Monod: mechanistic analysis of single-cell RNA sequencing count data
Methods and supplementary materials

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

June 12, 2022

S1 Data availability

The datasets released by the Allen Institute for Brain Science were obtained from `http://data.nemoarchive.org/biccn/grant/u19_zeng/zeng/transcriptome/scell/10x_v3/mouse/raw/MOp/`, and filtered according to the metadata annotations [63]. Throughout this report, we refer to the sequenced samples by their internal identifiers (B08, C01, F08, H12), which correspond to tissues from four individual mice (donor IDs 457911, 427378, 457909, and 426003, respectively). The cell type-filtered, processed *loom* files are available on Zenodo at [64]. The fits to all of the datasets are available on Zenodo at [64], as a single batch in the *Monod* format. Notebooks that reproduce all of the filtering, fitting, analysis procedures, and report further information about the genes presented in Table S10, are available at `https://github.com/pachterlab/monod_examples/tree/main/manuscript_computation`.

S2 Supported models

Monod implements four classes of biological models and three classes of technical noise models.

S2.1 Biological noise models

In the current section, we outline the biological models of interest and summarize their solutions. The models all have the following generic structure:

$$\varnothing \xrightarrow{K} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{S1}$$

where K is a stochastic process governing mRNA arrivals, β is the splicing rate, and γ is the degradation rate. For simplicity, $\mathcal{N}$ denotes nascent or unspliced mRNA species, whereas $\mathcal{M}$ denotes mature or spliced mRNA species. These concepts are not strictly synonymous – for example, "nascent" species may refer to mRNA in the process of transcription – and we adopt this convention to emphasize that the models are generic, as long as a two-step process accurately describes the underlying physics. For inference, we are interested in the probability generating function (PGF):

$$G(x_N, x_M) = \sum_{x_N, x_M=0}^{\infty} x_N^n x_M^m P(n, m), \tag{S2}$$

where $P(n, m)$ is the joint probability mass function of the nascent and mature counts. We consider the PGF in terms of shifted complex variables $u_z := x_z - 1$, where we introduce the generic transcript subscript $z \in \{N, M\}$.

S2.1.1 Bursty

The three-parameter bursty model describes the following biology:

$$\varnothing \xrightarrow{k} B \times \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{S3}$$

where B is a geometrically-distributed random variable on $\mathbb{N}_0$ with mean b, and k is the burst frequency, set to unity with no loss of generality at steady state. This system has the following PGF [31]:

$$\begin{aligned} U(u_N, u_M, s) &= u_M \frac{\beta}{\beta - \gamma} e^{-\gamma s} + \left(u_N - u_M \frac{\beta}{\beta - \gamma} \right) e^{-\beta s} \\ \ln G := \psi &= \int_0^{\infty} \left[\frac{1}{1 - bU(u_N, u_M, s)} - 1 \right] ds. \end{aligned} \tag{S4}$$

This model encodes the bursty production of mRNA, and is usually derived as the limit of a transcriptional process driven by a randomly switching promoter [20]. Specifically, if the promoter is described by a telegraph process with on rate k, off rate k_{off}, and transcription rate in the on state k_{init}, then in the physiological limit of rare, high-amplitude bursts, the CME reduces to the memoryless process in Equation S3 with $b := k_{init}/k_{off}$ (Section S1.3 of [41]).

S2.1.2 Constitutive

The two-parameter constitutive model describes the following biology:

$$\varnothing \xrightarrow{k} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{S5}$$

where k is set to unity with no loss of generality at steady state. This system has the following PGF [65]:

$$\psi = u_N \frac{1}{\beta} + u_M \frac{1}{\gamma}, \tag{S6}$$

i.e., an uncorrelated Poisson distribution. This model encodes the constitutive or unregulated production of mRNA.

S2.1.3 Extrinsic

The three-parameter extrinsic noise model describes the following biology:

$$\varnothing \xrightarrow{K} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{S7}$$

where K is a gamma distribution with shape α and scale 1. This system has the following PGF:

$$\psi = -\alpha \ln\left(1 - u_N \frac{1}{\beta} - u_M \frac{1}{\gamma}\right), \tag{S8}$$

which follows immediately from the properties of generating functions [41]. This model encodes driving by a slow, Gamma-distributed transcriptional process [9, 66–68].

S2.1.4 CIR-like

The three-parameter CIR-like noise model describes the following biology:

$$\varnothing \xrightarrow{K_t} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing, \tag{S9}$$

where K_t is, informally, the time derivative of a subordinating inverse Gaussian process [9]. This system emerges in the fast limit of transcription driven by a Cox–Ingersoll–Ross process, and has the following PGF:

$$\psi = \frac{1}{2} \int_0^\infty \left[1 - \sqrt{1 - 4bU(u_N, u_M, s)}\right] ds, \tag{S10}$$

where U is identical to the expression in Equation S4, whereas b is a burst size-like "gain" that appears in the definition of K_t. This model encodes driving by a fast chemical Langevin equation, which in turn represents a high-concentration limit of the constitutive birth-death process [9].

S2.2 Technical noise models

To account for the imperfect sequencing process, we define a set of technical noise models. All of the models assume that the reverse transcription, amplification, and identification of the *in vitro* mRNA pool is fundamentally an independent and identically distributed process across all cells and molecules of a single gene, although the identically distributed assumption is relaxed across genes. For inference, we define the PGF H, which is the function composition of G with the PGFs of the sequencing processes.

S2.2.1 Null process

In the simplest case, the null process, the sequencing is perfect and every mRNA is captured and reported as a unique molecule. The PGF of this process is x_z, yielding $H = G$.

The lower moments of this process take the form shown in Table S1. These moments induce the method of moments estimates listed in Table S2, where $\overline{Z}$ is the sample mean and S_z^2 is the sample variance. All of the moments and estimates were computed using the MATLAB R2022a Symbolic Math Toolbox [69, 70].

Variable	Bursty	Constitutive	Extrinsic	CIR-like
μ_N	$\frac{b}{\beta}$	$\frac{1}{\beta}$	$\frac{\alpha}{\beta}$	$\frac{b}{\beta}$
μ_M	$\frac{b}{\gamma}$	$\frac{1}{\gamma}$	$\frac{\alpha}{\gamma}$	$\frac{b}{\gamma}$
$(\sigma_N^2 - \mu_N)/\mu_N$	b	0	$\frac{1}{\beta}$	b
$(\sigma_M^2 - \mu_M)/\mu_M$	$\frac{b\beta}{\beta+\gamma}$	0	$\frac{1}{\gamma}$	$\frac{b\beta}{\beta+\gamma}$

Table S1: Lower moments of the null process.

Variable	Bursty	Constitutive	Extrinsic	CIR-like
$\hat{b}$	$S_N^2/\overline{N} - 1$			$S_N^2/\overline{N} - 1$
$\hat{\beta}$	$\hat{b}/\overline{N}$	$1/\overline{N}$	$\hat{\alpha}/\overline{N}$	$\hat{b}/\overline{N}$
$\hat{\gamma}$	$\hat{b}/\overline{M}$	$1/\overline{M}$	$\hat{\alpha}/\overline{M}$	$\hat{b}/\overline{M}$
$\hat{\alpha}$			$\frac{\overline{N}^2}{S_N^2 - \overline{N}}$	

Table S2: Method of moments biological parameter estimates for the null process.

