Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using *Monod*

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

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125
[2]Division of Biology and Biological Engineering; Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125
[*]lpachter@caltech.edu

April 17, 2023

Abstract

We present the Python package *Monod* for the analysis of single-cell RNA sequencing count data through biophysical modeling. *Monod* naturally "integrates" unspliced and spliced count matrices, and provides a route to identifying and studying differential expression patterns that do not cause changes in average gene expression. The *Monod* framework is open-source and modular, and may be extended to more sophisticated models of variation and further experimental observables.

The *Monod* package can be installed from the command line using `pip install monod`. The source code is available and maintained at `https://github.com/pachterlab/monod`. A separate repository, which contains sample data and Python notebooks for analysis with *Monod*, is accessible at `https://github.com/pachterlab/monod_examples/`. Structured documentation and tutorials are hosted at `https://monod-examples.readthedocs.io/`.

1 Introduction

The interpretation of single-cell transcriptomics data depends on the ability to distinguish between variation in gene expression due to technical noise arising from experimental artifact, and variability reflecting underlying biology. Thus, analysis of single-cell RNA sequencing (scRNA-seq) data begins with "depth normalization," a procedure whose purpose is to account for technical variation in the number of reads sequenced per cell due to the stochastic sampling of reads from cDNA libraries. Additionally, variance-stabilization transformations are applied to account for associations between variance and magnitude of gene expression. These transformations are also premised on technical artifact stemming from stochastic sampling of reads from cDNA libraries. Essentially all scRNA-seq analyses begin with these two steps. Other subsequent transformations and procedures to remove technical noise are also commonplace, e.g., dimensionality reduction by principal component analysis [1] and batch correction [2] being two examples.

Despite the omnipresence of normalization as a first step in single-cell RNA sequencing analysis, and extensive studies of its effectiveness in achieving variance-stabilization and uniformity in read depth per cell [3, 4], the question of whether normalization can inadvertently remove biological signal has not been thoroughly explored. This is a cause for concern, because investigation of technical noise removal in the context of batch correction [5] has shown that biological signal can be inadvertently removed in an attempt to account for technical artifact. Figure 1 shows that normalization can be similarly problematic. Using differences between cell types to bound biological variation, we find that not only normalization, but all the commonly applied transformations to single-cell RNA sequencing data remove biological signal, especially from highly expressed genes. The UMAP embedding step is particularly egregious; it adds large amounts of non-biological noise at the end of a process intended to remove it.

The removal of biological signal by transformations currently applied to single-cell RNA sequencing data is perhaps not surprising; current workflows are the culmination of experiments with heuristics, and the methods are not grounded in biophysical models. However, there is no reason that scRNA-seq analysis cannot be grounded in rigorous systems biology [6–8]. We propose the use of models of transcription with which technical and biological variation can be distinguished on the basis of mechanism. Our approach, via modeling with a chemical master equation (CME) [9], conforms with mechanistic approaches to quantitative biology that originated in the latter half of the twentieth century [10–13], and that substantially expanded over the past two decades [14–16], providing biophysical rationale for the biological component of variation observed in gene expression measurements.

In addition to relying on methods that provides an avenue to rationally thinking about biological stochasticity versus technical variation, our approach also addresses another vexing problem in single-cell RNA sequencing data analysis. Recently, interest in "RNA velocity," the inference of trajectory directions using the relative abundances of unspliced and spliced mRNA [17], has led to recognition that in addition to the standard count matrices produced in single-cell RNA sequencing pre-processing [18], reads aligned to non-coding sequences may also be informative [19]. Several software packages have been developed to quantify both "unspliced" and "spliced" modalities [17, 20, 21], but despite their widespread use for RNA velocity [22], a natural question they raise has not been addressed: how should spliced and unspliced count matrices be "integrated," or analyzed simultaneously, to obtain insights into gene expression beyond the context of trajectory inference? We show that mechanistic modeling also provides an answer to this question.

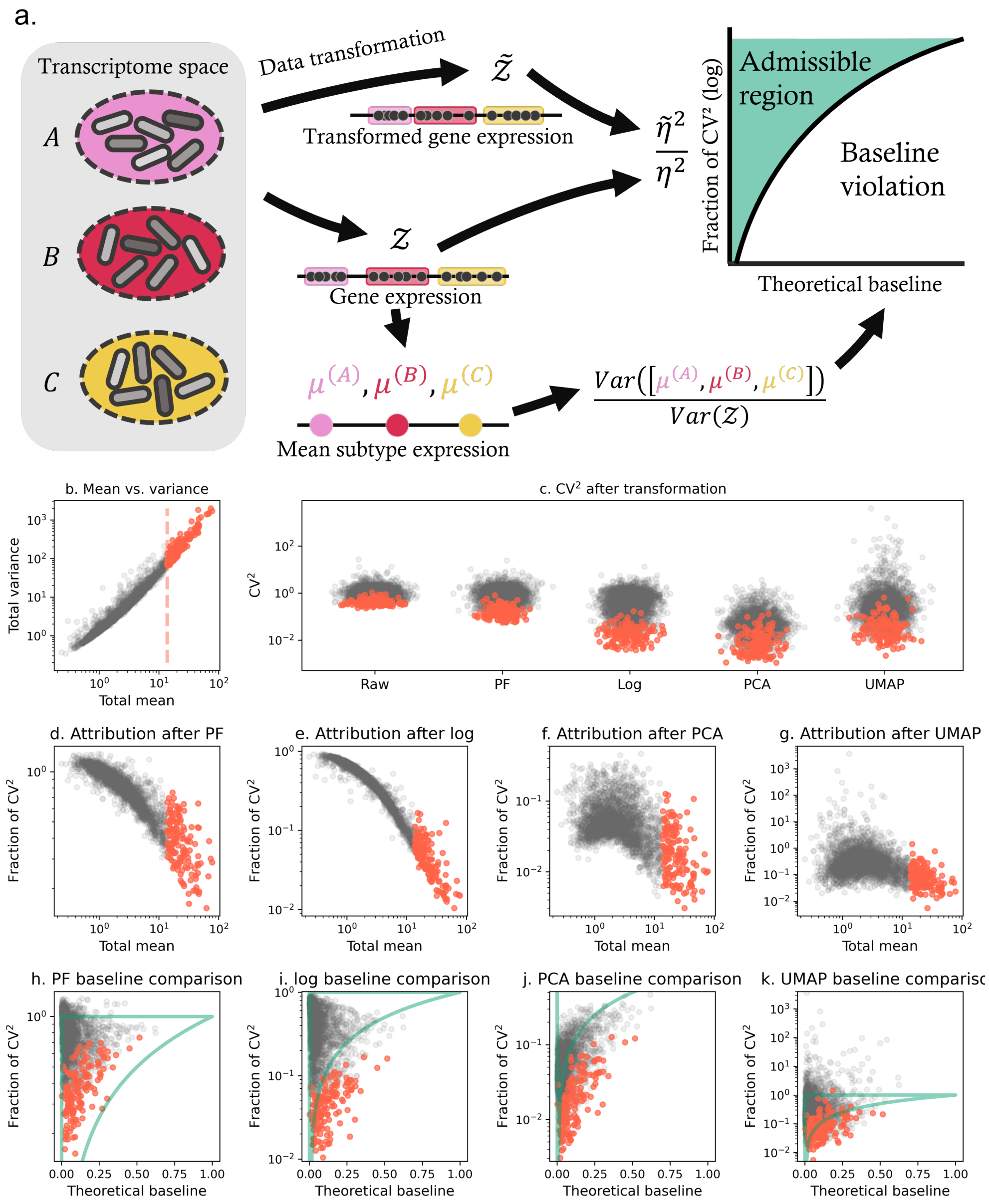

Figure 1: Normalization and dimensionality reduction distort and underestimate biological variation, especially in high-expression genes.

a. A proposed baseline for the analysis of residual variation after data transformation: the fraction of biological variability can be bounded by a theoretical baseline, which is computed from the variation in average subpopulation expression. If this baseline is violated, the data transformation has discarded some biophysically meaningful variation.

b. High-expression genes have high variance (gray points: genes below the 95th percentile by mature RNA expression; red points: genes above the 95th percentile by mean mature RNA expression, red line: percentile threshold).

c. Proportional fitting size normalization (PF), log-transformation (log), and principal component analysis (PCA) globally deflate the squared coefficient of variation (CV^2), whereas Uniform Manifold Approximation and Projection (UMAP) globally inflates it (gray and red points: as in **b**).

d.-g. All four of the steps substantially deflate high-expression genes' CV^2 relative to raw data, implicitly attributing their variability to nuisance technical effects (gray and red points: as in **b**).

h.-k. The deflation of variability results in the violation of the theoretical lower bound computed from cell subpopulation differences, particularly for high-expression genes (gray and red points: as in **b**; curved teal line: identity baseline, below which biological variability is removed; horizontal teal line: threshold, above which variability is inflated relative to raw data).

2 Results

2.1 Model definition

We focus on a class of bursty transcriptional models of mRNA expression:

$$\varnothing \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \varnothing. \tag{1}$$

This reaction schema encodes a continuous-time Markov chain on a bivariate discrete space of molecule counts for each gene, where $\mathcal{X}_N$ is a nascent mRNA species and $\mathcal{X}_M$ is a mature mRNA species. We identify the former with unspliced and the latter with spliced transcripts (Section 2.6 of [23]).The rate k is the characteristic burst frequency, such that the number of transcription events in any time interval of length τ is $Poisson(k\tau)$. B is the number of unspliced mRNA produced per transcriptional event, which is a generally a random variable. β is the splicing rate, such that a given molecule of $\mathcal{X}_N$ will become a molecule of $\mathcal{X}_M$ after an exponentially-distributed delay with rate β. Analogously, γ is the degradation rate, such that a given molecule of $\mathcal{X}_M$ will be eliminated after an exponentially-distributed delay with rate γ. At steady state, we fit the rate parameters in units of k, which is equivalent to imposing $k = 1$.

Typical fluorescence transcriptomics analyses use a model of random promoter switching [11] to define B as geometrically-distributed with mean b [15], inducing negative binomial-like [24] distributions of $\mathcal{X}_N$ and $\mathcal{X}_M$ [15, 25–31]. This model describes the dynamics at a promoter that randomly switches between a transcriptionally inactive, long-lived state and a highly active, short-lived state (Section S2.1.1); B is the number of molecules generated in the active state.

The chemistry of the process [32] suggests that a given mRNA molecule may be captured again after being reverse transcribed into cDNA once. Therefore, we assume that cDNA molecules are generated according to a Poisson birth process over unity time, such that:

$$\begin{aligned} \mathcal{X}'_N &\xrightarrow{\lambda_N} \mathcal{X}'_N + \mathcal{X}_N, \\ \mathcal{X}'_M &\xrightarrow{\lambda_M} \mathcal{X}'_M + \mathcal{X}_M, \end{aligned} \tag{2}$$

where λ_N and λ_M are the Poisson process rates for each species, $\mathcal{X}'_N$ and $\mathcal{X}'_M$ are unobserved *in vivo* molecules with dynamics following Equation 1, and $\mathcal{X}_N$ and $\mathcal{X}_M$ are UMIs observed after capture, amplification, sequencing, and alignment. We further assume $\lambda_N := C_N L$ for a constant C_N and variable gene length L: the dependence on L coarsely models the possibility of multiple priming at internal poly(A) stretches [33].

We use the bursty/Poisson model of biology and sequencing throughout our analysis. However, to facilitate comparisons with alternative descriptions of biological variation, we characterize and implement various other models of biology, with optional technical noise components, outlined in Section S2.

