

Genetic Interrogation of Expression Regulation

Thesis by
Maria T. Carilli

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Biochemistry and Molecular Biophysics



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended May 18, 2026

© 2026

Maria T. Carilli

ORCID: 0000-0002-8977-7224

Some rights reserved. This thesis is distributed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License

ACKNOWLEDGEMENTS

Firstly, I would like to thank to my advisor, Lior Pachter, whose kindness, wisdom, and passion for the truth were bright beacons throughout my graduate school journey. I will be forever amazed at his ability to provide the space and encouragement to let so many personalities, interests, and ideas flourish uniquely. He showed me how important it is to be a good and principled scientist; and as important to be a good and principled person.

I would like to thank also my committee chair, Rob Phillips, whose enthusiasm for science and sense of wonder leave an indelible mark. Few days go by that I don't look about the world and wonder at its working because of the philosophy I've learned from him. Thank you also to my committee members Yisong Yue and Barbara Engelhardt, who have provided thoughtful scientific guidance and advice on the longer-term prospects of career and life.

Over the course my Ph.D., I have had the immense privilege to work with and learn from curious and creative collaborators and mentors, whom I also call friends. I especially thank Gennady Gorin for his unconventional vision for what single-cell data analysis should be. I am thankful for his mentorship, discussions, and availability during and after his time at Caltech. I would also like to thank Ingilief B. Hallgrímsdóttir for introducing me to the world of genetics: I learned from her experience and insight at our every meeting and conversation. Thank you to Kayla Jackson, for companionship through tough classes and whose close collaboration on our recent project reminded me how enriching and enjoyable the pursuit of knowledge can be. Thank you to Tara Chari, for advice on projects and the ups and downs of grad school.

Thank you to all of the members of the Pachter lab, past and present, who provided an intellectually rich and personally supportive community in which to work: Laura Luebbert, Delaney Sullivan, Meichen Fang, Nikhila Swarna, Conrad Oakes, Joe Rich, Maya Caskey, Charlene Kim, Vera Beilenson, Catherine Felce, Rebekah Loving Ngo, Taleen Dilanyan, Anne Yeokyoung Kil, Kristján Eldjárn Hjörleifsson, Lambda Moses, Sina Boeshaghi. Thank you also to the members of the Mortazavi lab, particularly Ryan Weber.

I have deeply appreciated my years as part of the close-knit and brilliant Caltech community. Particularly, thank you to Tom Röscher who, as my mentor for my

first-year first rotation, made my entry into graduate school smooth and fun; and to Justin Bois, who is an example of what a great teacher should be. I also thank Elijah Cole from whom I learned about doing research in industry during a summer internship.

I am thankful for the mentorship of Deborah Wuttke, my undergraduate advisor, for her support and belief in me: without it I would never have come to Caltech.

Beyond the academic, thank you to Nikhila Swarna, Juliet Lee, and Samantha Chang, for our LA explorations, music making, and many girls' nights.

These past five years, I have run more miles than there are pages in this thesis (by at least one order of magnitude). For being companions on these adventurous endeavors on the trails of Southern California, I would like to thank my friends in the Caltech Alpine Club. I wish I could list all of their names here, but if I tried, I'd be writing all night, and I have to get for the 7 a.m. trail run.

I most especially want to thank my family, to have as which I am the luckiest person in the world. Michael, Joe and Tina, John and Kelsey, and my sweetest nieces, Kaia and Nika. They may be my siblings, but they are awfully cool humans. Thank you to my sister, Rose, to whom I could (and do) talk for hours and still have words for hours more. Though we've been on opposite coasts, the stresses of graduate school and the weights of life have been lessened by her friendship.

And, of course, my thanks to Jackson Owen Parker, my "shelter from the storm," for his constancy, his humor, and his goodness of heart.

Finally, I would like to thank my parents, Christopher and Adrianna Carilli, for filling our home with books and our heads with questions. Thank you is a pale phrase to offer in light of their sacrifices, originality, and love.

ABSTRACT

Understanding the effects of variation in the genome on organisms' phenotypes is a central goal of biology. For the past several decades, this has been done by performing large-scale association tests to identify links between genetic variants across the genome and particular diseases or traits. As most variants are in non-coding regions of the genome and act only in specific contexts, there arose the intermediate step of associating variants with gene expression patterns in particular tissues using bulk RNA-sequencing data and, with the recent widespread adoption of single-cell RNA-sequencing, particular cell types. However, associative tests are generally restricted to variants within some sequence distance of the gene, as testing tens of millions of distal variants against tens of thousands of genes in hundreds of contexts is computationally and statistically burdensome. Moreover, associating these proximal variants to changes in average gene expression using scRNA-seq data is far from pinpointing the cellular process they may affect. To move beyond analysis of averages, recent work has advocated for a mechanistic approach to analyzing scRNA-seq by modeling the underlying biological processes of transcription, splicing, and degradation.

In this thesis, we first show how biophysical models are useful for identifying changes in biophysical processes not accessible at the level of mean expression. We next develop accelerated inference procedures for these models to make feasible their application at the scale required for genetic association tests. We then propose a framework for testing for the presence or absence of proximal or distal gene regulation using homozygous crosses. Finally, we couple the genetic testing framework and biophysical models to identify regulatory strategies of biophysical processes across cell types in eight tissues of eight genetically diverse mouse strains.

The contents of this thesis are largely drawn from the following publications. Relevant publications are cited at the beginning of each chapter with the following abbreviations for author contributions: M.C.: Maria Carilli, K.J.: Kayla Jackson, L.P.: Lior Pachter, I.B.H: Ingilief B. Hallgrímsdóttir, G.G.: Gennady Gorin, T.C.: Tara Chari, Y.C.: Yongin Choi, J.J.V.: John J. Vastola.

PUBLISHED CONTENT AND CONTRIBUTIONS

- [1] Maria Carilli, Kayla Jackson, and Lior Pachter. *The Rayleigh Quotient and Contrastive Principal Component Analysis II*. Apr. 8, 2026. DOI: 10.64898/2026.04.08.717236. bioRxiv: 2026.04.08.717236. URL: <https://doi.org/10.64898/2026.04.08.717236>.
M.C., K.J., and L.P. developed the methods, produced the results and drafted the manuscript. M.C. worked on functional ρ PCA, while K.J. worked on spatial ρ PCA.
- [2] Ingileif B. Hallgrímsdóttir, Maria Carilli, and Lior Pachter. *Estimating cis and trans contributions to differences in gene regulation*. Feb. 18, 2026. DOI: 10.1101/2024.07.13.603403. bioRxiv: 2024.07.13.603403. URL: <https://doi.org/10.1101/2024.07.13.603403>.
I.B.H., M.C., L.P. participated in the conception of the project, models, and statistical tests, and writing the manuscript. M.C. performed analyses and implemented the method as a software package.
- [3] Ryan Weber, Maria Carilli, Elisabeth Rebboah, Ghassan Filimban, Heidi Yahan Liang, Diane Trout, Margaret Duffield, Parvin Mahdipoor, Erisa Taghizadeh, Negar Fattahi, Negar Mojgani, Romina Mojaverzargar, Shimako Kawauchi, Brian A. Williams, Grant R. MacGregor, Barbara J. Wold, Lior Pachter, Ingileif B. Hallgrímsdóttir, and Ali Mortazavi. “Hybrid crosses reveal a cell-type-specific landscape of mouse regulatory variation”. In: *bioRxiv* (2026). DOI: 10.64898/2026.04.02.716195.
A.M., I.B.H., L.P., and B.J.W. designed the project. S.K. acquired mice and oversaw their husbandry. G.F. and H.Y.L. dissected tissues under guidance from B.A.W., G.R.M., and S.K. M.D., P.M., E.T., N.F., N.M., and R.M. helped with sample collection as well as performed nuclei isolation, barcoding, and library preparation. D.T. uploaded data and metadata to the IGVF portal. M.C. and E.R. provided input and code for data analysis. R.W. performed data analysis, generated figures, and wrote and edited the manuscript with significant input from A.M.
- [4] Maria Carilli, Kayla Jackson, and Lior Pachter. *The Rayleigh Quotient and Contrastive Principal Component Analysis I*. Nov. 19, 2025. DOI: 10.1101/2025.11.19.689125. bioRxiv: 2025.11.19.689125. URL: <https://doi.org/10.1101/2025.11.19.689125>.
M.C., K.J., and L.P. developed the methods, produced the results and drafted the manuscript.
- [5] Gennady Gorin, Tara Chari, Maria Carilli, John J. Vastola, and Lior Pachter. “Monod: model-based discovery and integration through fitting stochastic transcriptional dynamics to single-cell sequencing data”. In: *Nature Methods* 22 (Nov. 7, 2025), pp. 2286–2300. DOI: 10.1038/s41592-025-02832-x.

URL: <https://doi.org/10.1038/s41592-025-02832-x>.

M.C. conceptualized and implemented biology discovery and model selection sections, while discussing all sections. All authors contributed extensively to the work presented in this paper. The first draft of the paper was conceptualized and written by G.G. and L.P. ‘Assumptions and approach of Monod’ was predominantly conceptualized by G.G., J.J.V. and L.P. ‘Monod generalizes differential expression analyses’ was predominantly conceptualized and executed by G.G., T.C. and L.P. ‘Monod identifies strategies of resistance and recovery’ and ‘Model selection and insight about gene regulation strategies’ were predominantly conceptualized and executed by M.C. and L.P. ‘Principled integration of multiple modalities’ was predominantly conceptualized and executed by G.G., J.J.V., M.C. and L.P. ‘Assessing loss of biological signal after preprocessing’ was predominantly conceptualized and executed by G.G., J.J.V. and L.P.

- [6] Laura Luebbert, Delaney K. Sullivan, Maria Carilli, Kristján Eldjárn Hjörleifsson, Alexander Vilorio Winnett, Tara Chari, and Lior Pachter. “Detection of viral sequences at single-cell resolution identifies novel viruses associated with host gene expression changes”. In: *Nature Biotechnology* 44 (Apr. 22, 2025), pp. 100–109. doi: 10.1038/s41587-025-02614-y. L.L. and L.P. conceptualized the project. L.L., D.K.S. and K.E.H. developed and implemented the kallisto translated search algorithm, including extensive testing. L.L. and L.P. designed the validation, benchmarking and experimental data analysis with input from all authors, and L.L. performed the analyses and visualized the results. M.C. designed the logistic regression models, with adjustments from L.L. and T.C. T.C. provided critical feedback on statistical methods. A.V.W. provided extensive validation data and critical feedback on the validation, manuscript and revisions. L.L. wrote the original draft of the manuscript and revisions. L.P. supervised the project. All authors contributed to revisions and reviewed, edited and approved the final manuscript.
- [7] Maria Carilli, Gennady Gorin, Tara Chari, and Lior Pachter. “Spectral neural approximations for models of transcriptional dynamics”. English. In: *Biophysical Journal* (May 2024). Publisher: Elsevier. ISSN: 0006-3495. doi: 10.1016/j.bpj.2024.04.034. URL: [https://www.cell.com/biophysj/abstract/S0006-3495\(24\)00314-X](https://www.cell.com/biophysj/abstract/S0006-3495(24)00314-X) (visited on 06/26/2024). M.C. implemented and adapted the proposed strategy, trained models, and performed analyses. G.G. conceptualized the neural approximation strategies and was involved in implementation, statistical analyses, and basis function design. T.C. and L.P. provided feedback, design suggestions, and trouble-shooting advice.
- [8] Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. “Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data”. In: *Nature Methods* 21.8 (2024), pp. 1466–

1469.

M. C., G.G. and Y. C. implemented the strategy. M.C. ran analyses. G. G. conceptualized the project and provided detailed feedback and project direction. T. C. and L.P. provided guidance and suggestions throughout.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Published Content and Contributions	vii
Table of Contents	ix
Chapter I: Introduction	1
Chapter II: Contrastive Dimensionality Reduction for Differential Variance	8
2.1 The Rayleigh Quotient for Genomics	8
2.2 Solving the Rayleigh Quotient for Functional Data	18
Chapter III: Biophysical Models for Stochastic Transcription	25
3.1 Chemical Master Equation Models	25
3.2 Application for Biological Insight	31
3.3 Model Selection for Biophysical Hypothesis Generation	34
Chapter IV: Neural Approximations for Intractable Biophysical Models	40
4.1 Existing Solvers and Limitations	42
4.2 Kernel Weight Regression and Extensions	45
Chapter V: Biophysical Models in Variational Autoencoders	57
5.1 Single-cell Variational Inference	57
5.2 Adapting Variational Autoencoders for Biophysical Models	59
Chapter VI: Disentangling <i>Cis</i> and <i>Trans</i> Gene Regulation	73
6.1 New Axes for New Insight	73
6.2 Single and Multi-Sample Hypothesis Testing	77
Chapter VII: Regulation of Biophysical Processes in Founder Mouse Strains	89
7.1 Adapting Modeling Strategies	89
7.2 Large-scale Trends and Summarization	96
7.3 Exploring the Processes Driving Allele Specificity	99
Chapter VIII: Conclusion and Perspectives	104
Bibliography	106

Chapter 1

INTRODUCTION

It requires indeed some courage to undertake a labor of such far-reaching extent; this appears, however, to be the only right way by which we can finally reach the solution of a question the importance of which cannot be overestimated in connection with the history of the evolution of organic forms.

GREGOR MENDEL

One of the most fundamental questions in biology is how chemical information stored in a sequence of DNA molecules encapsulated in tiny cell-chambers of living organisms gives rise to a plethora of biological behaviors and characteristics, differing across tissues within the same organism, organisms of the same species, and species ranging the tree of life. While today, unprecedentedly large-scale analyses associate variants in DNA sequences to phenotypic differences using tens of thousands of samples, it is valuable to remember how we arrived in the age of big data genomics as a way of guiding how to advance.

That discrete traits can be passed down from organism to organism, Mendel first observed in the 1860s by performing crosses of the unassuming garden pea, *Pisum sativum* [1]. Without knowledge of DNA, he described the laws governing “*elementen*,” or hereditary factors, that hybrid crosses inherited from each parent: functional units that later came to be described as *alleles* of *genes*. A few years later, in 1869, Miescher discovered a phosphorus-rich compound in the nuclei of cells that he called “*nuclein*,” composed mostly of what we now call deoxyribose nucleic acid (DNA) [2]. However, it was not until the work of Oswald Avery in the mid-1940s, building on Frederick Griffith’s earlier experiments, that the nucleic acid DNA was shown to “transform” bacteria from non-virulent to virulent and was thus discovered to be the substance carrying Mendel’s proposed genetic unit [3].

The 1950s saw the structure of DNA resolved, and the famous (and infamously misunderstood) proposition by Crick of directional information flow from DNA to RNA to protein, but not from protein to either protein or nucleic acid [4]. In the 1970s and 80s, nucleic acid sequencing technologies emerged, with the first RNA genome, bacteriophage MS2, sequenced in 1976 [5] and the first DNA genome, bacteriophage ϕ X174, in 1977 [6], both only several thousand nucleotides long. The first multicellular organism to have its 97 megabase (Mb) genome sequenced was *Caenorhabditis elegans*, published in 1998 [7], followed by the 2001 publication of the 30-times larger human genome as the culmination of the 3-billion-dollar¹ Human Genome Project (HGP) [8]. The initial publication, however, was incomplete, and only as recently as 2022 did the follow-up Telomere-to-Telomere consortium publish an end-to-end reference sequence that includes highly repetitive regions near centromeres and telomere ends [9]².

The creation of a human reference sequence was a foundational event in the ongoing history of characterizing the function of sequences of DNA. With it, genome-wide association studies (GWAS) began to link *variation* from that reference in the form of single-nucleotide polymorphisms (SNPs) to sample phenotypes like disease and other traits [11, 12, 13]. Today, as the cost of sequencing a genome has dropped from three billion to several hundred dollars, GWAS are performed with ever larger sample sizes [14], increasing the number of variants implicated in an extensive set of diseases and traits, including heart disease [15], psychiatric disorders [16, 17], multiple forms of cancer [18], height [19], and behavioral habits like tobacco use [20].

However, most of the SNPs identified as significant in GWAS are located in non-coding regions of the genome, making their functional mechanism hard to interpret. This is unsurprising, as only about 1.5% of the 3 billion haploid base pairs in the human genome codes for proteins. The remaining 98.5%, though, is far from inert: aiming to catalogue non-coding genomic regions and create an Encyclopedia of DNA Elements, the ENCODE project identified over 80% of the genome as biochemically active [21]. Many of these are important regulatory regions, such as promoters, enhancers, or regions of structured chromatin and histone modifications.

¹About one dollar per base pair!

²Arguably, there is no “human genome,” as every genome contains variation in the form of single-nucleotide differences, copy-number differences in repetitive regions, small inserts and deletions (indels), and larger scale structural variation. Now, there is a push towards the use of “pangenome graphs” rather than a single, linear reference genome to better capture a range of genetic diversity [10].

This showed that the vast majority of our DNA contains *regulatory information* that carefully controls the directed flow of information from DNA to “transcribed functional units” of RNA. The regulatory information governs the production (transcription) of nascent RNA molecules, as well as how they are altered (e.g., removing introns by splicing to leave exons comprising mature RNA molecules) before being exported from the nucleus and translated into protein. This process, how genes are turned on and off or modulated to produce dynamic distributions of RNA in different contexts and in response to stimuli, is called **gene expression regulation**.

To fully understand how the genome’s information content produces biological function, then, it is necessary to ascertain not only the precise sequence of DNA but the levels of intermediates between genome and protein: RNA transcripts. Genome-wide measurements of RNA expression were first made using microarrays, which contained thousands of probes pre-designed to hybridize to known transcript sequences [22]. In the 2000s, the development of high-throughput, massively parallel sequencing methods gave rise to a more unbiased quantification of gene products, RNA-seq [23]. Most RNA-seq technologies use poly(T) probes to capture polyadenylated tails of messenger RNA (mRNA), with initial applications to bulk tissue samples (bulk RNA-seq). The probes also contain priming sites for reverse transcriptase and DNA polymerase, which convert RNA to cDNA libraries via reverse transcription and subsequent amplification, and indices that bind to sequencing machine flow cells. Sequencing machines then perform many rounds of complementary strand synthesis with fluorescently labeled nucleotides to “read” cDNA sequences corresponding to the original RNA. These reads are then aligned to a reference genome, and the number that overlap with annotated genes are used to quantify the gene expression profile of input samples. With the proper annotation, reads can also be assigned at the sub-gene level: for example, as spliced/mature or unspliced/nascent if aligning to exonic or intronic regions of a gene, respectively [24, 25].

In the 2010s, these RNA capture technologies were combined with microfluidic and well-based approaches that separate individual cells [26, 27], and sequencing probes were designed with cell-specific barcodes and unique molecular identifiers (UMIs), to allow the quantification of unique RNA molecules across the full spectrum of protein coding genes in *individual cells*. This technology, single-cell RNA-seq (scRNA-seq), unlocked extensive observations of the wide diversity of gene expression profiles across cells: these differences are used to distinguish and classify cell

types (more on this in Chapter II). It also paved the way for the creation of comprehensive *cell atlases*: from identifying all the genes and functional elements in the human genome came the effort to classify all cell types and states in the human body [28].

Returning to the goal of GWAS to determine the association between SNPs and high-level phenotypes like disease, there was introduced the intermediate step of linking SNPs to gene expression (RNA levels) in particular biological settings, and then determining post hoc if the gene's protein product had any role related to the SNP-linked disease or trait. These variant to expression associations (expression Quantitative Trait Loci, or eQTL, studies) have been performed across tissues in the human body [29, 16], and, after the rise of scRNA-seq, even specific cell types [30]. These studies provide maps between genetic variants, the gene encoding the functional protein, the biological context, and particular diseases or traits. However, several issues remain.

Firstly, there is a combinatorics problem for identifying significant variant-gene associations. Testing for significant associations between hundreds of millions of SNPs in the human genome ($f \times 10^8$) [31] and tens of thousands of genes ($f \times 10^3$) in hundreds of cell types and contexts ($f \times 10^2$) is computationally and statistically burdensome ($f \times 10^{14}$ tests!). To reduce the number of tests, variant testing is split into two digestible conceptual categories: *cis*, or variants that are restricted to be within some distance (usually ± 1 Mb) of the gene; and *trans*, or any variants outside of this window. This reduces the number of *cis* tests from tens to thousands per gene, as opposed to tens of millions. *Trans* variant identification has not escaped the combinatoric encumbrance. To give a sense of the difficulty in identifying *trans* eQTLs, in a final release of the Gene-Expression-Tissue (GTEx) project that tested variant-gene associations across 54 human tissues, only 143 genes with *trans* regulation were identified, as opposed to 23,268 *genes* with *cis* eQTLs [29].

Another missing piece is mechanism: eQTL studies associate variants with a change in average gene expression, revealing nothing of the cellular *processes* that the variant could be affecting. While adaptations of eQTL studies for testing links between SNPs and ratios of spliced to unspliced RNA counts for a gene (splicing QTL, or sQTL, studies) give oblique insight into mechanism [32], the ratio of spliced to unspliced counts is a mere proxy for the actual process of splicing.

ScRNA-seq data could hold the key to linking SNP to disrupted cellular process, as they give access to mechanistic information. In parallel to the genomic revolution,

small scale investigations of transcription in individual cells using fluorescently labeled RNA molecules reported key parameters governing the process, such rates of transcriptional initiation and RNA decay. These were inferred by fitting particular biophysical models to the observed distribution of RNA counts, grounded in the observation of the systems’ stochasticity [33, 34, 35]. Crucial to these discoveries was the insight that the variance in RNA counts over genetically identical cells in identical conditions was not merely technical noise but contained information about the underlying biological system: the *distribution* is more informative than the mean. Count distributions from scRNA-seq can be similarly modeled; and, with the inclusion of spliced and unspliced information, used to learn about the lifecycle of RNA in different cell types and organisms [36, 37, 38].

Unfortunately, although common for decades in fluorescence studies, using a biophysical model to analyze scRNA-seq data has been slow to gain traction in the genomics field. It is by far outnumbered by heuristic choices and arbitrary data manipulations [39], with the final output being a list of “differentially expressed genes” (DEGs), or genes with different means in the different tested groups [40]. Rather than taking advantage of count variation across single cells, sc-eQTL pipelines attempt to remove it by porting over approaches from bulk eQTL pipelines. A recent review on best-practices for sc-eQTL pipelines tested whether median, mean, or sum of RNA counts over cells of a cell type would be “best,” as defined by finding the greatest number of significant variant-gene associations [41]. While this failure to account for the data generating process is a characteristic of the genomic data analysis philosophy belonging to the “algorithmic modeling” versus “stochastic data” modeling culture [42], it is also a problem of practicality. For all but the simplest biophysical models, there is no closed-form distribution solution [43], and tools to scale computationally costly inference procedures to the datasets required for genetic association studies have not been available.

So, in pursuit of understanding how DNA sequences lead to biological function, and how genetic variants alter this, we are left with challenges:

1. **Statistical and computational difficulties in identifying trans-acting variants:** Before we pinpoint exact *trans* variants, can we determine *to what extent and which genes are regulated in trans*?
2. **Resolution and context dependence:** How does regulation act at the tissue and cell type level, and *how do we make full use of single-cell data to explore*

this?

3. **Mechanism:** By what mechanisms do genetic variants act, i.e. *what is their effect on biophysical processes of the cell?*

These are the overarching/motivating questions I consider in my thesis, asking what we can gain from coupling genetic analyses to single cell biophysics, and how to scale this to modern-day genomic datasets. We develop the theory and tools that are the necessary building blocks for answering these questions, and then apply them to infer biophysical mechanisms and its regulation for tens of thousands of genes across eight tissues in eight founder mouse strains: the largest analysis of single-cell biophysical mechanism of which we are aware.

Before fitting biophysical models to full RNA count distributions, we showcase the utility of moving beyond comparison of its first moment (mean) to its second (variance) [44, 45]. We start with one of the first steps in single-cell analyses pipelines: dimensionality reduction. While most scRNA-seq pipelines use principal component analysis (PCA) to group similar cells into “types,” we demonstrate in Chapter II how it can be adapted to move beyond comparing average expression to identify patterns of variation that are different in a target or case versus a background or control genomic dataset.

We then describe the biophysical modeling formulation we use for scRNA-seq distributions in Chapter III [37]. We provide examples of how this can be used for biological discovery and hypothesis generation by giving access to mechanistic interpretation of biological responses to intervention. In Chapter IV and V, we show how distributions for intractable biophysical models can be efficiently approximated and scaled for inference using machine learning, expanding the practical scope of their applicability [46, 47].