S2.2.2 Bernoulli process

In the Bernoulli process, each molecule has probability p_z of being captured and reported; multiple priming is impossible. The PGF of this process is $G_{z,t} = p_z x_z + (1 - p_z)$. This yields:

$$\begin{aligned} G_{z,t} - 1 &= u_z p_z \\ U(G_{N,t}, G_{M,t}) &= p_M u_M \frac{\beta}{\beta - \gamma} e^{-\gamma s} + \left(p_N u_N - p_M u_M \frac{\beta}{\beta - \gamma} \right) e^{-\beta s}. \end{aligned} \tag{S11}$$

The moments (Table S3) are identical to those reported in Table S2, up to scaling. The method of moments parameter estimates, conditional on a particular $\{p_N, p_M\}$, are given in table S4. Evidently, bespoke routines are unnecessary:

- Given $\hat{b}$ obtained with the null model's equation, we find the Bernoulli estimate is $\hat{b}/p_N$.
- We can directly use $\hat{\alpha}$ obtained with the null model's equation.
- Given $\hat{\beta}$ obtained with the null model's equation and the appropriate $\hat{b}$ or $\hat{\alpha}$, we find the Bernoulli estimate is $p_N \hat{\beta}$.
- Given $\hat{\gamma}$ obtained with the null model's equation and the appropriate $\hat{b}$ or $\hat{\alpha}$, we find the Bernoulli estimate is $p_M \hat{\gamma}$.

S2.2.3 Poisson process

In the Poisson process, each molecule has a rate λ_z of being captured and reported; multiple priming is possible, but rare if λ_z is low. This PGF of this process is $G_{z,t} = e^{\lambda_z (x_z - 1)}$. This yields:

$$G_{z,t} - 1 = e^{\lambda_z u_z} - 1, \tag{S12}$$

Variable	Bursty	Constitutive	Extrinsic	CIR-like
μ_N	$\frac{bp_N}{\beta}$	$\frac{p_N}{\beta}$	$\frac{\alpha p_N}{\beta}$	$\frac{bp_N}{\beta}$
μ_M	$\frac{bp_M}{\gamma}$	$\frac{p_M}{\gamma}$	$\frac{\alpha p_M}{\gamma}$	$\frac{bp_M}{\gamma}$
$(\sigma_N^2 - \mu_N)/\mu_N$	bp_N	0	$\frac{p_N}{\beta}$	bp_N
$(\sigma_M^2 - \mu_M)/\mu_M$	$\frac{b\beta p_M}{\beta+\gamma}$	0	$\frac{p_M}{\gamma}$	$\frac{b\beta p_M}{\beta+\gamma}$

Table S3: Lower moments of the Bernoulli process.

Variable	Bursty	Constitutive	Extrinsic	CIR-like
$\hat{b}$	$(S_N^2/\overline{N} - 1)/p_N$			$(S_N^2/\overline{N} - 1)/p_N$
$\hat{\beta}$	$\hat{b}p_N/\overline{N}$	$p_N/\overline{N}$	$\hat{\alpha}p_N/\overline{N}$	$\hat{b}p_N/\overline{N}$
$\hat{\gamma}$	$\hat{b}p_M/\overline{M}$	$p_M/\overline{M}$	$\hat{\alpha}p_M/\overline{M}$	$\hat{b}p_M/\overline{M}$
$\hat{\alpha}$			$\frac{\overline{N}^2}{S_N^2 - \overline{N}}$	

Table S4: Method of moments biological parameter estimates for the Bernoulli process.

which cannot be simplified further. The moments are analogous, and given in Table S5. The corresponding method of moments estimates are given in Table S6. As before, we can recycle the null model estimates:

- Given $\hat{b}$ obtained with the null model's equation, we find the Bernoulli estimate is $\hat{b}/\lambda_N - 1$.
- We are required to compute $\hat{\alpha}$ using the adjusted estimate.
- Given $\hat{\beta}$ obtained with the null model's equation and the appropriate $\hat{b}$ or $\hat{\alpha}$, we find the Bernoulli estimate is $\lambda_N\hat{\beta}$.
- Given $\hat{\gamma}$ obtained with the null model's equation and the appropriate $\hat{b}$ or $\hat{\alpha}$, we find the Bernoulli estimate is $\lambda_M\hat{\gamma}$.

Variable	Bursty	Constitutive	Extrinsic	CIR-like
μ_N	$\frac{b\lambda_N}{\beta}$	$\frac{\lambda_N}{\beta}$	$\frac{\alpha\lambda_N}{\beta}$	$\frac{b\lambda_N}{\beta}$
μ_M	$\frac{b\lambda_M}{\gamma}$	$\frac{\lambda_M}{\gamma}$	$\frac{\alpha p_M}{\gamma}$	$\frac{b\lambda_M}{\gamma}$
$(\sigma_N^2 - \mu_N)/\mu_N$	$\lambda_N(1+b)$	λ_N	$\lambda_N\left(1+\frac{1}{\beta}\right)$	$\lambda_N(1+b)$
$(\sigma_M^2 - \mu_M)/\mu_M$	$\lambda_M\left(1+\frac{b\beta}{\beta+\gamma}\right)$	λ_M	$\lambda_M\left(1+\frac{1}{\gamma}\right)$	$\lambda_M\left(1+\frac{b\beta}{\beta+\gamma}\right)$

Table S5: Lower moments of the Poisson process.

S3 Statistical procedures

S3.1 Noise decomposition

The models afford analytical noise decompositions. For example, we can define the squared coefficient of variation $\mathrm{CV}_M^2 := \eta_M^2 = \sigma_M^2/\mu_M^2$ for the mature species under the bursty model. If we

Variable	Bursty	Constitutive	Extrinsic	CIR-like
$\hat{b}$	$(S_N^2/\overline{N}-1)/\lambda_N - 1$			$(S_N^2/\overline{N}-1)/\lambda_N - 1$
$\hat{\beta}$	$\hat{b}\lambda_N/\overline{N}$	$\lambda_N/\overline{N}$	$\hat{\alpha}\lambda_N/\overline{N}$	$\hat{b}\lambda_N/\overline{N}$
$\hat{\gamma}$	$\hat{b}\lambda_M/\overline{M}$	$\lambda_M/\overline{M}$	$\hat{\alpha}\lambda_M/\overline{M}$	$\hat{b}\lambda_M/\overline{M}$
$\hat{\alpha}$			$\frac{\overline{N}^2}{S_N^2-\overline{N}(1+\lambda_N)}$	

Table S6: Method of moments biological parameter estimates for the Poisson process.

assume there is no technical noise, we find:

$$\begin{aligned}\eta_M^2 &= \frac{1}{\mu_M}\left[1+\frac{b\beta}{\beta+\gamma}\right] \\ &= \varsigma_{M,int}^2 + \varsigma_{M,ext}^2, \\ \varsigma_{M,int}^2 &= \frac{1}{\mu_M} = \frac{\gamma}{b}, \\ \varsigma_{M,ext}^2 &= \frac{1}{\mu_M}\frac{b\beta}{\beta+\gamma} = \frac{\beta\gamma}{\beta+\gamma},\end{aligned} \tag{S13}$$

where $\varsigma_{M,int}^2$ is the noise contribution from the discrete nature of the process and $\varsigma_{M,ext}^2$ is the noise contribution due to the bursty transcriptional dynamics. We denote this noise component as *extrinsic*, as it emerges from variation in the transcription rate [9]. Interestingly, b does not explicitly occur in $\varsigma_{M,ext}^2$: the value of the burst size serves as a scaling factor and cancels out in the definition of CV^2. The two species' and four models' noise decompositions are given in Table S7.