2.2 Probabilistic investigation of normalization and dimensionality reduction

Due to the scale of scRNA-seq data, standard analyses heavily use data transformation and dimensionality reduction to produce a version of the data more amenable to statistics [18]. For example, a typical analysis of cell type heterogeneity may apply size normalization (e.g, proportional fitting or PF, which treats RNA counts as compositional quantities [34]), log-transformation, principal component analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP). Each of these steps has a specific purpose; for the four steps above, the purposes are, in turn, to remove variability due to technical heterogeneity, to obtain easily tractable normal-like log-abundance distributions, to select the latent data dimensions that contain the most variability, and to visualize the cell type structure [18]. These transformations rely on implicit assumptions about the structure of the data; these assumptions may be mutually contradictory, and their violation may produce results that range from suboptimal to catastrophically incorrect.

These limitations and failure modes have previously been investigated. Size normalization privileges relative, rather than absolute RNA species abundance; occasionally, this approach produces inconsistent results across the genome [35] and retains apparently technical variation [34, 36]. Log-transformation is optimal for homogeneous, high-expression, approximately negative binomial data [4, 18, 34], and relies on an arbitrary genome-wide "pseudocount" hyperparameter that can distort the distributions [4, 34, 35, 37]. PCA is optimal for multivariate normal data, and can be misled by the large zero fractions observed in single-cell data [37]. Finally, UMAP appears to be optimal for data with uniform, low-noise coverage of a latent manifold, with risk of distortions due to violated assumptions and stochastic initialization [22, 38, 39]. A comprehensive treatment of the distortions induced or ameliorated by each step appears, however, to be out of reach, as the transformations' results are heavily data-dependent and elude theoretical analysis.

In Figure 1a, we propose a procedure for the quantitative benchmarking of data transformations relative to an internal baseline. Each step transforms the data distribution, purportedly retaining relevant biological variability – such as cell type differences – and removing incidental or technical variability, quantified by the squared coefficient of variation (CV^2). Therefore, by removing some fraction of variability, a data transformation implies this component is immaterial to analysis, whereas the residual fraction of variation – the CV^2 ratio for the distribution after and prior to transformation, denoted by $\tilde{\eta}^2/\eta^2$ – is attributed to biology. However, this residual fraction should not vary arbitrarily; under mild assumptions, we can bound the biological fraction of CV^2 from below by the variability in cell subpopulation averages (Section 5.1.1).

To compare the results of the transformation procedures to this baseline, we analyzed a mouse glutamatergic neuron dataset [40], using pre-annotated subtypes to produce a lower bound. The details of the analysis are given in Section 5.1. We considered a set of 2,951 genes, emphasizing the top 5% by dataset-wide average; these high-variability genes are typically of most interest in single-cell analyses (Figure 1b). The iterative application of transformations up to PCA typically deflated the gene-specific CV^2 values, particularly for the high-expression genes and in the log-transformation step. However, the application of UMAP inflated CV^2 throughout. We found that the high-expression genes' variability was typically deflated relative to the raw data, suggesting that the data transformations attribute overdispersion to nuisance technical effects (Figure 1d-g).

Log-transformation, PCA, and UMAP violated the baseline computed from inter-subtype variation, particularly for the high-expression genes. In addition, a considerable fraction of genes demonstrated variability exceeding that of the original data after PF and UMAP. As shown in Figure S8, after computing the UMAP, more than a third of the genes in the dataset had, at some point in the analysis, gone below the lower bound. This result suggests that ubiquitous transformations efface meaningful biological signal. Despite the claim in [4] that the log-transform is "best," our analysis shows that it may be even better not to apply a transform which is agnostic to technical versus biological stochasticity. UMAP attempts to recover variance by inflating cell type differences; however, since this inflation is genome-wide, it does not restore the quantitative information lost in previous steps, and may generate false findings.

We propose that a mechanistic approach provides a more reliable avenue for the analysis of sequencing data. In this worldview, all assumptions about the noise behaviors are explicit rather than implicit; count data are not to be denoised, but fit to a first-principles model that includes biological and technical noise terms. Once a satisfactory parametric fit is available, the fractions of biological and technical variability follow immediately (Section 5.1.3). This approach is outlined schematically in Figure 2a: given annotations, we can separately fit cell subtypes, obtain their biophysical parameters, and aggregate them to obtain the fraction of biological variability.

The fit, implemented in *Monod*, attributes overdispersion in high-expression genes to biological variability (Figure 2b), in striking contrast to the non-parametric transformations (Figure S9). As a consequence, the inferred fraction of biological variability coheres with the baseline (Figure 2c). Interestingly, this agreement is not merely a consequence of independently fitting cell subtypes and aggregating the variance. By applying *Monod* to the entire glutamatergic dataset, introducing some error due to the neglect of subtype heterogeneity, we obtain similar results, with a single violation of the bound (Figure S10). This control suggests that the mechanistic procedure largely explains biological variability by transcriptional bursting, rather than subtype differences.

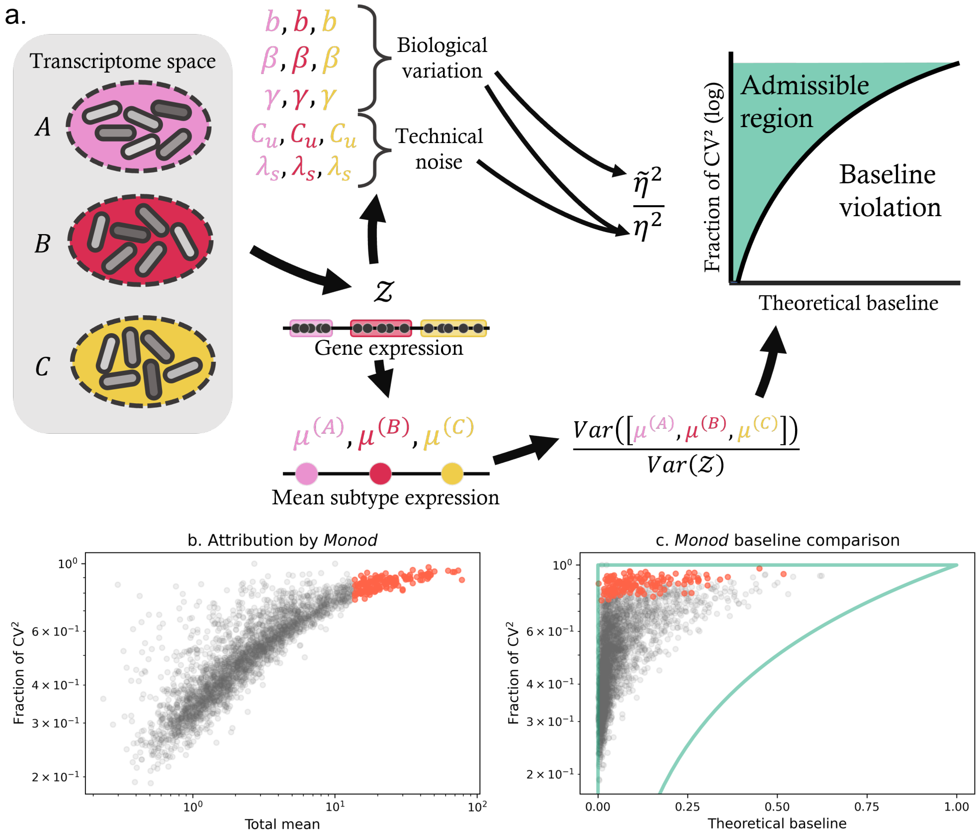

Figure 2: The *Monod* mechanistic analysis of biological and technical variability produces coherent results.

a. The baseline introduced in Figure 1a may be compared to point estimates of the biological variability fractions, which follow immediately from a fit to a parametric model of transcription and sequencing.

b. The *Monod* fits explicitly attribute the variability in high-expression genes to biological phenomena (gray and red points: as in Figure 1b).

c. The *Monod* results lie entirely within the admissible region (gray and red points: as in **b**; curved teal line: identity baseline, below which inferred biological variability is lower than inter-cell population variability; horizontal teal line: threshold, above which inferred biological variability exceeds that of raw data).

2.3 Principled integration of single-cell and single-nucleus data

In parallel with single-cell RNA sequencing [41], recent years have seen the rapid adoption of single-nucleus RNA sequencing (snRNA-seq) technologies [42,43]. As illustrated in Figure 3a, snRNA-seq releases and isolates the nuclei, whereas scRNA-seq requires tissue dissociation into individual cells. As discussed in [44], in spite of the loss of cytoplasmic signal, which limits the ability to detect certain relevant genes [45], single-nucleus sequencing provides technical advantages. Single-nucleus protocols require considerably simpler tissue handling; for example, they can be applied to frozen cell samples [46–48].

The interest in single-nucleus sequencing, as well as the recognition of systematic differences in the findings from the two technologies [45, 49–51], has motivated the analysis of these differences [52] and the development of more or less *ad hoc* data integration methods [50, 53]. Several recent reports [52–54] have found that the nuclear datasets exhibit a strong length bias, with longer genes being overrepresented in nuclei. This discrepancies, in turn, appears to stem from a fundamental methodological difference: single-cell analyses typically only use exonic reads, whereas single-nucleus combine intronic and exonic reads [42, 51]. Although the exonic molecule counts do not appear to exhibit a length bias, intronic ones do [33], likely due to internal priming in poly(A)-rich intronic regions [17]. Furthermore, even if all reads are included in the analysis of a single-cell dataset, the length bias may be attenuated due to the abundance of fully processed molecules. However, the appropriate way to correct for this effect is obscure. Previous reports have suggested [53, 54] or eschewed [55] normalizing by gene length; this scaling, if applied, prevents the application of discrete models.

We propose that scRNA-seq and snRNA-seq may be more analyzed in a more principled way through a mechanistic lens. This strategy treats nascent (intron-containing) and mature (exonic) molecules as distinct, in the spirit of [52, 54], and takes the distinction to its logical conclusion by defining a model with nuclear export (Section 5.2.1). Under a particular set of assumptions, this model reduces to the form in Section 2.1, with nuclear export taking the role of cytoplasmic degradation as the mechanism of mature RNA efflux. However, the nascent RNA dynamics – i.e., transcription and splicing – should be identical for the two technologies, as they are confined to the nucleus.

This axiom provides a foundation for the joint analysis of the technologies. We fully outline the approach in Section 5.2. For example, Figure 3b-c compares the average counts for 2,000 genes in scRNA-seq and snRNA-seq datasets generated from a single mouse brain tissue sample by 10x Genomics [56, 57]. Surprisingly, in spite of the depletion of cytoplasmic RNA, the mature count averages were visually similar, whereas the nuclear count averages were approximately half an order of magnitude higher in the single-nucleus dataset. Quantitatively, 83% of the mature and over 99% of the nascent averages were higher in the snRNA-seq sample. To explain this difference, we adopt the usual "marker gene" paradigm, i.e., that closely related cell types typically differ in the expression of a small number of genes [18], whereas the other genes have similar distributions. This assumption implies that incidental enrichment of certain cell subpopulations cannot explain the striking, widespread discrepancy, and immediately leads us to conclude that the difference is purely technical; due to the details of the nuclear sequencing protocol, the procedure retains considerably more RNA of both types. This assumption appears to be supported by Figure 3d-e: both species exhibited an overall decrease in the noise levels (66% of the mature and 98% of the nascent CV^2 values), which is consistent with decreased molecule loss. The difference in mature RNA amounts should, then, be explained by the combination of two competing effects: the depletion of cytoplasmic RNA, as well as more effective capture of remaining molecules, in the single-nucleus protocol.

To quantify the efflux rates, we fit the datasets using *Monod* and inferred the technical noise parameters for the single-cell dataset. Next, we identified the set of single-nucleus technical noise parameters (Figure S6) that provided the best match to the burst size and splicing rate parameters (Figure 3f-g); the discovered set of technical noise parameters had higher (more effective) sampling rates. The inferred efflux rates at this set were considerably higher for the single-nucleus dataset, both visually (Figure 3h) and statistically: the t-test $\{t, p\}$ values were $\{-2.7, 7.3 \times 10^{-3}\}$ for the burst size, $\{1.6, 0.11\}$ for the splicing rate, and $\{-11, 2.1 \times 10^{-27}\}$ for the efflux rate.