In Chapter VI, with a nod to Mendel and his peas, we present a new mathematical framework for determining the relative contributions of *cis* and *trans* regulation to differences in homozygous parental gene expression using their hybrid crosses [48]. This complements eQTL studies by providing a way to discover the presence and extent of *trans* regulation by induction (while we cannot determine the *trans* acting variant, we can say that it exists), restricting the search space for candidate genes.

Finally, in Chapter VII, we combine the threads of the thesis for an exciting demonstration of coupling genetics to single cell biophysics at scale. We fit biophysical

models for the largest analysis of biophysical parameters of which we know: 15 diverse genotypes, eight tissues, 92 unique cell types, and 30,763 genes. We next apply the framework developed in Chapter VI to determine gene regulatory strategies of biophysical mechanism between the diverse strains across genes, cell types, and tissues.

In Chapter VIII, we think ahead to the future of modeling approaches in light of ever expanding, dizzyingly large genomic datasets and computational revolutions: what kinds of questions may be possible to ask, and what are the role and capacity of the human computational biology researcher?

Chapter 2

CONTRASTIVE DIMENSIONALITY REDUCTION FOR DIFFERENTIAL VARIANCE

One of the first steps in a scRNA-seq workflow is to reduce the dimension of observations from tens of thousands of genes to a lower dimension, with the underlying assumption that cells can be described by modules of correlated genes that are co-expressed. These lower dimension cell representations can then be assigned to discrete groups that display similar gene expression profiles and classified as belonging to a certain cell type by the expression of known cell type marker genes [49]. Perhaps the most commonly employed method for dimensional reduction is principal component analysis (PCA), first proposed in the early twentieth century [50, 51], which finds linear combinations of features (e.g., genes) that best preserve the variance of the observations in the lower dimension. This takes advantage of the variance inherent in single-cell observations by preserving it: features highly weighted in PCA are those that display variability across cells.

However, *contrasting* the variability of observations, and the features driving it, in settings for which target and background datasets are available is a powerful way to move beyond the mean and suggest features as having differential variance between conditions. This chapter summarizes work on developing the theory and methods for contrastive principal component analysis of genomics data via the Rayleigh quotient and associated generalized eigenproblems, useful for single-cell genomics but also for data science more generally.

2.1 The Rayleigh Quotient for Genomics

This section summarizes content from [44] by M.C.*, K.J.* and L.P.*, where * denotes co-first authorship. M.C., K.J., and L.P. developed the methods, produced the results and drafted the manuscript.

Contrastive learning methods can be powerful tools for genomics, enabling the identification of signals in an experiment via dimension reduction while reducing noise using a control. One such approach is contrastive PCA (cPCA) [52] and its extensions [53, 54], which attempt to maximize variance of target samples while minimizing variance of background samples. However, the cPCA objective

involves a subtraction of target and background covariance matrices, which leads to non-physical negative variances. A more natural objective for contrastive principal component analysis that retains positive variance relationships involves the ratio of covariance matrices formulated via a Rayleigh quotient, analogous in form to PCA itself and many existing dimensional objectives.

Principal Component Analysis and Other Rayleigh Quotient Instantiations

Principal Component Analysis (PCA) refers to the identification of a sequence of subspaces of a Euclidean vector space that provide optimal low-dimensional linear approximations to a high-dimensional dataset [50, 51]. Formally, given centered data vectors $x_1, \dots, x_n \in \mathbb{R}^d$ with sample covariance matrix $\hat{\Sigma}$, PCA seeks unit vectors $v_1, \dots, v_k$ that maximize

$$\mathcal{R}_{\text{PCA}}(v) = \frac{v^\top \hat{\Sigma} v}{v^\top I v}, \quad (2.1)$$

The vectors $\{v_i\}_{i=1}^k$ that maximize $\mathcal{R}_{\text{PCA}}(v)$ are the eigenvectors of $\hat{\Sigma}$, and the corresponding eigenvalues represent the variances explained along each direction. The k -dimensional subspace spanned by the first k eigenvectors maximizes the total projected variance and equivalently minimizes the mean-squared reconstruction error among all k -dimensional linear subspaces.

The quantity $\mathcal{R}_{\text{PCA}}(v)$ is a special case of the Rayleigh quotient, which, for two square matrices A and B , is defined as

$$\mathcal{R}(v) = \frac{v^\top A v}{v^\top B v}. \quad (2.2)$$

When both A and B are symmetric and B is positive definite, the maxima of (2.2) satisfy the generalized eigenvalue problem $Av = \lambda Bv$ [55] (see the subsection on solving ρPCA below).

Many other dimension-reduction and signal-extraction problems can be formulated as a Rayleigh quotient maximization. For example, Fisher linear discriminant analysis (LDA; [56, 57]) maximizes the Rayleigh quotient

$$\mathcal{R}_{\text{LDA}}(v) = \frac{v^\top S_B v}{v^\top S_W v}, \quad (2.3)$$

where S_B is a between-class scatter matrix and S_W is a within-class scatter matrix. Maximizing $\mathcal{R}_{\text{LDA}}(v)$ yields directions that best separate classes by maximizing inter-class variance while minimizing intra-class variance.

Recently, contrastive PCA (cPCA; [52]) proposed to identify patterns present in one dataset but not another. The method seeks components that exhibit high variance in a target dataset while exhibiting low variance in a background dataset. The original publication defines cPCA as the eigendecomposition of a contrastive "covariance" matrix,

$$C_\alpha = \hat{\Sigma}_T - \alpha \hat{\Sigma}_B, \quad (2.4)$$

where $\hat{\Sigma}_T$ and $\hat{\Sigma}_B$ are the sample target and background covariance matrices, respectively, and $\alpha > 0$ is a tunable contrast parameter. Note that while the sum of covariance matrices is always a covariance matrix, the difference of covariance matrices may not be positive semidefinite; C_α can have negative eigenvalues, thus suggesting non-physical variances.

We propose rather maximizing the ratio of the target to background projection variances by solving the Rayleigh quotient problem

$$\mathcal{R}_{\rho\text{PCA}}(v) = \frac{v^\top \hat{\Sigma}_T v}{v^\top \hat{\Sigma}_B v}. \quad (2.5)$$

This yields non-negative variance ratios without requiring a tuning parameter α . The Rayleigh quotient is also scale invariant (scaling either covariance matrix by a positive constant rescales all eigenvalues but leaves the eigenvectors, and therefore the contrastive subspace, unchanged, i.e., for $a, b > 0$, $\mathcal{R}_{\rho\text{PCA}}(v; a\hat{\Sigma}_T, b\hat{\Sigma}_B) = \frac{a}{b} \mathcal{R}_{\rho\text{PCA}}(v)$); and monotonic (if $\hat{\Sigma}_T^{(1)} \succeq \hat{\Sigma}_T^{(2)}$ or $\hat{\Sigma}_B^{(1)} \preceq \hat{\Sigma}_B^{(2)}$, then $\mathcal{R}_{\rho\text{PCA}}^{(1)}(v) \geq \mathcal{R}_{\rho\text{PCA}}^{(2)}(v)$). This means that increasing the target variance or decreasing the background variance never decreases the Rayleigh quotient. Furthermore, the cPCA objective (2.4) can be derived as a Taylor approximation (under a strong assumption) of the ρPCA objective [44].

We thus propose the Rayleigh quotient objective as the correct method for contrastive principal component analysis. The Rayleigh quotient (2.2) is frequently referred to as ρ ; we therefore adopt the term ρPCA to highlight the fact that ρPCA extends PCA via a Rayleigh quotient where the matrix in the denominator is not necessarily the identity. Furthermore, the term ρPCA rather than just ρ is intended to convey that the method is more than just the solution of the quotient (computing generalized eigenvalues and eigenvectors of covariance matrices derived from data); ρPCA includes the projections of the data onto the subspaces spanned by the generalized eigenvectors. Below we provide details on how to solve the ρPCA objective.

Solving ρ PCA

The ρ PCA objective (2.5) seeks to find

$$\operatorname{argmax}_v \mathcal{R}_{\rho\text{PCA}}(v) = \operatorname{argmax}_v \frac{v^\top \hat{\Sigma}_T v}{v^\top \hat{\Sigma}_B v}, \quad (2.6)$$

where $\hat{\Sigma}_T$ and $\hat{\Sigma}_B$ denote the empirical target and background covariances, respectively, and $\hat{\Sigma}_B$ is assumed to be positive definite (if empirical target covariance matrices are not positive definite, regularization such as Tikhonov shrinkage can be applied [44]). That is, a direction v in which the target variance is maximized while the background variance is minimized. Because the quotient is homogeneous in v , scaling v by any nonzero constant does not change the objective. Therefore, we can impose a normalization constraint, leading to the equivalent constrained optimization problem

$$\max_{v \in \mathbb{R}^D} v^\top \hat{\Sigma}_T v \quad \text{subject to} \quad v^\top \hat{\Sigma}_B v = 1. \quad (2.7)$$

The constraint specifies unit variance in the background so that the maximization measures relative rather than absolute variance. To solve this constrained problem, we introduce a Lagrange multiplier λ and define the Lagrangian

$$\mathcal{L}(v, \lambda) = v^\top \hat{\Sigma}_T v - \lambda(v^\top \hat{\Sigma}_B v - 1). \quad (2.8)$$

Taking the gradient with respect to v and setting it equal to zero yields

$$\begin{aligned} \nabla_v \mathcal{L} &= 2\hat{\Sigma}_T v - 2\lambda\hat{\Sigma}_B v = 0, \\ \Rightarrow \hat{\Sigma}_T v &= \lambda\hat{\Sigma}_B v. \end{aligned} \quad (2.9)$$

Equation (2.9) is a generalized eigenvalue problem, and its solutions (v, λ) satisfy the stationarity condition of the Rayleigh quotient. The eigenvectors v corresponding to the largest generalized eigenvalues λ define the directions along which the target variance is maximized relative to the background variance. Each eigenvalue λ_i represents the ratio of target-to-background variance along its associated direction v_i .

This extends naturally to multiple dimensions. Let $V = [v_1, \dots, v_d] \in \mathbb{R}^{D \times d}$ denote the matrix whose columns are the top d generalized eigenvectors of $(\hat{\Sigma}_T, \hat{\Sigma}_B)$, and let $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ be the diagonal matrix of the associated eigenvalues. The multidimensional generalization of the Rayleigh quotient replaces the scalar ratio by a trace ratio, yielding the objective

$$\max_{V \in \mathbb{R}^{D \times d}} \operatorname{tr}((V^\top \hat{\Sigma}_B V)^{-1} V^\top \hat{\Sigma}_T V), \quad (2.10)$$

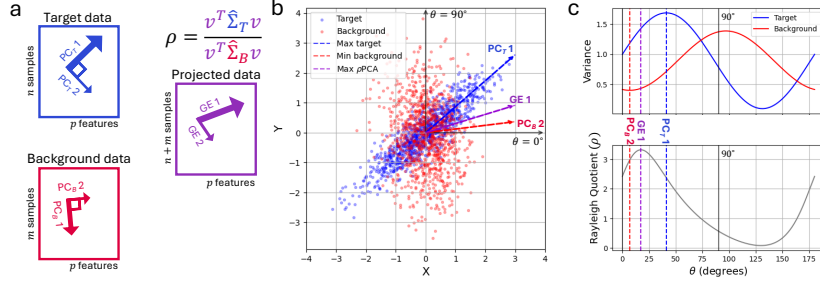


Figure 2.1: ρ PCA finds directions that maximize target variance while minimizing background variance. **a**, ρ PCA is useful in a setting in which there are target datasets of interest with axes of interesting variation and background datasets with different axes of variation that should be removed before analysis. In the following example, the sample size $n = m = 1,000$ and the feature size $p=2$. **b**, ρ PCA (black dashed line) finds the direction that maximizes the ratio of the target variance (direction of maximum target variance, dashed blue line) while minimizing the background variance (direction of minimum background variance, red dashed line). **c**, Variance of target (blue) and background (red) data projected on lines that pass through the origin from 0° to 180° , with the ratio of target to background variance (Rayleigh Quotient) shown below. Dashed lines correspond to the directions from **b**.

which is equivalent to maximizing $\text{tr}(V^\top \hat{\Sigma}_T V)$ subject to the constraint of background orthonormality $V^\top \hat{\Sigma}_B V = I_d$. Introducing a symmetric matrix of Lagrange multipliers Λ , the stationarity condition of the Lagrangian

$$\mathcal{L}(V, \Lambda) = \text{tr}(V^\top \hat{\Sigma}_T V) - \text{tr}(\Lambda(V^\top \hat{\Sigma}_B V - I_d)) \quad (2.11)$$

with respect to V gives

$$\nabla_V \mathcal{L} = 2\hat{\Sigma}_T V - 2\hat{\Sigma}_B V \Lambda = 0, \quad (2.12)$$

or equivalently,

$$\hat{\Sigma}_T V = \hat{\Sigma}_B V \Lambda. \quad (2.13)$$

This is the matrix form of the generalized eigenvalue equation: each column of V is an eigenvector of $(\hat{\Sigma}_T, \hat{\Sigma}_B)$ and each diagonal entry of Λ its corresponding eigenvalue. The optimal V consists of the d generalized eigenvectors associated with the largest eigenvalues, which together span the ρ PCA subspace. This subspace maximizes the total target variance relative to background variance, and the constraint $V^\top \hat{\Sigma}_B V = I_d$ ensures that the learned directions remain mutually orthogonal under the background inner product, $v_i^\top \hat{\Sigma}_B v_j = 0$ for $i \neq j$.

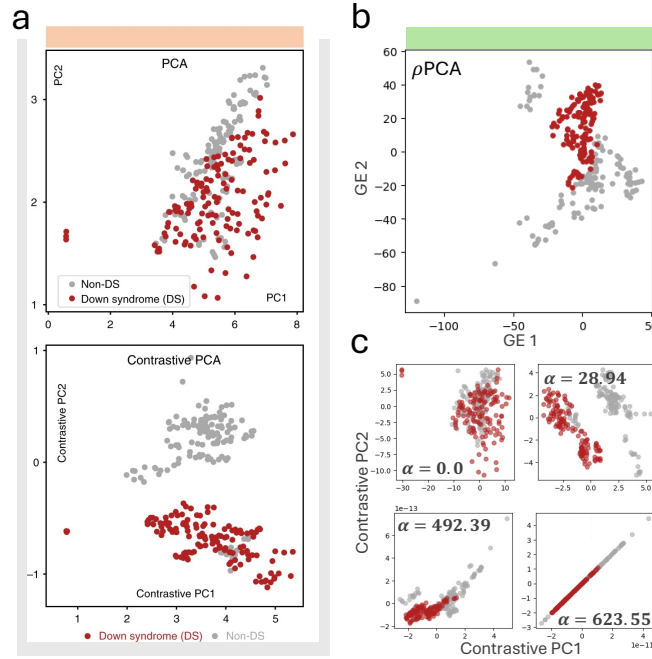


Figure 2.2: ρ PCA recovers structure in biological subgroups without requiring parameter tuning or prohibitively long runtimes. **a**, Original Fig. 3a from [52] showing PCA and cPCA projections of protein expression from mice exposed to shock therapy, some of which developed Down Syndrome (DS) and some of which did not (non-DS). **b**, ρ PCA recovers separation of the two subgroups without searching over the contrastive parameter α . **c**, cPCA by default outputs projections at four values of α , with no guidance on how to select the best value.

Comparing ρ PCA and cPCA

Given the heuristics and approximations underlying cPCA, we hypothesized that ρ PCA would be more effective at finding a projection that maximizes variance in a target dataset while minimizing variance in a background dataset.

To directly compare cPCA to ρ PCA on biological data, we applied ρ PCA to a dataset consisting of protein expression measurements from mice exposed to shock therapy [52], some of which developed Down Syndrome (DS) and some of which did not (non-DS). As reported in [52], performing PCA on the shock-treated mice data does not reveal distinct protein expression between those that developed Down Syndrome and those that did not. Introducing as a background protein expression measurements from non-DS mice that had not received shock therapy, ρ PCA finds an effective projection without any parameter tuning (Fig 2.2). cPCA projections reported in [52] as well as those obtained when running cPCA with default parameters produce different projections at different values of α , without clear guidance

given on how to choose the best value. Moreover, in an improvement with respect to cPCA, the ρ PCA projection shows different protein expression patterns between the DS and non-DS mice (Fig 2.2).

ρ PCA for Biological Discovery

To demonstrate the interpretability and scalability of ρ PCA, we performed an analysis of a large single-nucleus RNA-seq dataset profiling kidney in mice [58]. We applied ρ PCA to a target dataset of four female mice with a "background" of four male mice in order to identify cell type specific variation in female gene expression that is distinct from the variation in males. After filtering out low quality nuclei and selecting highly expressed and highly variable genes dataset, we analyzed 2,000 genes from 47,798 female nuclei and 42,142 male nuclei of 17 different cell types. While a typical analysis might focus on genes that are differentially expressed (mean or rank differences) between male and female cell types, ρ PCA provides a new interpretation of sexually dimorphic gene expression as genes that are *variable* primarily in the target group (female) and not in the background group (male).

The coefficients (loadings) of ρ PCA generalized eigenvectors are analogous to PCA loadings, but in this case highlight genes displaying variation among females that is absent from males (Fig. 2.3). These genes are not identified in a standard application of PCA: unsurprisingly, the genes with the highest loadings on GE 1 and GE 2 are not the genes with the highest loadings on PC 1 and PC 2 (Fig. 2.3a).

The variance of projected samples onto each generalized eigenvector also yields insights into the particular cell type driving the main axes of female specific variation. We projected male and female nuclei onto the first and second GEs and calculated the variance per cell type and sex (for all cell types, see Supplementary Fig. 2 in [44]). We then took the ratio of female to male variance, which is the Rayleigh quotient ρ of each GE. We find that GE 1 has the highest female to male Rayleigh quotient in proximal tubule (PT) epithelial cells, while GE 2 has the highest ratio in adipocytes (Fig. 2.3b,c, with absolute variances shown in Supp. Fig. 3 of [44]). As the first generalized eigenvector is the axis that explains the most variation in the target after having accounted for variation in the background, this suggests that of the tested cell types, females show the most unique variability in PT epithelial cells. This is consistent with previous studies: the PT is known to show strong sexually dimorphic gene expression patterns, reported to be regulated by hormone receptors (androgen and estrogen) [59, 60].

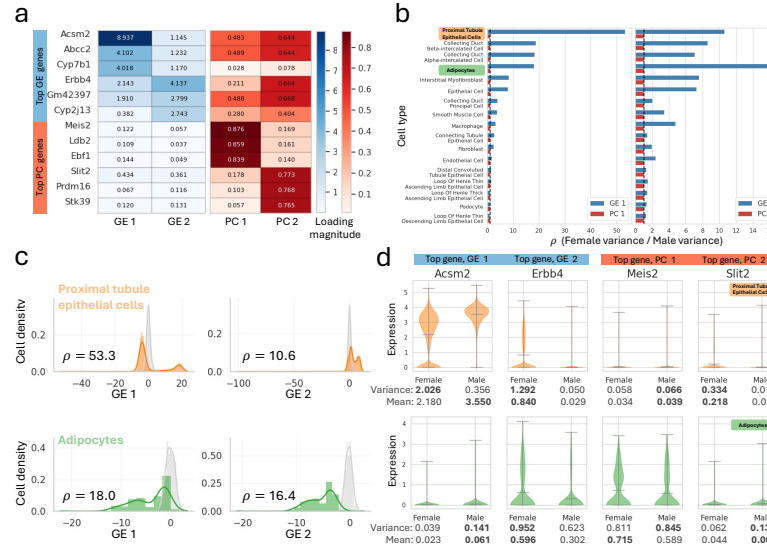


Figure 2.3: ρ PCA analysis of snRNA-seq data from male and female mice reveals genes driving female-specific variation across cell types that PCA fails to identify. **a**, Loadings for the three genes with the greatest magnitude loadings on the first and second generalized eigenvectors (GE 1 and GE 2 respectively) and the first and second principal components (PC 1 and PC 2 respectively): ρ PCA highly weights different genes than PCA. **b**, The ratios of variance of female to male nuclei projected onto the first and second GEs for 16 cell types in the kidney. GE 1 separates female proximal tubule epithelial cells, while GE 2 has the highest variance ratio for adipocytes. **c**, Histogram of male and female proximal tubule epithelial cells (orange) and adipocytes (green) projected onto the first and second generalized eigenvectors. **d**, *Acsm2* (gene with the highest magnitude on GE 1), *Erb4* (gene with the highest magnitude on GE 2), *Meis2* (gene with the highest magnitude on PC 1), and *Slit2* (gene with the highest magnitude on PC 2) expression in male and female nuclei, separated by cell type (proximal tubule epithelial cells, top row, and adipocytes, bottom row).

As expected, the male and female specific expression patterns for the genes driving GE 1 and GE 2 show strong differences in variance. The counts distributions for *Acsm2* (top gene for GE 1) and *Erb4* (top gene for GE 2) are indeed more variable in females than in males in PT epithelial cells and adipocytes, respectively (Fig. 2.3d), while the top genes for PC 1 (*Meis2*) and PC 2 (*Slit2*) do not show clear differences in variation. It is interesting that while *Acsm2* is more variable in females than in males in the PT, it has a lower mean expression in females than males. While this gene has been previously noted to have male-biased expression in PT cells [61], ρ PCA identifies that it displays female specificity in the second moment. This variance across cells in females is interesting in light of [62], which suggests that *Acsm2* is controlled by androgen in PT cells; each cell may be responding to differences in hormone concentration across the PT. While *Erb4*, the gene that codes for the

receptor protein-tyrosine kinase with roles in cell signaling and growth, is not as well studied, another growth-regulating receptor of the *ErbB* family, *ErbB1*, is known to have sex-biased expression in the kidney and to be sensitive to changes in sex hormones [63]. These examples illustrate the potential of ρ PCA as a computational approach for identifying genes whose cell-to-cell variation may have functional significance [64].

Discussion

Indeed, it is no accident that the ρ PCA objective has been used for dimension reduction since the introduction of LDA in 1936 [56], and has been rediscovered in various fields. For example, the extraction of common spatial patterns method (CSP) of [65] at its core involves maximizing a Rayleigh quotient, where the target and background covariance matrices are derived from two distinct signal conditions [66]. The resulting generalized eigenvectors define spatial filters that maximize variance in one condition while minimizing variance in the other. CSP has become one of the most popular algorithms for brain computer interface design [67]. One reason that maximizing the ρ PCA objective works well in practice is that ρ PCA can be viewed as a restricted Gaussian maximum-likelihood estimator analogous to classical LDA [68, 69]. In this view, target and background data arise from Gaussians whose covariances differ by a structured latent component, and the ρ PCA solution is the large-sample maximum-likelihood solution. This probabilistic interpretation differs fundamentally from the contrastive latent variable model of Severson, Ghosh, and Ng [70], in which both target and background distributions are generated by shared and condition-specific latent variables and the full joint likelihood is maximized. The models of [68] and [69] provide asymptotic insight into ρ PCA, in contrast to the finite-sample likelihood maximization of probabilistic PCA [71]. Moreover, the expectation-maximization (EM) algorithm of [69] offers an analog of the EM algorithm of [72] for probabilistic PCA.

The ρ PCA method is also powerful due to its efficiency and scalability, making it much more suited to genomics applications than cPCA. We found that ρ PCA is much faster than cPCA. For example, in an analysis of a dataset with 10,000 samples and 1,000 features, cPCA takes almost two minutes to run, while ρ PCA runs in only a few seconds on the same machine, a speedup of two orders of magnitude. The difference in runtime between ρ PCA and cPCA is even more dramatic as the number of samples and features increases (Fig. 2.4). Analyzing data at increasingly large scales, such as single-cell RNA sequencing data with hundreds of thousands, or

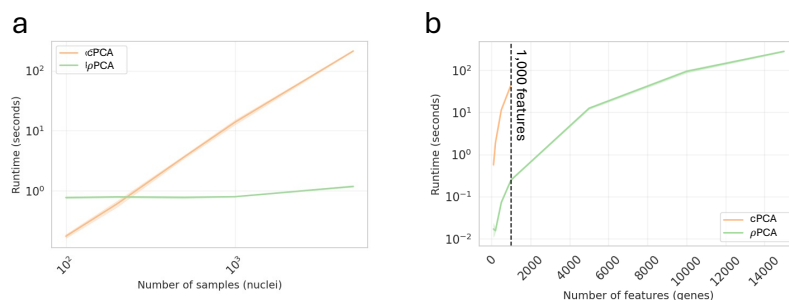


Figure 2.4: We use snRNA-seq data from mouse kidney to illustrate that ρ PCA is orders of magnitude faster than cPCA and remains efficient as **a**, number of samples (with fixed 2,000 features) and **b**, number of features (with fixed 1,000 samples in target and 1,000 samples in background) grow. Note that when samples have more than 1,000 features, cPCA first performs PCA to restrict the number of features to 1,000 (black dashed line).

even millions, of cells and genes, would require prohibitively long run times with cPCA.