Under the Bernoulli technical noise model, we find that the following behavior holds:

$$\eta_z^2 = p_z^{-1}\varsigma_{z,int}^2 + \varsigma_{z,ext}^2, \tag{S14}$$

i.e., the variation induced by imperfect sampling is multiplicative with respect to the intrinsic noise term. The noise components ς^2 listed in Equation S14 correspond to the values given in Table S7, and do not depend on p_z. We can find the fraction of variation attributable to each noise component:

$$\begin{aligned}f_{z,int} &= \frac{\varsigma_{z,int}^2}{p_z^{-1}\varsigma_{z,int}^2 + \varsigma_{z,ext}^2}, \\ f_{z,ext} &= \frac{\varsigma_{z,ext}^2}{p_z^{-1}\varsigma_{z,int}^2 + \varsigma_{z,ext}^2}, \\ f_{z,tech} &= 1 - f_{z,int} - f_{z,ext}.\end{aligned} \tag{S15}$$

Under the Poisson technical noise model, we find that the following behavior holds:

$$\eta_z^2 = \varsigma_{z,int}^2 + \varsigma_{z,ext}^2 + \varsigma_{z,tech}^2, \tag{S16}$$

Variable	Bursty	Constitutive	Extrinsic	CIR-like
$\varsigma^2_{N,int}$	$\frac{1}{\mu_N}$	$\frac{1}{\mu_N}$	$\frac{1}{\mu_N}$	$\frac{1}{\mu_N}$
$\varsigma^2_{M,int}$	$\frac{1}{\mu_M}$	$\frac{1}{\mu_M}$	$\frac{1}{\mu_M}$	$\frac{1}{\mu_M}$
$\varsigma^2_{N,ext}$	β	0	$\frac{1}{\alpha}$	β
$\varsigma^2_{M,ext}$	$\frac{\beta\gamma}{\beta+\gamma}$	0	$\frac{1}{\alpha}$	$\frac{\beta\gamma}{\beta+\gamma}$

Table S7: Noise decomposition of the null process. $\varsigma^2_{z,ext}$ denotes super-Poisson (overdispersed) noise attributable to bursting or extrinsic variation in transcription rates.

i.e., the variation induced by Poisson-like sampling is additive. This additive term is $\varsigma^2_{z,tech} = \frac{1}{\mu_z}$, which depends on λ_z. We can find the fraction of noise attributable to each noise component:

$$\begin{aligned} f_{z,int} &= \frac{\varsigma^2_{z,int}}{\eta^2_z}, \\ f_{z,ext} &= \frac{\varsigma^2_{z,ext}}{\eta^2_z}, \\ f_{z,tech} &= \frac{\varsigma^2_{z,tech}}{\eta^2_z}. \end{aligned} \tag{S17}$$

This approach is fundamentally parametric, and presupposes that the biological and technical noise models are correct and have been accurately fit. However, much of scRNA-seq data processing is non-parametric; for example, variance stabilization and normalization are conventionally used to remove extraneous noise. We can write down an analogous phenomenological noise decomposition:

$$\begin{aligned} \eta^2_z &= \frac{S^2_z}{\overline{Z}^2}, \\ \eta^2_{\phi(z)} &= \frac{S^2_{\phi(z)}}{\overline{\phi(Z)}^2}, \\ f_{z,ret} &= \frac{\eta^2_{\phi(z)}}{\eta^2_z}, \\ f_{z,disc} &= 1 - f_{z,ret}, \end{aligned} \tag{S18}$$

where ϕ is a fairly generic function (e.g., a combination of normalization and log-transformation [38]), $f_{z,ret}$ is the fraction of variance retained after the transformation, and $f_{z,disc}$ is the fraction of variance discarded as a result of the transformation. When performing this analysis, we make the tacit assumption that ϕ decreases the data variance.

S3.1.1 Cell subtype comparisons

Given a set of cell subtypes, indexed by κ, we can construct an estimate for the amount of inter-subtype variation. For a given gene and species Z, the sample mean of the expression $\overline{Z}$ is a

weighted sum of subtype averages $\overline{Z}_\kappa$:

$$\begin{aligned} c &= \sum_\kappa c_\kappa, \\ \overline{Z} &= \sum_\kappa \frac{c_\kappa}{c} \overline{Z}_\kappa, \end{aligned} \tag{S19}$$

where c_κ is the number of cells assigned to subtype κ and c is the total number of cells. By variance decomposition [71], we obtain a similar expression for the sample variance of the expression S^2 in terms of intra-subtype variances S^2_κ and means $\overline{Z}_\kappa$:

$$S^2 = \sum_\kappa \frac{c_\kappa}{c} \overline{S}^2_\kappa + \sum_\kappa \frac{c_\kappa}{c} \left(\overline{Z}_\kappa - \overline{Z}\right)^2. \tag{S20}$$

The expression above is an estimator of the true variance:

$$\sigma^2 = \mathbb{E}[\sigma^2_\kappa] + \mathrm{Var}(\mu_\kappa), \tag{S21}$$

where the expectations are taken with respect to the true categorical distribution over cell subtypes (in turn approximated by c_κ/c). We omit the species subscripts N, M for simplicity of notation. The subtype-specific statistics are, in turn, noise-corrupted versions of the unknown biological statistics $\tilde{\sigma}^2_\kappa$ and $\tilde{\mu}_\kappa$. Formally:

$$\begin{aligned} \tilde{\sigma}^2 &= \mathbb{E}[\tilde{\sigma}^2_\kappa] + \mathrm{Var}(\tilde{\mu}_\kappa), \\ \sigma^2 &= \mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa), \end{aligned} \tag{S22}$$

where Ξ_κ governs the variance change and ξ_κ governs the mean change induced by technical noise. We find that the fraction of biological noise takes the following form:

$$\begin{aligned} \eta^2 &= \frac{\sigma^2}{\mu^2} = \frac{\mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)}{\mathbb{E}[\xi \tilde{\mu}_\kappa]^2} = \frac{\mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)}{\xi^2 \tilde{\mu}^2}, \\ \tilde{\eta}^2 &= \frac{\tilde{\sigma}^2}{\tilde{\mu}^2}, \\ \frac{\tilde{\eta}^2}{\eta^2} &= \frac{\xi^2 \tilde{\sigma}^2}{\mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)} = \frac{\xi^2 \mathbb{E}[\tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)}{\mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)} \\ &\geq \frac{\mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)}{\mathbb{E}[\Xi_\kappa \tilde{\sigma}^2_\kappa] + \mathrm{Var}(\xi_\kappa \tilde{\mu}_\kappa)}. \end{aligned} \tag{S23}$$

This lower bound of the fraction of biological noise affords a consistent estimator:

$$f_{lb} = \frac{1}{S^2} \sum_\kappa \frac{c_\kappa}{c} \left(\overline{Z}_\kappa - \overline{Z}\right)^2. \tag{S24}$$