The procedure we have outlined has significant limitations: for example, we have neglected nuclear efflux in the single-cell data and cell type heterogeneity, both of which are physiologically important [40,58] likely contributors to deviations in Figure 3f-g. In addition, single-nucleus sequencing may harbor as of yet poorly-understood technical noise phenomena particular to the technology. Nevertheless, the model formulation provides a foundation for the incorporation of more sophisticated nuclear retention delays [44,59,60] jointly with technical noise. In addition, the strategy provides a principled solution to the dilemma of incorporating intronic reads: all the available data should be used, with species differences encoded in a multivariate mechanistic model. If its assumptions are explicitly formulated, the model can be fit, or extended to account for violations, based on experimental data.

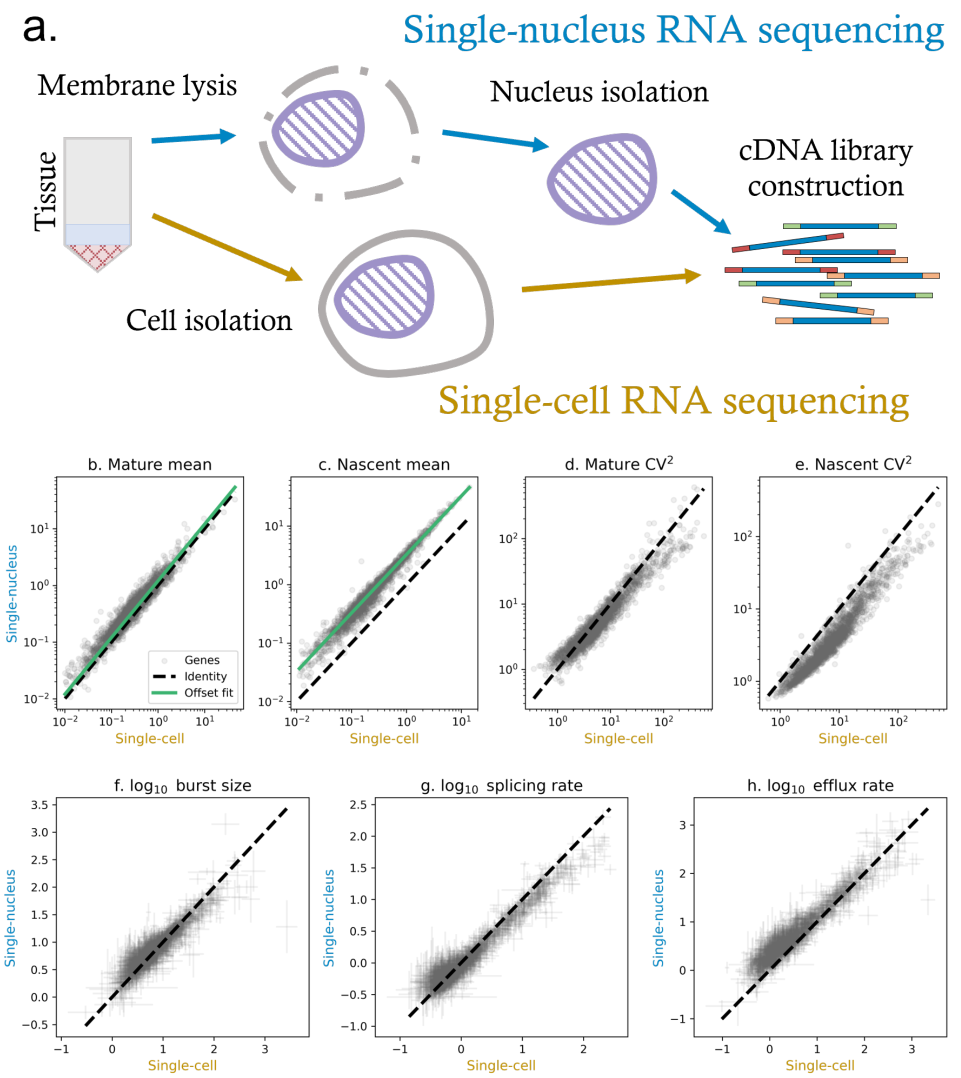

Figure 3: Single-cell and single-nucleus data can be conceptually "integrated" by defining distinct stochastic models that account for differences in mature RNA processing and technical noise, while maintaining the nascent RNA. **a.** Single-cell RNA sequencing protocols dissociate and isolate of individual cells, whereas single-nucleus RNA sequencing isolate nuclei and discard cytoplasmic molecules; however, the downstream sequencing steps may be identical.

b. Counterintuitively, representative paired mouse brain single-cell and single-nucleus datasets exhibit similar mature RNA levels (gray points: genes; dashed black line: line of identity; green line: the approximate average offset observed for single-nucleus data).

c. The single-nucleus dataset consistently has considerably higher nascent RNA counts, which suggests the presence of a technical effect between the two technologies (conventions as in **b**).

d. The single-nucleus dataset demonstrates slightly lower noise levels for mature count data (gray points: genes; dashed black line: line of identity).

e. The single-nucleus dataset demonstrates considerably lower noise levels for nascent count data (conventions as in **d**).

f.-g. By fitting mechanistic models to both datasets, we can identify technical noise parameters that produce consistent burst and splicing parameters between the technologies (points: maximum likelihood estimates for burst sizes and splicing rates; error bars: conditional 99% confidence intervals for inferred parameters; dashed black line: line of identity).

h. At the discovered technical noise parameters, the mature RNA efflux or turnover is considerably higher for the single-nucleus dataset, consistent with this parameter's interpretation as the rapid export from the nucleus (conventions as in **f**-**g**).

2.4 Mechanistic basis for differential expression analysis

In typical transcriptomics workflows, the determination of differences between cell types or conditions often reduces to the determination of differentially expressed (DE) genes, which exhibit statistically significant differences in their average copy numbers. However, the identification of DE genes requires careful accounting for technical covariates [18]. In addition, the data may exhibit *compensating* mechanistic effects that change the distribution while keeping the averages constant, which would not be identifiable by standard statistical methods [33, 61–63].

We propose that differential expression testing should be generalized to the identification of modulated parameters. We use the notation "DE-θ" to denote criteria using θ – which may be a data moment or an inferred biophysical parameter – as a test statistic. In particular, we stress the potential of multivariate data, which provides more statistical power [7] and enables the identification of parameter modulation patterns which would not otherwise be identifiable (Section 5.3.1).

For illustration, we revisit and extend an analysis performed in [33]: we used *Monod* to fit counts from pre-clustered glutamatergic and GABAergic cell types in four mouse brain datasets [40], then selected genes that appeared to be DE-θ for the burst size, splicing rate, or degradation rate. We further filtered for genes which were not DE-μ_M, i.e., had a low average difference in mature RNA expression between the cell types. The full analysis procedure is given in Section 5.3.2.

Based on the pervasive co-variation of splicing and degradation rate differences (Figure S39 of [33]), as well as physical considerations (Section 5.3.1), we suggest that this co-variation should properly be ascribed to *burst frequency* modulation, even though this parameter was not explicitly fit. Therefore, we summarized the findings further (Figure 4a) in terms of burst size and frequency differences, in the spirit of [26, 64]. The discovered genes are indicated according to the cell type differences' effect on noise: genes highlighted in red exhibit more overdispersion in the glutamatergic population, whereas genes highlighted in light teal exhibit more overdispersion in the GABAergic population. These genes exhibit only minor differences in average expression, and fall fairly close to the line of expression identity (solid diagonal line in Figure 4a), where an increase in burst size is precisely compensated by a decrease in the burst frequency. The differences in parameters are reflected in the data distributions and the model fits. For example, *Nin* and *Bach2*, which are involved in neuronal development, visually exhibit higher noise in the glutamatergic and GABAergic populations, respectively (Figure 4b). The mature count averages are, on the other hand, fairly close (*Nin* Glu: 1.7, GABA: 0.98; *Bach2* Glu: 0.87, GABA: 1.4).

In addition to identifying distributional differences in a small number of markers, the mechanistic approach also enables the summary of more far-reaching perturbations that move beyond the usual marker gene paradigm. For example, recent studies found that the introduction of a modified nucleotide (IdU) to a culture medium enhances transcriptional noise, but keeps average expression constant, hinting at a genome-wide mechanism for compensation [63, 65].

Although the model required to fully recapitulate the dynamics of DNA damage repair involved in this process is sophisticated, we found that we could characterize the effects of IdU using a simple bursty model. The analysis procedure is fully described in Section 5.3.3. In brief, we fit the nascent and mature data from control and IdU datasets using *Monod*. As in [33], the technical noise parameters were not readily identifiable from the 10x v2 sequencing data. We assumed the parameters were in a region we previously discovered for this technology (Figure 3e of [33]), and analyzed biophysical parameters under that assumption (Figure S7).

We found that the IdU-perturbed cell culture exhibited striking noise amplification, with very limited differences in mean expression (Figure 4c). This result strongly contrasts, e.g., Figure 4a and Section S7.10.4 of [33], which show fairly symmetric noise amplification and reduction between cell types. The asymmetry in the findings are consistent with the authors' conclusions and orthogonal validation, which likewise found that burst size increases and burst frequency decreases in the IdU condition [65].

We selected a set of well-fit genes that exhibited particularly high modulation and had average expression greater than 1 in at least one of the conditions for further analysis, identifying *Stx7*, *Washc5*, *Apod*, *Eif2ak2*, *Ubr2*, *Cnnm2*, *Dram2*, *Zfp110*, *Cul4a*, *Ddx19b*, and *Yap1* (red points in Figure 4c). Interestingly, two of these genes are directly related to the DNA damage activity of IdU: *Dram2* is involved in the autophagic response to DNA damage repair, whereas *Cul4a* is involved in the turnover of DNA repair proteins. Several other genes more generally mediate the cellular stress response: *Zfp110*, *Eif2ak2*, and *Yap1* regulate apoptosis, whereas *Ddx19b* may be active in stress granules. The role of the remaining genes is obscure: *Stx7* and *Washc5* are related to vesicular function, *Apod* is involved in lipid metabolism, *Ubr2* controls ubiquitination, and *Cnnm2* appears to be involved in ion transport [66].

We were able to partially compare our results for *Sox2*, *Nanog*, and *Mtpap*, whose transcriptional parameters were computed from fluorescence data in [63, 65]. We did not observe *Sox2* expression in either dataset. *Nanog* was rejected by our goodness-of-fit procedure. This is, in principle, consistent with the results in Table S2 of [63], which report gene on fractions near 30-55%; this regime violates the assumptions of the bursty model (gene on fraction tending to zero). The inferred signs for *Mtpap* parameter modulation agreed with Figure S4 of [65], although we obtained rather different magnitudes ($\log_2$ fold changes of ≈ -0.3 by smFISH vs. ≈ -1.5 by *Monod* for burst frequency; ≈ 2 by smFISH vs. ≈ 1.3 by *Monod* for burst size). Therefore, although the genome-wide trends broadly recapitulate the mechanistic explanations provided by the authors, and some of the high-noise genes appear to be implicated in DNA repair and stress, the quantitative comparison of fluorescence and sequencing data requires further analytical work.