In summary, the Rayleigh quotient provides an effective unifying principle for linear contrastive learning. This observation can lead to a consolidation of contrastive methods across fields.

2.2 Solving the Rayleigh Quotient for Functional Data

In the previous section, it was described how contrastive PCA problems can be written as solutions to generalized eigenvalue problems that maximize particular instantiations of the Rayleigh quotient. In this section, we discuss how understanding this framework leads to useful and creative extensions of ρ PCA; i.e., solving the Rayleigh quotient in the space of basis function coefficients (f- ρ PCA) to find modes of variation in functional data. This extension expands the scope of contrastive PCA while unifying disparate fields of contrastive dimensional reduction and functional methods within a single conceptual and mathematical framework. We showcase the utility of this extension with an analysis of gene expression in immune response to vaccination.

This section summarizes content from [45] by M.C.* , K.J.* and L.P.* , where * denotes co-first authorship. M.C., K.J., and L.P. developed the methods, produced the results and drafted the manuscript. M.C. worked on functional ρ PCA, while K.J. worked on spatial ρ PCA.

Functional PCA

Functional PCA (fPCA) is a technique that extends PCA to reduce the dimensionality of data that can be represented as curves or functions over a continuous domain (e.g., time or space) by finding dominant “modes of variation” [73, 74]. In this setting, n independent observations $X_1(t), \dots, X_n(t)$ are considered realizations of a square-integrable stochastic process $X(t)$ over the continuous variable t . Let $\mu(t) = E[X(t)]$ and $c(s, t) = \text{Cov}(X(s), X(t)) = E[(X(s) - \mu(s))(X(t) - \mu(t))]$ be the mean function and covariance function of the process $X(t)$, respectively. The covariance operator $C : L^2(\mathcal{T}) \rightarrow L^2(\mathcal{T})$ acts on square-integrable functions $f(s) \in L^2(\mathcal{T})$ as

$$(Cf)(t) = \int_{\mathcal{T}} c(s, t) f(s) ds.$$

As C is compact, self-adjoint (symmetric, or $c(s, t) = c(t, s)$), and positive semi-definite ($\int_{\mathcal{T}} \int_{\mathcal{T}} f(s) c(s, t) f(t) ds dt \geq 0$), by Mercer’s theorem, the function $c(s, t)$ can be expressed as

$$c(s, t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t),$$

where the ϕ_k are the eigenfunctions of C satisfying $C\phi_k = \lambda_k\phi_k$ and $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$. The orthonormality of eigenfunctions requires $\int_{\mathcal{T}} \phi_k(t)\phi_l(t)dt = 0$ for $k \neq l$ and $\int_{\mathcal{T}} \phi_k(t)\phi_l(t)dt = 1$ for $k = l$.

fPCA seeks to find orthogonal functions that capture the most variation in $X(t)$, or, analogously to PCA, maximize the variance of the integral inner product (the process projected onto the new function):

$$\begin{aligned} \int \int f(s)(X(s) - \mu(s))(X(t) - \mu(t))f(t) ds dt \\ = \int \int f(s)c(s, t)f(t) ds dt \\ = \langle Cf, f \rangle. \end{aligned}$$

Recalling the eigendecomposition of C , the functions that capture the most variation are the eigenfunctions $\phi_k(t)$, which form a basis for L^2 such that each process $X_i(t)$ can be written as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \theta_{ik} \phi_k(t).$$

Each observation i has weights θ_{ik} associated with each eigenfunction k that capture how strongly the eigenfunction contributes to the observation.

Practically, fPCA is implemented in one of two ways. In the first approach, observations are sampled on a grid of points over the continuous variable, and traditional PCA is employed on the resulting matrices. The eigenvectors can be interpolated between discrete features to produce eigenfunctions over a continuous feature space [74]. In the second approach, samples are first represented as combinations of a finite set of basis functions (e.g., B-splines, monomials, or Fourier coefficients). PCA is then performed on the coefficients of the bases (taking into account the non-orthogonality of basis functions if necessary) to find eigenvector coefficients for the set of basis functions [74, 75].

fPCA has been adapted to various settings. For example, it has been used to fit observations sampled at irregular or sparse time-points by smoothing covariance and mean functions [76], combined with mixed-effect models [77], and applied to adjust for known covariates [78]. A recent attempt to implement a contrastive version with

target and background functional data has been proposed that mimics the contrastive PCA method of [52] by subtracting from the target covariance function a weighted version of the background covariance function [79]. This approach to contrastive fPCA inherits the same problems as the [52] contrastive PCA method [44], including the issue that a difference of positive semi-definite matrices may not be positive semi-definite, and requiring a tunable contrastive parameter that can produce arbitrary eigenfunctions [45]. We show that contrastive functional dimension reduction can be achieved more naturally by performing ρ PCA in the space of basis coefficients, as shown in Figure 2.5A, leading to meaningful modes of variation present in target curves and absent in background curves.

Solving f- ρ PCA

The f- ρ PCA method solves the Rayleigh Quotient generalized eigenproblem for functional data. We consider n independent observations $X_i(t), \dots, X_n(t)$ of some square-integrable stochastic process $X(t)$, which we are interested in as a target process, and m independent observations $Y_i(t), \dots, Y_m(t)$ of a different square-integrable stochastic process $Y(t)$, which we treat as a background process.

Let $\mu_Z(t) = E[Z(t)]$ and $c_Z(s, t) = \text{Cov}(Z(s), Z(t)) = E[(Z(s) - \mu_Z(s))(Z(t) - \mu_Z(t))]$ be the mean and covariance functions of the processes $Z(t)$, where $Z \in \{X, Y\}$ indicates target or background, respectively. The covariance functions are kernels for target and background covariance operators: $(C_Z f)(t) = \int_{\mathcal{T}} c_Z(s, t) f(s) ds$. We seek to find functions $\phi(t)$ that maximize the Rayleigh quotient:

$$\operatorname{argmax}_{\phi} \frac{\langle C_X \phi, \phi \rangle}{\langle C_Y \phi, \phi \rangle},$$

requiring that C_Y is positive definite on its domain, or $\langle C_Y f, f \rangle > 0$ for non-zero functions f (e.g., covariance operators for non-degenerate Gaussian processes [80]).

Discrete domain: In practice, when functional observations have been measured at p discrete points, or $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{m \times p}$, target and background sample covariance matrices $\hat{\Sigma}_X$ and $\hat{\Sigma}_Y$ can be calculated and the discrete ρ PCA objective solved [44]:

$$\operatorname{argmax}_{\mathbf{v}} \frac{\mathbf{v}^T \hat{\Sigma}_X \mathbf{v}}{\mathbf{v}^T \hat{\Sigma}_Y \mathbf{v}}, \quad (2.14)$$

for discrete eigenvectors $v \in \mathbb{R}^p$. These eigenvectors can be interpolated between discrete features for continuous support [74].

Basis representation: Alternatively, and commonly when data is irregularly or sparsely sampled, the Rayleigh quotient can be solved using a basis function representation for samples, as in functional PCA [74]. Full details are included in the associated publication’s Supplemental Methods [45]. Briefly, $B(t) = (b_1(t), \dots, b_D(t))^T$, a set of linearly independent basis functions, are fit to each sample:

$$X_i(t) = \sum_{d=1}^D a_{id} b_d(t) = \mathbf{a}_i^T B(t).$$

All n observations can be written in matrix form, $X(t) = A_X B(t)$, where the matrix $A_X \in \mathbb{R}^{n \times D}$ contains n rows, one per observation, of coefficients for the D basis functions. The background can be similarly fit and represented as $Y(t) = A_Y B(t)$. The Rayleigh quotient is then optimized to find eigenvectors $\mathbf{w}$ in coefficient space, using the bases’ Gram matrix $G \in \mathbb{R}^{D \times D}$ with entries $G_{kl} = \int_{\mathcal{T}} b_k(t) b_l(t) dt$ to account for possible non-orthogonality of functions:

$$\operatorname{argmax}_{\mathbf{w}} \frac{\mathbf{w}^T G^{1/2} A_X^T A_X G^{1/2} \mathbf{w}}{\mathbf{w}^T G^{1/2} A_Y^T A_Y G^{1/2} \mathbf{w}}.$$

The resulting coefficient vectors $\mathbf{w}$ are then converted back to the original data space to find eigenfunctions $v(t) = \mathbf{w}^T G^{1/2} B(t)$.

f- ρ PCA Application to Longitudinal Bulk RNA-seq

Measurement of blood transcriptomes is an effective way to profile immunological responses before and after vaccination, with the goal of guiding the design of vaccination protocols for best conferred protection and reduction of adverse events [82]. To demonstrate the ability of f- ρ PCA (Fig. 2.5A) to extract functional patterns from longitudinal data, we applied it to bulk RNA-seq samples collected over a two-week time course before and after a first and second dose of COVID-19 mRNA vaccines [81]. In the study, blood from 23 subjects was sampled immediately before, for 9 days after and 14 days after vaccination (a total of 11 time points) for both an initial “primer” and a secondary “booster” vaccination (Fig. 2.5B). The original study performed gene set enrichment analyses on the first and second doses

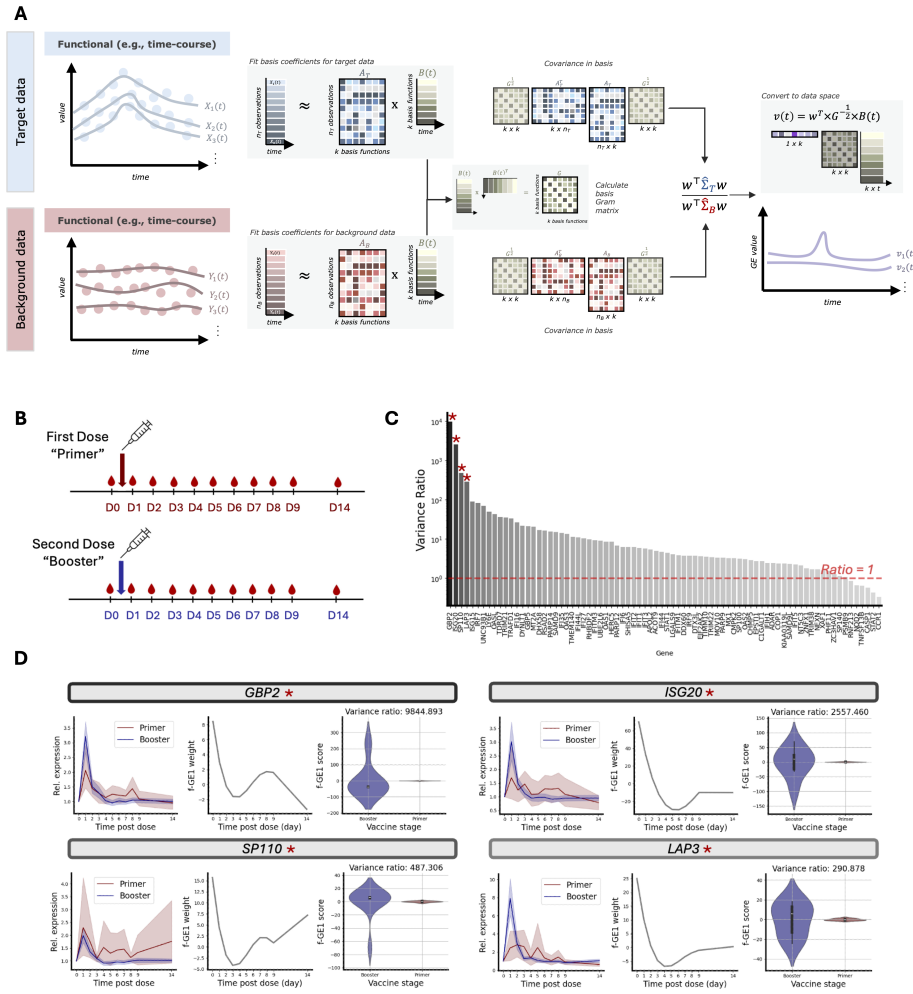


Figure 2.5: **A.** Diagram of functional ρ PCA (f- ρ PCA) using a basis representation. Target and background samples are measured at discrete time points, then fit to a set of basis functions to approximate the underlying continuous process. Functional ρ PCA calculates the covariance of the fit coefficients, accounting for non-orthogonality of basis functions using the Gram matrix of basis functions, then finds solutions to the Rayleigh quotient. Generalized eigenfunctions are recovered by converting the obtained generalized eigenfunctions back to original data space (e.g., time). **B.** We applied f- ρ PCA to longitudinal bulk RNA-seq conducted on patients before and after a first ("primer") and second ("booster") dose of a COVID-19 mRNA vaccine (diagram adapted from [81]). **C.** Ratio of the variance of target ("booster") samples projected onto the first found eigenfunction to background ("primer") samples projected onto the first eigenfunction for 86 interferon genes. **D.** For four genes with the largest target to background variance ratio, we show the gene expression values of primer and booster samples (mean line and shaded standard deviation over the samples indicated), the values over time of the first eigenfunction (f-GE1 weight), and violin plots of the distributions of target and background samples projected onto the first eigenfunction (f-GE1 score). Target samples' projections display higher variance than background samples.

separately. They subsequently compared the number and function of significant gene modules to understand how primer and booster responses differed. We applied f - ρ PCA to contrast the two time courses jointly, using the first dose as “background” and the second dose as “target,” and identified key genes that differ between the doses.

[81] report that a set of interferon genes (referred to as module A28 [83, 81]) have the most significant response in the first three days after primer administration. Focusing on this set, we obtained fit basis functions to the normalized expression profile for each gene. We then performed f - ρ PCA on the gene profiles that passed quality control for a total of 86 genes with contrastive analyses (Fig. 2.5C). The first eigenfunctions have high values at the points in time most variable in the booster that are not variable in the primer (Fig. 2.5D); for the analyzed interferon genes, there is a peak at the initial administration, consistent with the original paper’s observation that after the second dose “the interferon response was noticeably sharper in comparison to the response observed following the first dose and peaked on day 1 instead of day 2” [81]. We also projected fit samples onto the discovered first eigenfunctions for these genes and calculated the ratio of variance of booster samples to primer samples, finding that genes display a range of target and background sample separability (Fig. 2.5C).

The four genes with the highest booster to primer variance ratio, *GBP2*, *ISG20*, *SP110*, and *LAP3*, have previously been associated with SARS-CoV-2 (for expression profiles over all fit samples, see Supp. Fig. S4 of [45]). *GBP2*, which codes for the GTPase guanylate-binding protein 2, inhibits the cleavage of the SARS-CoV-2 spike protein, thus affecting membrane fusion and viral entry [84]. Interferon-stimulated gene 20 (coded for by *ISG20*) is an antiviral RNA exonuclease that has degraded RNA vectors derived from SARS-CoV-2 replicons [85]. *SP110* was identified as significantly associated with extreme COVID-19 phenotypes [86]; and *LAP3*, shown to have increased expression during COVID-19 infection [87], also exhibits poly(A)-tail elongation in patients with COVID-19 compared to controls [88]. These examples demonstrate that f - ρ PCA can be used to contrast two time courses of gene expression immune response in a single analysis, a complementary approach to performing post hoc comparisons of two separate analyses.

Discussion

We have shown that $f\text{-}\rho\text{PCA}$ identifies modes of variation that distinguish two groups of functional observations. Crucially, $f\text{-}\rho\text{PCA}$ can operate in the space of basis function coefficients in addition to on discrete measurements, producing eigenfunctions that are directly interpretable as temporal modes of variation. The result is that $f\text{-}\rho\text{PCA}$ is well-suited for experimental designs involving paired and sequential conditions, such as dose comparisons or longitudinal case-control studies. In cases where the goal is to identify functional differences in the response between groups, $f\text{-}\rho\text{PCA}$ unifies the comparative analysis rather than requiring pairwise post-hoc tests. Moreover, $f\text{-}\rho\text{PCA}$ can, in principle, be extended to joint analysis of several features, incorporated as multivariate functional observations rather than treated independently. However, such an extension requires a framework for performing tensor SVD that retains the spectral properties of its lower dimensional counterparts. This remains an area of active research [89].

The Rayleigh quotient formulation of ρPCA extends this approach by providing a mechanism for performing contrastive analysis and by unifying these problems analytically, showing that they can be interpreted as the same generalized eigenvalue problem. Finally, the examples presented in this work demonstrate that both methods can recover biologically relevant signals from complex genomic datasets. We anticipate that this practical, flexible approach will find broad utility across biological domains.

Chapter 3

BIOPHYSICAL MODELS FOR STOCHASTIC TRANSCRIPTION

As discussed in Chapter I, the richness of single-cell data lies in distributions over single cells and the access they give to the cellular processes that generated the counts. In this chapter, we discuss how biophysical models describing stochastic transcriptional systems unlock deeper insight not present at the level of mean count comparison. First, we describe the Chemical Master Equation formulation and how it can be used to appropriately model scRNA-seq count data. Then, we show how such this can identify strategies of cellular recovery from radiation treatment at particular stages of the RNA life-cycle. We finally show how different models can be fit and compared to generate testable hypotheses about the underlying system biophysics.

3.1 Chemical Master Equation Models

Transcription occurs due to the interactions of discrete, low-copy number molecular components, like two copies of a gene on homologous chromosomes and several transcription factors and RNA polymerases jostling in the nucleus. These systems can be modeled as discrete-valued, continuous time Markov chains (CTMCs), with the states defined as having a particular number of RNA molecules. The Chemical Master Equation (CME) formulation is appropriate to represent such a system as it tracks the probability over time of a system being in any given state, $P(x, t)$, given rates of transitions between states:

$$\frac{dP(x, t)}{dt} = AP(x, t), \quad (3.1)$$

where A is a matrix of rate transitions, and $P(x, t)$ is column vector of probabilities over the possible states. If the set of possible states is infinite, $P(x, t)$ is infinite dimensional. As a simple example, if there are two gene states, "on" and "off" corresponding to states $x = 1$ and $x = 0$, respectively, and the gene switches "on" at a rate k_{on} and "off" at a rate k_{off} , the probability of being in the two states can be tracked:

$$\frac{d[P(x=1, t), P(x=0, t)]}{dt} = \begin{bmatrix} -k_{\text{off}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{on}} \end{bmatrix} \begin{bmatrix} P(x=1, t) \\ P(x=0, t) \end{bmatrix}. \quad (3.2)$$

This can be solved at steady-state (or, as $t \rightarrow \infty$) for probabilities over states by setting the time-derivative equal to zero: $0 = AP_{ss}(x, t)$. For the example above, this can be trivially solved to find $P_{ss}(x=1) = \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$ and $P_{ss}(x=0) = \frac{k_{\text{off}}}{k_{\text{on}}+k_{\text{off}}}$.

For scRNA-seq count data, to modeling a single species of RNA from gene g , $\mathcal{N}_g$, we define the states are all discrete integers (including zero) of RNA molecules. The one-species model includes a production step and a degradation step



where γ is a Poisson rate of molecule decay and the dynamics of production are specifically left blank for now. For a model of one-species constitutive transcription (another Poisson process) with rate k model can be represented:



$\mathcal{N}_g$ represents the RNA species itself, while we will use n to represent the associated observation (random variables) in probabilities: $P(n)$. The constitutive transcription CME is written

$$\frac{dP(n, t)}{dt} = kP(n, t) - kP(n-1, t) + \gamma(n+1)P(n+1, t) - \gamma nP(n, t). \quad (3.5)$$

However, as discussed in Chapter I, scRNA-seq data inherently includes *two* (or more) count matrices: of nascent/unspliced¹ molecules that align to introns and mature/spliced molecules that align to exons². Our model schema can be expanded to include a Poissonian splicing step (β) from nascent $\mathcal{N}_g$ to mature $\mathcal{M}_g$ molecules, which are degraded:



¹Throughout this thesis, we use nascent and unspliced synonymously, as well as mature and spliced synonymously, although we note that the definition of a nascent or fully mature RNA molecule is not rigid.

²Technically, we can only assign reads that cross exon-exon boundaries as truly spliced. For the purposes of this thesis, molecules that ambiguously align to exons we assign to spliced counts, though their ambiguity can be modeled [90]

This results in a bivariate steady-state distribution over nascent and mature molecules: $P_{ss}(n, m)$.

If, instead, we assume nascent molecules are produced in bursts with average size b

To solve infinite state-space distributions such as the constitutive model (infinite number of RNA molecules) analytically, probability generating functions (PGFs) can be applied to convert infinite dimensional ODEs in time to finite dimensional PDEs. Generating functions represent infinite sequences of numbers as coefficients of a power series; for probability functions, this is:

$$G(z, t) = \sum_{x=0}^{\infty} z^x P(x, t), \quad (3.7)$$

or for nascent and mature RNA species,

$$G(z_n, z_m, t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} z_n^n z_m^m P(n, m, t). \quad (3.8)$$

For certain, simple systems (like constitutive transcription), these can be solved for exact analytical solutions using PGF rules of derivatives and partial derivatives [43, 91, 37] and comparing to PGFs of known probability distributions. For example, the above single species constitutive transcription model at steady-state results in the PGF $G(z) = e^{\frac{k}{\gamma}(z-1)}$, which is the PGF of the Poisson distribution, the well-known results that constitutive transcription leads to Poisson distributed counts.

It is often helpful to convert the PGF to factorial-cumulant generating function, or log-PGF (F), and introducing a change of variables: $F(u, v, t) = \ln G(1+u, 1+v, t)$ [43].

When the PGF or log-PGF cannot be solved analytically, the PGF can be solved on a large grid around the complex unit circle, then recover the probabilities by performing an inverse Fourier Transform (iFFT):

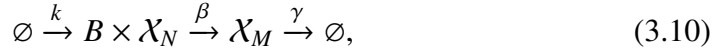
$$p(n, m) = \text{iFFT} G(e^{\frac{-2\pi i n}{N}}, e^{\frac{-2\pi i m}{M}}) \quad (3.9)$$

where N and M are chosen to be sufficiently larger integers (larger than the observed max of nascent and mature molecules) [43, 92].

To facilitate the fit of biophysical models, we built the software package *Monod* [37]: the following outlines the biophysical models supported by *Monod*, as well as how to include technical/sequencing noise. The following subsections are adapted from [37] Supplementary Sections 1.1 and 1.2.

Bursty

The three-parameter bursty model describes the following biology:



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 log-PGF [43]:

$$\begin{aligned} U(u_N, u_M, s) &:= u_N e^{-\beta s} + u_M \frac{\beta}{\beta - \gamma} \left(e^{-\gamma s} - e^{-\beta s} \right) \\ \log G &= \int_0^\infty \left[\frac{1}{1 - bU(u_N, u_M, s)} - 1 \right] ds. \end{aligned} \quad (3.11)$$

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 [33]. 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 high initiation and off rates, the Chemical Master Equation (CME) reduces to the process in Equation 3.10 with $b := k_{init}/k_{off}$ (Section S1.3 of [93]).

Constitutive

The two-parameter constitutive model describes the following biology:



where k is set to unity with no loss of generality at steady state. This system has the following PGF [94]:

$$\log G = u_N \frac{1}{\beta} + u_M \frac{1}{\gamma}, \quad (3.13)$$

i.e., the joint distribution is the product of two independent Poisson distributions. This model encodes the unregulated production of mRNA.

Extrinsic

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

$$\emptyset \xrightarrow{K} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \emptyset, \quad (3.14)$$

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

$$\log G = -\nu \log \left(1 - u_N \frac{1}{\beta} - u_M \frac{1}{\gamma} \right), \quad (3.15)$$

which follows immediately from the properties of generating functions [93]. This is the log-PGF of the bivariate negative binomial distribution, with negative binomial marginals [95]. This model encodes driving by a slow, Gamma-distributed transcriptional process [96, 97, 98, 99].

CIR-like

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

$$\emptyset \xrightarrow{K_t} \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\gamma} \emptyset, \quad (3.16)$$

where K_t is, informally, the time derivative of a subordinating inverse Gaussian process [99]. This system emerges in the fast limit of transcription driven by a Cox–Ingersoll–Ross process, and has the following log-PGF:

$$\log G = \frac{1}{2} \int_0^\infty \left[1 - \sqrt{1 - 4bU(u_N, u_M, s)} \right] ds, \quad (3.17)$$

where U is identical to the expression in Equation 3.11, 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 [99].