Analogous results for the mechanistic noise models are straightforward to derive using the expressions in Tables S1 and S5. For example, if we are interested in the results for mature mRNA produced by the bursty model under Poisson noise, we find:

$$
\begin{aligned}
\tilde{\mu}_\kappa &= \frac{b}{\gamma}, \\
\tilde{\sigma}_\kappa^2 &= \tilde{\mu}_\kappa \left(1 + \frac{b\beta}{\beta + \gamma}\right), \\
\mu_\kappa &= \tilde{\mu}_\kappa \lambda_\kappa, \\
\sigma_\kappa^2 &= \mu_\kappa \left[1 + \lambda_\kappa \left(1 + \frac{b\beta}{\beta + \gamma}\right)\right], \\
\tilde{\eta}^2 &= \frac{\mathbb{E}[\tilde{\sigma}_\kappa^2] + \mathrm{Var}(\tilde{\mu}_\kappa)}{\tilde{\mu}^2}, \\
\eta^2 &= \frac{\mathbb{E}[\sigma_\kappa^2] + \mathrm{Var}(\mu_\kappa)}{\mu^2},
\end{aligned}
\tag{S25}
$$

which can be computed by plugging in parameter MLEs and taking the expectations with respect to the usual categorical measure.

S3.2 Differential parameter value identification

We would like to identify potential differences in parameter values between different cell types of experimental conditions. The inferred parameter MLEs demonstrate a linear relationship with unity slope for the bulk of the genes and conspicuous deviations for a small subset. However, the linear relationship is not necessarily identity – around the optimum, uncertainties in estimating the sampling parameters $\{\lambda_N, \lambda_M\}$ translate to a bias in the MLEs of biological parameters $\{b, \beta, \gamma\}$. We propose the following model for the MLEs $\hat{\theta}_j$ obtained for datasets indexed by 1 and 2:

$$
\begin{aligned}
\hat{\theta}_1 &= \theta_1 + o_1 + \varepsilon_{1,j}, \\
\hat{\theta}_2 &= \theta_2 + o_2 + \varepsilon_{2,j},
\end{aligned}
\tag{S26}
$$

where o_1 and o_2 are genome-wide offsets or biases, and $\varepsilon_{1,j}$ and $\varepsilon_{2,j}$ are zero-centered multivariate random normal variables. Assuming now that the true θ_j match:

$$
\hat{\theta}_2 + \varepsilon_{2,j} = \hat{\theta}_1 + o_t + \varepsilon_{1,j}, \tag{S27}
$$

where o_t is the total offset and $\varepsilon_{t,j}$ is the total error. Therefore, we obtain the following residuals:

$$
\begin{aligned}
\hat{\theta}_1 - \hat{\theta}_2 &= \theta_1 + o_1 + \varepsilon_{1,j} - \theta_2 - o_2 - \varepsilon_{2,j}, \\
\hat{\theta}_1 - \hat{\theta}_2 &= -o_t + \varepsilon_{1,j} - \varepsilon_{2,j} = -o_t + \varepsilon_{t,j}, \\
\hat{\theta}_1 - \hat{\theta}_2 + o_t &= \varepsilon_{t,j}.
\end{aligned}
\tag{S28}
$$

To identify systematic differences in parameters – i.e., cases where θ_1 is unlikely to be identical to θ_2 – we need to quantify the aleatory, or unsystematic, baseline variation. To do this, we perform the following fixed-point iteration:

1. Fit Equation S27 over all genes using orthogonal distance regression (ODR) with Fisher information matrix-derived uncertainty [29].

2. Use the obtained estimate to compute the de-biased residual $\varepsilon_{t,j} = \hat{\theta}_1 - \hat{\theta}_2 + o_t$.

3. Fit the resulting values of $\varepsilon_{t,j}$ to the nearest normal approximator, producing a (near-zero) mean μ and aggregated standard deviation σ.

4. Discard genes whose Z-statistic $z_j = (\varepsilon_{t,j} - \mu)/\sigma$ yields $2[1 - \Phi(|z_j|)] < p$, where Φ is the standard normal cumulative distribution function.

5. Repeat from the start, using the remaining genes for steps 1-3. Continue until convergence or until a preset number of iterations.

This approach appears to effectively fit the central Gaussian-like peak and identify the residual tails, putatively corresponding to differentially transcribed genes. As a comparison based on the lower moments, we implement a t-test and an analogous fixed-point iteration procedure with log-means instead of MLEs and unity-slope linear least squares instead of ODR.

A formal analysis of the procedure using simulations is a useful target for future work, but falls outside the scope of the current report. However, informally, the procedure is susceptible to false negatives due to small effect sizes (insufficiently high $|z_j|$), and may be susceptible to false positives due to the lack of multiple comparisons correction. Further, the approach may lose interpretability as the cell types are *a priori* defined based on gene expression. We anticipate a greater interest in this class of models will lead to the development and implementation of more rigorous statistical analyses.

S3.2.1 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 in the current report 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 $\log_{10}\frac{\gamma}{k}$ reported in Figure 4 are 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:

$$\begin{aligned}
\Delta \log_{10} \frac{\beta}{k} &= \log_{10} \frac{\beta_2}{k_{i,2}} - \log_{10} \frac{\beta_1}{k_{i,1}}, \\
\Delta \log_{10} \frac{\gamma}{k} &= \log_{10} \frac{\gamma_2}{k_{i,2}} - \log_{10} \frac{\gamma_1}{k_{i,1}}, \\
\Delta \log_{10} k &\approx -\Delta \log_{10} \frac{\beta}{k} = \log_{10} k_{i,2} - \log_{10} k_{i,1} \\
&\approx -\Delta \log_{10} \frac{\gamma}{k} = \log_{10} k_{i,2} - \log_{10} k_{i,1}.
\end{aligned} \tag{S29}$$

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.

S4 mRNA quantification and filtering

Raw FASTQ files were obtained for all datasets. We generated count matrices with *kallisto* |*bustools* (kb), using the `--lamanno` setting to separately quantify unspliced and spliced molecules. The procedure is described in greater detail in [7]. We obtained and processed the datasets listed in Table S8 and Section S1.

The data were filtered with respect to cells twice: first, with the default kb barcode filter; then, by removing cells with fewer than 10^4 spliced mRNA counts, computed over all genes (Figure S1). We split the filtered datasets according to pre-existing cell type annotations [63]. For the analysis of normalization procedures, we extracted barcodes corresponding to all annotated subtypes (Lamp5, Sncg, Vip, Sst, Pvalb for GABAergic; L2/3 IT, L5 IT, L6 IT, L6 IT Car3, L5 ET, L5/6 NP, L6 CT, L6b for glutamatergic). Sncg and L6 IT Car3 contained fewer than thirty barcodes and were removed from further investigation. For the differential expression analysis, we extracted barcodes corresponding to the GABAergic and glutamatergic types, omitting those assigned to the Sncg and L6 IT Car3 subtypes.

S5 Analysis parameters and summaries

We input the *loom* files generated by the quantification pipeline to the *Monod* pre-processing script and removed two sets of genes: those with very low expression ($\overline{N} \leq 0.01$, $\overline{M} \leq 0.01$, $\max N \leq 3$, $\max M \leq 3$), and those with excessively high expression, which are too computationally intensive to fit ($\max N \geq 400$, $\max M \geq 400$). As in [29], $\max Z$ denotes the maximum copy number observed for species Z. This procedure produced a set of 2,130 genes that met the thresholds in all of the datasets.