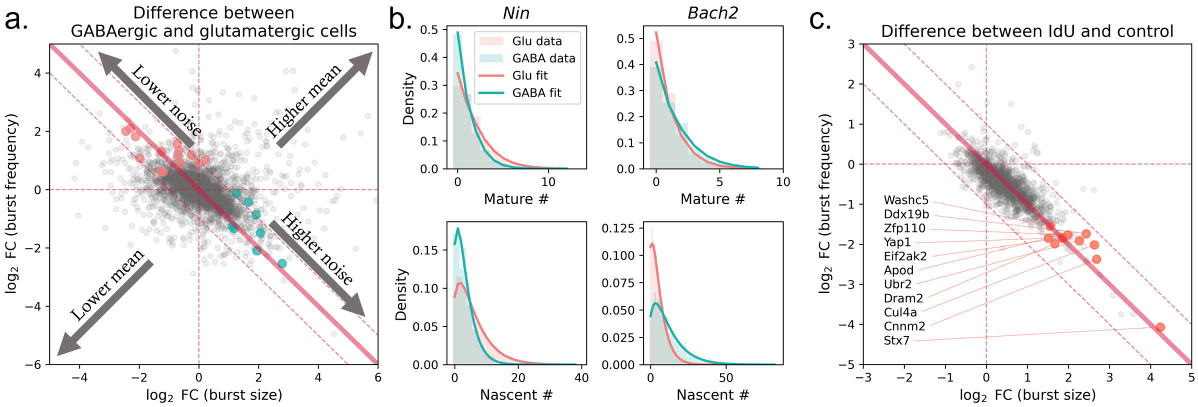

Figure 4: The inference of mechanistic parameters with *Monod* allows us to generalize differential expression testing to the identification of genes with distributional differences, without requiring substantial changes in average expression.

a. The differences between mouse glutamatergic and GABAergic cell types, computed from four independent replicates, include genes with substantial noise enhancement but little to no change in average expression, which may reflect biophysically important compensation mechanisms (light red points: genes with significantly higher noise in glutamatergic cells; light teal points: genes with significantly higher noise in GABAergic cells; gray points: all other genes; solid diagonal line: parameter combinations where burst size and frequency differences compensate to maintain a constant average expression; dashed diagonal lines: $\pm 1 \log_2$ expression fold change region about the constant-average expression line; vertical and horizontal lines: parameter combinations where burst size and frequency, respectively, do not change; fits originally performed in [33]).

b. Differences in inferred noise behaviors reflect differences in distribution shapes (light red: glutamatergic cell type; light teal: GABAergic cell type; histograms: raw counts; lines: *Monod* fits; top row: mature RNA marginal; bottom row: nascent RNA marginal).

c. Perturbation by IdU, which triggers DNA damage and repair, rarely changes expression levels, but induces genome-wide noise enhancement [65] detectable by *Monod* (lines and gray points: as in **a**; red points and labels: well-fit, moderate-expression genes identified as highly noise-enhanced).

3 Discussion

Our Python software *Monod* facilitates mechanistic inference from multimodal scRNA-seq data. At this time, it is restricted to a narrow set of transcriptional models (tractable by quadrature), technical noise models (catalysis or drop-out), modalities (nascent and mature RNA), correlation structure (no inter-gene relationships), and heterogeneity structures (a single homogeneous cell type). While some of these assumptions may be simplistic, the current approach to single-cell RNA sequencing analysis corresponds to an even more unrealistic model, which makes contradictory implicit assumptions and violates fundamental constraints. With even a basic mechanistic model for integrating nascent and mature RNA counts, we have demonstrated the possibility for interesting discovery. Technical noise may be described in a self-consistent fashion; single-nucleus and single-cell sequencing data can be described in a common framework; subtle distributional differences between pre-annotated cell types can be identified and ascribed to biophysical phenomena. We anticipate that *Monod* can be extended to utilize multimodal data to parametrize more complex mechanistic models.

4 Acknowledgments

We thank Sina Booeshaghi, Maria Carilli, Tara Chari, and Meichen Fang for useful discussions in the course of developing *Monod*. The *Monod* package uses algorithms implemented in the *NumPy* [67], *SciPy* [68], and *numdifftools* [69] Python packages. G.G. and L.P. were partially funded by NIH U19MH114830 and NIH 5UM1HG012077-02.

5 Methods

5.1 Normalization and dimensionality reduction

We would like to characterize the performance of typical data processing and dimensionality reduction techniques. To do so, we need a meaningful baseline for "good" performance. For the purposes of illustration, we focus on the methods' effects on cell type differences, whose characterization is a commonplace application of single-cell analyses [18]. In this section, we seek to outline and apply a framework for investigating these methods' implicit assumptions and distortive effects.

We essentially have three options for constructing a baseline for studying heterogeneity, which have different trade-offs. First, we can define a stochastic model under a particular set of hypotheses, simulate from it, and compare the algorithm performance to the underlying ground truth. However, this approach may be overly simplistic, as the simulation may not accurately represent all features of the underlying data-generating process. Second, we can obtain datasets collected from distinct tissues, concatenate them, and treat them as a single dataset. However, this approach is somewhat artificial and divorced from typical use cases, which treat a single tissue. In addition, there may be hard-to-characterize technical batch effects between datasets. Third, we can obtain a pre-annotated dataset from a single tissue, and perform the analysis conditional on the assumption that the annotations are sufficiently accurate for our purposes. Although this approach necessarily represents a compromise, we use it for simplicity.

5.1.1 Variance decomposition baseline

Given a generic set of cell populations, indexed by κ, we can construct an estimate for the amount of biological variation. Under mild assumptions about technical noise, the overall biological variation is bounded from below by the inter-cell population variability. Quantitatively, this property holds for the squared coefficient of variation (CV^2), which we denote by η^2 in our derivations.

First, we construct a categorical distribution $\boldsymbol{\pi}$ that contains the fractional cell type abundances π_κ. From the laws of total expectation and variance [70], we obtain the sample-wide mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$ of a particular gene's counts, prior to corruption by technical noise:

$$\begin{aligned} \tilde{\mu} &= \mathbb{E}_{\boldsymbol{\pi}}[\tilde{\mu}_\kappa] = \sum_\kappa \pi_\kappa \tilde{\mu}_\kappa \\ \tilde{\sigma}^2 &= \sum_\kappa \pi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2, \end{aligned} \tag{3}$$

where $\tilde{\mu}_\kappa$ is the mean and $\tilde{\sigma}_\kappa^2$ is the variance of each discrete cell population.

The observed statistics contain contributions from technical noise:

$$\begin{aligned} \mu_\kappa &= \xi_\kappa \tilde{\mu}_\kappa \\ \sigma_\kappa^2 &= \Xi_\kappa \tilde{\sigma}_\kappa^2. \end{aligned} \tag{4}$$

To relate the observations to the biological processes, we make two assumptions about the form of the technical noise. First, we assume that the sequencing process uniformly samples cells from the underlying population. In other words, we suppose that the observed cell type proportions match the biological proportions. This allows us to write down an analogous decomposition:

$$\begin{aligned} \mu &= \sum_\kappa \pi_\kappa \mu_\kappa = \sum_\kappa \pi_\kappa \xi_\kappa \tilde{\mu}_\kappa, \\ \sigma^2 &= \sum_\kappa \pi_\kappa \sigma_\kappa^2 + \sum_\kappa \pi_\kappa (\mu_\kappa - \mu)^2 \\ &= \sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\xi_\kappa \tilde{\mu}_\kappa - \mu)^2. \end{aligned} \tag{5}$$

Second, we assume that $\xi_\kappa = \xi$ for all κ. In other words, we suppose that, for a particular gene and on average, all cell types are chemically and statistically identical with respect to the sequencing process. We find that the lower moments of the observed distributions can be rewritten in terms of the lower moments of the biological distributions:

$$\begin{aligned} \mu &= \xi \sum_\kappa \pi_\kappa \tilde{\mu}_\kappa = \xi \tilde{\mu}, \\ \sigma^2 &= \sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\xi \tilde{\mu}_\kappa - \xi \tilde{\mu})^2 \\ &= \sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2. \end{aligned} \tag{6}$$

Under this set of assumptions, we find that the ratio of variances with and without technical noise takes the following form:

$$\frac{\tilde{\sigma}^2}{\sigma^2} = \frac{\sum_\kappa \pi_\kappa \tilde{\sigma}_\kappa^2 + \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}{\sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}, \tag{7}$$

which cannot be easily manipulated or constrained in the absence of ground truth statistics. However, if we instead use the ratio of coefficients of variation, we find:

$$\begin{aligned} \frac{\tilde{\eta}^2}{\eta^2} = \frac{\mu^2}{\tilde{\mu}^2}\frac{\tilde{\sigma}^2}{\sigma^2} = \frac{\xi^2 \tilde{\mu}^2}{\tilde{\mu}^2}\frac{\tilde{\sigma}^2}{\sigma^2} = \xi^2 \frac{\tilde{\sigma}^2}{\sigma^2} &= \frac{\xi^2 \sum_\kappa \pi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}{\sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2} \\ &\geq \frac{\xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}{\sum_\kappa \pi_\kappa \Xi_\kappa \tilde{\sigma}_\kappa^2 + \xi^2 \sum_\kappa \pi_\kappa (\tilde{\mu}_\kappa - \tilde{\mu})^2}, \end{aligned} \tag{8}$$

i.e., the fraction of biological variability (as quantified by the CV^2) is at least as high as the fraction of variability attributable to the inter-population mean differences.

5.1.2 Non-mechanistic model definition

Intuitively, we should hope that transformations commonly applied to "denoise" scRNA-seq data retain the biological variability of interest. For example, if we would like to preserve the quantitative relationships between cell types, we seek to keep the per-gene noise post-transformation above the lower bound derived in Section 5.1.1; if this lower bound is violated, some cell type differences have been degraded.

Transformations are iteratively applied to an entire data matrix $\mathcal{Z}$, with entries $\mathcal{Z}_{ij}$ indexing over cells $i = 1, \ldots, c$ and genes $j = 1, \ldots, g$. We conceptualize a transformation as some function

$\Phi_\ell(\mathcal{Z})$, such that $\Phi_\ell = \phi_\ell \circ \cdots \circ \phi_1$. Thus, for example, "proportional fitting" count normalization followed by log-transformation (log1pPF [34]) would be represented by a composition of $\ell = 2$ transformations, with

$$\begin{aligned} C_{\mathcal{Z}} &= \frac{1}{c} \sum_i \left[\sum_j \mathcal{Z}_{ij} \right] \text{ being the size factor,} \\ \phi_1(\mathcal{Z})_{ij} &= \frac{\mathcal{Z}_{ij}}{\sum_j \mathcal{Z}_{ij}} \times C_{\mathcal{Z}}, \text{ and} \\ \phi_2(\mathcal{Z})_{ij} &= \log(1 + \mathcal{Z}_{ij}). \end{aligned} \tag{9}$$

This formulation assumes that each ϕ is a function that maps a $c \times g$ matrix to another $c \times g$ matrix. However, some transformations, such as principal component analysis (PCA), accomplish dimensionality reduction, and map a $c \times g$ matrix to a $c \times g'$ matrix, with $g' \leq g$. Such a projection ψ places the data onto a lower-dimensional manifold within g. We can characterize how much variance is retained by such a projection by applying an inverse transformation, such that a dimensionality-reducing step's $\phi = \psi^{-1} \circ \psi$. This inverse transformation is typically not unique, and may not be deterministic. However, if the cell populations largely lie along the low-dimensional manifold, we should expect the "denoising" steps to have a minimal effect on the variance thus removed.

5.1.3 Mechanistic model definition

The approach outlined in Section 5.1.1 makes fairly mild assumptions about the distributions to obtain a limit on the fraction of biological variability. By making stronger assumptions, we can obtain point, rather than region estimates, at the cost of potential model misspecification or inaccuracy in estimated parameters.