Bursty with delayed splicing

The three-parameter bursty model with delayed splicing describes the following biology:

$$\emptyset \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\tau} \mathcal{X}_M \xrightarrow{\gamma} \emptyset, \quad (3.18)$$

where τ is the deterministic retention time for $\mathcal{X}_N$ and k is set to unity with no loss of generality at steady state. This model has the following log-PGF:

$$\log G = \frac{\tau b u_N}{1 - b u_N} - \frac{1}{\gamma} \log (1 - b u_M). \quad (3.19)$$

Since this equation is separable in u_N and u_M , the PGF is the product of two independent marginal PGFs; the nascent marginal is geometric-Poisson [100, 101] whereas the mature marginal is negative binomial.

Bursty with delayed degradation

The three-parameter bursty model with delayed degradation describes the following biology:

$$\emptyset \xrightarrow{k} B \times \mathcal{X}_N \xrightarrow{\beta} \mathcal{X}_M \xrightarrow{\tau} \emptyset, \quad (3.20)$$

where τ is the deterministic retention time for $\mathcal{X}_M$ and k is set to unity with no loss of generality at steady state. This model has the following log-PGF [101]:

$$\begin{aligned} U'(u_N, u_M, s) &:= u_M + (u_N - u_M)e^{-\beta s} \\ \log G &= \frac{1/\beta}{1 - b u_M} \log \left(\frac{b U'(u_N, u_M, \tau) - 1}{b u_N - 1} \right) \\ &\quad + \frac{\tau b u_M}{1 - b u_M} - \frac{1}{\beta} \log (1 - b U'(u_N, u_M, \tau)). \end{aligned} \quad (3.21)$$

Technical noise models

In this section, we define the technical noise models of interest through their generating function definitions and solutions. These models account for the artifacts and biases introduced during the sequencing process.

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 $G_{z,t}$ of the sequencing processes:

$$H(u_N, u_M) = G(G_{N,t}(u_N) - 1, G_{M,t}(u_M) - 1) \quad (3.22)$$

Null process

In the simplest case, the null process, the sequencing is perfect and every mRNA is captured and reported as a unique molecule:

$$\mathcal{X}'_z \rightarrow \mathcal{X}_z \text{ with probability } 1. \quad (3.23)$$

The PGF of this process is $G_{z,t}(g_z) = g_z = u_z + 1$, yielding $G_{z,t}(u_z) - 1 = (u_z + 1) - 1 = u_z$, i.e., $G = H$.

Bernoulli process

In the Bernoulli process, each molecule has probability p_z of being captured and reported; multiple priming is impossible:

$$\mathcal{X}'_z \rightarrow \mathcal{X}_z \text{ with probability } p_z. \quad (3.24)$$

The PGF of this process is $G_{z,t}(g_z) = p_z g_z + (1 - p_z) = p_z u_z + 1$, yielding $G_{z,t}(u_z) - 1 = p_z u_z$.

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:

$$\mathcal{X}'_z \rightarrow \mathcal{X}'_z + \mathcal{X}_z \text{ at rate } \lambda_z. \quad (3.25)$$

This PGF of this process over unity time is $G_{z,t} = e^{\lambda_z(g_z - 1)}$, i.e., $G_{z,t} - 1 = e^{\lambda_z u_z} - 1$, which cannot be simplified further.

3.2 Application for Biological Insight

In the following section, we show how the CME models of transcriptional processes can be applied to scRNA-seq data to discover modulated molecular signatures beyond mean expression. In particular, we use the *Monod* package, which

This section summarizes section *Monod identifies strategies of resistance and recovery* of [37] by G.G., T.C., M.C., J.J.V., and L.P. M.C. identified the application dataset, performed the analysis, and interpreted results, with discussions and suggestions from G.G., T.C., J.J.V., and L.P.

The expanded capabilities of biophysical DE analysis also allow the exploration more complex biological settings, where it is less clear which mechanisms are expected to change and which genes provide actionable targets for followup experiments. For example, understanding cellular strategies of recovery post-treatment aids in the selection of relevant therapies and promotion of desired cell behaviors [102].

For example, the intestine is one of the organs most sensitive to radiation damage: understanding how it heals following radiotherapy is important for promoting healthy recovery and survival of patients treated for abdominal, gastrointestinal, pelvic, and retroperitoneal tumors [102]. Particularly interesting is how immune cells are affected by and respond to radiation treatment. T cells, which are essential actors in the body's natural response to tumor control [103, 104], are required for successful radiotherapy but are also believed to be damaged during treatment. While studies have investigated the transcriptomic landscapes of T cells following radiotherapy [103, 102], we suggest that a fuller understanding of the regulation of key genes, as revealed by biophysical parameter changes, could inform how therapeutics may be designed and more efficiently administered to enhance recovery from radiotherapy.

We re-analyze data from a study of the transcriptomic response of the intestinal microenvironment to radiation-induced intestinal injury (RIII) in wild-type C57BL/6J male mice [102] (Figure 3.1a). Mice were treated with 15 Gy of abdominal radiation, and intestinal samples were collected pre-radiation treatment (day 0) and at several time points following treatment (days 1, 3, 7 and 14 post treatment). We selected T cells based on the original study's annotations, and inferred biophysical parameters for samples (where each sample is a unique mouse) with at least 100 T cells, leaving 11 samples across days 0, 1, 3 and 14, with at least two samples per day. We used the bursty model with Poisson technical sampling noise, and after fitting 3,000 genes shared across samples, subset to the 1,182 that passed goodness-of-fit testing in all samples, further correcting inferred parameters for sample-specific offsets (using one sample as the reference, we fit an orthogonal distance regression model over all genes for other samples and removed the per sample bias).

We first show that analysis through the lens of a biophysical model reveals more genes with differences in behavior across the timecourse of recovery. To assess the significance of parameter changes, we performed a t -test with unequal variance on the $\log_{10}$ of each corrected parameter per gene for day 0 versus all other days. Specifically, we tested for changes in burst size (b), relative splicing rate (β), relative degradation rate (γ) and normalized mature counts (μ_M), correcting p -values for

the number of tested genes using the Benjamini Hochberg procedure (separately per parameter) to obtain false discovery rates (FDRs). Figure 3.1b shows that many genes that do not show significant differences in mature counts are detectable via parameter changes: testing the significance of inferred parameters in addition to means allows for the exploration of more genes whose significant difference between days or conditions may lie in regulatory strategy rather than mean expression change. For example, comparing day 1 to day 0, there were 380 genes with significant differences in a parameter but not in mature expression, 127 genes with differences in both, and 157 genes with significant differences in mature expression but not parameters (Figure 3.1b). We further emphasize that the analysis of parameters does not preclude testing for changes in mean, with our method adding to the list of possible genes for follow-up investigation.

In addition to identifying genes with parameter changes that are not detectable using standard mature expression comparisons, for genes which *do* show significant differences in expression, *Monod* can suggest what cellular regulatory strategy is responsible (e.g., an increase in burst size or a decrease in degradation rate, Figure 3.1c). For example, we show in Figure 3.1c (left) five genes which had significant increases between day 1 and day 0 in both burst size and mature RNA counts. As these genes also show increased degradation rate, the observed increase in mature counts is likely due to an increase in burst size (that exceeds the increase in degradation rate). These genes, which are upregulated early in the radiation-response timecourse (day 1 post treatment), have been studied as promising targets in cancer immunotherapy. *Cbl* encodes the E3 ubiquitin ligase Casitas B lymphoma, which regulates immune cells and whose homologue Casitas B lymphoma-b (CBL-b) has been shown to promote immunosuppressive tumor microenvironments (TME) [105]. *Nt5e* encodes 5'-nucleotidase (or Cd73), an immunosuppressive protein whose high expression is also associated with poor outcome of cancer patient's TMEs [106] and is currently emerging as a novel target for immunotherapy [107]. Other genes whose high expression has been correlated with lower overall survival in cancers, are *Ncapd3*, which is associated with reduced infiltration of immune cells into glioma [108], and the gene that codes for insulin-like growth factor 2 receptor (*Igrf2*), which was found to be an indicator of poor prognosis in patients with triple-negative breast cancer [109]. On the other hand, the expression of *Mpp7*, which codes for a scaffolding protein, has been shown to have a positive correlation with increased infiltration of T cells in melanomas [110].

Other notable genes show increases in mature RNA expression best explicable by decreases in degradation rate (Figure 3.1c, right). *Ets1* encodes a transcription factor that is a proto-oncogene that promotes aggressive tumor behavior in a variety of cell types [111]. While not directly associated with cancer progression, the proteins encoded by *Pak2* and *Git2* are known to interact [112], and are crucial for the stability and suppressive function of regulatory T cells [113]. With knowledge about what specific point in the RNA lifecycle (e.g., bursty transcription or degradation) may lead to increased expression of a gene, researchers could design strategies to discourage known tumor promoters and encourage known tumor suppressors for increased efficacy and complementarity of radio- and immunotherapy.

Finally, we can identify genes whose parameters show similar dynamics (fold changes) over the timecourse of recovery from radiation treatment as compared to known tumor suppressing genes. For example, we found the three genes with fold change patterns most similar to *Znrf3* (Figure 3.1), whose encoded protein negatively influences cancer progression by inhibiting key proliferative (WNT) signaling pathways [114, 115]; and another three with patterns most similar to *Tgfb1*, which potentially suppress tumorigenesis at early cancer stages but promote tumorigenesis at advanced cancer stages [116]. Of these genes, some have been reported as important in cancer: for example, *Gch1* is a known immunosuppression gene with upregulation previously linked to reduced survival in breast cancer [117]; *Lcp1* encodes a prognostic biomarker in gastric cancer [118]; and *Arhgap15* codes for a protein that is reported to slow colorectal tumor growth [119]. Other identified genes have not been as well-characterized in cancer studies: *Otulin* encodes an immune-response peptidase [120], and *Fus* and *Bbx* are more general DNA damage-response genes [121, 122]. Analysis of biophysical parameters allows one to explore how genes that have been previously implicated in cancer prognoses are experiencing similar regulatory changes in response to radiation treatment, as well as to identify genes that have *not* been previously implicated as targets for follow up therapeutic development.

3.3 Model Selection for Biophysical Hypothesis Generation

This section summarizes section *Model selection and insight about gene regulation strategies* of [37] by G.G., T.C., M.C., J.J.V., and L.P. M.C. identified the application dataset, performed the analysis, and interpreted results, with discussions and suggestions from G.G., T.C., J.J.V., and L.P.

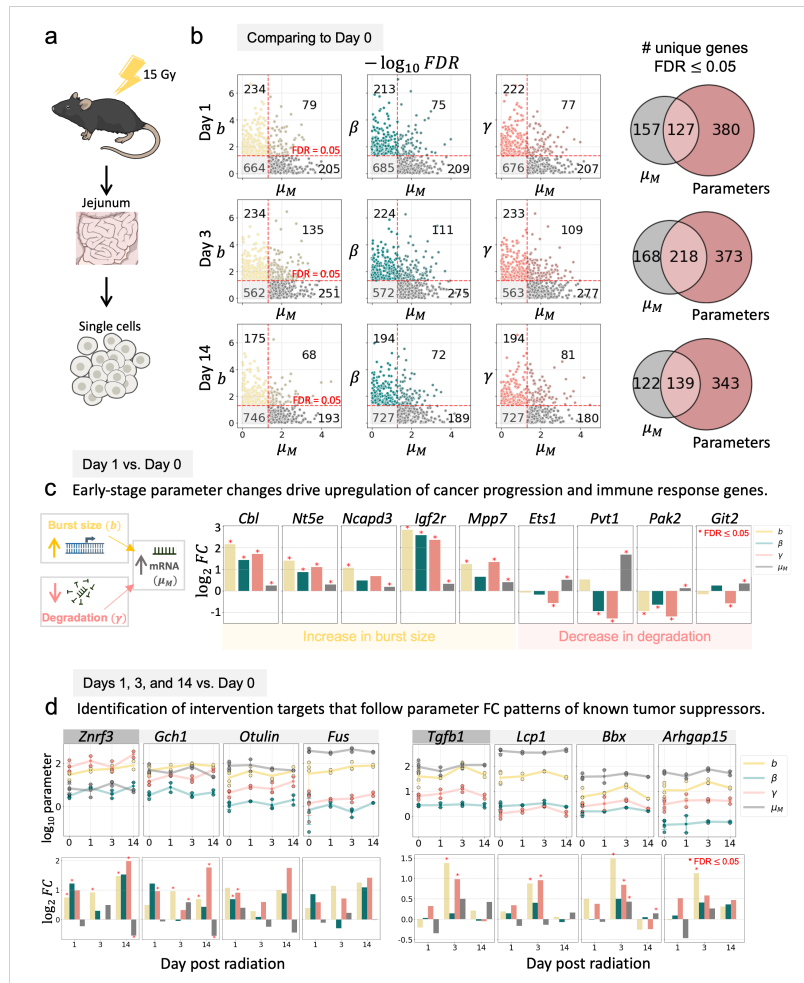


Figure 3.1: **Monod analysis of mechanisms of T cell recovery post radiation treatment.**

a. Single-cell samples were taken from the intestines of mice exposed to 15 Gy of abdominal radiation before and at several time points after treatment. Illustrations adapted from NIAID NIH BIOART Source [123, 124, 125]. **b.** Two-sided t -tests comparing T cells at days 1 ($n=2$), 3 ($n=3$) and 14 ($n=2$) post treatment to T cells before treatment (day 0, $n=4$) flag more genes as significant (false discovery rates, $FDR \leq 0.05$) when comparing inferred parameters (b, β, γ) than when merely comparing mature RNA counts (μ_M). In the scatter plots, red dashed lines indicate a significance threshold of an FDR of 0.05. **c.** The $\log_2$ fold change (FC) of parameters (comparing day 1 to day 0) show that statistically significant changes in the mature mean expression of genes related to the progression of cancer and immune response can be attributed to statistically significant differences in parameters (i.e., burst size or degradation rate). A red asterisk indicates a $FDR \leq 0.05$. **d.** Comparing genes' parameter changes over the timecourse of recovery can identify similarly regulated genes. The top row shows the $\log_{10}$ of parameters and mean mature RNA expression at days 0, 1, 3 and 14. Each outlined point indicates a separate sample (mouse) with a smaller point at the mean over the $\log_{10}$ of samples' parameters per day. Error bars indicate the interquartile range of the samples' parameters per day. For certain genes, the $\log_2$ FCs of each parameter and mean mature expression for days 1, 3 and 14 from day 0 exhibit similar patterns as known tumor suppressors (bottom row).

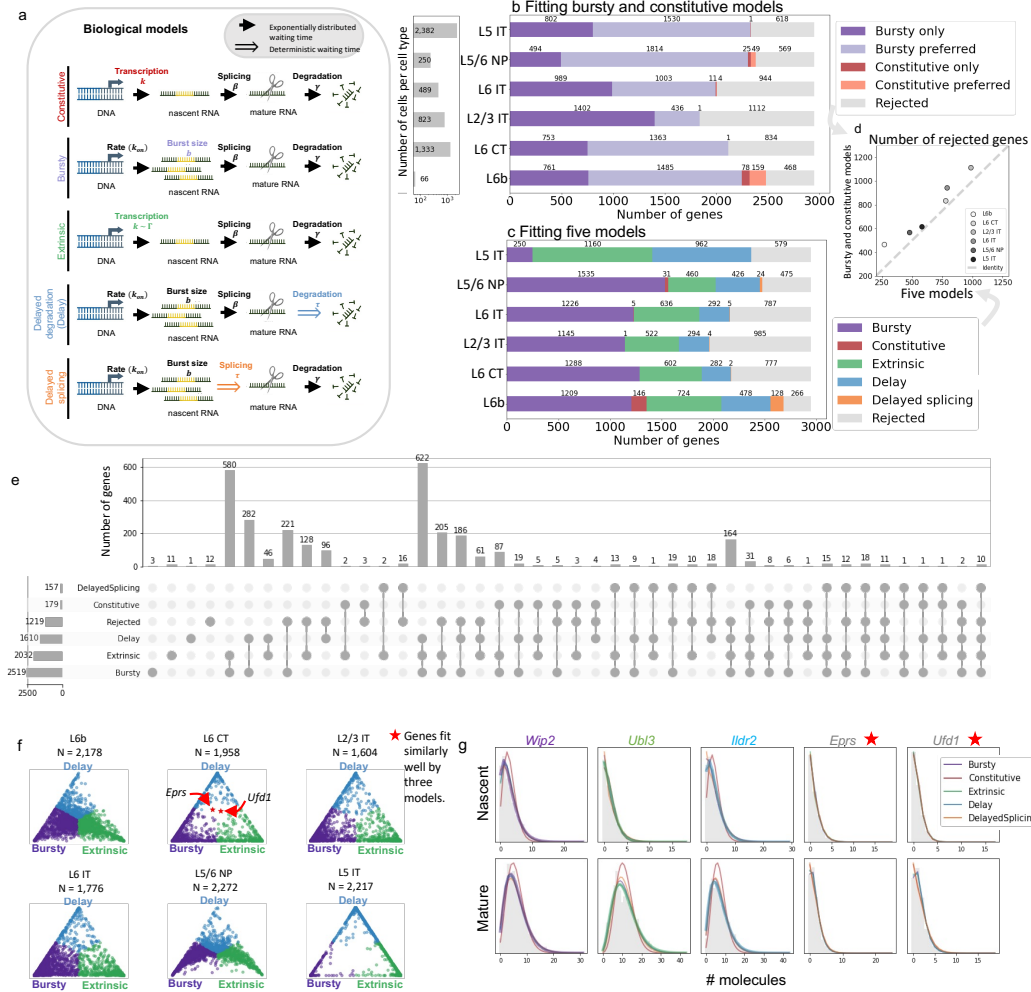


Figure 3.2: *Monod* facilitates the comparison of different biophysical models on a per gene basis for thousands of genes across cell types. Results are shown for Allen sample B08 [126]. **a.** Diagrams of five of the biological models implemented in *Monod*, varying in transcription and processing kinetics. **b.** Of the 2,951 fit genes in each of six cell types, most genes display behavior more consistent with the bursty model than with the constitutive model using the Akaike information criterion (AIC) and two goodness-of-fit metrics (Hellinger distance from empirical distributions and chi-squared test). **c.** For the same six cell types and genes, the number of genes that are best fit by each of the five considered biological models, colored by model. Gray indicates genes for which all models were rejected). **d.** Fewer genes are rejected when fitting and comparing five models than fitting and comparing only the bursty and constitutive models. **e.** The same gene can exhibit different behavior in different cell types: the upset plot shows the number of genes that are best fit by a given model or combination of models across the six cell types. **f.** Ternary diagram of genes' normalized AIC weights for the three most commonly selected models, colored by the model with the largest AIC weight. Only genes for which none of these models were rejected are displayed, with the number indicated. **g.** Predicted marginal nascent and mature distributions for each model plotted over normalized empirical histograms of molecule counts for the indicated genes in layer 6 neurons (L6 CT). *Eprs* and *Ufd1*, which had similar AIC weights for the bursty, extrinsic, and delay models, are noted with a red star in the ternary diagrams.

Monod currently supports 6 sophisticated yet tractable models of transcription, previously shown to describe and distinguish between features of experimental data [99, 101, 36] for thousands of genes across multiple cell types. We emphasize that *Monod* supports *modular* model comparison, in the sense that one can freely modify assumptions about a specific model component (e.g., whether degradation occurs with a deterministic or stochastic waiting time) to obtain a new model to test against data. We recommend using *Monod* with the bursty model by default, as it captures observed overdispersion [127] and known biophysics [33, 128] while only minimally complexifying the constitutive model. But the bursty model may not *always* fit the data best, and it is worth testing to what extent a less complex, more complex, or simply different model performs better in practice.

We use *Monod* to analyze mouse brain data with 2,951 genes and six cell types (Allen sample B08 [126]) and show that the bursty model almost always outperforms the simpler constitutive model, and tends to perform fairly well against alternative models (see Figure 3.2a for schematics of the 5 models we consider in this section). Moreover, we show that it is useful to consider a richer space of transcription models, since specific models may fit specific genes or cell types better or worse, which suggests *Monod*-derived fits yield insight about gene- and cell-type-specific gene regulation strategies. We reject models (for a given cell type and gene) if they fail to pass the goodness-of-fit thresholds of Hellinger distance to empirical distributions and chi-square tests.

We first compare the constitutive model to the bursty model. For many genes, only the bursty model passes goodness-of-fit thresholds. When neither the bursty nor the constitutive models are rejected for a gene, the bursty model is usually a better fit according to Akaike information criterion (AIC, Figure 3.2b). It is extremely rare that the constitutive model fits the data better, perhaps unsurprisingly as the constitutive model is a special case of the bursty model and the overdispersion of single-cell count data is well known, but diverging from previous studies that advocate the Poisson model [129].

Although the transcriptional behavior of many genes is consistent with the bursty model, comparing all five of the (comparably complex) proposed models reveals that the extrinsic, delayed degradation, and delayed splicing models are better fits for some genes in different glutamatergic cell types (Figure 3.2c). This comparison also suggests differences between cell types' transcriptional regulation strategies, with more genes best fit by the extrinsic model in layer 6 intratelencephalic (L6 IT)

neurons than in layer 5/6 near-projecting (L5/6 NP) neurons, for example (Figure 3.2c). Further, including the additional models increases the number of genes in each cell type for which one or more models pass goodness-of-fit thresholds: fewer genes are rejected by all models when considering more models (Figure 3.2d). As all models were compared in a standard way using the same data, fitting pipeline, and statistical rejection criteria, the discrepancies between proposed models could be evidence of real biological variety in transcriptional and post-transcriptional regulation mechanisms.

While cell type comparisons show distinctions between the number of genes best fit by each model in a given cell type, *Monod* also provides the tools to investigate how genes exhibit expression patterns consistent with different transcriptional models across cell types (Figure 3.2e). We found the model with the largest AIC weight for every gene in every cell type, then for each gene collapsed models over cell types to a set of unique models. The upset plot in Figure 3.2e shows how many genes exhibited each combination of models preferentially across the six cell types. For example, only three genes exhibited *only* bursty behavior preferentially, while 580 genes were best fit by the bursty model in at least one cell type and by the extrinsic model in at least one other cell type. All but 27 genes exhibited expression trends described best by at least two unique models, with the most common set (622 genes) being bursty, extrinsic, and delayed degradation. Interestingly, no genes were best fit by only the delayed splicing or only the constitutive model across the six cell types (Figure 3.2g).

While these trends may reflect biological differences, they may also highlight differences in the mutual identifiability of the proposed models. For example, the bursty, extrinsic, and delayed degradation model all account for correlation between nascent and mature counts, while in the constitutive and delayed splicing model, nascent and mature counts are uncorrelated. Thus the genes that are best fit by the three models or some subset may be in fact displaying similar behavior across different cell types, with the correlation between their nascent and mature counts, ruling out the constitutive and delayed splicing models, as the salient feature driving the fits (similarly to Gorin et al. [101]).

To explore this question, we can assess the identifiability of models, or the extent to which the “best” fit model is “better” than other models. While we can assign a gene to the model with the largest AIC weight, there is not always a clear distinction between model fits. To illustrate this, we show a simplex of normalized AIC weights

for the bursty, extrinsic, and delayed degradation models, colored by the assigned regulatory model: the weights for the “best” fit models are sometimes arbitrarily close to the other two models’ weights (Figure 3.2f). In this way, genes with clear preference for a given model can be distinguished from those which are similarly fit by two or more models. These results can also be verified by visualizing model-based probabilities and normalized empirical count histograms. Figure 3.2g shows genes that were not rejected by only the bursty model (*Wipf2*), only the extrinsic model (*Ubl3*), and only the delayed degradation model (*Ildr2*) in layer 6 neurons (L6 CT). *Eprs* and *Ufd1*, indicated with red stars in Figure 3.2f, were fit similarly well under the three models. In the histograms for *Wipf2*, *Ubl3*, and *Ildr2*, we underline the (Poisson) constitutive model’s poor ability to recapitulate the overdispersed data, consistent with Figure 3.2b-c.

Monod makes fitting subtly distinct mechanistic models of transcription for thousands of genes straightforward, thereby leading to increased ease of model comparison and selection. Performing goodness-of-fit and likelihood tests on *Monod* fits provides a route for discerning between biophysical models that may have generated the data. However, we note that definitive model selection is beyond the scope of *Monod*: rather than to come to definite conclusions, *Monod* should be used to explore data and to compare and generate hypotheses that can later be tested experimentally.

Chapter 4

NEURAL APPROXIMATIONS FOR INTRACTABLE BIOPHYSICAL MODELS

As discussed in Chapter III, biophysical models formalized via the Chemical Master Equation (CME) are appropriate and potent models to represent mRNA counts captured from single cells. Not only do they invalidate heuristic manipulations of data, but in doing so produce testable hypotheses relating to transcriptional rates and parameters. However, solutions for all but the simplest of such models have no closed-form and require computationally intensive solvers, limiting the application of these biophysical models to the large-scale, genome-wide datasets commonly produced. In this Chapter, we discuss an efficient approach to approximate solutions using neural networks and statistical analysis of system moments.