S5.1 Inference

For all models and datasets, we set up a 20×21 grid over the $\{\log_{10} C_N, \log_{10} \lambda_M\}$ domain listed in Table S9. These bounds were chosen according to the results previously obtained for aggregated datasets B01, C01, A08, and B08 from the same study [63] (as reported in Table S4 of [29]).

At each grid point, we iterated over the 2,130 genes, using gradient descent to identify the conditional maximum likelihood estimate of $\{\log_{10} b, \log_{10} \beta, \log_{10} \gamma\}$, where the rates are defined

Table S8: Summary of datasets used in the analysis. All datasets originate from *M. musculus* [63]. B08, C01, F08, and H12 are biological replicates.

Dataset	Cell type	Cell subtype	Cells	Genes kept	Runtime (h)
B08	—	—	7584	—	—
B08	GABAergic	—	1551	7489	1.6
B08	Glutamatergic	—	4394	7929	1.8
B08	Glutamatergic	L2/3 IT	621	5864	0.79
B08	Glutamatergic	L5 IT	1702	6759	1.35
B08	Glutamatergic	L6 IT	471	5773	0.70
B08	Glutamatergic	L5 ET	57	2978	0.31
B08	Glutamatergic	L5/6 NP	191	4268	0.46
B08	Glutamatergic	L6 CT	1359	6599	0.89
B08	Glutamatergic	L6b	34	3000	0.42
C01	—	—	8655	—	—
C01	GABAergic	—	1704	6284	0.79
C01	Glutamatergic	—	4674	7049	1.1
F08	—	—	8150	—	—
F08	GABAergic	—	1621	7191	1.5
F08	Glutamatergic	—	4820	7858	1.8
H12	—	—	6837	—	—
H12	GABAergic	—	1452	6414	0.92
H12	Glutamatergic	—	3335	6699	0.98

in units of k. We used the conditional method of moments estimate ("Bursty" column of Table S6) as the starting point and performed 10 steps of gradient descent. The procedure was parallelized over sixty processors (Intel Xeon Gold 6152, 2.10GHz). Runtimes varied between 0.3 hours for the smallest datasets and 1.8 for the largest.

To identify the optimum in sampling parameters, we identified the grid point with the lowest total Kullback-Leibler divergence, computed over all genes. To ensure we obtained the true optimum under the bursty model, we performed ten rounds of fixed-point iteration, rejecting a subset of genes by the chi-squared test with $p = 0.05$ and a Bonferroni correction, and recalculating the optimum based on the remaining data. This procedure did not change the optimum for any of the datasets. Further, we investigated the stability of the optimum under gene subsampling, and found it to be stable and consistent. We computed the standard errors of the gene-specific biological parameter vectors, conditional on the sampling parameter value at the optimal grid point, by inverting the Hessian of the KL divergence at the MLE.

S5.2 Noise decomposition

To perform the nonparametric noise decomposition, we used the log1pPF transformation described in a recent report [38] as the function $\phi(z)$ applied to the spliced count data. Specifically, given a data matrix $\mathcal{Z}_{ij}$ for species Z, with i indexing over cells $1, ..., c$ and j indexing over genes $1, ..., g$,

Table S9: Search parameter bounds for grid scans and gradient descent.

Parameter	Lower bound ($\log_{10}$)	Upper bound ($\log_{10}$)
C_N	-7.5	-5.5
λ_M	-2	0
b	-1	4.2
β	-1.8	2.5
γ	-1.8	3.5

the function $\phi(z)$ corresponding to log1pPF performs the following transformation:

$$\begin{aligned} C_z &= \frac{1}{c}\sum_i \left[\sum_j \mathcal{Z}_{ij}\right] \\ \mathcal{Z}'_{ij} &= \frac{\mathcal{Z}_{ij}}{\sum_j \mathcal{Z}_{ij}} \times C_z \\ \mathcal{Z}''_{ij} &= \log\left(\mathcal{Z}'_{ij} + 1\right) \\ \mathcal{Z}'' &:= \phi(\mathcal{Z}), \end{aligned} \tag{S30}$$

i.e., we compute the mean "cell size" C_z, scale each cell's total counts to C_z, then apply a log-transformation. Upon obtaining an array for each gene, we plug the result into Equation S18.

To perform the mechanistic noise decomposition, we evaluated Equation S17 at the MLE. As shown in Figure S2, the cell type sampling parameters were consistent, but the subtypes with few cells deviated from the cell type estimate. To control for this uncertainty, we used the gene-specific parameter estimates fit at the maximum likelihood sampling parameter estimate for the entire glutamatergic dataset B08.

S5.3 Differential parameter value identification

To identify genes with differential expression in the parameter values (DE-b, DE-β/k, and DE-γ/k), we used the procedure in Section S3.2, with a p-value threshold of 0.001 for the Z-test. To identify genes with differential expression in the mean (DE-$\mu_{\phi(\mathcal{M})}$), we applied a t-test to the transformed data obtained by the procedure listed in Section S5.2, with a p-value threshold of 0.1 and the Bonferroni correction. The results are shown in Table S10. Each column reports the number of genes identified as DE according to a particular test statistic; for example, $\neg\mu \cap b$ consists of all genes that do not exhibit DE in $\mu_{\phi(\mathcal{M})}$ but do exhibit DE in b. On the basis of Figure 4, we assigned all genes with simultaneous β/k and γ/k modulation to the category k. This category generally accounted for a large fraction of identified genes in each individual category. We found that compensated modulation, reported in the $\neg\mu \cap b$ and $\neg\mu \cap k$ categories, was fairly common, but considerably more consistent for burst sizes rather than burst frequencies.

Next, we validated these findings over the combination of the four datasets, To identify differential expression in the mean, we applied the t-test to $\log_2$ spliced means for each cell type, without normalization. To identify DE in the parameter values, we applied the t-test to the $\log_{10}$ biological parameter values for each cell type. These values were debiased by fitting a linear model

with an offset and unity slope to account for slight uncertainty in the sampling parameters. Both of the t-tests used in validation were computed with a p-value threshold of 0.1 and the Bonferroni correction.

Table S10: Counts of differentially expressed genes according to each test statistic. Per Figure 4, $k := \beta \cap \gamma$. Single-dataset results were computed using a t-test on transformed spliced counts for μ and fixed-point iteration for biophysical parameters. Validation results were computed using a t-test on mean $\log_2$ spliced counts for each dataset for μ, and inferred $\log_{10}$ values for the biophysical parameters.

