Monod: mechanistic analysis of single-cell RNA sequencing count data

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

Abstract

We present the Python package *Monod* for the analysis of single-cell RNA sequencing count data through chemical master equation models. *Monod* can effectively identify biological and technical components of noise, enabling insights into potential pitfalls of standard normalization techniques. By parameterizing multidimensional distributions with biophysical variables, it 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.

1 Introduction

The vast majority of single-cell RNA-sequencing (scRNA-seq) datasets [1] are used to identify cell types, differences in gene expression between experimental conditions, and differentiation trajectories, based on analysis of a gene count matrices [2]. Gene count matrices are obtained by counting the number of molecular barcodes with reads aligning to the coding portions of genes, i.e., the spliced mRNA transcripts. More recently, interest in "RNA velocity," the inference of trajectory directions using the relative abundances of unspliced and spliced mRNA [3], has led to recognition that reads aligned to non-coding sequences may also be informative [4]. Several software packages have been developed to quantify these modalities [3,5,6], but despite their widespread use for RNA velocity [7], 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?

One approach to "integrating" spliced and unspliced matrices in a mechanistically coherent way is to interpret them as realizations of a stochastic system described by a chemical master equation (CME) [8]. This entails fitting the stationary distributions of observed species and investigating the

modulation of those models' parameters to draw conclusions about the mechanisms of transcription and differential gene expression in biophysical terms. A biophysical, rather than phenomenological, worldview offers the dual advantages of accommodating more complex experimental observations and physical phenomena as modular model components, as well as providing a framework for the design of new experiments [9]. This approach has roots in early 20th-century work on the modeling and inference of stochastic processes [10]. Augmented by the mid-century theoretical work of Jacob and Monod [11,12], as well as the late-century advances in the quantification of single molecules [13–15], the stochastic modeling of microscopic biological processes has become a standard tool in fluorescence transcriptomics [16–20]. Unfortunately, the scRNA-seq field has not adopted this approach to any appreciable extent [7]. Several articles have applied the popular Poisson-Beta and Poisson-Gamma (negative binomial) models to scRNA-seq [21–25], explicitly contextualizing them as the *result* of a particular hypothesis about the underlying physics. However, a typical single-cell analysis takes the negative binomial distribution as a convenient statistical convention or a model for *technical* variation [26–28], without discussing any formal basis in stochastic transcription. We contend that such formalization is desirable for analysis and mandatory for the development of large-dataset quantitative biology in the long term.

To facilitate computational application and interpretation of CME models, we present *Monod*, an extended and packaged standalone version of the algorithms underlying [29]. The *Monod* workflow, which generalizes the previous study to multiple datasets and models, is outlined in Figure 1. First, a set of raw data files (in the *loom* [3], *adata* [30], or *mtx* formats), obtained from alignment or pseudoalignment software [3, 5, 6], are loaded in. These files contain spliced and unspliced mRNA counts for a set of cells and genes. The preprocessing step filters for genes with moderate spliced and unspliced abundance (violet points) and outputs a random subset (red points) as a list of genes stored on disk. The data extraction step loads the raw data files and stores the filtered data matrices and their lower moments to disk.

To infer parameters, it is necessary to define a model object, which characterizes the biological and technical components of variation. To account for inter-gene coupling through sequencing, the inference procedure iterates over a grid of technical noise parameters and computes a conditional maximum likelihood estimate (MLE) for each gene's biological noise parameters. This approach provides analytical expressions for the amount of latent variation induced by biology and the experiment. By applying the inference and analysis workflow to several datasets, or separate pre-selected cell types, it is possible to compare the distribution shapes and identify differential effects on mRNA distributions. We use this procedure to investigate the effect of standard normalization procedures and probe patterns of mechanistically grounded differential expression among neural cell types.

We implement these methods in the *Monod* package, which 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/`.