For example, if we are interested in the CV^2 values for mature RNA with and without technical noise, we can immediately exploit the analytical statistics reported in Section S2.4:

$$\begin{aligned} \tilde{\mu}_\kappa &= \frac{b}{\gamma}, \\ \tilde{\sigma}^2_\kappa &= \tilde{\mu}_\kappa \left(1 + \frac{b\beta}{\beta + \gamma}\right), \\ \mu_\kappa &= \tilde{\mu}_\kappa \lambda, \\ \sigma^2_\kappa &= \mu_\kappa \left[1 + \lambda \left(1 + \frac{b\beta}{\beta+\gamma}\right)\right], \\ \tilde{\eta}^2 &= \frac{\mathbb{E}_{\boldsymbol{\pi}}[\tilde{\sigma}^2_\kappa] + \mathrm{Var}_{\boldsymbol{\pi}}(\tilde{\mu}_\kappa)}{\mathbb{E}_{\boldsymbol{\pi}}[\tilde{\mu}_\kappa]^2}, \\ \eta^2 &= \frac{\mathbb{E}_{\boldsymbol{\pi}}[\sigma^2_\kappa] + \mathrm{Var}_{\boldsymbol{\pi}}(\mu_\kappa)}{\mathbb{E}_{\boldsymbol{\pi}}[\mu_\kappa]^2}. \end{aligned} \tag{10}$$

To lighten the notation, we elide the population-specific subscripts κ on the parameters b, β, γ, and λ. Therefore, the fraction of biological variability under this model can be computed by separately fitting the cell types, then substituting in the maximum likelihood estimates to obtain $\tilde{\eta}^2/\eta^2$.

5.1.4 Data processing

To investigate the effect of common data transformations and compare them to the bound from Section 5.1.1, we used the glutamatergic cell subtypes reported for the mouse sample B08, originally generated by the Allen Institute for Brain Science [40].

Preprocessing. We used *kallisto* | *bustools* 0.26.0 to pre-process data. We downloaded a pre-built *M. musculus* genome from `https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest` (mm10, 2020-A version). To build intronic and exonic references, we used the `kb ref` function with the `--lamanno` option. We obtained the raw FASTQ files for the dataset B08 [40, 71], which was generated using the 10x v3 single-cell chemistry. Next, we used the `kb count` function with the `--lamanno` option, as well as the `-x 10xv3` whitelist option to quantify the datasets, producing unspliced (intron-containing) and spliced (non-intron-containing) RNA count matrices [72, 73].

Filtering. We filtered the dataset to remove "low-quality" cells or empty droplets. First, we removed all barcodes that did not pass the default *kallisto* | *bustools* filter. Next, we removed all cell barcodes that were associated with fewer than 10^4 molecular barcodes, computed over all genes, corresponding to standard knee plot filtering procedures (Figure S2).

Next, we split the dataset by cell type and subtype annotations [40]. We extracted seven classes: six glutamatergic subtypes (L2/3 IT, L5 IT, L6 IT, L5/6 NP, L6 CT, and L6b) and their union ("glutamatergic"). We omitted low-abundance cell types (L6 IT Car3 and L5 ET, with fewer than ten barcodes) from analysis and inclusion in the "glutamatergic" category.

Next, we used *Monod* 0.2.6.0 to extract genes with moderate to high expression. We removed two sets of genes: those with very low observed average and maximum expression ($\overline{X}_N \leq 0.01$, $\overline{X}_M \leq 0.01$, $\max X_N \leq 3$, $\max X_M \leq 3$), and those with excessively high observed maximum expression, which are too computationally intensive to fit ($\max X_N \geq 400$, $\max X_M \geq 400$). We use the notation X_z to denote the observed distribution of species $\mathcal{X}_z$ (nascent or mature) for a particular gene; $\overline{X}_z$ is the observed mean and $\max X_z$ is the observed maximum. This procedure produced a set of 2,951 genes that met the thresholds in all of the cell populations.

As high-expression, high-variability genes are typically of most interest in single-cell analyses, we further selected the top 5% of genes by expression, and colored them orange in all visualizations. These genes tended to have the highest variance in the dataset. This selection procedure is shown in Figure 1b.

Baseline computation. To calculate the baseline introduced in Section 5.1.1 for each gene, we used a plug-in estimate for the lower bound in Equation 8, using only the mature RNA data. Specifically, the bound affords the consistent estimator

$$f_{\text{baseline}} = \frac{1}{S^2} \sum_{\kappa} \frac{c_\kappa}{c} \left((\overline{X}_M)_\kappa - \overline{X}_M \right)^2, \tag{11}$$

where S^2 is the sample variance over all glutamatergic cells, $(\overline{X}_M)_\kappa$ is the average expression in cell subtype κ, and c_κ is the number of cells in that subtype. We used the existing annotations to extract the κ-indexed variables.

To represent the admissible region, wherein the retained fraction CV^2 after transformation is no lower than the fraction of CV^2 attributable to cell type differences, we plotted it as a teal line. In addition, we plotted the location at which the CV^2 after transformation exceeds that of the original dataset. Violations of this upper bound are not necessarily a cause for concern. For example, a transformation may, in principle, effectively inflate differences between cell types to make them more distinguishable.

Computation of variability retained by transformation. We considered the effects of four transformations of the glutamatergic mature RNA count matrix: proportional fitting normalization, log-transformation, principal component analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP) [74]. These transformations take place in sequence: log-transformation is applied to normalized data; PCA is applied to log-transformed data; UMAP is applied to PCA-transformed data. This series of transformations is associated with four ϕ and Φ functions, as well as five $c \times g$ $\mathcal{Z}$ matrices:

$$\begin{aligned}
&\mathcal{Z}, \text{ the raw mature count matrix,}\\
\mathcal{Z}_1 &= \Phi_1(\mathcal{Z}) = \phi_1(\mathcal{Z}), \text{ the normalized mature count matrix,}\\
\mathcal{Z}_2 &= \Phi_2(\mathcal{Z}) = \phi_2(\mathcal{Z}_1), \text{ the log1pPF transformed mature count matrix,}\\
\mathcal{Z}_3 &= \Phi_3(\mathcal{Z}) = \phi_3(\mathcal{Z}_2), \text{ its reconstructed PCA projection, and}\\
\mathcal{Z}_4 &= \Phi_4(\mathcal{Z}) = \phi_4(\mathcal{Z}_3), \text{ its reconstructed UMAP projection.}
\end{aligned} \tag{12}$$

The first two transformations, ϕ_1 and ϕ_2, corresponding to normalization and log-transformation, are reported in Equation 9.

The reconstructed PCA projection is given by $\psi_3^{-1} \circ \psi_3$, where ψ_3 is implemented through the *Scikit-learn* 1.0.1 function `transform` and ψ_3^{-1} is implemented through the function `inverse_transform`, both associated with a `sklearn.decomposition.PCA` object [75]. We used 50 components to compute the PCA projection.

The reconstructed UMAP projection is given by $\psi_3^{-1} \circ \psi_4^{-1} \circ \psi_4 \circ \psi_3$, where ψ_4 is implemented through the *umap-learn* 0.5.1 function `transform` and ψ_4^{-1} is implemented through the function `inverse_transform`, both associated with a `umap.UMAP` object [74]. We used the 50-dimensional PCA projection and default parameters to compute the UMAP projection.

To compute the amount of retained CV^2 for each gene j and each step indexed by l, we computed the ratio of coefficients of variation prior to and after transformation:

$$
\begin{aligned}
\overline{\mathcal{Z}}_j &= \frac{1}{c}\sum_i \mathcal{Z}_{ij} \\
S_j^2 &= \frac{1}{c}\sum_i (\mathcal{Z}_{ij} - \overline{\mathcal{Z}}_j)^2 \\
\eta_j^2 &= \frac{S_j^2}{\overline{\mathcal{Z}}_j^2} \\
(\overline{\mathcal{Z}}_l)_j &= \frac{1}{c}\sum_i (\mathcal{Z}_l)_{ij} \\
(S_l)_j^2 &= \frac{1}{c}\sum_i ((\mathcal{Z}_l)_{ij} - (\overline{\mathcal{Z}}_l)_j)^2 \\
(\eta_l)_j^2 &= \frac{(S_l)_j^2}{(\overline{\mathcal{Z}}_l)_j} \\
f_{j,l,ret} &= \frac{(\eta_l)_j^2}{\eta_j^2}.
\end{aligned}
\tag{13}
$$

To characterize the overall effect of the cumulative application of transformations, we plotted the distributions of $(\eta_l)_j^2$ – i.e., the transformed data CV^2 – after each transformation. This analysis is shown in Figure 1c.

To characterize the relationship between the average expression and the fraction of variation attributed to biological variability of interest, we plotted $f_{j,l,ret}$ against the mean $\overline{\mathcal{Z}}_j$ for each gene. The normalization and dimensionality reduction procedures attempt to eliminate noise while maintaining biological "signal," and this visualization reveals whether an increase in variability is implicitly attributed to the former or the latter. This analysis is shown in Figure 1d-g.

To understand the transformations' effect on distributions, we plotted the $f_{j,l,ret}$ value for each step l against the baseline value $f_{j,\text{baseline}}$, computed using Equation 11 separately for each gene j. To easily compare the results to the threshold, we plotted the admissible region. The lower bound represents the reduction of inter-cell type variability, whereas the upper bounds represents the inflation of overall variability above its original value. This analysis is shown in Figure 1h-k. In addition, we computed the fraction of genes violating the bound at each step: 0%, 8.2%, 35.2%, and 7.6% after proportional fitting, log-transformation, PCA, and UMAP respectively.

As the transformations are applied cumulatively, the distribution at step l may fall within the admissible region, but still be quantitatively degraded because it fell outside it at some step $l' < l$. To characterize the loss of quantitative information about cell type relationships, we plotted the same data points as above, and colored them according to the history of the analysis. If a gene has ever violated the lower bound, we plotted it in a violet color. This analysis is shown in Figure S8. In addition, we computed the fraction of genes identified after each step. We found that 0%, 8.2%, 35.2%, and 35.4% of the genes had at some point violated the bound after proportional fitting, log-transformation, PCA, and UMAP respectively.

Computation of biological variability under a mechanistic model. To fit a mechanistic model, we used *Monod* 0.2.6.0. We set up a 20 × 21 grid over the $\{\log_{10} C_N, \log_{10} \lambda_M\}$ domain listed in Table S7. These bounds were chosen according to the results previously obtained for mouse brain datasets, as reported in Figure S24 of [33].

At each grid point, we iterated over the 2,951 genes, using gradient descent to identify the conditional maximum likelihood estimate of $\{\log_{10} b,\ \log_{10}\beta,\ \log_{10}\gamma\}$, where the rates β and γ are defined in units of burst frequency k (Section S4.3.2 of [33]). We used the conditional method of moments estimate (Table S6, "Bursty") as the starting point and performed 15 steps of gradient descent. The procedure was parallelized over up to sixty processors (Intel Xeon Gold 6152, 2.10GHz). Runtimes varied between 33 minutes for the smallest dataset and 2.7 hours for the largest.

To identify the optimal sampling parameters, we identified the grid point with the lowest total Kullback-Leibler divergence, computed over all genes. To ensure we obtained the true optima under the bursty model, we performed four rounds of fixed-point iteration. First, we rejected a subset of genes if they were detected by the chi-squared test with $p = 0.001$ with a Bonferroni correction, and their Hellinger distance from the data distribution exceeded 0.05. Next, we recalculated the optimum based on the remaining data (Section S4.3.5 of [33]), and repeated the procedure. This procedure did not change the optimum for any of the datasets. Further, we investigated the stability of the optima under gene subsampling, and found them to be stable and consistent (Section S4.3.5 of [33]).

Although the optima discovered for the cell subtypes were fairly close, they were not identical, with smaller datasets showing striking deviations (Figure S5). From physical considerations, we assumed the subtypes, which originate from a single technical sample, have the same set of sampling parameters. For simplicity, we assumed the parameter set inferred from the entire glutamatergic dataset provided a sufficiently accurate estimate for all of its constituent subtypes, and analyzed the data under that set of $\{\log_{10} C_N, \log_{10}\lambda_M\}$.