This section summarizes [46] by M.C.*, G.G.*, T.C., and L.P. G.G., where * denotes co-first authorship. M.C. implemented and adapted the strategies, trained models, and performed analyses. G.G. conceptualized the neural approximation strategies and was involved in implementation, statistical analyses, and basis function design. T.C. and L.P. provided feedback, design suggestions, and trouble-shooting advice.

The advent of high-throughput transcriptomics provides an opportunity to advance mechanistic understanding of transcriptional processes and their connections to cellular function at an unprecedented, genome-wide scale. These transcriptional systems, which involve discrete, stochastic events, are naturally modeled using Chemical Master Equations (CMEs), which can be solved for probability distributions to fit biophysical rates that govern system dynamics. While CME models have been used as standards in fluorescence transcriptomics for decades to analyze single species RNA distributions, there are often no closed-form solutions to CMEs that model multiple species, such as nascent and mature RNA transcript counts. This has prevented the application of standard likelihood-based statistical methods for analyzing high-throughput, multi-species transcriptomic datasets using biophysical models. Inspired by recent work in machine learning to learn solutions to complex dynamical systems, we leverage neural networks and statistical understanding of

system distributions to produce accurate approximations to a steady-state bivariate distribution for a model of the RNA life-cycle that includes nascent and mature molecules. The steady-state distribution to this simple model has no closed-form solution and requires intensive numerical solving techniques: our approach reduces likelihood evaluation time by several orders of magnitude. We demonstrate two approaches, where solutions are approximated by (1) learning the weights of kernel distributions with constrained parameters, or (2) learning both weights and scaling factors for parameters of kernel distributions. We show that our strategies, denoted by kernel weight regression (KWR) and parameter scaled kernel weight regression (psKWR), respectively, enable broad exploration of parameter space and can be used in existing likelihood frameworks to infer transcriptional burst sizes, RNA splicing rates, and mRNA degradation rates from experimental transcriptomic data.

Background

The production, processing, and degradation of RNA molecules are governed by a set of stochastic events that lead to variable levels of RNA transcripts and proteins between genetically identical cells in identical conditions [34]. The inherent stochasticity and resulting variability in RNA transcript levels play an essential role in cellular function, development, and response, but must also be tightly regulated to produce deterministic functional outcomes at precise times. Extracting the rates that underlie distributions of RNA molecules is essential for understanding the intricate balance between stochastic gene expression and precise cellular responses required for life.

To explore the dynamics of transcriptional systems, specific biophysical models can be proposed that are explicitly parameterized by relevant biological rates, such as rates of transcriptional initiation, RNA splicing, or RNA degradation. These stochastic systems are naturally modeled by discrete-valued, continuous-time Markov chains (CTMCs). The evolution of a system’s microstate probabilities under a CTMC model of biology is given by a chemical master equation (CME) [130]. For example, a CME may track the changes in probability P of there being a certain number of nascent (n) and mature (m) RNA molecules at time t given certain biophysical parameters θ , or $P(n, m, t; \theta)$. This can be used to obtain insights into transcriptional biophysics [99] by fitting θ most consistent with experimental distributions of molecule numbers [131].

Until recently, mechanistic modeling and insight into the underpinnings of tran-

scriptional processes have been tied to intensive, temporal, fluorescence datasets which generally measure the production of one species of RNA molecules in one cell line [34, 132, 133]. Recent high-throughput transcriptomics methods, such as single-cell RNA sequencing (scRNA-seq), offer discrete molecular readouts at genome-wide scale across diverse cell populations [134], thus making possible the exploration of transcriptional parameters across distinct cell-types, tissues, and organisms. In addition to increased cell-type coverage, the diversity of molecular species captured, such as the simultaneous quantification of unspliced and spliced RNA molecules [135, 136, 137], proteins [138], and chromatin accessibility [139], has the potential to deepen our understanding of transcriptional system dynamics [140]. While single-cell RNA-seq data has been modeled using CME solutions to extract biophysical rates of transcriptional initiation and burst sizes, mechanistic modeling has not been adopted in recent analyses of multi-species transcriptomic experiments [141]. One major challenge is that closed-form solutions to CMEs that model multi-species systems with several interacting species are rarely available [99, 142], and current methods to solve them numerically for probabilities are prohibitive to the recursive calculations necessary for maximum likelihood inference of biophysical parameters.

4.1 Existing Solvers and Limitations

The methods currently used to circumvent the lack of closed-form solutions to CMEs are computationally expensive and compromise accuracy. For example, CTMCs are often simulated using Gillespie’s stochastic simulation algorithm [143], which generates statistically realistic molecular trajectories over possible state-spaces, but the number of simulations required for accurate distributional approximation is inefficient for parameter space exploration. Alternatively, an approach called finite state projection (FSP) [144, 145], which uses matrix methods to solve the CME over a truncated state space, has been applied for inference [146, 133, 131, 140] but is extremely inefficient when the state space is much larger than the number of data points. In certain cases, CMEs can be solved using generating functions that transform infinite-dimensional CMEs to complex partial differential equations (PDEs). However, this approach requires performing an inverse Fast Fourier Transform over a finite, truncated array [147]. Both FSP and generating function methods require state-space truncation, which leads to error in calculated probabilities, with a trade-off between higher accuracy and increased evaluation time with increased grid size. Further, these methods are not amenable for efficient calculation of likelihood

gradients necessary for widely used machine-learning based numerical inference procedures [148, 149]. In addition, the challenge of approximation suffers from the so-called “curse of dimensionality”: the size of the state-space grows exponentially with the number of modeled species.

In the fields of machine learning and data-driven modeling, recent work has demonstrated the ability of sparse function approximations to distributions [150] and learned mappings between function spaces rather than between finite-dimensional Euclidean spaces [151] to learn generalizable solutions that govern the dynamics of complex systems. For example, [151] has introduced a neural operator formulation that models fluid dynamics in turbulent flows governed by such PDEs as the Navier-Stokes equations several orders of magnitude faster than traditional solvers.

Inspired by these approaches, we present a computationally efficient approach that uses neural networks coupled to statistical analysis of system distributions to approximate multidimensional solutions to CME models of transcriptional dynamics. In particular, we investigate a CME system involving transcriptional bursting, splicing and degradation (Fig. 4.1a). The system of bursty transcription has been experimentally observed in bacteria for a single gene [152] and more recently in mammalian cells for thousands of genes [153, 154]. While our proposed bursty system does not explicitly model complex genetic regulatory networks and feedback loops, its inclusion of bursty transcription and splicing is an improvement on simpler, more commonly applied models of constitutive transcription that lack splicing. However, the proposed bursty system with splicing does not yield a closed-form steady-state solution and requires costly numerical evaluation procedures [43]. We alleviate this problem by using features of the system’s univariate conditional distributions to define a statistical approximation for conditional nascent RNA distributions in terms of a mixture of negative binomial kernel functions. The weights for these kernel functions are learned by a neural network (Fig. 4.1B). We then multiply by marginal mature RNA probability to reconstruct full joint distributions (Fig. 4.1C). We denote this procedure kernel weight regression (KWR).

To increase flexibility of the neural approximation while further decreasing evaluation time, we also implement an approximation approach by learning scaling factors for pre-calculated mean and variances of fewer negative binomial distribution kernels in addition to weights. The joint probability is again approximated by applying conditional probability. We call this approach parameter scaled kernel weight regression (psKWR). Both procedures reconstruct joint distributions by approxi-

imating conditional distributions for mature RNA molecules with neural-network scaled kernel distributions, then reconstructing full joint distributions using an analytical solution for marginal nascent RNA distribution. KWR and psKWR also provide approximations on an unbounded support and are compatible with gradient calculations required for machine-learning based likelihood optimization strategies, thereby facilitating inference of system parameters at increased resolution [47].

Our work complements other recent reports that leverage neural networks to approximate CME solutions for single-cell biophysics models. [155] use recurrent neural networks to directly learn stochastic system distributions given some system-governing input parameters, but probabilities are learned for a predefined, discretized state space. [156] have proposed neural approximations for continuous-valued statistics of CME systems but cannot be used to perform inference. Most pertinently, [157] have independently proposed a method for approximating CME solutions with a set of negative binomial mixture functions, implemented using the Nessie framework. However, Nessie learns mixture function locations and weights, while our approach simplifies the learning process by learning weights and location scaling factors, initially placing kernel functions by exploiting the unimodality and moments of the system’s CME solution. Further, Nessie is designed to fit univariate distributions; we present a strategy for learning multivariate distributions, an essential feature for the analysis of increasingly available single-cell transcriptomic data with multiple species.

Transcriptional system

We consider the system illustrated in Fig. 4.1A, which encodes a standard model of the mRNA life-cycle [43]. Transcriptional events arrive according to a Poisson process with rate α , and generate B nascent mRNAs (N), where B is a geometric distribution with mean b . After an exponentially-distributed delay with rate β , nascent mRNAs are spliced into mature mRNAs (M). Finally, after an exponentially-distributed delay with rate γ , mature mRNAs are degraded. This system induces a stationary distribution with the probability mass function (PMF) $P(n, m; b, \beta, \gamma)$, where β and γ are considered in units of α with no loss of generality at steady state. Although the system does not afford a closed-form solution, its joint PMF can be approximated to high precision through generating functions evaluated over a state-space grid (the larger the state-space, the greater the accuracy) [43]. Further, its nascent marginal $P(n)$ is negative binomial [43]. A more complete discussion is available in Sections S1 and S2 with notation for the considered statistical and

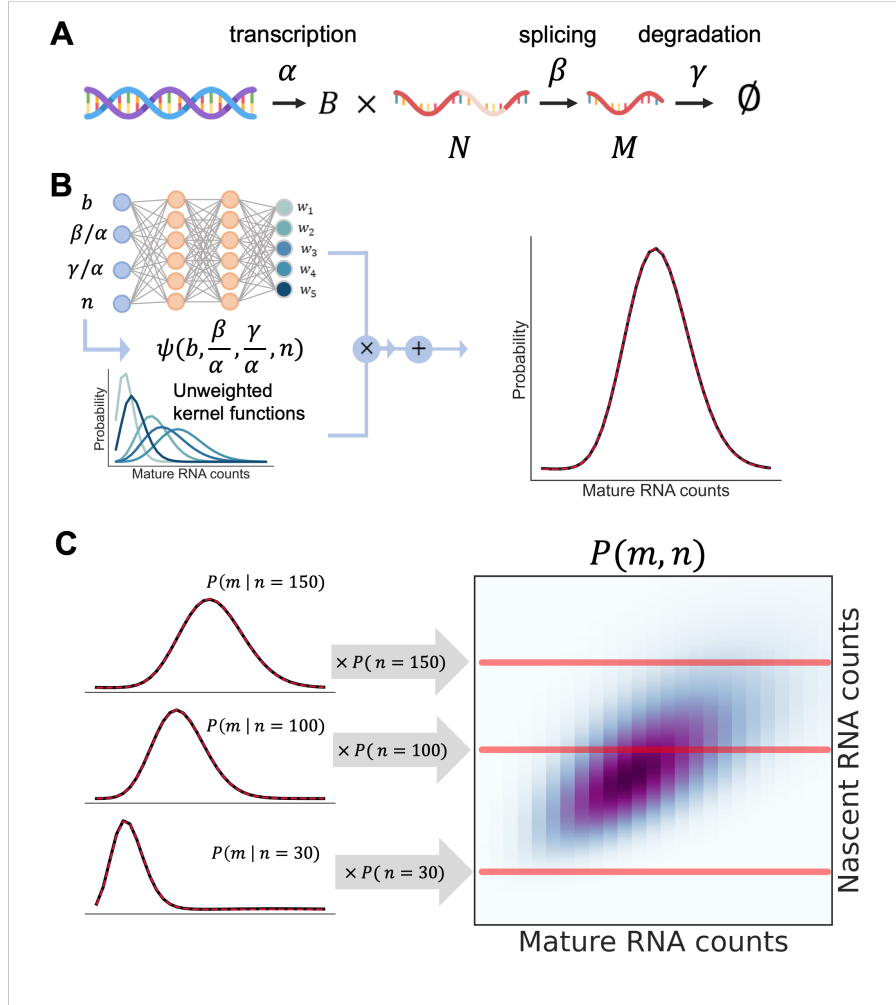


Figure 4.1: A schematic representation of the KWR approximation procedure. **A.** The biological process under consideration: bursty transcription of nascent RNA, followed by splicing to produce mature RNA and mRNA degradation. **B.** Univariate conditional distributions are approximated by summing a set of kernel functions with neural network learned weights. **C.** Joint distributions are reconstructed by multiplying conditional mature RNA probabilities by marginal nascent RNA probabilities.

biological quantities listed in Table S1.

4.2 Kernel Weight Regression and Extensions

Kernel weight regression (KWR) approximation overview

To approximate full joint distributions, we train a neural network to approximate the distributions of mature RNA molecules conditional on a nascent copy number n (Fig. 4.1B), i.e., the univariate PMFs $P(m|n; b, \beta, \gamma)$. We first generate rate vectors $\theta := \{b, \beta, \gamma\}$ and compute high-quality generating function solutions over state-space grids to obtain joint PMFs $P(m, n; \theta)$ following the solution in [43].

These solutions are denoted QV20 (as they are evaluated on grids with limits 20 standard deviations above nascent and mature marginal means) and treated as ground truth. We then divide by nascent probabilities $P(n; \theta)$ to extract conditional mature probabilities $P(m|n; \theta)$.

To construct a basis for the conditional distribution at the nascent copy number n , we use a deterministic function $\psi(\theta, n)$ to calculate kernel means μ_k and standard deviations σ_k of negative binomial basis kernels P_k , where k indexes the negative binomial kernels (Section S3). The function takes advantage of the marginal means and variances of the system, constructing an initial basis by moment-matching to an analytically tractable lognormal distribution (Section S3). Next, we use a neural network to map $\tilde{\theta}_{\text{KWR}}$, a transformation of θ and n , to a set of weights $\mathbf{w}$ and a kernel width scaling factor h . We train the network to minimize truncated Kullback-Leibler divergence (KL divergence, Section S4) between the conditional true distributions $P(m|n; \theta)$ and the conditional kernel approximations $Q_{\text{KWR}}(m|n; \theta)$:

$$Q_{\text{KWR}}(m|n; \theta) = \sum_k w_k P_k(m; \mu_k, h\sigma_k). \quad (4.1)$$

Finally, to reconstruct the joint distributions, we multiply each $Q_{\text{KWR}}(m|n; \theta)$ by the analytical nascent marginal probability $P(n; \theta)$, as in Fig. 4.1C:

$$Q_{\text{KWR}}(m, n; \theta) = Q_{\text{KWR}}(m|n; \theta)P(n; \theta). \quad (4.2)$$

We call this method kernel weight regression (KWR), with a full description in Section S7 and Table S2. Model architecture choice is motivated in Fig. S1. Fig. S2 depicts final model architecture and Table S3 lists final model parameters.

Parameter scaled kernel weight (psKWR) approximation overview

To further allow for flexibility of distribution approximation, we present an alternative strategy, which scales the pre-calculated kernel means and variances. Here, we use the same function $\psi(\theta, n)$ for a given n to estimate kernel distribution means μ_k and standard deviations σ_k . We then use a neural network to map $\tilde{\theta}_{\text{psKWR}}$, a slightly modified transformation of θ and n , to weights $\mathbf{w}$ and scaling factors $\mathbf{c}_\mu$ and $\mathbf{c}_\sigma$ to construct the approximate conditional distribution $Q_{\text{psKWR}}(m|n; \theta)$:

$$Q_{\text{psKWR}}(m|n; \theta) = \sum_k w_k P_k(m; c_{\mu_k} \mu_k, c_{\sigma_k} \sigma_k), \quad (4.3)$$

where k again indexes the kernel. The joint probabilities are then reconstructed by multiplying each $Q_{\text{psKWR}}(m|n; \theta)$ by marginal $P(n; \theta)$:

$$Q_{\text{psKWR}}(m, n; \theta) = Q_{\text{psKWR}}(m|n; \theta)P(n; \theta). \quad (4.4)$$

We again train the network to minimize the truncated KL divergence between the joint true distributions $P(m|n; \theta)$ and the joint approximations $Q_{\text{psKWR}}(m|n; \theta)$.

We call this approach parameter scaled kernel weight regression (psKWR). It is fully described in Section S8, with model architecture choice motivated in Fig. S3, final model architecture illustrated in Fig. S4, and final model parameters listed in Table S4.

Data generation and generating function solutions

To generate training, validation, and testing distributions, we sampled vectors of biophysical rate vectors θ within the following biologically motivated bounds:

$$\begin{aligned} b &\in [1, 300], \\ \beta &\in [0.05, 50], \\ \gamma &\in [0.05, 50], \end{aligned} \quad (4.5)$$

where the sampling measure was log-uniform. As described previously, β and γ are unitless quantities, scaled to the value of α . We generated a vector of three random numbers within the bounds. Next, we computed μ_N and μ_M , the average nascent and mature copy numbers (Table S1). To ensure that training distributions matched the low-copy number regime of experimental scRNA-seq data, we restricted to vectors with μ_N and μ_M both less than 1,000. These bounds are well under numbers of RNA molecules per gene in single cells reported in fluorescent transcriptomic studies [153, 154]. Distributions of training, validation, and testing rate vectors that passed these thresholds are shown in Fig. S5. For each θ , we evaluated the previously derived generating function solution to the system, given in [43], up to the bounds $\mu_z + 20\sigma_z$ (where z is either nascent n or mature m value) using adaptive quadrature (implemented in the `scipy` function `integrate.quad_vec` [158]) and the inverse real fast Fourier transform (implemented in the `scipy` function `fft.irfft2` [159]). We denote this high-precision procedure by QV20 and treat it as ground truth. We further validated the accuracy of QV20 by showing its concordance with distributions obtained using Gillespie's stochastic simulation algorithm in Fig. S6 [143].

We further generated lower-order generating function solutions for benchmarking. We calculated generating function solutions using adaptive quadrature on grids of 10, 4, and 1 standard deviations above nascent and mature means, or within the bounds $\mu_z + 10\sigma_z$, $\mu_z + 4\sigma_z$, and $\mu_z + \sigma_z$ in each dimension $z \in \{N, M\}$. We refer to these solutions as QV10, QV4, and QV1 in the text and Fig. 4.2.

We trained KWR models on normalized conditional values and psKWR models on joint microstate probabilities. To generate normalized conditional training data for KWR, we extracted the entries corresponding to each n and divided them by the analytical solution for the marginal $P(n; \theta)$ (see Section S2, implemented in the `scipy` function `stats.nbinom.pmf` [158] with $r = \beta^{-1}$ and $p = (1 + b)^{-1}$ [160]). The resulting conditional distributions were stored along with their values of θ and n .

Model training and architectures

KWR models were trained to minimize the truncated KL divergence (see Section S4) between approximated conditional distributions $Q_{\text{KWR}}(m|n; \theta)$ and normalized QV20 conditional distributions $P(m|n; \theta)$ for mature RNA transcripts.

Architectures of varying number of layers, nodes per layer, and number of kernel functions were trained and compared, with testing results for different models shown in Fig. S1. The architecture chosen for final comparisons is a MLP comprising 2 hidden layers of 256 nodes the output of which is passed to one layer of 256 nodes that produces weights for 10 basis functions, and a different layer of 256 nodes that produces the scaling factor h (described in S7). This final model was trained for 35 training epochs with a learning rate of 0.001 with 1,536 training rate vectors θ (corresponding to 211,515 vectors $\tilde{\theta}_{\text{KWR}}$ conditional vectors) and 512 validation rate vectors θ (corresponding to 72,906 vectors $\tilde{\theta}_{\text{KWR}}$ conditional vectors). All models were implemented in PyTorch with the Adam optimizer [161]. Full descriptions of training and architectures are included in Section S7, Fig. S2, and Table S3.

psKWR models were trained to truncated KL divergence or Hellinger distance between conditional approximations $Q_{\text{psKWR}}(m|n; \theta)$ and QV20 conditional microstate probabilities $P(m|n; \theta)$. We again compared model architectures of different numbers of layers, nodes per layers, and basis functions to obtain a final model (Section S7 and Fig. S1). Models trained using truncated KL divergence and Hellinger distance as loss achieved similar results (Fig. S3); we use a model trained with truncated KL divergence as the final model for better comparison with

the final KWR model. The final model chosen for psKWR approximation is a MLP with 3 hidden layers (L1, L2, and L3) of 256 nodes per layer that produce weights $\mathbf{w}$ for five kernel functions. It further includes one layer of 256 nodes that takes outputs from L2 to produce scaling factors $\mathbf{c}_\mu$ for kernel means and another layer of 256 nodes that takes outputs from L2 to produce scaling factors $\mathbf{c}_\sigma$ for kernel standard deviations. It was trained for 100 epochs on 1,280 rate vectors $\boldsymbol{\theta}$ (corresponding to 211,515 conditional vectors $\tilde{\boldsymbol{\theta}}_{\text{psKWR}}$) and 512 validation rate vectors $\boldsymbol{\theta}$ (corresponding to 72,906 conditional vectors $\tilde{\boldsymbol{\theta}}_{\text{psKWR}}$). The model was implemented in PyTorch and optimized using the Adam optimizer with a learning rate of 0.001 [161]. A full description of training is included in Section S8 with final model architecture diagrammed in Fig. S4 and final model parameters listed in Table S4.

Inference of biophysical parameters in mouse brain cells

We implemented psKWR and KWR neural approximation strategies into an existing maximum likelihood inference framework *Monod* [162] to infer burst sizes and rates of splicing and degradation for 200 genes in GABAergic cells in the mouse brain [126]. We obtained data from mouse sample C01 (donor ID 427378) from the Allen Institute Mouse Primary Motor Cortex (MOp) dataset [126]. Raw FASTQ files for the data were obtained and processed using `kallisto | bustools` as described in [162]. Briefly, raw reads were aligned to a reference of unspliced and spliced transcripts (generated using the flag `--lamanno`) to obtain counts for unspliced (intron containing, treated as nascent transcripts) and spliced transcripts (non-intron containing, treated as mature transcripts). Cells were filtered to remove low-quality cells (as in [162] by removing those that did not pass `kallisto | bustools` filtering and those with fewer than 10^4 unique molecular identifier counts).

We then restricted our analysis to cells of a single cell sub-type, GABAergic neurons, according to cell type annotations from the original study [126]. For these 1,704 subset GABAergic cells, gene selection was performed using *Monod* to obtain genes with unspliced and spliced means (across all cells) of at least 0.01, maximum of unspliced and spliced counts (across all cells) of at least 3, and maximum spliced counts (across all cells) no greater than 400. We then selected the top 200 of remaining genes by expression (as in [162]).

For all 200 genes, inference was performed to obtain burst size b , relative splicing rate β/α , and relative degradation rate γ/α , according to the bursty model of tran-

scription by maximizing data likelihood under the model’s steady-state distribution using *Monod* [162]. For each gene, *Monod* identifies maximum likelihood estimates of burst size b , splicing rate β , and degradation rate γ (with splicing and degradation rates in units of transcription rate α) with the function `scipy.optimize.minimize` [158]. This minimizes the Kullback-Leibler divergence between predicted distributions and empirical data distributions (equivalent to maximizing data likelihood) using gradient descent and the L-BFGS-B algorithm for bounded optimization [158], with termination when either the projected gradient falls below the default tolerance of 10^{-6} or the maximum number of iterations has been reached. Parameters were initiated to their method of moment estimates and optimized within the bounds of $[-1.0, 4.2]$, $[-1.8, 2.5]$, $[-1.8, 3.5]$ for $\log_{10}b$, $\log_{10}\frac{\beta}{\alpha}$, and $\log_{10}\frac{\gamma}{\alpha}$, respectively. Inference was performed with a maximum of 15 iterations and a single search (no restarts).

Results and discussion

To benchmark the algorithm’s performance, we investigated the accuracy and runtime for PMF grid evaluation for a set of testing rate vectors (Fig. 4.2A). To quantify accuracy, we calculated the Hellinger distance, which measures the discrepancy between probability distributions P and Q and takes the form $\frac{1}{\sqrt{2}}\|\sqrt{P} - \sqrt{Q}\|_2$, using the high-quality QV20 solutions as ground truth.