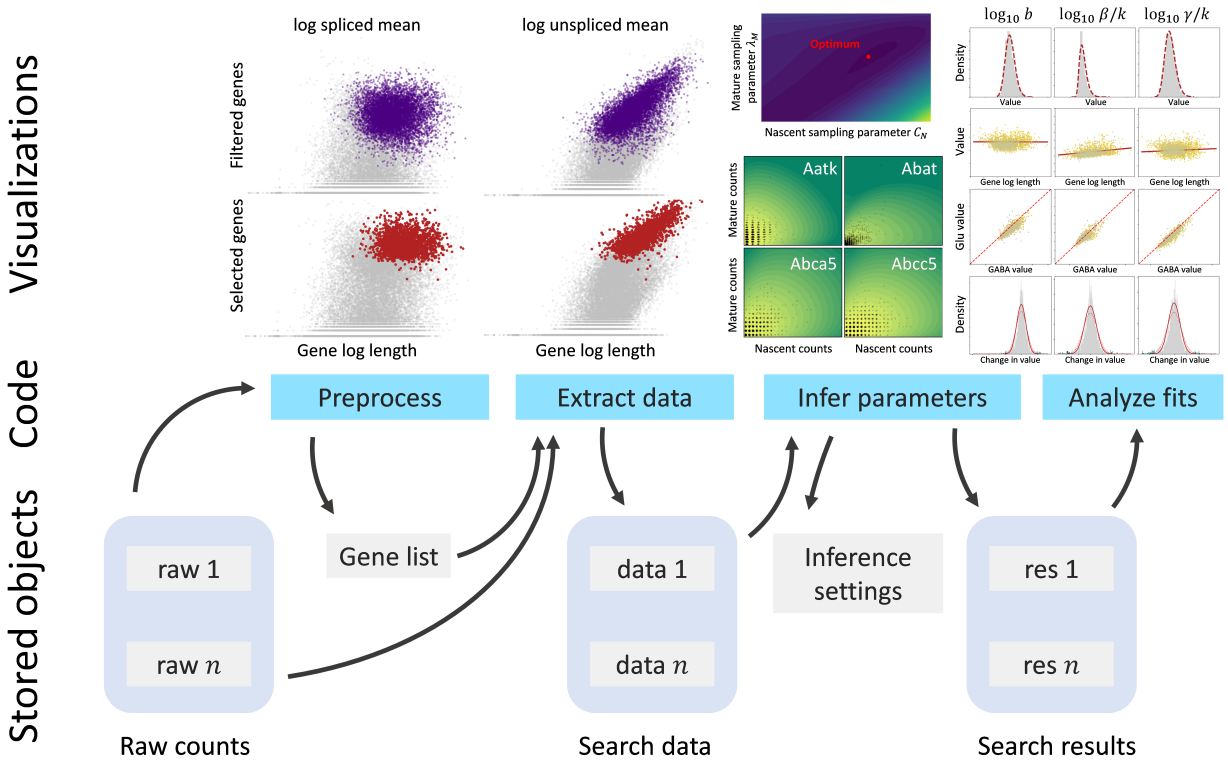

Figure 1: A schematic representation of the *Monod* workflow. Center: steps of the algorithm. Top: visualizations produced for the user. Bottom: data objects stored on disk.

2 Results

2.1 Model definition

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

$$\varnothing \xrightarrow{k} B \times \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing. \tag{1}$$

This diagram encodes a continuous-time Markov chain on a bivariate discrete space of molecule counts for each gene, where $\mathcal{N}$ is an unspliced (or nascent) mRNA species and $\mathcal{M}$ is a spliced (or mature) mRNA species. 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 $\mathcal{N}$ will become a molecule of $\mathcal{M}$ after an exponentially-distributed delay with rate β. Analogously, γ is the degradation rate, such that a given molecule of $\mathcal{M}$ 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 [13] to define B as geometrically-distributed with mean b [20], inducing negative binomial-like [31] distributions of $\mathcal{N}$ and $\mathcal{M}$ [18–20, 32–36]. 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 [37] 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:

$$\begin{aligned} \mathcal{N}' &\xrightarrow{\lambda_N} \mathcal{N}, \\ \mathcal{M}' &\xrightarrow{\lambda_M} \mathcal{M}, \end{aligned} \tag{2}$$

where λ_N and λ_M are the Poisson process rates for each species, $\mathcal{N}'$ and $\mathcal{M}'$ are unobserved *in vivo* molecules, and $\mathcal{N}$ and $\mathcal{M}$ 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 [29].