Dataset	μ	b	β	γ	k	$\neg\mu \cap b$	$\neg\mu \cap k$	$k \cap b$
Allen B08	36	53	67	71	56	46	42	22
Allen C01	13	59	46	61	42	56	39	16
Allen F08	27	60	48	63	43	54	32	18
Allen H01	19	41	48	63	44	38	37	12
All four	10	16	27	28	22	14	13	2
Validation	4	22	16	16	6	22	6	1
Validation + any of four	1	15	15	15	6	14	4	1

S6 Supplementary figures

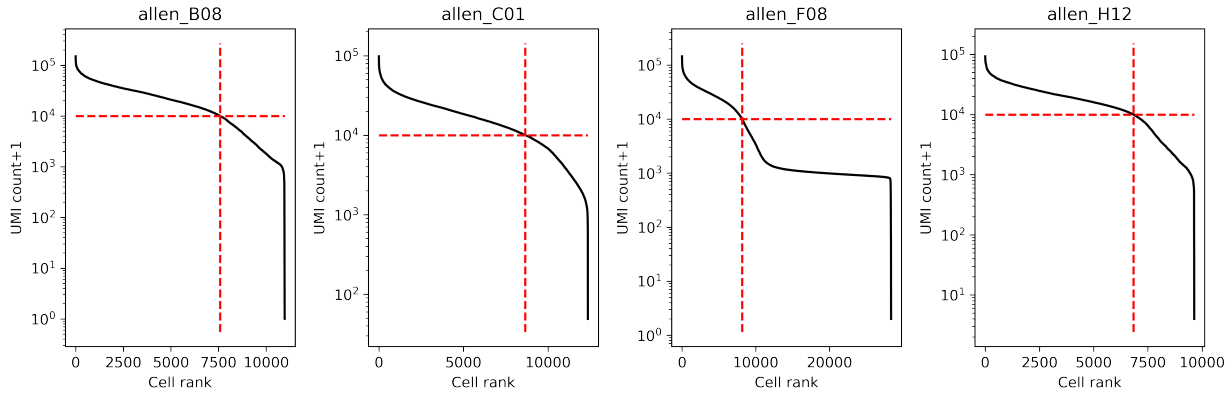

Figure S1: Knee plots for raw data (subplots: Allen 10X primary mouse cortex sequencing samples B08, C01, F08, and H12; black: cell rank/UMI relationship; red lines: location of the filtering threshold, set to 10^4 UMIs per cell).

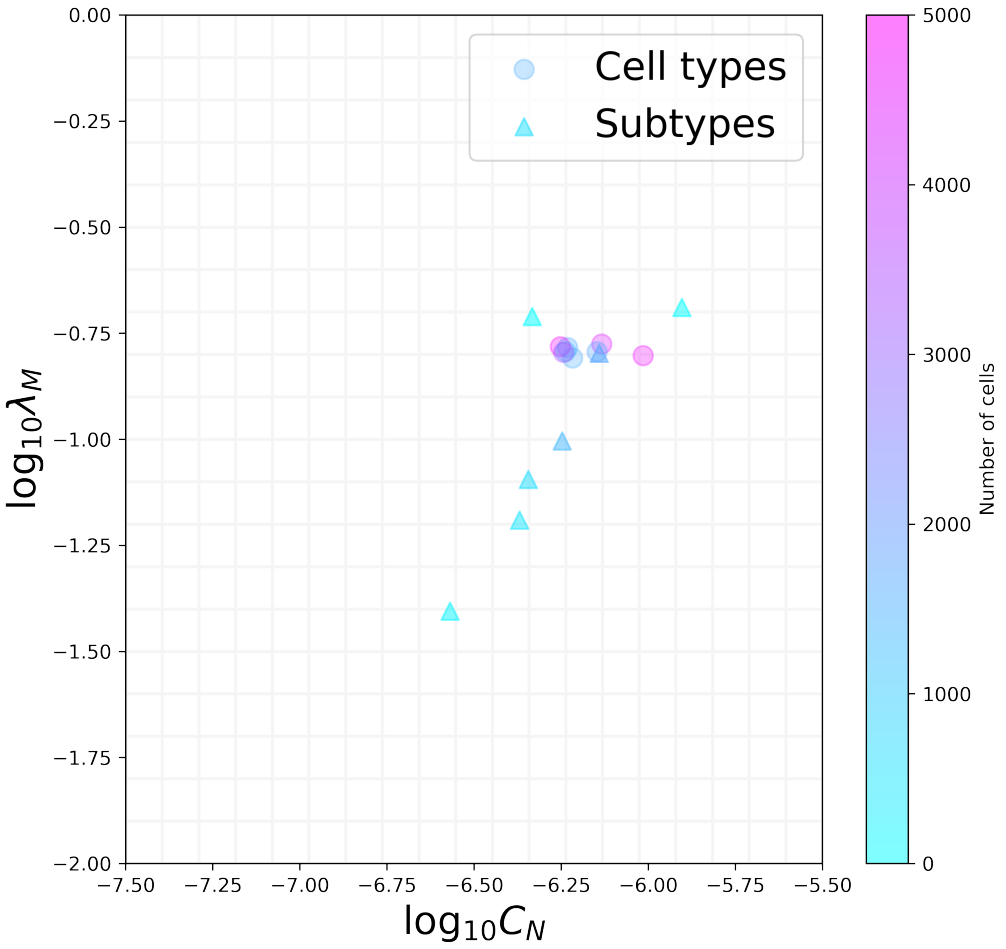

Figure S2: Inferred locations of sampling parameters on a grid are concordant between datasets, but show deviations at low cell numbers (circles: GABAergic and glutamatergic cell type datasets; triangles: glutamatergic subtypes from dataset B08; color: number of cells in the dataset; grid indicates the points evaluated during the inference process; Gaussian jitter added).

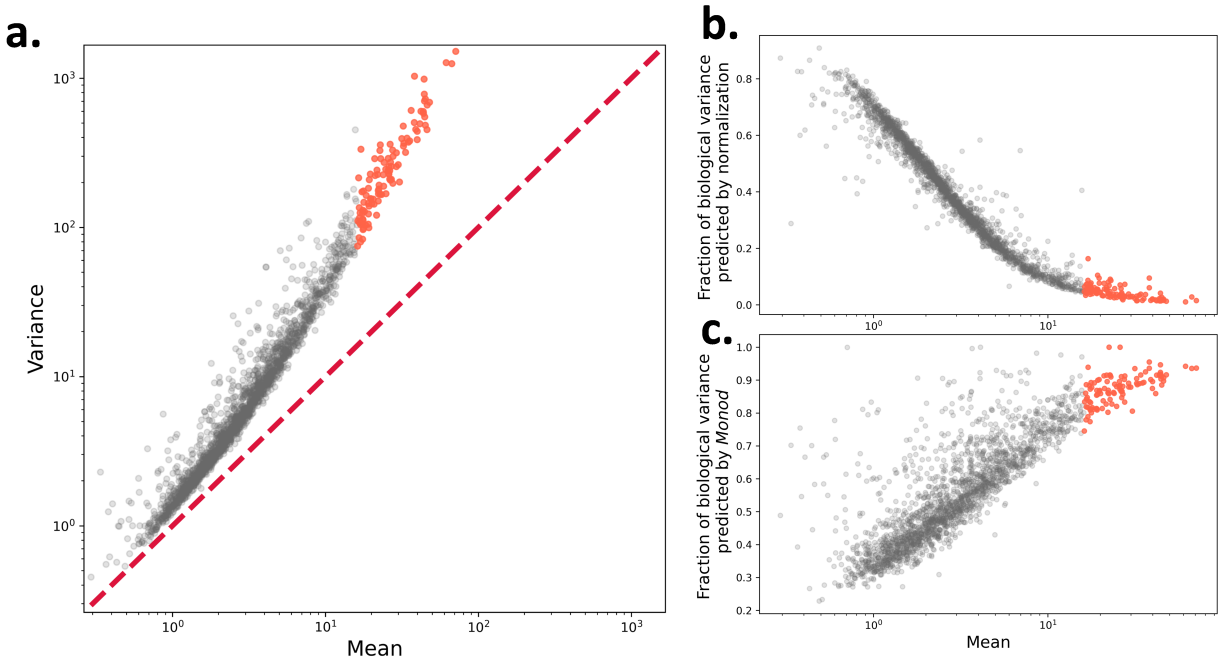

Figure S3: Parametric and parameter-free models ascribe drastically different explanations to the overdispersion observed in high-expression genes. **a.** The glutamatergic cell type demonstrates ubiquitous overdispersion, which is particularly pronounced for high-expression genes. **b.** Normalizing the data eliminates the vast majority of variation in high-expression genes, attributing only a small fraction to biology. **c.** A *Monod* mechanistic analysis suggests that the variation in high-expression genes is dominated by biological stochasticity (scarlet line: identity; gray points: genes below the 95th percentile by mean spliced expression; red points: genes above the 95th percentile by mean spliced expression).