To compute the fraction of biological variability, we used the identities in Equation 10 using the parameters inferred for each gene j:

$$f_{j,bio} = \frac{\tilde{\eta}_j^2}{\eta_j^2}. \tag{14}$$

To characterize the relationship between the average expression and the fraction of variation attributed to biological variability of interest, we plotted $f_{j,bio}$ against the mean $\overline{\mathcal{Z}}_j$ for each gene. This visualization reveals whether an increase in variability is attributed to biological or technical effects. This analysis is shown in Figure 2b.

To understand the fits' sensibility, we plotted the $f_{j,bio}$ value against the baseline value $f_{j,\text{baseline}}$, computed using Equation 11 separately for each gene j. To easily compare the results to the threshold, we plotted the admissible region. This analysis is shown in Figure 2c.

If the fit is sufficiently good, Equation 10 naturally enforces the bound in Equation 11. To understand whether a mechanistic analysis provides actionable information, or merely exploits external information about cell types, we used the fit to the entire dataset to repeat the analysis. This approach introduces some error, as we neglect intra-cell type differences altogether, and do not use goodness-of-fit testing to omit genes that show subtype heterogeneity. If the trends look substantially similar, the results suggest that the *Monod* procedure attributes the vast majority of CV^2 to intrinsic and bursting noise, with only a minor fraction ascribed to inter-subtype differences. In other words, if we do not use $\boldsymbol{\pi}$ at all, but still obtain similar results, the agreement with the bound is not an incidental consequence of the expectation over $\boldsymbol{\pi}$ in the last two lines of Equation

10. To compute $\tilde{\eta}^2/\eta^2$ in this scenario, we used the mean and variance identities in the first four lines of Equation 10. These results are shown in Figure S10a-b.

Comparisons between transformation and the mechanistic model. To compare the attributions implicit in the transformation procedures and the mechanistic fit, we plotted them against each other. If the two methods agree, they should lie on the line of identity. The results are shown in Figure S9 for the *Monod* subtype analysis and Figure S10c-f for the analysis of the entire dataset.

5.2 Nuclear data integration

We would like to coherently integrate single-cell and single-nucleus RNA sequencing data. To do so, we need to specify the relationship between the two modalities. We can establish such a relationship from first principles by making assumptions regarding the underlying biophysical processes. For example, by proposing that nascent RNA are restricted to the nucleus, we can reasonably assume that the nascent RNA dynamics should be identical between the two modalities. On the other hand, the mature RNA distributions and dynamics may have substantial differences, as nuclei are depleted in this species relative to the entire cell. In this section, we propose a possible foundation for the integration of these modalities.

5.2.1 Mechanistic model definition

To describe the stochastic dynamics and sampling in a single-cell dataset, we use the formulation given in Section 2.1 and outlined in more detail in [33]. To connect this model to nuclear data, we note that formally, it can arise from the following model:

$$\varnothing \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_{M,nuc} \xrightarrow{\gamma_e} \mathcal{X}_{M,cyt} \xrightarrow{\gamma_c} \varnothing. \tag{15}$$

where $\mathcal{X}_{M,nuc}$ and $\mathcal{X}_{M,cyt}$ are nuclear and cytoplasmic mature RNA species, respectively. The rate γ_e describes the efflux of nuclear RNA, whereas the rate γ_c describes the degradation of cytoplasmic RNA.

In the limit $\gamma_c \ll \gamma_e$ or $\gamma_c \gg \gamma_e$, the model in Section 2.1 approximately holds for cytoplasmic data: if one of these stages is considerably longer-lived, the two-stage processing of mature RNA can be effectively described by a one-stage model. In this case, γ can be interpreted as the lower rate. We typically assume that the first limit is most relevant, although orthogonal data suggest that the details are highly gene- and tissue-dependent [58]. We note that it is, in principle, straightforward [59] implement a model that explicitly incorporates both parameters; however, for computational facility, we use the simpler reduced model and discard genes that fail to fit it. On the other hand, for nuclear data, the model holds for $\gamma = \gamma_e$.

5.2.2 Data processing

To compare the distributions of single-cell and single-nucleus datasets and explain them using a mechanistic argument, we used mouse neuron datasets generated by 10x Genomics.

Preprocessing. We used *kallisto* | *bustools* 0.26.0 to pre-process data. We downloaded a pre-built *M. musculus* genome from `https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest` (mm10, 2020-A version). To build intronic and exonic references, we used the `kb ref` function with the `--lamanno` option. We obtained the raw FASTQ files for the "Brain 4" and "Brain Nuclei 4" datasets from two multiplexing experiments, both generated using the 10x v3 chemistry. We selected these datasets because they had the highest average molecule counts per cell in both technologies. Next, we used the `kb count` function with the `--lamanno` option, as well as the `-x 10xv3` whitelist option to quantify the datasets, producing unspliced (intron-containing) and spliced (non-intron-containing) RNA count matrices [72, 73].

Filtering. We filtered the dataset to remove "low-quality" cells or empty droplets. First, we removed all barcodes that did not pass the default *kallisto* | *bustools* filter. Next, we removed all cell barcodes that were associated with fewer than T molecular barcodes ($T = 3 \times 10^3$ for sc and 6×10^3 for sn), computed over all genes, corresponding to standard knee plot filtering procedures (Figure S3). In addition, we removed cells with more than 10^5 barcodes, as they may reflect obscure technical noise sources unique to single-nucleus data.

Next, we used *Monod* 0.2.6.0 to extract genes with moderate to high expression. We removed two sets of genes: those with very low observed average and maximum expression ($\overline{X}_N \leq 0.01$, $\overline{X}_M \leq 0.01$, $\max X_N \leq 3$, $\max X_M \leq 3$), and those with excessively high observed maximum expression, which are too computationally intensive to fit ($\max X_N \geq 400$, $\max X_M \geq 400$). This procedure produced a set of 5,690 genes that met the thresholds in all of the cell populations. We randomly selected 2,000 genes for further analysis.

Inference and analysis of biophysical parameters. To fit a mechanistic model, we used *Monod* 0.2.6.0. We set up a 20×21 grid over the $\{\log_{10} C_N, \log_{10} \lambda_M\}$ domain listed in Table S7.

At each grid point, we iterated over the 2,000 genes, using gradient descent to identify the conditional maximum likelihood estimate of $\{\log_{10} b,\ \log_{10} \beta,\ \log_{10} \gamma\}$, where the rates β and γ are defined in units of burst frequency k. We used the conditional method of moments estimate as the starting point and performed 15 steps of gradient descent. The procedure was parallelized over up to fifteen processors (Intel Xeon Gold 6152, 2.10GHz). Runtimes varied between 2.2 hours for the whole-cell dataset and 3.8 hours for the nuclear dataset.

To identify the optimal sampling parameters, we identified the grid point with the lowest total Kullback-Leibler divergence, computed over all genes. To ensure we obtained the true optima under the bursty model, we performed four rounds of fixed-point iteration. First, we rejected a subset of genes if they were detected by the chi-squared test with $p = 0.01$ with a Bonferroni correction, and their Hellinger distance from the data distribution exceeded 0.05. Next, we recalculated the optimum based on the remaining data, and repeated the procedure. This procedure did not change the optimum for any of the datasets. Further, we investigated the stability of the optima under gene subsampling, and found them to be stable and consistent.

The optimum discovered for the single-nucleus dataset demonstrated noticeably higher molecule observation probabilities (orange points, Figure S6). This observation was supported by basic observations of the dataset statistics: despite the depletion of cytoplasmic RNA, the single-nucleus dataset had as much mature RNA as the single-cell dataset, and approximately half an order of magnitude more nascent RNA (Figure 2b-c). To illustrate these trends, we computed the offset from the ratio of the dataset-wide means. In addition, the single-nucleus dataset appeared to exhibit lower noise levels (Figure 2d-e). To obtain a quantitative understanding of the average and noise behaviors, we computed the fraction of genes that lay above the line of identity.

From physical considerations, the two independent experiments, performed using different technologies, should not necessarily have the same sampling parameters. However, as the samples were taken from the same tissue, they should have the same physics of transcription and splicing. Therefore, we somewhat arbitrarily assumed that the single-cell optimum was sufficiently accurate, and chose a set of single-nucleus sampling parameters that provided the lowest squared errors for the $\log_{10} b$ and $\log_{10} \beta$ parameters. As shown by the blue points in Figure S6, the optimum so discovered lay approximately half an order of magnitude above the optimum for the single-cell data, and within the top 5th percentile for the sampling parameter likelihood landscape (hatched region).

We analyzed the single-nucleus data under that set of parameters, recomputing the goodness-of-fit statistics accordingly.

To illustrate the differences between the datasets, we plotted the inferred parameters and the identity line. To quantify uncertainty in the parameters, we exploited the Fisher information matrix as described in Section S4.3.4 of [33]; we visualized the error bars, which represent the 99% confidence intervals for the biological parameters, conditional on the sampling parameter values. Finally, we applied a t-test, implemented through `scipy.stats.ttest_ind` [68], to the pairs of single-cell and single-nucleus parameter estimates. We omitted genes rejected by goodness-of-fit procedures from these computations and visualizations.

5.3 Mechanistic differential expression

We seek to move beyond averages and explain the differences between single-cell samples and cell types in terms of biophysical distribution parameters, in the spirit of [76]. In this section, we propose and apply an approach for the identification of cell type differences which would be poorly detectable using standard average-based procedures, and demonstrate its performance using an experiment studying transcriptional noise amplification.

5.3.1 Signatures of frequency modulation

We fit the rate parameters $\log_{10}\beta$ and $\log_{10}\gamma$, setting the burst frequency k to unity. This is formally equivalent to fitting $\log_{10}\frac{\beta}{k}$ and $\log_{10}\frac{\gamma}{k}$: at steady state, the system is characterized by three independent parameters, which cannot be distinguished based on a single dataset.

The models we present are not natively adapted to detect changes in k: to unambiguously distinguish between modulation of upstream and downstream processes, time-resolved data are mandatory. However, the high correlation between the magnitudes of changes in $\log_{10}\frac{\beta}{k}$ and $\log_{10}\frac{\gamma}{k}$ (e.g., as shown in Section S7.10.3 of [33]) highly suggestive of the hypothesized frequency modulation.

We propose that the modulation of k can be motivated by biological argument. β and γ, the rates of splicing and degradation, use a one-step, first-order, memoryless reaction as a highly simplified representation of a series of chemical transformations effected in tandem with a spliceosome or a ribonuclease (RNase) complex respectively. However, spliceosomes and RNases are promiscuous, whereas transcription is highly regulated. Therefore, we hypothesize that targeted modulation of the burst frequency upstream at the gene locus is more mechanistically plausible than the synchronized and targeted modulation of the downstream processes.

If we *assume* β and γ are constant between conditions or cell types, we can compute an estimate of k modulation between population 1 and population 2:

$$\begin{aligned}
\Delta\log_{10}\frac{\beta}{k} &= \log_{10}\frac{\beta_2}{k_2} - \log_{10}\frac{\beta_1}{k_1},\\
\Delta\log_{10}\frac{\gamma}{k} &= \log_{10}\frac{\gamma_2}{k_2} - \log_{10}\frac{\gamma_1}{k_1},\\
\Delta\log_{10}k &\approx -\Delta\log_{10}\frac{\beta}{k} = \log_{10}k_2 - \log_{10}k_1\\
&\approx -\Delta\log_{10}\frac{\gamma}{k} = \log_{10}k_2 - \log_{10}k_1.
\end{aligned} \tag{16}$$

Therefore, if the approximate equality $\Delta\log_{10}\frac{\beta}{k}\approx\Delta\log_{10}\frac{\gamma}{k}$ holds, we can propose that $\Delta\log_{10}k$ has a similar magnitude, but the opposite sign. We average the two to estimate the burst frequency modulation:

$$\Delta\log_{10}k \approx -\frac{1}{2}\left(\Delta\log_{10}\frac{\beta}{k} + \Delta\log_{10}\frac{\gamma}{k}\right) \tag{17}$$

5.3.2 Data processing: mouse neurons

To illustrate the approach, we compared the parameters for glutamatergic and GABAergic cell types from four mouse datasets (B08, C01, F08, and H12) generated by the Allen Institute for Brain

Science [40, 71]. We previously performed the fits and identified genes that suggested substantial parameter modulation [33]. Here, we revisit the fits and summarize the key findings.