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 the constitutive model (with $B = 1$) and two other overdispersed models, with optional technical noise components, outlined in Section S2.

2.2 Normalization and noise decomposition

The workflow and its mathematical foundations allow us to summarize the noise behaviors implicit in mRNA distributions. Specifically, the combined models of biological and technical noise determine the fractions of variation that correspond to intrinsic noise resulting from the discrete numbers of molecules, extrinsic noise resulting from stochasticity in the transcription rate (i.e., variable burst size) between cells, and technical noise, resulting from stochastic sequencing of cDNA libraries. We describe these noise decompositions in Section S3.1.

This mechanistic approach to noise decomposition affords a mechanistically grounded alternative to standard sequencing analysis workflows, which normalize and log-transform data to reduce or re-scale variation [2]. Such standard workflows are premised on empirical observations of overdispersion with respect to a Poisson distribution in observed data (typically modeled using a negative binomial distribution), and an assumption that this variance is technical [28].

Formally, in standard workflows the steps are typically conceptualized as accounting for technical variation under the following model:

$$M_i = Multinomial(X_i, p_1, ..., p_g), \tag{3}$$

where i is the index of a cell, g is the total number of genes, $p_1, ..., p_g$ are the probabilities of sampling each gene, assumed uniform across cells, and X_i is a stochastic scaling factor that represents the number of mRNA captured in each cell [2]. This model incorporates a simple description of variability in the sequencing chemistry, but omits all variability in the underlying biology, which may be substantial. With *Monod*, it is possible to investigate whether this conceptualization holds, and whether normalization actually removes technical noise. As normalization is targeted at stabilizing the distributions of high-expression, noticeably overdispersed genes (Figure S3a), we focus on its impact on this subset.

First, we used the log1pPF procedure [38, 39], described in Section S5.2, and computed the fraction of variance retained by applying it to the spliced count matrix, relative to the raw data.

For this analysis, we used the entire glutamatergic neuron population from a single mouse sample consisting of cells from the primary motor cortex. This fraction, calculated according to the procedure in Section S3.1, putatively represents the amount of per-gene biological variation under the model in Equation 3. The results are shown in Figure S3b: the amount of variation smoothly decreases with the average expression. Therefore, the procedure tends to attribute overdispersion to technical effects.

Next, we extracted glutamatergic cell subtypes, fit the bursty model with technical noise outlined in Section 2.1 to each subtype using *Monod*, obtained the parameter maximum likelihood estimates, and applied the analysis outlined in Section S3.1.1. The results are shown in Figure S3c, and demonstrate an increase in biological noise with average expression: the procedure tends to attribute overdispersion to biological effects. Applying the noise decomposition reported in Equation S17 to compute the fraction of per-gene biological variation based on the entire glutamatergic dataset produced largely the same behavior.

The results for the two procedures are compared in Figure S4; the noise attributions are starkly discordant. Even though the true amount of biological variation is unknown, we can compute a lower bound on this variation by evaluating the fraction of inter-subtype variance in average expression relative to total variance in the cell type (Section S3.1.1). Comparisons to this lower bound are shown in Figure 2: the mechanistic analysis is concordant with the lower bound for all genes, whereas the normalization procedure consistently violates it for high-expression genes of interest. This result suggests that caution in the interpretation of normalized data is warranted.

2.3 Identification of differential expression patterns between cell types

Coarse cell types are defined with reference to a set of known, high-expression marker genes, which are leveraged to obtain clusters with statistically identifiable differences in average marker expression [2]. We propose that a more interpretable picture of differential expression patterns can be gleaned through using distributions parameterized by biophysical variables. Here, we take the first step using a *post hoc* analysis: we fit a model to pre-clustered cell type data, then select genes that appear to have substantial deviation in biological parameters between cell types. The details of the procedure are described in Section S3.2.

