured and reverse transcribed multiple times, with the possibility of such false positives growing with the length of molecule and the number of poly(A) capture sites (Fig. S2). Although Poisson models for capture have been proposed before (as outlined in section S5.1), their derivation is largely ad hoc, and their implications for the reliability of sequencing data have not been examined in detail.

We fit the proposed model to a variety of data sets and discovered that the parameter values, and thus entire mRNA distributions, are consistent for sets of technical and biological replicates. Furthermore, the parameter values themselves (Fig. S29) were concordant with previous reports. Average burst sizes in the technical noise model were in the range $(10^{0.5}, 10^{1.5})$ (58, 59) rather than $(10^{-0.5}, 10^{0.5})$ in the noise-free model (section S7.4). Degradation rates γ/k were in the range $(10^{0}, 10^{1})$, roughly consistent with fluorescence-based genome-wide results (5). Finally, the splicing rates β/k were relatively slow and largely fell within the range $(10^{-0.5}, 10^{0.5})$, i.e., on the order of 100 min. This result suggests that β is best interpreted as the rate of an abstracted, multiintron process, as a single intron takes minutes to tens of minutes to splice (22,23,25). We discuss potential refinements of this model in section S5.2.

By fitting the model to closely matched data sets, we investigated technical and biological differences between conditions. We considered the differences between 10x v2 and v3 scRNA-seq data sets and found that the lower-quality v2 data sets can be described in a biophysically consistent way by proposing lower values for the parameters describing the sequencing process. Further, we applied the model to characterize cell-type differences at the level of transcriptional parameters. Although this procedure relies on preexisting annotations and inherits their limitations, it provides a principled way to interrogate the biophysical basis of cell-type differences. With this approach, we have demonstrated the possibility for interesting discovery. For example, it is possible to identify distributional differences that are not accompanied by substantial expression changes. These differences appear to be associated with compensatory mechanisms and motivate further study of the role of noise in biophysical systems.

DATA AND CODE AVAILABILITY

https://github.com/pachterlab/GP_2021_3 contains a Python notebook that can be used to reproduce all figures. The same repository contains all scripts used to make references, quantify transcripts, and process the resulting count matrices through the inference pipeline. The raw data and all search results have been deposited in Zenodo (60).

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpr.2022.100097.

AUTHOR CONTRIBUTIONS

G.G. and L.P. designed the study, performed the research. and wrote the manuscript. G.G. processed the sequencing data, developed and solved the model, and implemented the statistical procedures.

ACKNOWLEDGMENTS

G.G. and L.P. were partially funded by NIH U19MH114830. The DNA and RNA illustrations used in Fig. 2 were derived from the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. A part of the reported work was performed during a Data Sciences Co-op with Celsius Therapeutics, Inc. We thank Lambda Moses, Tara Chari, Meichen Fang, and Sina Booeshaghi for useful discussions in the course of conceptualizing the current work and developing *Monod*. The *Monod* package uses algorithms implemented in the *NumPy* (61), *SciPy* (26), and *numdifftools* (62) Python packages.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zheng, G. X. Y., J. M. Terry, …, J. H. Bielas. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049.
2. Peccoud, J., and B. Ycart. 1995. Markovian Modeling of Gene-Product Synthesis. *Theor. Popul. Biol.* 48:222–234.
3. Munsky, B., B. Trinh, and M. Khammash. 2009. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* 5:318.
4. Munsky, B., G. Neuert, and A. van Oudenaarden. 2012. Using Gene Expression Noise to Understand Gene Regulation. *Science.* 336:183–187.
5. Dar, R. D., B. S. Razooky, …, L. S. Weinberger. 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. USA.* 109:17454–17459.
6. Gorin, G., and L. Pachter. 2022. Modeling bursty transcription and splicing with the chemical master equation. *Biophys. J.* 121:1056–1069.
7. Singh, A., and P. Bokes. 2012. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophys. J.* 103:1087–1096.
8. La Manno, G., R. Soldatov, …, P. V. Kharchenko. 2018. RNA velocity of single cells. *Nature.* 560:494–498.
9. Bergen, V., M. Lange, F. J. Theis…, 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* 38:1408–1414.
10. Luecken, M. D., and F. J. Theis. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15:e8746.