Comparison to generating function grid evaluations

We first compared KWR to three lower-order generating function-based solutions, denoted QV1, QV4, and QV10, which evaluate the numerical steady-state solution (Section S6) to varying accuracies based on grid size. We found that the KWR approximation achieves accuracy comparable to lower-order generating function solutions while reducing evaluation time (Fig. 4.2A). For example, the average Hellinger distances for testing examples (normalized by grid size) were 3.085×10^{-5} and 3.370×10^{-6} for QV4 and KWR, respectively, but KWR approximation was more than 8 times faster (even over a grid) with an average time of 8.690×10^{-5} seconds for QV4 evaluation and 9.870×10^{-6} seconds for KWR approximation (Fig. 4.2A). KWR was both faster and almost an order of magnitude more accurate than generating function solution QV1, which had the larger average evaluation time (normalized by grid size) of 5.389×10^{-5} seconds and the greater Hellinger distance of 2.637×10^{-4} (both normalized by grid size, Fig. 4.2A). While psKWR was slightly less accurate than KWR, its accuracy was still comparable to the generating function

solutions of QV4 and more accurate than QV1 with a 4.469×10^{-5} average Hellinger distance. psKWR even further reduced evaluation time with an average normalized grid evaluation runtime of 5.296×10^{-6} seconds, an order of magnitude faster than QV1 (Fig. 4.2A). This benchmark used 256 testing examples, with Hellinger distances and runtimes normalized by the grid size. While we use Hellinger distance in Fig. 4.2, we show similar trends assessing distribution concordance with KL divergence in Fig. S7. In addition, while these 256 testing rate vectors were sampled within the same bounds as those used to train psKWR and KWR models, we demonstrate in Fig. S8 and Fig. S9 that psKWR and KWR models can extrapolate beyond the training regime to produce accurate approximations for rate vectors sampled outside of training bounds.

Comparison to statistical controls and baseline models

As a control to assess how learning kernel function weights contributes to reconstruction accuracy, we investigated approximations obtained using random weights (RW) $\mathbf{w}$ in Equation 4.1. RW employs the same function $\psi_{KWR}(\boldsymbol{\theta}, n)$ to generate kernel means and variances as KWR, and identically scales variances by learned factor h , but weights kernels with random weights that sum to unity. As shown quantitatively in Fig. 4.2B, RW does not accurately reconstruct joint probabilities, motivating the use of learned weights.

To explore how much accuracy improves when using a neural network to scale means and variances rather than using mere statistical approximations, we also benchmarked against a non-neural, statistical approximation (moment-matched negative binomial or MMNB). The MMNB approach (Section S10) generates statistically plausible means and variances for mature RNA conditional distributions (using the same function $\psi_{KWR}(\boldsymbol{\theta}, n)$). These means and variances parameterize a single negative binomial distribution, which are then multiplied by the analytical nascent marginal probability to produce joint probability approximations. MMNB underperforms KWR and psKWR in Hellinger distance accuracy over a grid, evidence that the neural procedure is necessary to achieve highly accurate approximations.

To assess the utility of kernel functions, we also compared KWR and psKWR to a method of direct approximation that uses a fully connected multilayer perceptron to map from an input $\{\boldsymbol{\theta}, n, m\}$ to a single-valued joint probability $P(m, n; \boldsymbol{\theta})$. The network was trained to reduce the mean squared error between numerically calculated and predicted joint probabilities. We refer to this method as direct

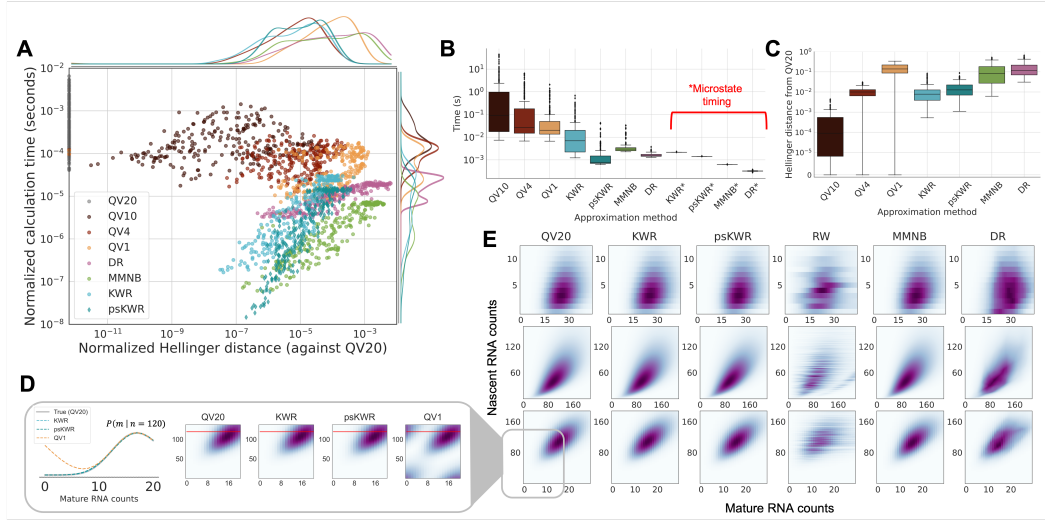


Figure 4.2: KWR and psKWR approximations accurately reconstruct joint probability distributions, with low runtime. **A.** Time to generate joint distributions over grid versus Hellinger distance, both normalized by grid size, for 256 testing rate vectors, comparing three generating function based methods (QV10, QV4, and QV1), direct regression (DR), a moment-matched negative binomial (MMNB) approximation, kernel weight regression (KWR), and parameter scaled kernel weight regression (psKWR) to ground truth (QV20). **B.** Time (unnormalized) to calculate probability for 768 testing rate vectors, where the first seven bars are grid-evaluation and an asterisk indicates method runtime for a single microstate probability. Boxes show the first ($Q1$) and third quartiles ($Q3$) with central lines at the second quartile and whiskers extending to $Q1 - 1.5 \times (Q3 - Q1)$ and $Q3 + 1.5 \times (Q3 - Q1)$. **C.** Hellinger distance (unnormalized) over grid for 768 testing rate vectors. Box plots conventions are as in **B.** **D.** Conditional probabilities and joint probabilities calculated using KWR, psKWR, and QV1 on a grid one standard deviation above nascent and mature means for a single testing rate vector show that significant error (as compared to QV20) is introduced on a small grid by truncation for quadrature methods but not neural approximation methods. **E.** Three example distributions approximated by KWR, psKWR, and several controls (RW, MMNB, and DR) demonstrate qualitative agreement between neural approximation methods and high-quality numerical solutions (QV20) but not all controls.

regression (DR), described in Section S9 with testing results in Fig. S10 and network parameters in Table S5. DR was considerably less accurate than the KWR and psKWR approximations (Fig. 4.2A,B). This highlights that basis function learning methods like KWR and psKWR, which output parametric forms rather than single probabilities, are desirable and efficient for this application.

Evaluation time for microstate probabilities

In the context of statistical inference from data, optimizing likelihoods involves the evaluation of a single microstate $P(n, m; \theta)$. Here, KWR and psKWR outperform generating function-based methods by design by bypassing grid evaluation.

Generating function evaluation over grids can take tens of seconds (QV10, Fig. 4.2B), extremely prohibitive for inference processes in high-throughput genomics in which likelihoods must be iteratively calculated for thousands of genes. Evaluating single microstate probabilities using the neural approximations, by contrast, took 2.170×10^{-3} and 1.397×10^{-3} seconds on average, decreasing evaluation time by four orders of magnitude (Fig. 4.2B) with similar accuracies (Fig. 4.2A,C). While DR and MMNB slightly outperform KWR on microstate timing (Fig. 4.2B), their accuracy is substantially lower (Fig. 4.2A,C). Further, psKWR has evaluation time comparable to DR and MMNB and is more accurate (Fig. 4.2A,B,C).

Joint distribution reconstruction

Some examples of probability distributions obtained by the approximation procedures are shown in Fig. 4.2D, where we again treat QV20 as ground truth. Both KWR and psKWR approximations are qualitatively indistinguishable from ground truth, suggesting that they are appropriate for moderate-precision inference. While MMNB results look qualitatively similar, Hellinger distances from ground truth are substantially greater (Fig. 4.2A,C), suggesting that while it is appropriate as a coarse approximation, it may not be sufficient for precise inference. The random weight control recapitulates the general shape of the distribution, but introduces considerable error into microstate probabilities, confirming that the learned weights are essential for an accurate approximation (Fig. 4.2D, fourth column). Finally, the DR results noticeably distort the distribution shapes, further evidencing the utility of the kernel function approach (Fig. 4.2D, sixth column).

KWR and psKWR use in biophysical parameter inference

KWR and psKWR can be used in methods in an existing maximum likelihood framework [162] to infer biophysical parameters from experimental data. We used both neural approximation strategies and the default generating function solver to infer burst sizes (denoted QV in Fig. 4.2), relative splicing rates, and relative degradation rates for 200 highly expressed genes in GABAergic cells from the mouse primary motor cortex [126].

For the 200 fit genes, we found that inference results using KWR and psKWR approaches for data likelihood evaluation highly correlate with results using the standard generating function approach (Fig. 4.3A), with better defined likelihood landscapes (Fig. 4.3B). Pearson correlations of greater than 0.9 were obtained for $\log_{10}$ of inferred burst sizes, relative splicing rates, and relative degradation rates

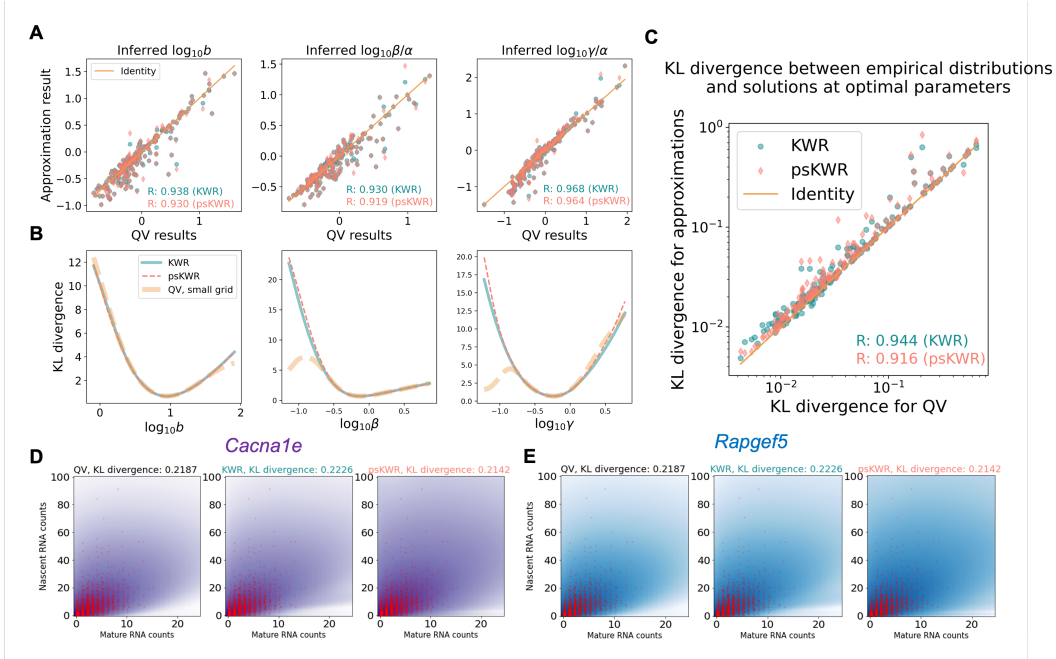


Figure 4.3: Use of KWR and psKWR in inference frameworks results in likelihoods and parameters consistent with those obtained using standard numerical solutions, with better behaved likelihood landscapes. **A.** Parameters for the bursty model of transcription (burst size b , splicing rate β , and degradation rate γ , with splicing and degradation rate considered in units of transcription rate α) highly correlate with results using adaptive quadrature ($N = 200$). **B.** For the gene *Cacna1e*, Kullback-Leibler (KL) divergence between data and biophysical model likelihoods calculated at two of the optimal parameters with the indicated third varied. Likelihood landscapes for biophysical parameters evaluated using KWR and psKWR show one clear local and global minimum, while truncation necessary for numerical solution QV introduces off-optimum decreases in KL divergence. **C.** The log₁₀ of lowest KL divergence obtained between data distributions and neural approximation strategies and data distributions and QV solutions at their respective optimal parameters highly correlate ($N = 200$). **D.** Joint count distributions for *Cacna1e* and **E.** *Rapgef5*, both empirical (red scatter points) and evaluated using QV, KWR, and psKWR (colored density) at the optimal parameters they respectively produced.

for KWR against QV and psKWR against QV (Fig. 4.3A). KL divergence between empirical and model distributions at optimal inferred parameters also agree between QV and both neural approximation strategies (Fig. 4.3C). In addition, distributions produced by all methods capture empirical RNA distributions, with example genes *Cacna1e* and *Rapgef5* shown in Fig. 4.3D and E. These results suggest that KWR and psKWR are reliable alternatives to numerical solutions for parameter inference.

Discussion

After training on a relatively small set of example rate vectors, the KWR and psKWR neural approximation approaches generalize to produce highly accurate and unbounded solutions to a mechanistic model of the lifecycle of RNA in single cells (Fig. 4.2). They are as accurate as lower-order generating function methods, but bypass quadrature and Fourier inversion, achieving orders of magnitude faster runtimes. Thus, these approaches can greatly reduce the amount of time necessary to explore steady-state distributions across a wide range of transcriptional parameters for inference, parameter sensitivity investigations, and experimental design [163]. In particular, we have recently employed KWR application in a popular variational autoencoder framework to infer biophysical parameters for thousands of cells and genes, made possible because KWR, unlike generating function solutions, is compatible with gradient descent algorithms [47, 148].

The current implementations of KWR and psKWR are tailored to the bursty model of transcription. That is, while other approaches, such as [157] and [164], are more flexible in their distributional shapes, able to learn multimodal distributions with unconstrained means and variances, KWR kernel functions' means and variances are constrained based on statistical moments of the underlying system. Moments for psKWR kernel functions, while allowed to vary, are initialized at the preset means and variances, which simplifies and accelerates the learning process. This is one of the key advantages of our methods: we leverage statistical insight about a specific class of systems to develop a reliable solver. While this limits the applicability of the current KWR and psKWR models, the principle behind the algorithms can be extended to other systems about which we have statistical knowledge (e.g., employing a lognormal mixture distribution to find kernel moments for multimodal systems). Thus while a strength of KWR and psKWR algorithms is that results will be regulated or initialized by known system statistics, they may be less suited for systems which are difficult to analyze statistically or that have undefined moments.

A second notable innovation of psKWR and KWR is that they produce bivariate probability approximations for two RNA species, while related methods learn distributions for a single RNA species [157, 164]. They do so by taking advantage of the tractability of nascent distributions and applying the law of conditional probability. This bivariate approach of KWR and psKWR could be adapted for measurements of chromatin accessibility and RNA counts, jointly available in single cells using ATAC-seq [165].

In addition to bivariate systems, KWR and psKWR can be naturally extended for multivariate systems by repeated application of conditional probability: $P(A, B, C) = P(A|B, C) \times P(B|C) \times P(C)$. For example, recent advances in simultaneous quantification [166] and modeling [167] of mRNA and proteins as trivariate unspliced mRNA – spliced mRNA – protein data have been made, but the corresponding stochastic system does not afford an analytical solution. The statistical neural approach of KWR and psKWR could facilitate inference of transcription, splicing, and translation rates using such trivariate data. As single-cell assays with multiple molecular species become increasingly more common, so increases the need for computationally efficient methods for modeling and inference that go beyond statistical summaries to yield mechanistic insight into system dynamics and distributions.

Chapter 5

BIOPHYSICAL MODELS IN VARIATIONAL AUTOENCODERS

In the previous Chapter, we discussed how biophysical models can be efficiently approximated for high-throughput inference of transcriptional parameters at the scale of modern day scRNA-seq data (millions of cells and tens of thousands of genes). Here, we combine this approximation strategy with deep learning architectures for data integration, dimension reduction, and parameter inference. We present *biVI*, which combines the variational autoencoder framework of *scVI* with CME models describing the transcription and splicing kinetics of RNA molecules. We demonstrate on simulated and experimental scRNA-seq data that *biVI* retains the variational autoencoder’s ability to capture cell type structure in a low-dimensional space while enabling genome-wide exploration of the biophysical mechanisms, such as system burst sizes and degradation rates, that underlie observations.

This chapter summarizes [47] by M.C. *, G.G*, Y.C., T.C., and L.P. M. C., G.G. and Y. C. implemented the strategy. M.C. ran analyses. G. G. conceptualized the project and provided detailed feedback and project direction. T. C. and L. P. provided guidance and suggestions throughout. M. C., G. G., T. C. and L. P. wrote the manuscript.

5.1 Single-cell Variational Inference

Advances in experimental methods for single-cell RNA sequencing (scRNA-seq) allow for the simultaneous quantification of multiple cellular species, such as nascent and mature transcriptomes [136, 137], surface [168, 138, 166] and nuclear [169] proteomes, and chromatin accessibility [165]. While these datasets enable insight into cell type and state in development and disease, joint analyses of distinct modalities remain challenging. We show that principled biophysical “integration” of multimodal datasets can be achieved through parameterization of interpretable mechanistic models [99], scalable to thousands of genes across tens of thousands of cells [170].

Recent approaches to integrate and reduce the dimensionality of multimodal single-cell genomics data have leveraged advances in machine learning [149, 171, 172].

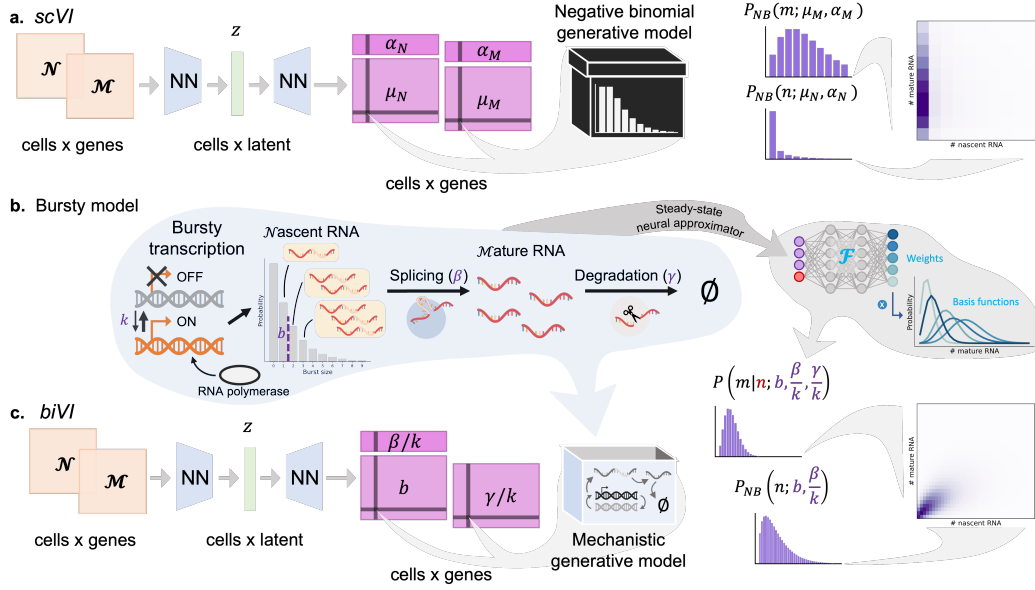


Figure 5.1: *biVI* reinterprets and extends *scVI* to infer biophysical parameters. **a.** *scVI* can take in concatenated nascent ($\mathcal{N}$) and mature ($\mathcal{M}$) RNA count matrices, encode each cell to a low-dimensional space z , and learn per-cell parameters and per-gene parameters for independent nascent and mature count distributions. This is not motivated by a biophysical model. **b.** The telegraph model of transcription: a gene locus has the on rate k , the off rate k_{off} , and the RNA polymerase binding rate k_{RNAP} . Nascent RNA molecules are produced in geometrically distributed bursts with mean $b = k_{RNAP}/k_{off}$, which are spliced and degraded at rates β and γ , respectively. The model’s steady-state distribution can be approximated by a pre-trained neural network $\mathcal{F}$ and a set of basis functions. **c.** *biVI* intakes nascent and mature count matrices, produces a low-dimensional representation for each cell, and outputs per-cell and per-gene parameters for a mechanistically motivated joint distribution of nascent and mature counts.

For example, the popular tool *scVI* is a variational autoencoder that uses neural networks to encode scRNA-seq counts to a low-dimensional representation. This is decoded by another neural network to cell- and gene- specific parameters for conditional likelihood distributions of observed counts [148]. These distributions are chosen *post hoc* to be consistent with the discrete, over-dispersed nature of scRNA-seq counts, but can be derived from biophysical models. Extensions of *scVI* for protein [149] and chromatin measurements [173] jointly encode data modalities to a single latent space, then employ two decoding networks to produce parameters for *independent* conditional likelihoods specific to each datatype. Nascent and mature transcripts, available by realigning existing scRNA-seq reads [136, 137], could be similarly treated (Figure 5.1a). However, using independent conditional likelihoods for bimodal measurements derived from the same gene ignores their inherent causality and has no biophysical basis: the generative model is a “black

box” representation to summarize data.

Nevertheless, good causal model candidates are available (Figure 5.1b). For example, Figure 5.1b illustrates the extensively validated [33, 35, 174] bursty model of transcription. While the joint steady-state distribution induced by the bursty model is analytically intractable [43], we have previously shown that it can be approximated by a set of basis functions with neural-network learned weights [46].

5.2 Adapting Variational Autoencoders for Biophysical Models

We introduce *biVI*, a strategy that adapts *scVI* to work with well-characterized stochastic models of transcription. We propose models, formalized by chemical master equations (CMEs), for RNA lifecycles, then use the bivariate, CME-derived distribution as the conditional data likelihood distribution for nascent and mature counts (Figure 5.1c). The inferred conditional likelihood parameters thus have biophysical interpretations as part of a mechanistic model of transcription, moving beyond associational analyses to fit biophysical values that parameterize causal relationships based on known transcriptional dynamics [43]. The likelihood distributions cannot be obtained solely from the data distributions but require some “knowledge of the data-generating process” [175].

We used simulated data to show that *biVI* models can recapitulate ground-truth distributions, produce latent representations that preserve cell type structure, (see Supplementary Figures 3, 4, and 5 of [47]) and recover marker genes (Supplementary Figure 6 of [47]). Several clustering metrics are slightly lower for *biVI* than for *scVI*, suggesting some trade-off between latent feature representation and improved interpretability (see Supplementary Figures 3, 4, and 5 of [47]). Average inferred parameters were relatively robust to error in nascent versus mature count quantification (Supplementary Figure 7 of [47]). To validate the reliability of *biVI* inferred parameters, we showed that *biVI* inferred burst sizes and relative degradation rates correlate with those obtained in fluorescence assays and time-series labeling data (Supplementary information and Supplementary Figures 8, 9, and 10 of [47]) [154, 176].