In the broadest sense, the task of identifying differentially expressed (DE) genes can be viewed as a statistical hypothesis test. This hypothesis test identifies differences in a test statistic of interest. Therefore, we find it useful to introduce the notation DE-θ to denote tests using θ as a test statistic. In this notation, the typical approach uses a DE-$\mu_{\phi(\mathcal{M})}$ test, i.e., it identifies differences in the mean μ of some transformation ϕ (typically, normalization followed by a logarithm) of the spliced count matrix $\mathcal{M}$. This procedure can identify changes in the location parameter, but leaves an important lacuna: some genes may show differences in spliced and unspliced count distribution shapes that are not accompanied by a readily identifiable change in mean spliced expression.

We filtered a set of mouse brain datasets, extracting GABAergic and glutamatergic cell populations and storing them in separate *loom* files. By applying a t-test to spliced count data, which demonstrate some deviations in mean expression between cell types (Figure S5a), we identified a set of putative DE-$\mu_{\phi(\mathcal{M})}$ genes (Figure S5b). Upon fitting the bursty transcription model, we observed a strongly positive relationship between the cell types' inferred parameters, albeit with conspicuous off-diagonal genes (Figure S6a). We constructed a Gaussian model for aleatory variation in the *residuals* and used fixed-point iteration to detect outliers, putative DE-b, DE-β/k, and DE-γ/k genes (Figure S6b).

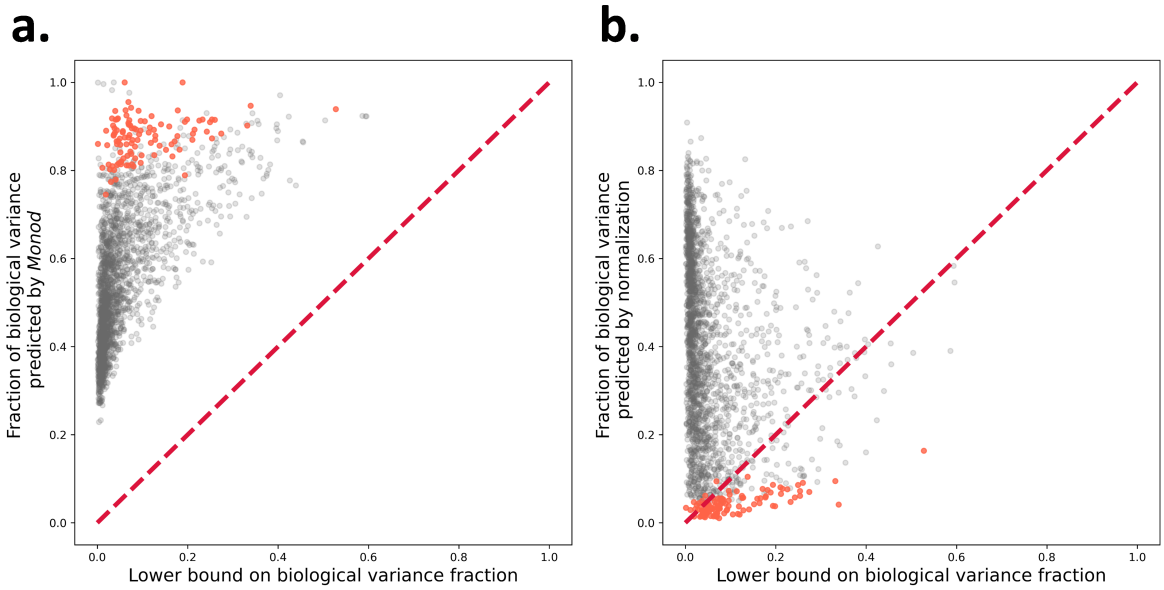

Figure 2: Normalization underestimates biological variation, especially in high-expression genes. **a.** The predicted fraction of biological variance under a *Monod* mechanistic model is consistent with the lower bound on biological variance obtained from inter-cluster variation. **b.** The fraction of biological variance retained after normalization is less than the lower bound (scarlet line: identity lower bound on biological variance fraction; gray points: genes below the 95th percentile by mean spliced expression; red points: genes above the 95th percentile by mean spliced expression).