First, we used the fits to identify the differentially expressed genes for each parameter (Section S4.6.2 of [33]). We use the notation DE-θ to indicate that the log-parameter θ exhibited a Bonferroni-corrected p-value lower than 0.1 and mean $\log_2$ fold change higher than 1. The $\log_2$ fold changes were defined as the difference between the parameter values in the GABAergic and glutamatergic cell types. We omitted data points that were discarded by goodness-of-fit testing. With this procedure, we identified a set of DE-b, DE-β, and DE-γ genes, separated according to the sign of the $\log_2$ fold change.

We computed the mean $\log_2$ fold difference in mature RNA averages between the cell types, and selected the set of identified DE-θ genes with a magnitude lower than 1. In other words, these genes have detectably large differences in biophysical parameters, but do not, on average, exhibit large differences in mature RNA averages μ_N.

Next, we averaged the mean $\log_2$ fold changes in β and γ to obtain an estimate of the $\log_2$ fold change in the burst frequency, as in Equation 17. We plotted the resulting aggregated fold changes in burst size and burst frequency against each other, highlighting genes that were DE-θ for some biological θ, but not DE-μ_N.

We colored these genes by the effect on noise. It is elementary to show that, if the mean remains constant, a decrease in b compensated by an increase in k – equivalently, decrease in β/k and γ/k – leads the joint distribution of nascent and mature RNA to become bivariate Poisson. For example, if a gene was found to exhibit significantly higher b, β, or γ in GABAergic cells, we assigned it to the GABAergic set, as it suggests relative noise amplification in this cell type. This analysis is shown in Figure 4a.

The structure of this plot bears further discussion, as it provides a convenient summary of useful statistical properties. The solid diagonal line denotes the set of b and k combinations that yield a constant mean (all other parameters held equal). The dashed diagonal lines are offset by unity, and show the range of parameters that give averages with a lower than twofold change in the mean. The dashed vertical and horizontal lines correspond to no change in b and k, respectively. Qualitatively, moving toward the top right corresponds to increasing the mean; moving toward the bottom left corresponds to decreasing the mean; moving toward the top left corresponds to decreasing the noise to the Poisson limit; moving to the bottom right corresponds to increasing the noise.

To demonstrate the qualitative impact of noise modulation, we visualized the distributions, as well as the fits, in both cell types, based on data from the B08 dataset. We selected the genes *Nin* and *Bach2*, which are associated with neuronal development, as discussed in [33]. In addition, we computed these genes' mature count averages in each cell type. This demonstration is given in Figure 4b.

5.3.3 Data processing: mouse embryonic stem cells

To demonstrate the potential of this approach for detecting broad trends in transcriptional modulation without replicates, we considered the transcriptomes of mouse embryonic stem cells with and without 5'-iodo-2'-deoxyuridine (IdU) perturbations. This dataset was generated by Desai et al. [63, 65] to investigate the effect of IdU incorporation on transcriptional bursting properties; the authors found that the perturbation appeared to increase the noise genome-wide, but did not affect averages.

Preprocessing. We used *kallisto* | *bustools* 0.26.0 to pre-process data. We downloaded a pre-built *M. musculus* genome from `https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest` (mm10, 2020-A version). To build intronic and exonic references, we used the `kb ref` function with the `--lamanno` option. We obtained the raw FASTQ files for the DMSO (control) and IdU datasets, both of which were generated using the 10x v2 chemistry. Next, we used the `kb count` function with the `--lamanno` option, as well as the `-x 10xv2` whitelist option to quantify the datasets, producing unspliced (intron-containing) and spliced (non-intron-containing) RNA count matrices [72, 73].

Filtering. We filtered the dataset to remove "low-quality" cells or empty droplets. First, we removed all barcodes that did not pass the default *kallisto* | *bustools* filter. Next, we removed all cell barcodes that were associated with fewer than 4×10^3 molecular barcodes, computed over all genes, corresponding to standard knee plot filtering procedures (Figure S4).

Next, we used *Monod* 0.2.6.0 to extract genes with moderate to high expression. We removed two sets of genes: those with very low observed average and maximum expression ($\overline{X}_N \leq 0.01$, $\overline{X}_M \leq 0.01$, $\max X_N \leq 3$, $\max X_M \leq 3$), and those with excessively high observed maximum expression, which are too computationally intensive to fit ($\max X_N \geq 400$, $\max X_M \geq 400$). This procedure produced a set of 4,373 genes that met the thresholds in both cell populations. We randomly selected 2,000 genes for further analysis, ensuring that the genes analyzed in the previous publications (*Nanog*, *Sox2*, *Pou5f1*, *Klf4*, *Wdr83*, *Stx7*, *Hif1an*, *Mtpap*, *Farsa*, *Wipi2*, and *Snd1*) were included.

Inference and analysis of biophysical parameters. To fit a mechanistic model, we used *Monod* 0.2.6.0. We set up a 20×21 grid over the $\{\log_{10} C_N, \log_{10} \lambda_M\}$ domain listed in Table S7.

At each grid point, we iterated over the 2,000 genes, using gradient descent to identify the conditional maximum likelihood estimate of $\{\log_{10} b,\ \log_{10} \beta,\ \log_{10} \gamma\}$, where the rates β and γ are defined in units of burst frequency k. We used the conditional method of moments estimate as the starting point and performed 15 steps of gradient descent. The procedure was parallelized over up to eighty processors (Intel Xeon Gold 6152, 2.10GHz). Runtimes varied between sixteen and seventeen minutes.

To identify the optimal sampling parameters, we identified the grid point with the lowest total Kullback-Leibler divergence, computed over all genes. To ensure we obtained the true optima under the bursty model, we performed four rounds of fixed-point iteration. First, we rejected a subset of genes if they were detected by the chi-squared test with $p = 0.01$ with a Bonferroni correction, and their Hellinger distance from the data distribution exceeded 0.05. Next, we recalculated the optimum based on the remaining data, and repeated the procedure. This procedure did not change the optimum for any of the datasets. Further, we investigated the stability of the optima under gene subsampling, and found them to be stable and consistent.

The discovered optima were not consistent between datasets (orange points, Figure S6), and the likelihood landscapes were rugged and inconclusive (hatched region, Figure S6). This observation accords with our previous analyses of 10x v2 datasets (e.g., panels a. of figures in Section S7.6 of [33]): the older v2 technology does not appear to provide enough information to identify the technical noise parameters. Therefore, we somewhat arbitrarily used the grid point closest to $\log_{10} C_N = -6.5$, $\log_{10} \lambda_M = -1.2$, near the optimum discovered for a mouse neuron dataset in Figure 3e of [33]. We analyzed the datasets under that set of parameters, recomputing the goodness-of-fit statistics accordingly.

We computed the mean $\log_2$ fold change in burst size and burst frequency (Equation 17), and plotted them against each other, using the conventions in Figure 4. We omitted data points that were discarded by goodness-of-fit testing. Finally, we identified all genes with $\log_2$ change higher than 1.5 in b as well as k, which demonstrated significant noise amplification. To focus on genes with biologically interesting effects, we selected only those which had a mature RNA mean greater than unity in at least one of the conditions, and reported them.

References

[1] Florian Wagner, Dalia Barkley, and Itai Yanai. Accurate denoising of single-cell rna-seq data using unbiased principal component analysis. *BioRxiv*, page 655365, 2019.

[2] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

[3] A Sina Booeshaghi, Ingileif B Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, pages 2022–05, 2022.

[4] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell rna-seq data. *Nature Methods*, pages 1–8, 2023.

[5] Scott R Tyler, Supinda Bunyavanich, and Eric E Schadt. Pmd uncovers widespread cell-state erasure by scrnaseq batch correction methods. *bioRxiv*, pages 2021–11, 2021.

[6] Oleg Lenive, Paul D W Kirk, and Michael P H Stumpf. Inferring extrinsic noise from single-cell gene expression data using approximate bayesian computation. *BMC systems biology*, 10(1):1–17, 2016.

[7] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13(1):7620, December 2022.

[8] Chen Jia and Ramon Grima. Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model. *Iscience*, 26(1):105746, 2023.

[9] Crispin Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer, third edition, 2004.

[10] C. W. Gardiner and S. Chaturvedi. The poisson representation. I. A new technique for chemical master equations. *Journal of Statistical Physics*, 17(6):429–468, December 1977.

[11] Jean Peccoud and Bernard Ycard. Markovian Modeling of Gene Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[12] Donald A. McQuarrie. Kinetics of Small Systems. I. *The Journal of Chemical Physics*, 38(2):433–436, January 1963.

[13] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, June 1961.

[14] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, December 2005.

[15] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, September 2006.

[16] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, 2006.

[17] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[18] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019.

[19] Simone Tiberi. DifferentialRegulation, April 2022.

[20] Páll Melsted, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, July 2021.

[21] Charlotte Soneson, Avi Srivastava, Rob Patro, and Michael B. Stadler. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology*, 17(1):e1008585, January 2021.

[22] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLOS Computational Biology*, 18(9):e1010492, September 2022.

[23] Maria T. Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Mechanistic modeling with a variational autoencoder for multimodal single-cell RNA sequencing data. Preprint, bioRxiv: 2023.01.13.523995, January 2023.

[24] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[25] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty Gene Expression in the Intact Mammalian Liver. *Molecular Cell*, 58(1):147–156, April 2015.

[26] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, October 2012.

[27] Adam M Corrigan, Edward Tunnacliffe, Danielle Cannon, and Jonathan R Chubb. A continuum model of transcriptional bursting. *eLife*, 5:e13051, February 2016.

[28] A. Sanchez and I. Golding. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(6163):1188–1193, December 2013.

[29] Damien Nicolas, Nick E. Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290, 2017.

[30] Joseph Rodriguez and Daniel R. Larson. Transcription in Living Cells: Molecular Mechanisms of Bursting. *Annual Review of Biochemistry*, 89(1):189–212, June 2020.

[31] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–368, July 2016.

[32] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, April 2017.

[33] Gennady Gorin and Lior Pachter. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophysical Reports*, 3(1):100097, March 2023.

[34] A. Sina Booeshaghi, Ingileif B. Hallgrímsdóttir, Angel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. Preprint, bioRxiv: 2022.05.06.490859, May 2022.

[35] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20:296, December 2019.

[36] Michael B. Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, 8(4):315–328.e8, April 2019.

[37] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, December 2019.

[38] Shamus M. Cooley, Timothy Hamilton, J. Christian J. Ray, and Eric J. Deeds. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. Preprint, bioRxiv: 689851, September 2020.

[39] Tara Chari, Joeyta Banerjee, and Lior Pachter. The Specious Art of Single-Cell Genomics. Preprint, bioRxiv: 2021.08.25.457696, September 2021.