After validations, we applied *biVI* to experimental datasets from mouse brain tissue [126], Supplementary Figures 11 and 12 of [47]) for Allen sample B08 using the well-attested [33, 35, 174] bursty model of transcription. *biVI* can recapitulate empirical distribution shapes better than *scVI* (Figure 5.2a-b) as measured by both Kullback-Leibler divergence and Hellinger distance (Supplementary Figure 13 of

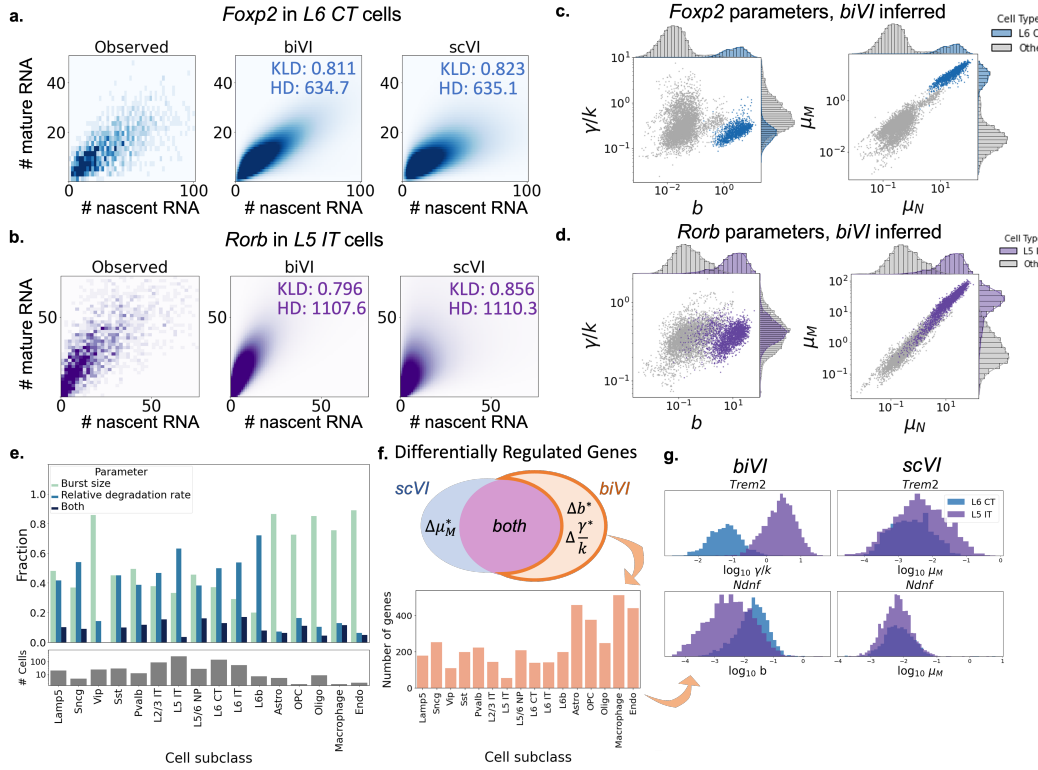


Figure 5.2: *biVI* fits single-cell data from mouse primary motor cortex (Allen sample B08 [126]) and suggests the biophysical basis for expression differences. **a.-b.** Observed, *scVI* and *biVI* reconstructed distributions of *Foxp2*, a marker gene for L6 CT (layer 6 corticothalamic) cells, and *Rorb*, a marker gene for L5 IT (layer 5 intratelencephalic) cells, restricted to respective cell type. **c.-d.** Cell-specific parameters inferred for *Foxp2* and *Rorb* demonstrate identifiable differences in means and parameters in the marked cell types (1,333 L6 CT cells and 2,395 L5 IT cells colored of 6,398 total illustrated cells). **e.** Cell subclasses show different modulation patterns, especially pronounced in non-neuronal cells (top: fractions of 2,000 highly variable genes that exhibit differences in each parameter; bottom: number of cells in each subclass). **f.** *biVI* allows the identification of cells which exhibit differences in burst size or relative degradation rate, without detectable differences in mature mean expression. **g.** Histograms of *biVI* parameters and *scVI* mature means for genes that exhibit modulation in *biVI* parameters (degradation rate in L5 IT cells for *Trem2*, top; burst size in L6 CT cells for *Ndnf*, bottom) but no identifiable mature mean modulation.

[47]), while allowing for interpretation of cell-specific parameters to determine *how* genes are regulated. For example, in Figure 5.2c-d, we illustrate how the upregulation of markers *Foxp2* and *Rorb* can be ascribed to an increase in burst size. Figure 5.2e shows the fraction of genes in each cell subclass that exhibited significant differences in burst size, relative degradation rate, or both. Interesting trends across cell subclasses emerge: neuronal cells appear to regulate gene expression via a mix of regulatory strategies, while non-neuronal cells seem to preferentially modulate burst size.

Finally, *biVI* can identify distributional differences which do not result in mean expression changes (Figure 5.2g). To identify biophysical parameters with statistically significant differences between cell subclasses, we performed Bayes factor hypothesis testing, which accounts for the uncertainty in parameter estimates by comparing multiple samples from cells' latent space posterior distributions (see subsection "Bayes factor hypothesis testing for differential expression" and Supplementary Figure 14 of [47]). Certain cell subclasses had several hundred genes that displayed differential burst size or relative degradation rate but not differential spliced mean expression. For example, the gene *Ndnf*, which codes for the neuron derived neurotrophic factor NDNF, demonstrated a significant difference in the *biVI* inferred burst size, but not *scVI* inferred mature mean, in the neuronal subclass L6 CT (Figure 5.2g, bottom row, with empirical count distributions in Supplementary Figure 15 of [47]). NDNF promotes the neuronal health [177]; characterizing its regulatory patterns could help elucidate its role in neuronal maintenance. As another example, the relative degradation rate of the gene coding for the triggering receptor expressed on myeloid cells-2 (TREM2), was found to be greater in the neuronal L5 IT subclass than in other subclasses (Figure 5.2g, top row, with empirical count distributions in Supplementary Figure 15 of [47]). TREM2 variants are associated with increased risk of Alzheimer's disease [178]: though known to be highly expressed in microglia, finer-grained analyses could uncover cell type specific effects on Alzheimer's development. Such mechanistic description provides a framework for characterizing the connection between a gene's role and a cell's regulatory strategies beyond observational changes in mean expression [131].

We have demonstrated that bivariate distributions arising from mechanistic models can be used in variational autoencoders for principled integration of nascent and mature RNA-seq data. In addition to the bursty model, *biVI* implements constitutive and extrinsic noise models previously discussed in the literature (see Chapter III, Section "Biophysical Models for Stochastic Transcription") [179, 99, 180]. While all three biophysical models are available, one limitation of the current framework is that only one model can be fit to all genes of a input dataset at a time, requiring *post hoc* model comparison (Supplementary Figure 13 of [47]). In addition, sources of technical noise, such as length biases shown to distort nascent RNA capture [181], are not explicitly parameterized in the current version. Finally, while the conditional data likelihood distribution has been replaced by one derived from an interpretable biophysical hypothesis, the neural encoder and decoder still perform opaque mappings between data dimensions. Complementary improvements in in-

interpretability can be achieved using single-layer linear decoders [182] to directly link latent variables with gene parameters via layer weights (Supplementary Figure 16 of [47]). The current *biVI* framework can be naturally extended to model more modalities (e.g., protein counts and chromatin accessibility) and more complex regulatory hypotheses, and, as single-cell technologies expand in scale and resolution, used to build a more comprehensive picture of biophysical processes in living cells.

***biVI* Implementation and Fits**

In order to extend the *scVI* method to work with multimodal molecule count data in a way that is coherent with biology, we define bivariate likelihood functions that (i) encode a specific, precedented mechanistic model of transcriptional regulation and (ii) are admissible under the assumptions made in the standard *scVI* pipeline. On a high level, our method entails the following steps:

1. Choose one of the *scVI* univariate generative models (under ‘*scVI* models’), including the functional form of its likelihood and any assumptions about its distributional parameters.
2. Identify a one-species chemical master equation (CME) that produces this distribution as its steady state, and translate assumptions about distributional parameters into assumptions about the biophysical quantities that parameterize the CME (under ‘Master equation models interpreted for *biVI*’). The one-species system and its assumptions will typically not be uniquely determined.
3. Identify a two-species CME and derive assumptions about parameter values consistent with the one-species system (under ‘Master equation models interpreted for *biVI*’). There will typically be multiple ways to preserve the assumptions but only a single CME.
4. Modify the autoencoder architecture to output the variables that parameterize the CME solution under the foregoing assumptions, and use this solution as the generative model (see ‘*biVI* modifications to *scVI*’).

Statistical preliminaries

We use the standard parameterization of the Poisson distribution:

$$P_{\text{Pois}}(x; \mu) = \frac{\mu^x e^{-\mu}}{x!}. \quad (5.1)$$

We use the shape-mean parameterization of the univariate negative binomial distri-

bution:

$$P_{\text{NB}}(x; \alpha, \mu) = \frac{\Gamma(\alpha + x)}{x! \Gamma(\alpha)} \left(\frac{\alpha}{\alpha + \mu} \right)^\alpha \left(\frac{\mu}{\alpha + \mu} \right)^x. \quad (5.2)$$

We use mean parameterization of the geometric distribution on $\mathbb{N}_0$:

$$P_{\text{Geo}}(x; b) = \left(\frac{b}{b+1} \right)^x \left(\frac{1}{b+1} \right). \quad (5.3)$$

***scVI* models**

A brief summary of the generative process of the standard, univariate *scVI* pipeline is useful to contextualize the options and constraints of the bivariate model. In the Bayesian model, each cell has some posterior probability $p_c(z_c)$ over a low-dimensional space and can be represented as a sample z_c from that posterior. *scVI* uses the “decoder” neural network to map from realizations z_c to quantities ρ_{cg} , which describe the compositional abundance of gene g in cell c as a function of z_c , such that $\sum_g \rho_{cg} = 1$. Furthermore, a cell-specific “size factor” ℓ_c is sampled from a lognormal distribution parameterized by either fit or plug-in estimates of mean and variance such that the mean expression of a gene in a given cell is $\mu_{cg} = \rho_{cg} \ell_c$.

The univariate workflow provides the options of three discrete generative models: Poisson with mean μ_{cg} , negative binomial with mean μ_{cg} and gene-specific dispersion parameter α_g , and zero-inflated negative binomial, with an additional Bernoulli mixture parameter. We report the master equation models consistent with the first two generative laws below, and discuss a potential basis for and reservations about the zero-inflated model in ‘The zero-inflated negative binomial model and its multi-state interpretation.’

Due to the intractability of the posterior probability $p_c(z_c)$, *scVI* uses variational inference to infer an approximate posterior $q_z(z_c)$, which is a multivariate Gaussian. Models are trained via stochastic optimization of the Evidence Lower Bound, or ELBO, which minimizes the Kullback-Leibler divergence between the approximate posterior and a prior and maximizes the expected value of the conditional likelihood over the approximate posterior. The Gaussian form of the approximate posterior makes possible a reparameterization trick to calculate gradients of the ELBO over expectation estimates made by Monte Carlo sampling from the approximate posterior. Further, the encoding network amortizes inference by learning a map between data to parameters of the approximate posterior [171, 148].

Master equation models interpreted for *biVI*

The one-species CMEs encode reaction schema of the following type:



where X is a generic transcript species used to instantiate a univariate *scVI* generative model, γ is the transcript's Markovian degradation rate, and the specific dynamics of the transcription process (first arrow) are deliberately left unspecified for now. Such systems induce univariate probability laws of the form $P(x)$.

The two-species CMEs encode reaction schema of the following type:



where $\mathcal{N}$ denotes a *nascent* species, $\mathcal{M}$ denotes a *mature* species, and β denotes the nascent species' Markovian conversion rate. Such systems induce bivariate probability laws of the form $P(n, m)$. We typically identify the nascent species with unspliced transcripts and the mature species with spliced transcripts. We use the nascent/mature nomenclature to simplify notation and emphasize that this identification is natural for scRNA-seq data, but not mandatory in general.

Formalizing a model in terms of the CME requires specifying the precise mechanistic meaning of ρ_{cg} and ℓ_c . Previous reports equivocate regarding the latter [149], appealing either to cell-wide effects on the biology (in the spirit of [179, 180]) or technical variability in the sequencing process (in the spirit of [183]). For completeness, we treat both cases.

Below, we present the theoretical results, including the biophysical models, the functional forms of bivariate distributions consistent with the standard *scVI* models, and the consequences of introducing further assumptions. The full derivations are given in Supplementary information of [47].

Constitutive: The Poisson model and its mechanistic basis

The Poisson generative model can be recapitulated by the following schema:



where k is a constant transcription rate. This process converges to the bivariate Poisson stationary distribution, with the following likelihood:

$$P(n, m; \mu_N, \mu_M) = P_{\text{Pois}}(n; \mu_N) P_{\text{Pois}}(m, \mu_M), \quad (5.7)$$

where $\mu_N = k/\beta$ and $\mu_M = k/\gamma$. If we suppose each gene's β and γ are constant across cell types, the likelihoods involve a single compositional parameter ρ_{cg} , such that

$$\begin{aligned}\mu_N &= \frac{\gamma_g}{\beta_g} \rho_{cg} \ell_c \\ \mu_M &= \rho_{cg} \ell_c,\end{aligned}\tag{5.8}$$

where $\gamma_g/\beta_g \in \mathbb{R}^+$ is a gene-specific parameter that can be fit or naively estimated by the ratio of the unspliced and spliced averages. On the other hand, if the downstream processes' kinetics can also change between cell types, we must use two compositional parameters:

$$\begin{aligned}\mu_N &= \rho_{cg}^{(N)} \ell_c \\ \mu_M &= \rho_{cg}^{(M)} \ell_c.\end{aligned}\tag{5.9}$$

We refer to this model as ‘‘Poisson,’’ reflecting its functional form, or ‘‘constitutive,’’ reflecting its biophysical basis.

Extrinsic: The negative binomial model and a possible mixture basis

The negative binomial generative model can be recapitulated by the following schema:

$$\emptyset \xrightarrow{k \sim K} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \emptyset,\tag{5.10}$$

where k is the transcription rate, a realization of K , a gamma random variable with shape α , scale η , and mean $\langle K \rangle = \alpha\eta$. This process converges to the bivariate negative binomial (BVNB) stationary distribution, with the following likelihood:

$$P_{\text{extrinsic}}(n, m; \alpha, \mu_N, \mu_M) = \frac{\Gamma(\alpha + n + m)}{n!m!\Gamma(\alpha)} \left(\frac{1}{\alpha + \mu_N + \mu_M} \right)^{\alpha+n+m} \alpha^\alpha \mu_N^n \mu_M^m,\tag{5.11}$$

where $\mu_N = \langle K \rangle/\beta$ and $\mu_M = \langle K \rangle/\gamma$. If we suppose that cell type differences only involve changes in the transcription rate scaling factor η , with constant α , β , and γ , the likelihoods involve a single compositional parameter ρ_{cg} . The mean parameters are identical to Equation 5.8, with an analogous parameter γ_g/β_g , as well as a gene-specific shape parameter α_g . On the other hand, if the downstream processes' kinetics can also change between cell types, we must use two compositional parameters, as in Equation 5.9.

We refer to this model as ‘‘extrinsic’’ to reflect its biophysical basis in extrinsically stochastic rates of transcriptional initiation.

Bursty: The negative binomial model and a possible bursty basis

The negative binomial generative model may be recapitulated by the alternative schema [43]:

$$\emptyset \xrightarrow{k} B \times \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \emptyset, \quad (5.12)$$

where k is the burst frequency and B is a geometric random variable with mean b (Equation 5.3). This system converges to the following stationary distribution:

$$P_{\text{bursty}} \left(n, m; \frac{k}{\beta}, \mu_N, \mu_M \right) = P_{\text{NB}} \left(n; \frac{k}{\beta}, \frac{kb}{\beta} \right) P \left(m|n; b, \frac{\beta}{k}, \frac{\gamma}{k} \right), \quad (5.13)$$

where system moments can be calculated $\mu_N = kb/\beta$, $\mu_M = kb/\gamma$, and α is arbitrarily set to k/β , the shape parameter of the nascent marginal, for simplicity.

Although the nascent marginal is known to be negative binomial, the joint $P(n, m)$ and conditional $P(m|n)$ distributions are not available in closed form. For a given set of parameters, the joint distribution can be approximated over a finite microstate domain $n, m \in [0, S_N - 1] \times [0, S_M - 1]$, with total state space size $S_N \times S_M$. This approach is occasionally useful, if intensive, for evaluating the likelihoods of many independent and identically distributed samples. The numerical procedure entails using quadrature to calculate values of the generating function on the complex unit sphere, then performing a Fourier inversion to obtain a probability distribution [43]. However, this strategy is inefficient in the variational autoencoder framework, where each observation is associated with a distinct set of parameters. Furthermore, it is incompatible with automatic differentiation.

In the previous chapter, [46], we demonstrated that the numerical approach can be simplified by approximating $P(m|n)$ with a learned mixture of negative binomial distributions: the weights are given by the outputs of a neural network, whereas the negative binomial bases are constructed analytically. The neural network is trained on the outputs of the generating function procedure. Although the generative model does not have a simple closed-form expression, it is represented by a partially neural, pre-trained function that is *a priori* compatible with the VAE.

If we suppose cell type differences only involve changes in the burst size b , with constant k, β and γ , we use Equation 5.13 to evaluate likelihoods. These likelihoods involve a single compositional parameter ρ_{cg} , with mean parameters identical to Equation 5.8, with an analogous parameter γ_g/β_g , as well as a gene-specific shape parameter α_g . On the other hand, if kinetics of the degradation process can also

change between cell types, we must use two compositional parameters, as in Equation 5.9. There is no admissible way to allow modulation in the burst frequency.

We refer to this model as “bursty,” reflecting its biophysical basis.

***biVI* bursty generative model**

Following the notation of *scVI* [148], *biVI*’s generative process for the bursty hypothesis models expression values of x_{cn} and x_{cm} of nascent and mature counts, respectively, in cell c as:

$$\begin{aligned}
 z_c &\sim \text{Normal}(0, I) \\
 \ell_c &\sim \log \text{normal}(\ell_\mu, \ell_{\sigma^2}) \\
 \rho_{cg}^{(N)}, \rho_{cg}^{(M)} &= f(z_c, s_c) \\
 \mu_{cn}, \mu_{cm} &= \rho_{cg}^{(N)} \ell_c, \rho_{cg}^{(M)} \ell_c \\
 x_{cn}, x_{cm} &\sim P_{\text{bursty}}(n, m; \alpha, \mu_{cn}, \mu_{cm}),
 \end{aligned} \tag{5.14}$$

with a standard, multivariate normal prior on the latent space z vector. Here, the batch of cell c is denoted as s_c and $\ell_\mu, \ell_{\sigma^2}$ are by default observed mean and variance in log-sequencing depth (‘log-library size’ in *scVI*) across a cell’s batch, although they can be learned. Further, as in *scVI*, f is neural network that produces fraction of sequencing depth parameters $\rho_{cg}^{(N)}, \rho_{cg}^{(M)}$ for nascent and mature counts. The sum of nascent and mature fractions is constrained to be 1 over a cell c by a softmax applied to the network output: $\sum_{n,m=0}^G (\rho_{cg}^{(N)}, \rho_{cg}^{(M)}) = 1$, where G is the number of genes. $\alpha \in \mathbb{R}^G$ is a network parameter jointly optimized across all cells during the variational inference procedure. To recover biophysical parameters, α is arbitrarily set to k/β . Burst size b and relative degradation rate k/γ can be recovered according to the following conversions:

$$\begin{aligned}
 b &= \frac{\mu_{cn}}{\alpha} \\
 \gamma/k &= \frac{\mu_{nc}}{\mu_{cm}\alpha}
 \end{aligned} \tag{5.15}$$

We further set $k = 1$ with no loss of generality at steady-state. Generative processes for constitutive and extrinsic noise models are discussed in Supplementary information under ‘*biVI* constitutive generative process’ and ‘*biVI* extrinsic generative process’ of [47].

biVI* modifications to *scVI

Our code is built upon *scVI* version 0.18.0 [171]; the following outlines the modifications we made for *biVI*. The *scVI* framework already supports the constitutive model. By setting conditional likelihood to “poisson,” no modification of *scVI* architecture is necessary. The conditional data likelihood distribution is the product of two Poisson distributions (Equation 5.7). Explicitly, unspliced and spliced count matrices can be concatenated along the cell axis to produce a matrix of shape C by $2G$, where C is the number of cells and G the number of genes. *scVI* will then produce $2G$ Poisson mean parameters for the two Poisson distributions of Equation 5.7.

For the extrinsic and bursty models, mean parameters for nascent and mature counts, μ_N and μ_M , and a single shape parameter α are necessary. The default *scVI* architecture returns two independent parameters for nascent and mature counts of the same gene. *biVI* thus modifies the *scVI* architecture to update vectors $\alpha \in \mathbb{R}_{\geq 0}^G$ rather than $\alpha \in \mathbb{R}_{\geq 0}^{2G}$, where G is the number of genes. For the extrinsic model, the conditional data likelihood distribution is set to the extrinsic likelihood $P_{\text{extrinsic}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 5.11). For the bursty model, the conditional data likelihood distribution is set to the bursty likelihood $P_{\text{bursty}}(n, m; \alpha, \mu_N, \mu_M)$ (Equation 5.13). These models also intake concatenated unspliced and spliced matrices of shape C by $2G$. We implemented *biVI* in an environment with Python 3.8.10 and PyTorch 1.12.1+cu113 [161].

Preprocessing Allen sample B08

Raw 10x v3 single-cell data were originally generated by the Allen Institute for Brain Science [126]. The raw reads in FASTQ format [184] and cluster metadata [185] were obtained from the NeMO Archive. We selected mouse library B08 (donor ID 457911) for analysis in the main text.

To obtain spliced and unspliced counts, we first obtained the pre-built mm10 mouse genome released by 10x Genomics (<https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest>, version 2020-A). We used *kallisto|bustools* 0.26.0, a wrapper for *kallisto* 0.46.2 and *bustools* 0.40.0, [137] to build an intronic/exonic reference (`kb ref` with the option `--lamanno`). Next, we pseudoaligned the reads to this reference (`kb count` with the option `--lamanno`) to produce unspliced and spliced count matrices. We used the outputs produced by the standard *bustools* filter. This filter was relatively per-

missive: all (8,424) barcodes given cell type annotations in the Allen metadata were present in the output count matrix (10,975 barcodes).

Based on previous clustering results, we selected cells that were given cell type annotations, and omitted “low quality” or “doublet” barcodes [126], for a total of 6,418 cells. Although any choice to retain or omit cells from analysis is arbitrary, our work models the generating process that produced cells’ nascent and mature counts by presupposing each barcode corresponds to a single cell. Therefore, we propose that cells identified as low-quality (empty cells) or as doublets (two cells measured in one observation) [126] have a fundamentally different data-generating process than individual single-cells, and therefore remove them before fitting VAE models. However, we stress that the stochastic nature of transcription and sequencing, the intrinsic uncertainties associated with read alignment, and the numerical compromises made in clustering large datasets mean that previous annotations are not “perfect,” merely a reasonable starting point for comparing alternative methods.

We used Scanpy 1.7.2 [186] to restrict our analysis to the most variable genes, which presumably reflect the cell type signatures of interest. The spliced count matrix for the 6,418 retained cells was normalized to sum to 10,000 counts per cell, then transformed with $\log_1 p$. The top 2,000 most highly variable genes were identified using `scanpy.pp.highly_variable_genes` on spliced matrices with minimum mean of 0.0125, maximum mean of 3, and minimum dispersion of 0.5 [186]. Spliced and unspliced matrices were subset to include only the 2,000 identified highly variable genes, then concatenated along the cell axis in the order unspliced, spliced to produce a count matrix of size 6,418 by 4,000.

Fitting Allen sample B08

We applied *biVI* with the three generative models (bursty, constitutive, and extrinsic) and *scVI* with negative binomial likelihoods to the concatenated unspliced and spliced count matrix obtained by the filtering procedures outlined above. We made the key assumption that unspliced and spliced counts could be treated as the nascent and mature species of the bursty generative model (see discussion in Supplementary information under ‘Connecting the models to transcriptome data’ of [47]). 4,622 cells were used for training with 513 validation cells, and 1,283 cells were held out for testing performance. All models were trained for 400 epochs with a learning rate of 0.01. Encoders and decoder consisted of 3 layers of 128 nodes, and each model employed a latent dimension of 10. For each model, the loss was the Evidence

Lower Bound (ELBO) using a multivariate standard normal prior on the latent space as in standard *scVI* and each described generative likelihood (bursty, constitutive, extrinsic for *biVI* models and two independent negative binomial distributions for nascent and mature RNA for the *scVI* model) as the likelihood term in the ELBO.

Bayes factor hypothesis testing for differential expression

After fitting the VAE models, we sought to identify meaningful statistical differences that distinguish cell types. We excluded cell subclasses “L6 IT Car3,” “L5 ET,” “VLMC,” and “SMC” from this analysis, as they contained fewer than ten annotated cells and may require more sophisticated statistical models to account for small sample sizes. The following analysis thus considers 6,398 cells in 16 unique subclasses. We only computed differential expression metrics under the bursty model.

Differential parameter values were tested for each assigned subclass label (as annotated in [126]) versus all others using a Bayes factor hypothesis test following [149]. We reproduce Equations (18) - (21) of [149] below for clarity.

Estimating differential values of any parameter θ^g of gene g in cells a and b can be done according to the following Bayesian framework. First, as in Equation (18) of [149], the log fold change (LFC) of θ^g between two cells a and b can be calculated as follows:

$$\text{LFC}_{a,b}^g := \log_2 \theta_b^g - \log_2 \theta_a^g. \quad (5.16)$$

Then, as in Equation (19) of [149], the probability that the magnitude of the LFC is greater than some effect threshold T can be found by evaluating $\text{LFC}_{a,b}^g$ over the posterior distributions of each cell:

$$P(|\text{LFC}_{a,b}^g| \geq T | a, b) \approx \int \mathbb{1}(|\text{LFC}_{a,b}^g| \geq T) q_a(z_a) q_b(z_b) dz_a dz_b, \quad (5.17)$$

where, in practice, the integral is approximated with many Monte Carlo samples from the two cells’ posteriors. Two hypotheses are tested: H_1 , or that the magnitude of the LFC is greater than or equal to threshold T , and H_0 , or the null hypothesis that the magnitude of the LFC is less than T . A Bayes factor for gene g between cells a and b ($BF_{a,b}^g$) is calculated to compare the two hypotheses, as in Equation (20) of [149]:

$$\text{BF}_{a,b}^g = \frac{P(|\text{LFC}_{a,b}^g| \geq T|a, b)}{P(|\text{LFC}_{a,b}^g| < T|a, b)}. \quad (5.18)$$

Extending this to test differential expression between two groups of cells A and B amounts to “aggregating the posterior,” as in Equation (21) of [149], or evaluating the same $P(|\text{LFC}_{A,B}^g| \geq T|A, B)$ over

$$\frac{1}{|A|} \frac{1}{|B|} \sum_{a \in A} q_a(z_a) \sum_{b \in B} q_b(z_b). \quad (5.19)$$

In other words, a random sample z_a can be taken from the approximate posterior of any cell belonging to group A and decoded to produce parameter θ_a^g ; likewise a random sample z_b can be taken from the approximate posterior of any cell belonging to group B and decoded to produce parameter θ_b^g . The LFC between the two parameters can then be calculated. Repeating this for many Monte Carlo samples over the aggregate posteriors allows estimation of the Bayes factor between two groups.