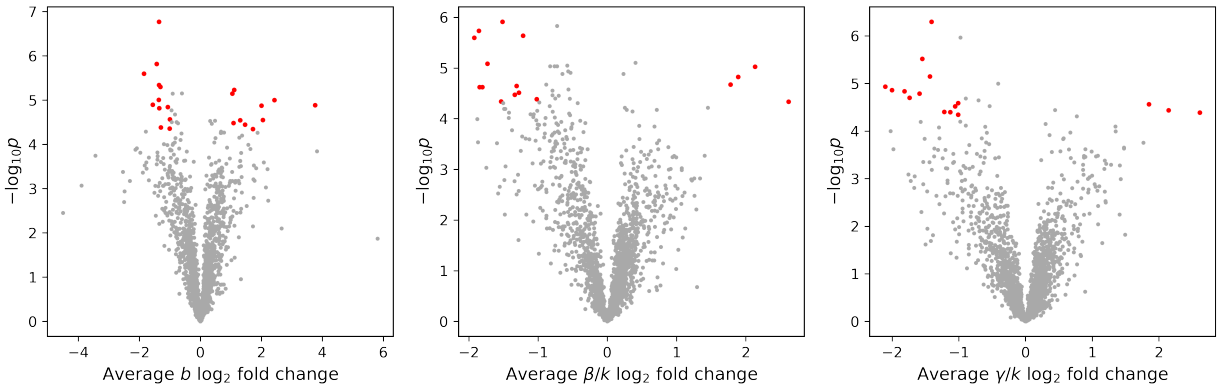

Figure 3: Differential expression analysis with *Monod* identifies genes that exhibit consistent inter-cell type parameter modulation across four biological replicates.

To validate the method for mechanistic DE analysis, we combined the inferred parameters from four biological replicates and performed a t-test (Figure 3). This procedure identified fewer genes, which were largely contained in the single-dataset fits (Table S10).

The mechanistic DE approach is reminiscent of negative binomial regression methods previously proposed for scRNA-seq data [26, 39, 40], but has several key distinctions. The model is not constrained to be closed-form: the shape of the distribution may be modified to encode hypotheses about the physics of transcription and sequencing [9]. The data are not assumed to be univariate: we fit bivariate distributions, with an eye to higher-dimensional datasets which may become available with improvements to isoform resolution [41] and protein quantification [42, 43].

Most importantly, the parameters have a specific physical interpretation, which can provide a natural route for analysis. For example, the co-modulation of the β/k and γ/k parameters (Figure 4, computed from four replicates), suggests that the DE-β/k and DE-γ/k patterns should properly be ascribed to *burst frequency* modulation, even though k was not explicitly fit (Section S3.2.1). In a longer-term perspective, this interpretability can motivate experiments and interventions targeting the physical phenomena identified through the model.

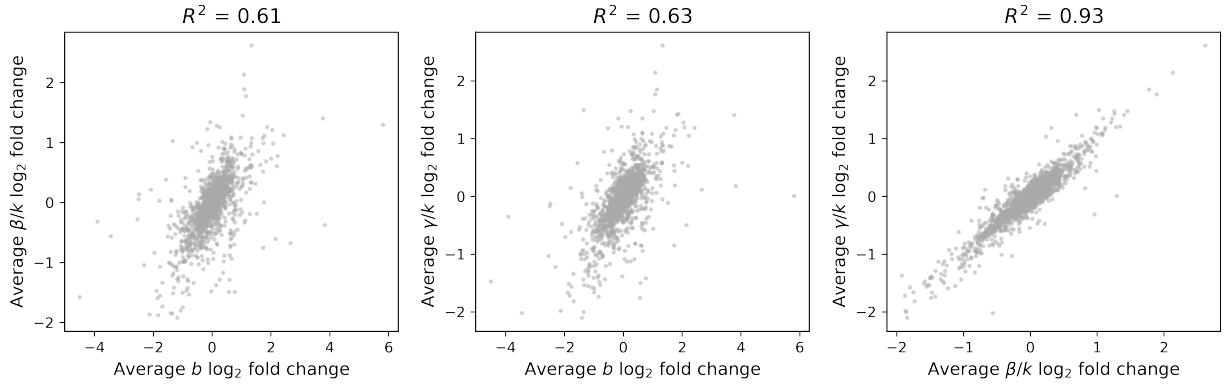

Figure 4: Non-dimensionalized splicing rate and degradation rate modulation are highly correlated, suggesting a transcription rate frequency modulation mechanism.