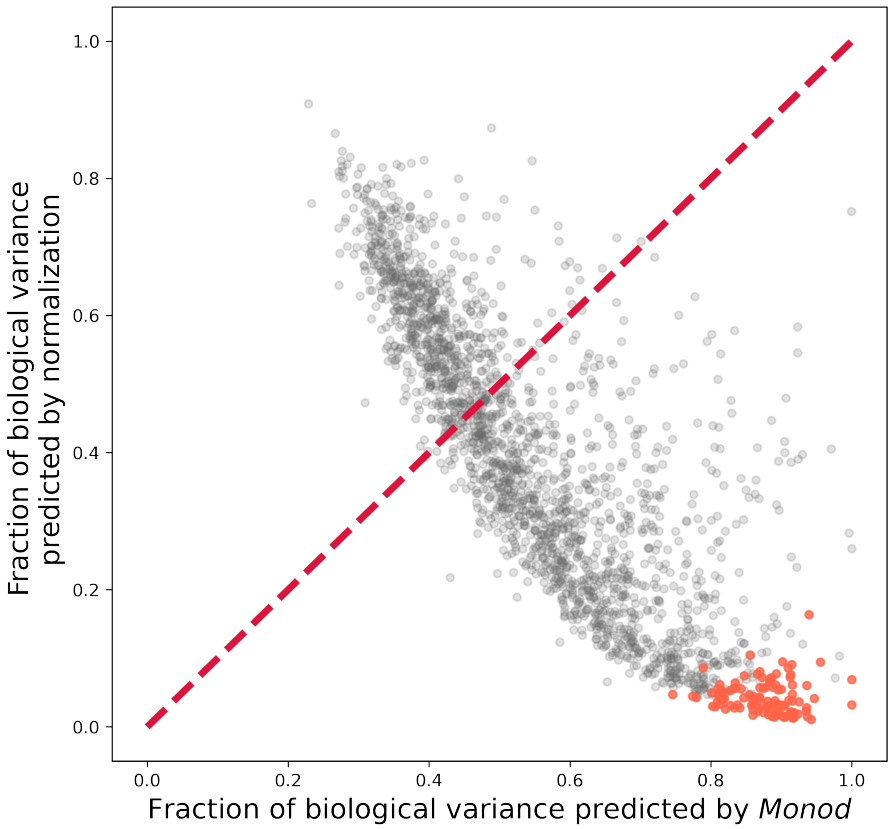

Figure S4: The amounts of relative biological variance attributed by normalization and a *Monod* mechanistic analysis are inversely correlated (scarlet line: identity; gray points: genes below the 95th percentile by mean spliced expression; red points: genes above the 95th percentile by mean spliced expression).

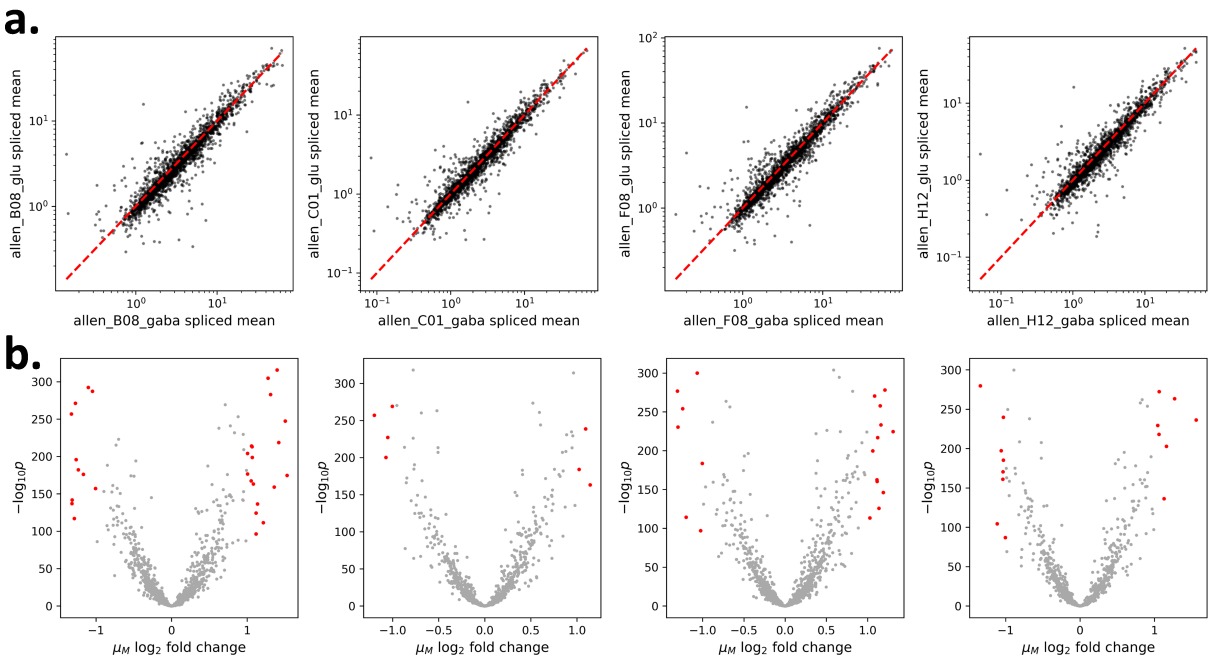

Figure S5: A demonstration of t-test differential expression analysis on the spliced mRNA count matrix. **a.** Sample mean spliced expression of genes in GABAergic and glutamatergic cell types is highly correlated, but exhibits some off-diagonal behavior (points: genes; red line: identity). **b.** By applying a t-test to the spliced count matrix, we can identify the modulation of average expression (red and gray points: genes selected and not selected by t-test with a minimum two-fold change, respectively).