11. Phipson, B., L. Zappia, and A. Oshlack. 2017. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* 6:595.
12. Gupta, A., F. Shamsi, …, A. Streets. 2021. Characterization of transcript enrichment and detection bias in single-nuclei RNA-seq for mapping of distinct human adipocyte lineages. Preprint at bioRxiv. https://doi.org/10.1101/2021.03.24.435852v1.
13. Nam, D. K., S. Lee, …, S. M. Wang. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. USA.* 99:6152–6156.
14. Bokes, P., J. R. King, …, M. Loose. 2012. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *J. Math. Biol.* 64:829–854.
15. Howe, K. L., P. Achuthan, …, P. Flicek. 2021. Ensembl 2021. *Nucleic Acids Res.* 49:D884–D891.
16. Melsted, P., A. S. Booeshaghi, …, L. Pachter. 2021. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* 39:813–818.
17. Yao, Z., H. Liu, …, E. A. Mukamel. 2021. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature.* 598:103–110.
18. Booeshaghi, A. S., Z. Yao, …, L. Pachter. 2021. Isoform cell-type specificity in the mouse primary motor cortex. *Nature.* 598:195–199.
19. Gorin, G., and L. Pachter. 2022. *Monod*: mechanistic analysis of single-cell RNA sequencing count data. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.11.495771.
20. Larsson, A. J. M., P. Johnsson, …, R. Sandberg. 2019. Genomic encoding of transcriptional burst kinetics. *Nature.* 565:251–254.
21. Sharova, L. V., A. A. Sharov, …, M. S. Ko. 2009. Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Res.* 16:45–58.
22. Alpert, T., L. Herzel, and K. M. Neugebauer. 2017. Perfect timing: splicing and transcription rates in living cells. *WIREs. RNA.* 8:e1401.
23. Drexler, H. L., K. Choquet, and L. S. Churchman. 2020. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol. Cell.* 77:985–998.e8.
24. Pandya-Jones, A., and D. L. Black. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA.* 15:1896–1908.
25. Singh, J., and R. A. Padgett. 2009. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* 16:1128–1133.
26. Virtanen, P., R. Gommers, …, Y. Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods.* 17:261–272.
27. Nicolas, D., N. E. Phillips, and F. Naef. 2017. *Mol. Biosyst.* 13:1280–1290.
28. Luo, X., F. Qin, …, G. Cai. 2022. BISC: accurate inference of transcriptional bursting kinetics from single-cell transcriptomic data. *Briefings Bioinf.* 23:bbac464.
29. Ham, L., R. D. Brackston, and M. P. Stumpf. 2020. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Phys. Rev. Lett.* 124:108101.
30. Kim, J., and J. C. Marioni. 2013. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* 14:R7.
31. Campbell, K. R., and C. Yau. 2019. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics.* 35:28–35.
32. Shapiro, E., T. Biezuner, and S. Linnarsson. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14:618–630.
33. Wang, J., M. Huang, …, N. R. Zhang. 2018. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA.* 115:E6437.
34. Kowara, R., M. Ménard, …, B. Chakravarthy. 2007. Co-localization and interaction of DPYSL3 and GAP43 in primary cortical neurons. *Biochem. Biophys. Res. Commun.* 363:190–193.
35. Scott, H. J., M. J. Stebbing, …, A. M. Turnley. 2006. Differential effects of SOCS2 on neuronal differentiation and morphology. *Brain Res.* 1067:138–145.
36. Jin, J., P. Ravindran, D. Di Meo, and A. W. Püschel. 2019. Igf1R/InsR function is required for axon extension and corpus callosum formation. *PLoS One.* 14:e0219362.
37. Thongkorn, S., S. Kanlayaprasit, …, T. Sarachana. 2021. Sex differences in the effects of prenatal bisphenol A exposure on autism-related genes and their relationships with the hippocampus functions. *Sci. Rep.* 11:1241.
38. Shim, K. S., M. Rosner, …, M. Hengstschläger. 2006. Bach2 is involved in neuronal differentiation of N1E-115 neuroblastoma cells. *Exp. Cell Res.* 312:2264–2278.
39. Gao, J., Y. Ma, …, W.-L. Jin. 2016. Non-catalytic roles for TET1 protein negatively regulating neuronal differentiation through srGAP3 in neuroblastoma cells. *Protein Cell.* 7:351–361.
40. Suzuki, M., Y. Hara, …, N. Ueno. 2011. *MID1* and *MID2* are required for *Xenopus* neural tube closure through the regulation of microtubule organization. *Development.* 138:385.
41. Baird, D. H., K. A. Myers, …, P. W. Baas. 2004. Distribution of the microtubule-related protein ninein in developing neurons. *Neuropharmacology.* 47:677–683.
42. Lee, S., E. Nakamura, …, S. Schlisio. 2005. Neuronal apoptosis linked to EglN3 prolyl hydroxylase and familial pheochromocytoma genes: Developmental culling and cancer. *Cancer Cell.* 8:155–167.
43. Imbault, V., C. Dionisi, …, M. Pandolfo. 2022. Cerebrospinal Fluid Proteomics in Friedreich Ataxia Reveals Markers of Neurodegeneration and Neuroinflammation. *Front. Neurosci.* 16:885313.
44. Okamoto, T., K. Imaizumi, and M. Kaneko. 2020. The Role of Tissue-Specific Ubiquitin Ligases, RNF183, RNF186, RNF182 and RNF152, in Disease and Biological Function. *Int. J. Mol. Sci.* 21:3921.
45. Chow, S. Y. A., K. Nakayama, …, Y. Ikeuchi. 2022. Human sensory neurons modulate melanocytes through secretion of RGMB. *Cell Rep.* 40:111366.
46. Samad, T. A. 2004. DRAGON: A Member of the Repulsive Guidance Molecule-Related Family of Neuronal- and Muscle-Expressed Membrane Proteins Is Regulated by DRG11 and Has Neuronal Adhesive Properties. *J. Neurosci.* 24:2027–2036.
47. Li, F., X. Tian, …, H. Pang. 2012. Dysregulated expression of secretogranin III is involved in neurotoxin-induced dopaminergic neuron apoptosis. *J. Neurosci. Res.* 90:2237–2246.
48. Ernst, W. L., Y. Zhang, …, J. L. Noebels. 2009. Genetic Enhancement of Thalamocortical Network Activity by Elevating 1G-Mediated Low-Voltage-Activated Calcium Current Induces Pure Absence Epilepsy. *J. Neurosci.* 29:1615–1625.
49. Tissir, F., I. Bar, …, C. Lambert De Rouvroit. 2002. Expression of the ankyrin repeat domain 6 gene (Ankrd6) during mouse brain development. *Dev. Dynam.* 224:465–469.
50. Alldred, M. J., K. E. Duff, and S. D. Ginsberg. 2012. Microarray analysis of CA1 pyramidal neurons in a mouse model of tauopathy reveals progressive synaptic dysfunction. *Neurobiol. Dis.* 45:751–762.
51. De Benedictis, C. A., C. Haffke, …, A. M. Grabrucker. 2021. Expression Analysis of Zinc Transporters in Nervous Tissue Cells Reveals Neuronal and Synaptic Localization of ZIP4. *Int. J. Mol. Sci.* 22:4511.