We found that transcription frequency modulation accounts for a large fraction of genes identified as DE in the two rate parameters, although not to the same extent: targeted modulation of γ/k appears to be more prevalent than that of β/k. The regulation of degradation has been implicated in buffering noise [44], and appears to be a standard bioregulatory motif. However, the specificity and kinetics of these processes on a genome-wide scale are poorly characterized, and it is not yet clear that the simple one-step model of mRNA processing we explore here is sufficient to motivate comparisons to the literature on splicing.

Most interestingly, we observed several genes that consistently exhibited transcriptional parameter modulation without being detected by the t-test on mean expression. This category included *Cacnb2*, *Camk2a*, *Extl3*, *Garnl3*, *Kcnh3*, *Lpgat1*, *Ndst1*, *Nhsl2*, *Nrip1*, *Prkce*, *Sdk2*, *Sv2b*, *Syngap1*, and *Tle4* for b modulation and *A230006K03Rik*, *Dlg1*, *Sv2b*, and *Satb2* for k modulation.

Of these, *A230006K03Rik*, *Cacnb2*, *Camk2a*, *Dlg1*, *Extl3*, *Kcnh3*, *Lpgat1*, *Ndst1*, *Nhsl2*, and *Prkce*, *Satb2*, and *Syngap1* exhibited a lower than twofold average expression change across the four datasets. We visualize the raw distributions and their fits for a single dataset in Figure S7. We omit the discussion of *A230006K03Rik*, *Camk2a*, *Satb2*, *Sv2b*, and *Prkce*, as their fits are quite

poor. *Camk2a* demonstrates considerable multimodality, likely due to cell subtype localization or incorrect assignment of Camk2a neurons to the GABAergic class [45]. *Sv2b* may show multimodality due to sequencing or alignment ambiguities between its two major isoforms [46]. In addition, *A230006K03Rik* and *Prkce* are uncharacterized. We note the multimodality of *Satb2* within the GABAergic subtype is best evident in the unspliced marginal, providing further motivation for the quantification of multiple mRNA species.

The identified genes are largely related to cellular structure. *Extl3* [47] and *Ndst1* [48, 49] participate in the synthesis of heparan sulfate, a polysaccharide localized to the membrane and extracellular matrix, but may also have signaling properties. Similarly, *Dlg1* codes for a scaffolding protein with emerging roles in signaling and vesicle trafficking [50]. *Lpgat1* is involved in the process of lipid biosynthesis [51, 52]. *Nhsl2* appears to participate in actin cytoskeleton assembly [53]. The role of the other genes is more ambiguous, although they are not inconsistent with cellular structure regulation. *Cacnb2* codes for a calcium channel [54] "required for dendritic development" [55]. *Syngap1* encodes a regulatory protein, one of whose roles involves mediating the development of dendritic spines [56]. *Kcnh3* is poorly characterized, but a recent report suggests its variants are associated with microcephaly [57]. However, the import and basis of cell type differences in variation rather than average expression is as of yet obscure. The mechanism may involve expression compensation explored using theoretical tools [58] and recently observed under DNA repair stress [59].

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 (driven by a memoryless process), technical noise models (Bernoulli- and Poisson-like, with a possible length bias), modalities (spliced and unspliced mRNA), correlation structure (only intra-gene relationships), and heterogeneity structures (a single cell type). While some of these assumptions may be simplistic, the current approach to single-cell RNA-seq analysis, particularly direct use of spliced count estimates (derived from UMI counts), corresponds to an even more unrealistic model which omits all biological variation (Equation 3). With even a basic mechanistic model for integrating spliced and unspliced counts, we have demonstrated the possibility for interesting discovery. For example, the identification of subtle distributional differences between cell types is possible by extracting internally homogeneous cell types based on previous annotations. Furthermore, we anticipate that *Monod* can be extended to utilize multimodal data to parametrize more complex mechanistic models.

4 Acknowledgments

We thank Tara Chari, Meichen Fang, and Sina Booeshaghi for useful discussions in the course of developing *Monod*. The *Monod* package uses algorithms implemented in the *NumPy* [60], *SciPy* [61], and *numdifftools* [62] Python packages. G.G. and L.P. were partially funded by NIH U19MH114830.