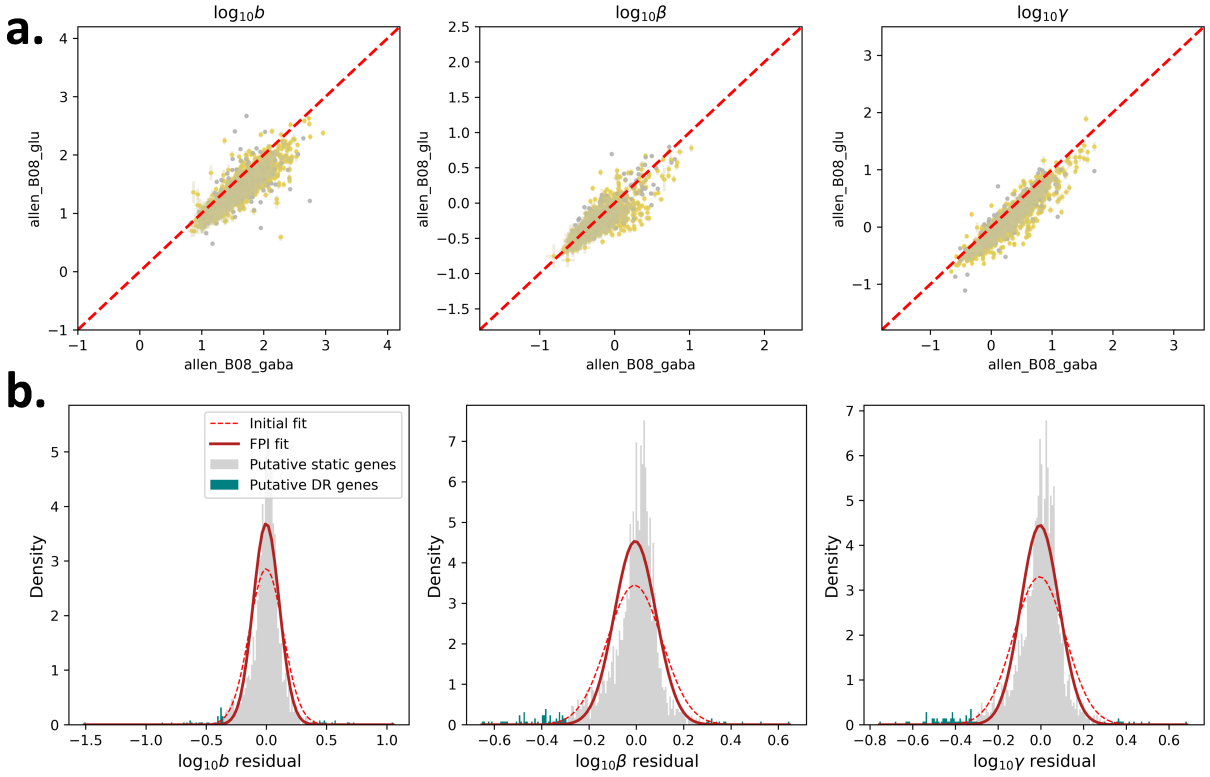

Figure S6: A demonstration of fixed-point iteration (FPI) Z-test mechanistic differential expression analysis with *Monod*. **a.** Inferred parameters for genes in GABAergic and glutamatergic cell types are highly correlated, but exhibit some off-diagonal behavior (points: genes; gold and gray: genes retained and rejected after chi-squared testing, respectively; error bars: 99% confidence intervals; red line: identity). **b.** By iteratively applying a Z-test, we can detect outliers, which may be attributable to systematic regulation of biophysical parameters (teal and gray histograms: genes selected and not selected by the test procedure, respectively; dashed red line: initial fit to the entire dataset; solid red line: fit after twenty iterations of Z-testing and outlier detection).

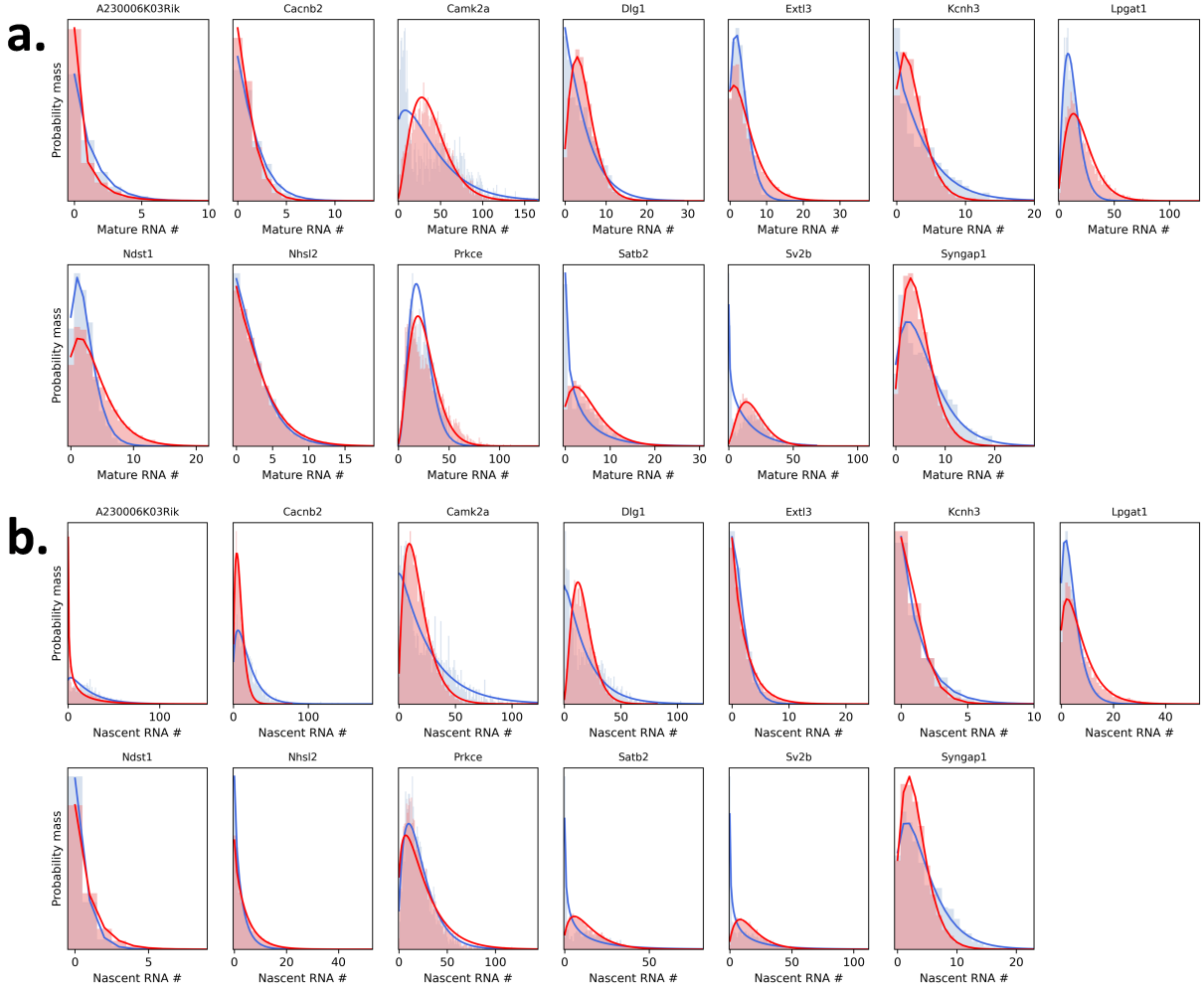

Figure S7: Several genes identified as differentially expressed according to k or b test statistics demonstrate differences in distribution shapes which do not induce a change in mean expression. **a.** Spliced mRNA distributions. **b.** Unspliced mRNA distributions (blue: GABAergic cell type; red: glutamatergic cell type; histograms: raw data; lines: fit).