52. Malgapo, M. I. P. 2019. Structure and function of the palmitoyltransferase dhhc20 and the acyl coa hydrolase mblac2, PhD Dissertation. Cornell, Ithaca, NY.

53. Mazille, M., K. Buczak, …, O. Mauger. 2022. Stimulus-specific remodeling of the neuronal transcriptome through nuclear intron-retaining transcripts. *EMBO J.* 41:e110192.

54. Adams, A. K., S. D. Smith, …, J. R. Gruen. 2017. Enrichment of putatively damaging rare variants in the DYX2 locus and the reading-related genes CCDC136 and FLNC. *Hum. Genet.* 136:1395–1405.

55. Rubió Ferrarons, L. 2020. Insights into the CREB-regulated transcription coactivators (CRTCs) in neurons and astrocytes, PhD Dissertation. Universitat Autònoma de Barcelona.

56. Badimon, A., H. J. Strasburger, …, A. Schaefer. 2020. Negative feedback control of neuronal activity by microglia. *Nature.* 586:417–423.

57. Desai, R. V., X. Chen, …, L. S. Weinberger. 2021. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science.* 373:eabc6506.

58. Sanchez, A., and I. Golding. 2013. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science.* 342:1188–1193.

59. Suter, D. M., N. Molina, …, F. Naef. 2011. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science.* 332:472–474.

60. Gorin, G., and L. Pachter. 2022. Supporting data for GP_2021_3. Zenodo Data: https://doi.org/10.5281/zenodo.7388133.

61. Harris, C. R., K. J. Millman, …, T. E. Oliphant. 2020. Array programming with NumPy. *Nature.* 585:357–362.

62. P. A. Brodtkorb and J. D'Errico, "numdifftools," (2021).