References

[1] 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.

[2] 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.

[3] 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.

[4] Simone Tiberi. DifferentialRegulation, April 2022.

[5] 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.

[6] 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.

[7] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. Preprint, bioRxiv: 2022.02.12.480214, February 2022.

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

[9] 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. Preprint, bioRxiv: 2021.09.06.459173, September 2021.

[10] Péter Érdi and Gábor Lente. *Stochastic chemical kinetics: theory and (mostly) systems biological applications*. Springer Complexity. Springer, New York, 2014.

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

[12] Mitchell Lewis. A tale of two repressors – a historical perspective. *Journal of molecular biology*, 409(1):14–27, May 2011.

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

[14] M S Ko, H Nakauchi, and N Takahashi. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *The EMBO Journal*, 9(9):2835–2842, September 1990.

[15] Minoru S H Ko. A Stochastic Model for Gene Induction. *Journal of Theoretical Biology*, 153(2):181–194, 1991.

[16] 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.

[17] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. *eLife*, 5:e12175, January 2016.

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

[19] 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.

[20] 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.

[21] Jong Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14:R7, 2013.

[22] 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.

[23] Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. Preprint, bioRxiv: 657619, June 2019.

[24] Lucy Ham, Rowan D. Brackston, and Michael P. H. Stumpf. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Physical Review Letters*, 124(10):108101, March 2020.

[25] Chen Jia. Kinetic Foundation of the Zero-Inflated Negative Binomial Model for Single-Cell RNA Sequencing Data. *SIAM Journal on Applied Mathematics*, 80(3):1336–1355, January 2020.

[26] 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.

[27] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell UMI data. *Genome Biology*, 21:196, December 2020.

[28] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, February 2020.

[29] Gennady Gorin and Lior Pachter. Length Biases in Single-Cell RNA Sequencing of pre-mRNA. Preprint, bioRxiv: 2021.07.30.454514, July 2021.

[30] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. anndata: Annotated data. Preprint, bioRxiv: 2021.12.16.473007, December 2021.

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

[32] 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.

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

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

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

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

[37] 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.

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

[39] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.

[40] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

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

[42] Gennady Gorin, Valentine Svensson, and Lior Pachter. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, 21:39, February 2020.

[43] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017.

[44] Antoine Baudrimont, Vincent Jaquet, Sandrine Wallerich, Sylvia Voegeli, and Attila Becskei. Contribution of RNA Degradation to Intrinsic and Extrinsic Noise in Gene Expression. *Cell Reports*, 26(13):3752–3761.e5, March 2019.

[45] Sruti Rayaprolu, Sara Bitarafan, Juliet V. Santiago, Ranjita Betarbet, Sydney Sunna, Lihong Cheng, Hailian Xiao, Ruth S. Nelson, Prateek Kumar, Pritha Bagchi, Duc M. Duong, Annie M. Goettemoeller, Viktor János Oláh, Matt Rowan, Allan I. Levey, Levi B. Wood, Nicholas T. Seyfried, and Srikant Rangaraju. Cell type-specific biotin labeling in vivo resolves regional neuronal and astrocyte proteomic differences in mouse brain. *Nature Communications*, 13:2927, May 2022.

[46] M Bajjalieh, D Fran, M Weimann, K McConnell, and RH Scheller. Differential Expression of Synaptic Vesicle Protein 2 (SV2) lsoforms. *The Journal of Neuroscience*, 14(9):5223–5235, September 1994.

[47] Shuhei Yamada. Specific functions of Exostosin-like 3 (EXTL3) gene products. *Cellular & Molecular Biology Letters*, 25:39, December 2020.

[48] Kay Grobe, Masaru Inatani, Srinivas R. Pallerla, Jan Castagnola, Yu Yamaguchi, and Jeffrey D. Esko. Cerebral hypoplasia and craniofacial defects in mice lacking heparan sulfate *Ndst1* gene function. *Development*, 132(16):3777–3786, August 2005.