[40] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S. Adkins, Andrew I. Aldridge, Seth A. Ament, Anna Bartlett, M. Margarita Behrens, Koen Van den Berge, Darren Bertagnolli, Hector Roux de Bézieux, Tommaso Biancalani, A. Sina Booeshaghi, Héctor Corrada Bravo, Tamara Casper, Carlo Colantuoni, Jonathan Crabtree, Heather Creasy, Kirsten Crichton, Megan Crow, Nick Dee, Elizabeth L. Dougherty, Wayne I. Doyle, Sandrine Dudoit, Rongxin Fang, Victor Felix, Olivia Fong, Michelle Giglio, Jeff Goldy, Mike Hawrylycz, Brian R. Herb, Ronna Hertzano, Xiaomeng Hou, Qiwen Hu, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Yang Eric Li, Jacinta D. Lucero, Chongyuan Luo, Anup Mahurkar, Delissa McMillen,

Naeem M. Nadaf, Joseph R. Nery, Thuc Nghi Nguyen, Sheng-Yong Niu, Vasilis Ntranos, Joshua Orvis, Julia K. Osteen, Thanh Pham, Antonio Pinto-Duarte, Olivier Poirion, Sebastian Preissl, Elizabeth Purdom, Christine Rimorin, Davide Risso, Angeline C. Rivkin, Kimberly Smith, Kelly Street, Josef Sulc, Valentine Svensson, Michael Tieu, Amy Torkelson, Herman Tung, Eeshit Dhaval Vaishnav, Charles R. Vanderburg, Cindy van Velthoven, Xinxin Wang, Owen R. White, Z. Josh Huang, Peter V. Kharchenko, Lior Pachter, John Ngai, Aviv Regev, Bosiljka Tasic, Joshua D. Welch, Jesse Gillis, Evan Z. Macosko, Bing Ren, Joseph R. Ecker, Hongkui Zeng, and Eran A. Mukamel. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, October 2021.

[41] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018.

[42] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746, June 2020.

[43] Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958, October 2017.

[44] Gennady Gorin, Shawn Yoshida, and Lior Pachter. Transient and delay chemical master equations. Preprint, bioRxiv: 2022.10.17.512599, October 2022.

[45] Nicola Thrupp, Carlo Sala Frigerio, Leen Wolfs, Nathan G. Skene, Nicola Fattorelli, Suresh Poovathingal, Yannick Fourne, Paul M. Matthews, Tom Theys, Renzo Mancuso, Bart de Strooper, and Mark Fiers. Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Reports*, 32(13):108189, September 2020.

[46] Alan Selewa, Ryan Dohn, Heather Eckart, Stephanie Lozano, Bingqing Xie, Eric Gauchat, Reem Elorbany, Katherine Rhodes, Jonathan Burnett, Yoav Gilad, Sebastian Pott, and Anindita Basu. Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Scientific Reports*, 10:1535, January 2020.

[47] Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, Emily R. Nadelmann, Kenny Roberts, Liz Tuck, Eirini S. Fasouli, Daniel M. DeLaughter, Barbara McDonough, Hiroko Wakimoto, Joshua M. Gorham, Sara Samari, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Giannino Patone, Joseph J. Boyle, Hongbo Zhang, Hao Zhang, Anissa Viveiros, Gavin Y. Oudit, Omer Ali Bayraktar, J. G. Seidman, Christine E. Seidman, Michela Noseda, Norbert Hubner, and Sarah A. Teichmann. Cells of the adult human heart. *Nature*, 588(7838):466–472, December 2020.

[48] Clayton P. Santiago, Megan Y. Gimmen, Yuchen Lu, Minda M. McNally, Leighton H. Duncan, Tyler Creamer, Linda Orzolek, Seth Blackshaw, and Mandeep Singh. Comparative analysis of

single-cell and single-nucleus RNA-sequencing in a rabbit model of retinal detachment-related proliferative vitreoretinopathy. Preprint, bioRxiv: 2022.11.07.515504, 2022.

[49] Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, Emma Garren, Jeff Goldy, Lucas T. Graybuck, Matthew Kroll, Roger S. Lasken, Kanan Lathia, Sheana Parry, Christine Rimorin, Richard H. Scheuermann, Nicholas J. Schork, Soraya I. Shehata, Michael Tieu, John W. Phillips, Amy Bernard, Kimberly A. Smith, Hongkui Zeng, Ed S. Lein, and Bosiljka Tasic. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE*, 13(12):e0209648, December 2018.

[50] Tallulah S. Andrews, Jawairia Atif, Jeff C. Liu, Catia T. Perciani, Xue-Zhong Ma, Cornelia Thoeni, Michal Slyper, Gökcen Eraslan, Asa Segerstolpe, Justin Manuel, Sai Chung, Erin Winter, Iulia Cirlan, Nicholas Khuu, Sandra Fischer, Orit Rozenblatt-Rosen, Aviv Regev, Ian D. McGilvray, Gary D. Bader, and Sonya A. MacParland. Single-Cell, Single-Nucleus, and Spatial RNA Sequencing of the Human Liver Identifies Cholangiocyte and Mesenchymal Heterogeneity. *Hepatology Communications*, 6(4):821–840, 2022.

[51] Elena Denisenko, Belinda B. Guo, Matthew Jones, Rui Hou, Leanne de Kock, Timo Lassmann, Daniel Poppe, Olivier Clément, Rebecca K. Simmons, Ryan Lister, and Alistair R. R. Forrest. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology*, 21(1):130, December 2020.

[52] John T. Chamberlin, Younghee Lee, Gabor T. Marth, and Aaron R. Quinlan. Variable RNA sampling biases mediate concordance of single-cell and nucleus sequencing across cell types. Preprint, bioRxiv: 2022.08.01.502392, August 2022.

[53] Blue B. Lake, Simone Codeluppi, Yun C. Yung, Derek Gao, Jerold Chun, Peter V. Kharchenko, Sten Linnarsson, and Kun Zhang. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Scientific Reports*, 7(1):6031, July 2017.

[54] Anushka Gupta, Farnaz Shamsi, Nicolas Altemose, Gabriel F. Dorlhiac, Aaron M. Cypess, Andrew P. White, Nir Yosef, Mary Elizabeth Patti, Yu-Hua Tseng, and Aaron Streets. Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Research*, 32(2):242–257, February 2022.

[55] Trygve E. Bakken, Nikolas L. Jorstad, Qiwen Hu, Blue B. Lake, Wei Tian, Brian E. Kalmbach, Megan Crow, Rebecca D. Hodge, Fenna M. Krienen, Staci A. Sorensen, Jeroen Eggermont, Zizhen Yao, Brian D. Aevermann, Andrew I. Aldridge, Anna Bartlett, Darren Bertagnolli, Tamara Casper, Rosa G. Castanon, Kirsten Crichton, Tanya L. Daigle, Rachel Dalley, Nick Dee, Nikolai Dembrow, Dinh Diep, Song-Lin Ding, Weixiu Dong, Rongxin Fang, Stephan Fischer, Melissa Goldman, Jeff Goldy, Lucas T. Graybuck, Brian R. Herb, Xiaomeng Hou, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Baldur van Lew, Yang Eric Li, Christine S. Liu, Hanqing Liu, Jacinta D. Lucero, Anup Mahurkar, Delissa McMillen, Jeremy A. Miller, Marmar Moussa, Joseph R. Nery, Philip R. Nicovich, Sheng-Yong Niu, Joshua Orvis, Julia K. Osteen, Scott Owen, Carter R. Palmer, Thanh Pham, Nongluk Plongthongkum, Olivier Poirion, Nora M. Reed, Christine Rimorin, Angeline Rivkin, William J. Romanow,

Adriana E. Sedeño-Cortés, Kimberly Siletti, Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Xinxin Wang, Fangming Xie, Anna Marie Yanny, Renee Zhang, Seth A. Ament, M. Margarita Behrens, Hector Corrada Bravo, Jerold Chun, Alexander Dobin, Jesse Gillis, Ronna Hertzano, Patrick R. Hof, Thomas Höllt, Gregory D. Horwitz, C. Dirk Keene, Peter V. Kharchenko, Andrew L. Ko, Boudewijn P. Lelieveldt, Chongyuan Luo, Eran A. Mukamel, António Pinto-Duarte, Sebastian Preissl, Aviv Regev, Bing Ren, Richard H. Scheuermann, Kimberly Smith, William J. Spain, Owen R. White, Christof Koch, Michael Hawrylycz, Bosiljka Tasic, Evan Z. Macosko, Steven A. McCarroll, Jonathan T. Ting, Hongkui Zeng, Kun Zhang, Guoping Feng, Joseph R. Ecker, Sten Linnarsson, and Ed S. Lein. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, 598(7879):111–119, October 2021.

[56] 10x Genomics. 30k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Cells Multiplexed, 12 CMOs, Brain 4, March 2021.

[57] 10x Genomics. 30k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Nuclei Multiplexed, 12 CMOs, Brain Nuclei 4, March 2021.

[58] Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662, December 2015.

[59] Gennady Gorin and Lior Pachter. Modeling bursty transcription and splicing with the chemical master equation. *Biophysical Journal*, 121(6):1056–1069, February 2022.

[60] Tatiana Filatova, Nikola Popovic, and Ramon Grima. Statistics of Nascent and Mature RNA Fluctuations in a Stochastic Model of Transcriptional Initiation, Elongation, Pausing, and Termination. *Bulletin of Mathematical Biology*, 83(1):3, January 2021.

[61] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, December 2016.

[62] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336(6078):183–187, April 2012.

[63] Ravi V. Desai, Xinyue Chen, Benjamin Martin, Sonali Chaturvedi, Dong Woo Hwang, Weihan Li, Chen Yu, Sheng Ding, Matt Thomson, Robert H. Singer, Robert A. Coleman, Maike M. K. Hansen, and Leor S. Weinberger. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science*, 373(6557):eabc6506, August 2021.

[64] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Asa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019.

[65] Giuliana P Calia, Xinyue Chen, Binyamin Zuckerman, and Leor S Weinberger. Comparative analysis between single-cell RNA-seq and single-molecule RNA FISH indicates that the pyrimidine nucleobase idoxuridine (IdU) globally amplifies transcriptional noise. Preprint, bioRxiv: 2023.03.14.532632, March 2023.

[66] National Library of Medicine. Gene [Internet], 2004.

[67] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[68] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020.

[69] Per A. Brodtkorb and John D'Errico. numdifftools, June 2021.

[70] Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability.* Texts in Statistical Science. CRC Press, Taylor & Francis Group, 2015.

[71] Allen Institute for Brain Science. FASTQ files for Allen v3 mouse MOp samples, February 2020.

[72] Páll Melsted, Vasilis Ntranos, and Lior Pachter. The barcode, UMI, set format and BUStools. *Bioinformatics*, page btz279, 2019.

[73] Kristján Eldjárn Hjörleifsson, Delaney K. Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. Preprint, bioRxiv: 2022.12.02.518832, December 2022.

[74] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint, arXiv: 1802.03426v2, December 2018. arXiv: 1802.03426.

[75] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[76] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[77] Allen Institute for Brain Science. Analyses for Allen v3 mouse MOp samples, February 2020.

[78] Pavol Bokes, John R. King, Andrew T. A. Wood, and Matthew Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of Mathematical Biology*, 64(5):829–854, April 2012.

[79] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54:1–26, September 2006.

[80] K. Jayakumar and Davis Antony Mundassery. On Moran's Bivariate Gamma and Bivariate Negative Binomial Distribution. *Calcutta Statistical Association Bulletin*, 59(1-2):15–28, March 2007.

[81] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 2020.09.25.312868, September 2020.

[82] Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics*, 152(14):144106, April 2020.

[83] Lucy Ham, Marcel Jackson, and Michael P.H. Stumpf. Pathway dynamics can delineate the sources of transcriptional noise in gene expression. Preprint, bioRxiv: 2020.09.30.319814, September 2020.

[84] Qingchao Jiang, Xiaoming Fu, Shifu Yan, Runlai Li, Wenli Du, Zhixing Cao, Feng Qian, and Ramon Grima. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nature Communications*, 12(1):2618, December 2021.

[85] The MathWorks. MATLAB R2022a Symbolic Math Toolbox, 2022.

[86] The MathWorks. MATLAB R2022a, 2022.