[49] Srinivas R. Pallerla, Yi Pan, Xin Zhang, Jeffrey D. Esko, and Kay Grobe. Heparan sulfate Ndst1 gene function variably regulates multiple signaling pathways during mouse development. *Developmental Dynamics*, 236(2):556–563, February 2007.

[50] Laurence Walch. Emerging Role of the Scaffolding Protein Dlg1 in Vesicle Trafficking. *Traffic*, 14(9):964–973, September 2013.

[51] Xiaolin Yu, Xiaomei Zhang, and Yanxia Zhang. Identification of a 5-Gene Metabolic Signature for Predicting Prognosis Based on an Integrated Analysis of Tumor Microenvironment in Lung Adenocarcinoma. *Journal of Oncology*, 2020:5310793, June 2020.

[52] Michael T. Traurig, Julieanna I. Orczewska, Daniel J. Ortiz, Li Bian, Alejandra M. Marinelarena, Sayuko Kobes, Alka Malhotra, Robert L. Hanson, Clint C. Mason, William C. Knowler, Clifton Bogardus, and Leslie J. Baier. Evidence for a role of *LPGAT1* in influencing BMI and percent body fat in Native Americans. *Obesity*, 21(1):193–202, January 2013.

[53] Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y. S. Kim, Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, and Manuel D. Leonetti. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, March 2022.

[54] Marcel Meissner, Petra Weissgerber, Juan E. Camacho Londoño, Jean Prenen, Sabine Link, Sandra Ruppenthal, Jeffery D. Molkentin, Peter Lipp, Bernd Nilius, Marc Freichel, and Veit Flockerzi. Moderate Calcium Channel Dysfunction in Adult Mice with Inducible

Cardiomyocyte-specific Excision of the *cacnb2* Gene. *Journal of Biological Chemistry*, 286(18):15875–15882, May 2011.

[55] H.C. Martins, A.Ö. Sungur, C. Gilardi, M. Pelzl, S. Bicker, F. Gross, J. Winterer, T.M. Kisko, N. Malikowska-Racia, M.D. Braun, K. Brosch, I. Nenadic, F. Stein, S. Meinert, R.K.W Schwarting, U. Dannlowski, T. Kircher, M. Wöhr, and G. Schratt. Bipolar-associated miR-499-5p controls neuroplasticity by downregulating the $Ca_v1.2$ L-type voltage gated calcium channel subunit CACNB2. Preprint, bioRxiv: 2021.06.09.447782, June 2021.

[56] Nallathambi Jeyabalan and James P. Clement. SYNGAP1: Mind the Gap. *Frontiers in Cellular Neuroscience*, 10:32, February 2016.

[57] Regie Lyn P. Santos-Cortez, Valeed Khan, Falak Sher Khan, Zaib-un-Nisa Mughal, Imen Chakchouk, Kwanghyuk Lee, Memoona Rasheed, Rifat Hamza, Anushree Acharya, Ehsan Ullah, Muhammad Arif Nadeem Saqib, Izoduwa Abbe, Ghazanfar Ali, Muhammad Jawad Hassan, Saadullah Khan, Zahid Azeem, Irfan Ullah, Michael J. Bamshad, Deborah A. Nickerson, Isabelle Schrauwen, Wasim Ahmad, Muhammad Ansar, and Suzanne M. Leal. Novel candidate genes and variants underlying autosomal recessive neurodevelopmental disorders with intellectual disability. *Human Genetics*, 137(9):735–752, September 2018.

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

[59] 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.

[60] 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.

[61] 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, İlhan 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.

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

[63] A. Sina Booeshaghi, Zizhen Yao, Cindy van Velthoven, Kimberly Smith, Bosiljka Tasic, Hongkui Zeng, and Lior Pachter. Isoform cell-type specificity in the mouse primary motor cortex. *Nature*, 598(7879):195–199, October 2021.

[64] Gennady Gorin and Lior Pachter. Monod supporting data (10.5281/zenodo.6612727), June 2022.

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

[66] 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.

[67] 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.

[68] 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.

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

[70] The MathWorks. MATLAB R2022a, 2022.

[71] Audrey Qiuyan Fu and Lior Pachter. Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Statistical Applications in Genetics and Molecular Biology*, 15(6), January 2016.