For the results shown in Figure 5.1, we used cutoffs of $T \geq 1.0$, or a magnitude LFC of ≥ 2 , and a Bayes factor threshold of 1.5. The Bayes factors were calculated on normalized burst size and means for *biVI*, i.e., the fractional inferred burst size or inferred means (before scaling by sampled sequencing depth for that cell), and normalized means for *scVI*. This controlled for differences in parameters due to sequencing depth that were not biologically meaningful. Relative degradation rate γ/k is independent of sequencing depth: hypothesis tests were performed directly on inferred relative degradation rates. While batch identity can also be integrated over to compare groups of cells from different batches, our analysis did not require this as all cells were from the same batch.

Reconstructing gene distributions

Let $\theta_{\kappa g}$ be mechanistic model parameters for gene g in cell type κ . In simulated data, and for consistency, let us assume it to be true of biological cell types, gene parameters are the same for all cells across a cell type, *biVI* and *scVI* infer unique parameters for every cell and gene: θ_{cg} , where c indexes over cells and g indexes over genes. To reconstruct distributions for a given gene in a specific cell type κ , we sample once from the posterior distribution $q_c(z)$ of each cell $c \in \kappa$ to obtain point-estimates of conditional parameters $\theta_{c\kappa g}$, where conditional refers to a single

sampling from a cell's posterior, or a particular realization of z_c . We then average over the cell-specific conditional probabilities for the gene to produce a cell type marginal distribution:

$$\hat{P}_{\kappa g}(n, m) = \frac{1}{n_\kappa} \sum_{c_\kappa=1}^{n_\kappa} P(n, m; \theta_{c_\kappa g}), \quad (5.20)$$

where n_κ is the total number of cells in cell type κ , and c_κ indexes over all cells in that cell type. This identity follows immediately from defining the cell type's distribution as the mixture of the distributions of its constituent cells. In the case of *biVI*, we plug in Equation 5.7, 5.11, or 5.13 for $P(n, m; \theta_{c_\kappa g})$. In the case of *scVI*, we use a product of two independent negative binomial laws:

$$P(n, m; \theta_{c_\kappa g}) = P_{\text{NB}}(n; \alpha_g^N, \mu_N) P_{\text{NB}}(m; \alpha_g^M, \mu_M), \quad (5.21)$$

where μ_N and μ_M are cell- and gene-specific, whereas α^N and α^M are fit separately and take on different values. For simplicity, this comparison omits uncertainty associated with θ_{cg} , which is formally inherited from the uncertainty in the latent representation z for each cell c .

Thus, Equation 5.21 is an approximation to the posterior predictive distribution, or marginal distribution of data given the approximated posterior, if we assume Monte Carlo sampling from the approximate posterior distributions of cells within that cell type as a reasonable proxy for sampling from the cell type's posterior distribution. The posterior predictive, or marginal, distribution is:

$$P_{\kappa, g}(n, m) \approx \int P(n, m|z) q_\kappa(z) dz$$

where $q(z)$ is the approximate posterior. We further note that conditional data likelihood and the marginal distribution are *not* necessarily of the same form (for example, if the conditional data likelihood distribution is negative binomial, the marginal distribution of genes is not necessarily negative binomial).

Chapter 6

DISENTANGLING *CIS* AND *TRANS* GENE REGULATION

This chapter discusses a different kind of variation: genetic variation, and how we can use single-cell and bulk transcriptomic data to infer gene regulatory strategy differences between homozygous strains that display genetic variability. As Mendel crossed homozygous strains and used their hybrids' features to determine hereditary laws [1], we discuss how hybrid crosses can be used to disentangle the regulatory differences between gene expression in parental strains. Again, setting up a model guided by biological hypotheses, although at a different level of abstraction, provides crucial insight into the mechanisms of regulation. By developing a framework for testing whether strains differ in gene expression due to *cis* (local) or *trans* (distal) effects, it became clear that the model would change if the *trans* regulation acted in a dominant, log-additive (multiplicative), or different way to alter gene expression: setting up a model forced us to more clearly define what we mean by *cis* and *trans* regulation.

This chapter summarizes work from [48]. I.B.H., M.C., L.P. participated in the development of the models and statistical tests, and writing the manuscript. M.C. performed analyses and implemented the method as a software package. I.B.H. proposed the linear transformation and use of GLMs.

6.1 New Axes for New Insight

In 1961, Jacob and Monod developed a theory of gene regulation in which they distinguished local effects (*cis*) from distal regulation (*trans*) [187]. Their work immediately raised the question of the relative contributions of these two regulation modalities [188, 189]. One approach for assessing whether *cis* or *trans* regulation is responsible for differences in gene expression between strains or species is to compare differences in expression of genes in parents to allele-specific differences in F_1 hybrids. This approach was explored in [190], who used crosses of C57BL/6J and CAST/Ei mice to study regulatory mechanisms that could explain differences in gene expression between parental strains. Their approach was developed in [191, 192], who used pyrosequencing to study the regulation differences between *D. melanogaster* and *D. simulans*. With the advent of RNA-seq, genome-wide scans

were possible, and [193] examined RNA-seq from F_1 crosses of C57BL/6J and CAST/EiJ to tease apart *cis* and *trans* contributions to gene expression differences between the parental strains. Similarly, [194] performed such an RNA-seq analysis using *Drosophila* lines.

Formally, the idea of using crosses to study *cis* and *trans* contributions to differences in gene expression between strains or species is as follows: consider a gene with expression X_{P1} in a homozygous strain 1, X_{P2} in a homozygous strain 2, and expression X_{H1} for the haplotype from strain 1 in the F_1 cross of 1 and 2, and expression X_{H2} for the haplotype from strain 2 in the F_1 cross of 1 and 2. Let $R_P = \log_2 \left(\frac{X_{P1}}{X_{P2}} \right)$ and $R_H = \log_2 \left(\frac{X_{H1}}{X_{H2}} \right)$. That is, R_P corresponds to the log2 fold change difference in expression between the two parental strains, and R_H to the log2 fold change difference between the expression of the hybrid haplotypes in the F_1 offspring. The connection between R_P , R_H , and regulation is as follows: consider that a gene can be regulated via *cis*, *trans*, or both (Fig. 1A). Gene expression measurements in the parents and hybrid (Fig. 1B), can be used to infer the nature of regulation underlying the difference in expression in the parental strains (Fig. 1C). Specifically, amending the classification of [192], we have:

- **conserved:** No change in gene expression indicating there has been no change in regulation, i.e. $R_P = 0$ and $R_H = 0$, which implies $R_P - R_H = 0$.
- **cis:** The relative difference in gene expression between the parents is the same as between the haplotypes in the hybrid indicating that the difference in parents is due to local *cis* effects, i.e. $R_H \neq 0$, $R_P \neq 0$, and $R_P = R_H$ which implies $R_P - R_H = 0$, arises from changes only in *cis*-regulatory elements.
- **trans:** Gene expression from the two haplotypes in the hybrid is the same, indicating that differences between the parents resulted from non-local *trans* regulation, i.e. $R_H = 0$ and $R_P \neq 0$, which implies $R_P - R_H \neq 0$, arises from a change only in *trans*-regulatory elements.
- **cis + trans:** $R_H \neq 0$ and $R_P \neq R_H$ with $\text{sgn}(R_H) = \text{sgn}(R_P - R_H)$ arises as a result of change in both *cis*- and *trans*-regulatory elements with changes in *cis* and *trans* contributing to changes in gene expression between strains in the same direction.
- **cis $\times$ trans:** $R_H \neq 0$ and $R_P \neq R_H$ with $\text{sgn}(R_H) \neq \text{sgn}(R_P - R_H)$ arises as a result of compensatory change in both *cis*- and *trans*-regulatory elements with

changes in *cis* and *trans* contributing to changes in gene expression between strains in the opposite direction.

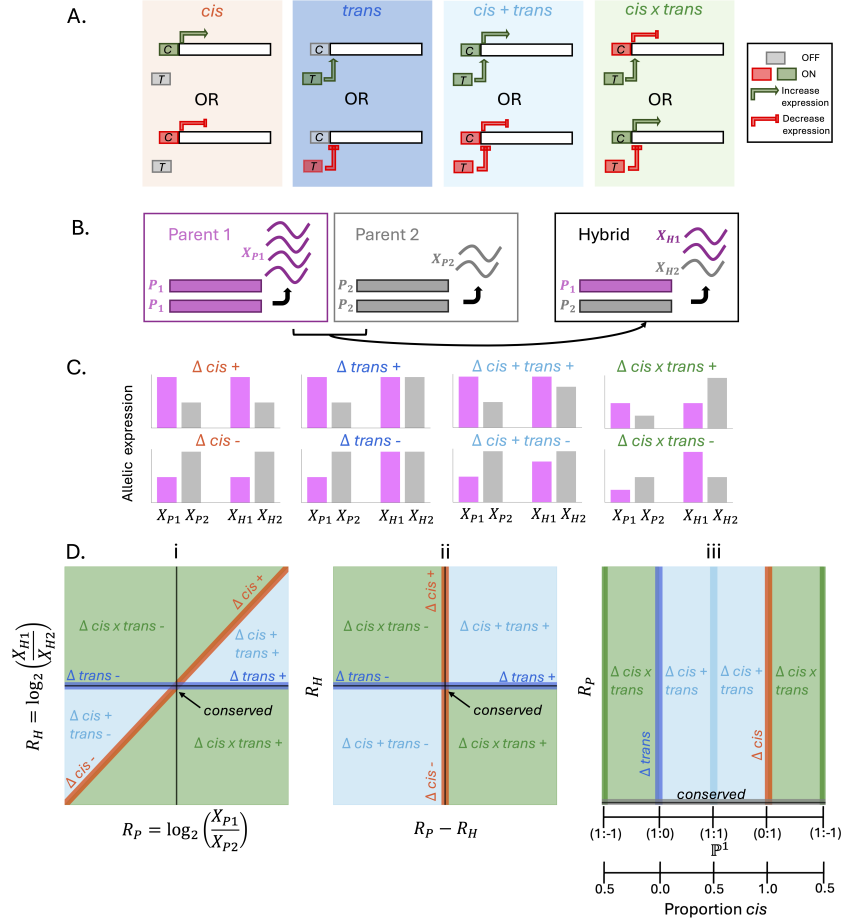


Figure 6.1: Geometry of parental and hybrid expression ratios can be used to assess regulatory differences between homozygous crosses. A) Diagram of the *cis*, *trans*, *cis + trans*, and *cis x trans* types of regulatory differences. B) Homozygous parents with haplotypes P_1 and P_2 , counts of RNA molecules for a gene (X_{P1} and X_{P2} respectively), and the haplotypes in an F_1 hybrid along with counts for the gene (X_{H1} and X_{H2}). C) Differences in regulation between the parents are reflected in distinct ratios between counts X_{P1} , X_{P2} and X_{H1} , X_{H2} . D. i) Illustration of how regulation differences emerge in log-fold changes R_P and R_H , D. ii) Linear transformation of $R_P = \log_2 \left(\frac{X_{P1}}{X_{P2}} \right)$ and $R_H = \log_2 \left(\frac{X_{H1}}{X_{H2}} \right)$ to yield orthogonal *cis* and *trans* coordinates, D. iii) Representation of proportion *cis* in real projective space $\mathbb{P}^1$.

This classification corrects [192], which fails to properly assign regulation differences to *cis + trans* when both $R_H < 0$ and $R_P < 0$. The relationships between R_P and R_H in general, can be visualized as lines and regions in a 2D plot [191], as illustrated in Fig. 1D(i). While this direct representation of R_P and R_H is useful, a quantitative assessment of the gene regulatory modalities reflected in R_P and R_H

requires a biologically meaningful notion of distance between points in Fig. 1D(i). Consider, for example, the situation where $X_{P1} = 2X_{P2}$ and $X_{H1} = 2X_{H2}$, i.e. a 2-fold change in gene expression between the parents due solely to *cis* regulation which corresponds to the point (1, 1) in Fig. 1D(i). The distance from this point to the origin (conserved), is $\sqrt{2}$, whereas the same 2-fold difference in gene expression in the parents due solely to *trans* regulation with $R_P = 1$ and $R_H = 0$ is distance 1 from the origin. This imbalance can be corrected via a linear transformation.

Geometry

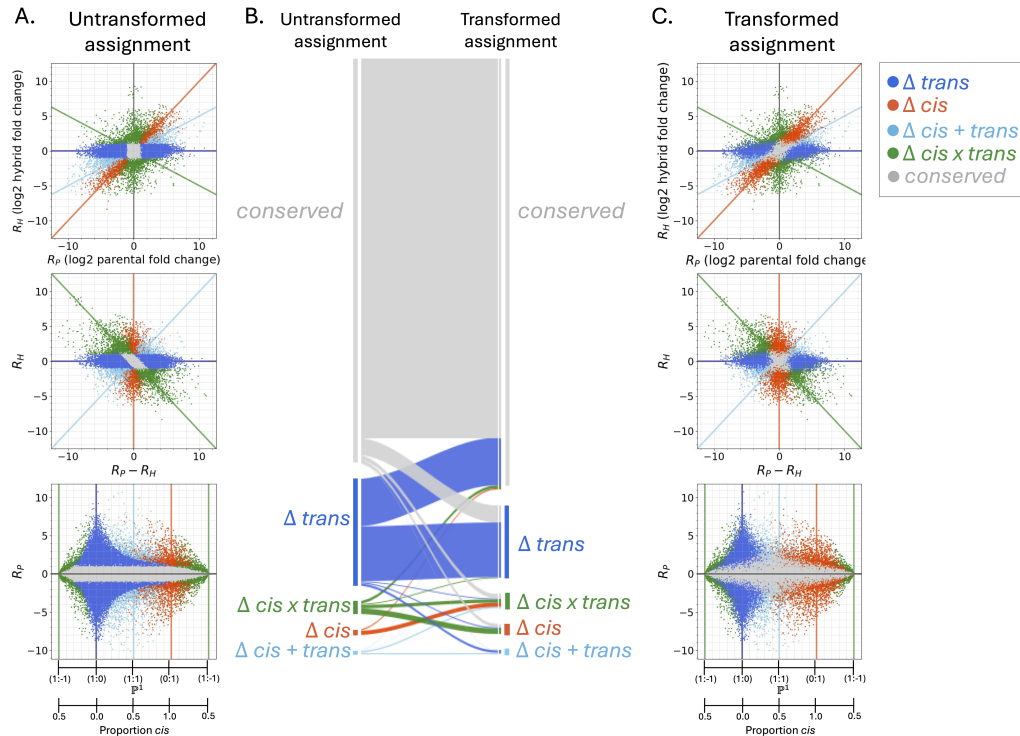


Figure 6.2: Contributions of *cis* and *trans* regulation to differences in gene expression between yeast strains, with data from [195] ($n = 285,777$ gene-cross combinations from 179 unique yeast parent crosses). A) Untransformed ratios, transformed ratios, and log2 of parental fold change versus proportion *cis*, colored by regulatory assignments from [195]. B) Comparison of the results of [195] to the assignment of genes based on the proposed hypothesis testing and geometric assignment procedure. C) Reanalysis of the 285,777 gene-cross expression data from [195] using the proposed hypothesis testing and geometric assignment procedure.

To decouple the effects of *cis* and *trans* regulation on R_P and R_H we begin by noting that if the difference in parental expression is solely due to *cis* regulation, then $R_P = R_H$, or equivalently $R_P - R_H = 0$ (orange vertical line in Fig. 1D(ii)). If the difference in parental expression is solely due to *trans* regulation, then $R_H = 0$

(blue horizontal line in Fig. 1D(ii)). Therefore, the transformation from Fig. 1D(ii) to Fig. 1D(i) is obtained by

$$\begin{pmatrix} R_P \\ R_H \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R_P - R_H \\ R_H \end{pmatrix}. \quad (6.1)$$

Thus, the inverse transformation from the coordinate system in Fig. 1D(i) to the coordinate system in Fig. 1D(ii) is given by

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad (6.2)$$

i.e., the transformation inverts the mapping of the vertical axis to the diagonal *cis* line.

The determination of whether a difference in parental gene expression is due to *cis* or *trans* can now be understood to be an assessment of whether the line passing through the origin and a point $(R_P - R_H, R_H)$ is a perturbation (due to noise in gene expression measurement) of the line Δ *trans*, the line Δ *cis* or sufficiently far away from the axes in Fig. 1D(ii) to merit a designation of Δ *cis* + *trans* or Δ *cis* $\times$ *trans* (the designations are enumerated in Supp. Table 1 of [48]). In other words, the sufficient statistic is a point in real projective space $\mathbb{P}^1$ (Fig. 1D(iii)), and the proportion of the difference in gene expression between parents that can be attributed to *cis* can be understood to be a scaling of the angle of the line through the origin corresponding to the point in $\mathbb{P}^1$, i.e.,

$$\text{proportion } cis = \frac{1}{\pi} \left| \text{atan2}(R_H, R_P - R_H) \right|. \quad (6.3)$$

6.2 Single and Multi-Sample Hypothesis Testing

The determination of whether a measurement (R_P, R_H) reflects a difference in gene expression between parents due to *cis* or *trans* regulation, or both, requires a statistical assessment [196]. Specifically, hypothesis tests can be used to reject a null hypothesis of a difference in gene expression being due solely to *trans* or solely to *cis*. Geometrically, as evident from the linear transformation on log-fold changes, these two tests correspond to testing whether one can reject the null hypotheses that $(R_P - R_H, R_H)$ is located on the *x*- and *y*- axes respectively.

We first developed a framework for hypothesis testing in this case where there are no replicates of the gene expression measurements, as in [195], with two hypothesis

tests (one for *cis* regulation, the binomial test, and one for *trans* regulation, the two sample binomial test), which are outlined below.

Binomial test

To test the null hypothesis that there is no difference between expression of parental alleles in hybrids (or that the difference in regulation is purely *trans*) was performed on rounded integer counts from [195] using the function `scipy.stats.binomtest` [197]. This tests for the probability of having observed a value at least as extreme as k successes given probability p of success and a total of N trials by summing over binomial probabilities:

$$P(k; N, p) = \binom{N}{k} p^k (1 - p)^{N-k}. \quad (6.4)$$

If X_{H1} is the allelic expression in the hybrid of one parental allele and X_{H2} is the allelic expression in the hybrid of the other parental allele, the binomial test was performed with $k = X_{H1}$, $N = X_{H1} + X_{H2}$, $p = 0.5$ and a two-sided alternative hypothesis.

Two sample binomial ratio test

To test the null hypothesis that there is no difference in the ratio of expression of alleles in parents and allelic expression in hybrids (or that the difference in regulation between parents is purely *cis*), we performed a two sample binomial ratio test in which we assumed both X_{P1} and X_{H1} are sampled from a binomial distribution with the same probability of success p_s . Using

$$p_s = \frac{X_{P1} + X_{H1}}{X_{P1} + X_{P2} + X_{H1} + X_{H2}}, \quad (6.5)$$

we calculated the grid of probabilities over possible P_1 and H_1 values:

$$P(X_{P1} = x_{P1}, X_{H1} = x_{H1}; N_P, N_H, p_s) \quad (6.6)$$

$$= \binom{N_P}{x_{P1}} p_s^{x_{P1}} (1 - p_s)^{N_P - x_{P1}} \cdot \binom{N_H}{x_{H1}} p_s^{x_{H1}} (1 - p_s)^{N_H - x_{H1}}, \quad (6.7)$$

where $N_P = X_{P1} + X_{P2}$ is the total number of counts from parents and $N_H = X_{H1} + X_{H2}$ is the total number of counts from the hybrid. We then calculated the probability of having observed values at least as extreme as the observed X_{P1} and X_{H1} .

To account for multiple testing, the significance values can be corrected using the Benjamini-Hochberg method to obtain false discovery rates [198]. In all of the reported results, rejected the null hypothesis for tests with false discovery rates less than 0.05.

Generalized linear models for multiple samples

In the case where replicates have been obtained, a generalized linear model (GLM), as is commonly used for differential expression in bulk RNA-seq [199, 200], can be adapted to utilize gene expression variance estimates within parents and hybrids, providing estimates of *no cis* and *no trans* regulation, both overall and condition-specific.

Our framework differs from previous applications of GLMs for *cis/trans* distinction [201] by virtue of using one GLM for all data (parental counts, hybrid ASE values, with all conditions represented), rather than fitting three separate models to different subsets of the data (e.g., only parental counts for one test or only hybrid ASE values for another). We do this by developing a model with weights for overall and condition-specific *cis* and *trans* regulatory effects, as well as including condition intercepts to properly attribute observed allele expression patterns to regulation or condition. In addition, our approach allows us to propose different hypotheses about the nature of the *trans* regulatory effect: if it is log-additive, dominant, or free (described below). To properly account for the overdispersion and discreteness of RNA-seq counts, we fit a GLM with a negative binomial likelihood function and log link function.

One condition: In the simplest case, we set the geometric mean of parental expression to be the intercept β and include weights for the *cis* regulatory difference (β_C) and *trans* regulatory difference (β_T) between parental strains. This allows the normalized expression from the parental alleles (X_{P1}, X_{P2}) and normalized allelic expression from the hybrids (X_{H1}, X_{H2}) to be modeled under three different hypotheses of *trans* regulation.

The first hypothesis (**log-additive**) assumes that the *trans* expression difference between parental strains at each allele is influenced by both sets of parental chromosomes multiplicatively, and is thus reduced to the square root of the full parental effect in the hybrids. The design matrix is accordingly

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1} \\ X_{H2} \end{bmatrix} = \begin{bmatrix} 1 & -0.5 & -0.5 \\ 1 & 0.5 & 0.5 \\ 1 & -0.5 & 0 \\ 1 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \end{bmatrix}. \quad (6.8)$$

A second hypothesis (**dominant**) assumes that the *trans* expression difference between parental strains is the same regardless of if both sets or one set of parental chromosomes are present: it is the same in the parents and hybrids, and the design matrix is

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1} \\ X_{H2} \end{bmatrix} = \begin{bmatrix} 1 & -0.5 & -0.5 \\ 1 & 0.5 & 0.5 \\ 1 & -0.5 & 0.5 \\ 1 & 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \end{bmatrix}. \quad (6.9)$$

Finally, and most analogous to previous hypothesis testing methods [201], *trans* expression differences between parental strains can have an unconstrained relationship (**free**) with their effects in the hybrids by allowing each hybrid to have individual specific effects (denoted below with the subscript i indexing the hybrid samples). The resulting design matrix is

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1i} \\ X_{H2i} \end{bmatrix} = \begin{bmatrix} 1 & -0.5 & -0.5 & 0 \\ 1 & 0.5 & 0.5 & 0 \\ 1 & -0.5 & 0.5 & 1 \\ 1 & 0.5 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \\ \beta_{H_i} \end{bmatrix}. \quad (6.10)$$

For all models, we can test the null hypothesis of *no cis* differences between parental strains by testing if $\beta_C = 0$ and the null hypothesis of *no trans* differences by testing if $\beta_T = 0$.

Multiple conditions: As more conditions are added to the experimental design, they can be naturally incorporated into the GLM framework, with condition-specific *cis* and *trans* weights as well as intercepts and interaction terms. For example, we fit data for liver and BAT samples for mice grown in warm and cold environments (four separate organ/environment) using the following model under the log-additive *trans* hypothesis:

$$\log \mathbb{E}[X] \sim \beta + \beta_C + \beta_T + \beta_{CO} + \beta_{TO} + \beta_{CE} + \beta_{TE} + \beta_{COE} + \beta_{TOE} + \beta_O + \beta_E + \beta_{OE}. \quad (6.11)$$

The subscripts C , T , O and E refer respectively to differential *cis* regulation between strains, differential *trans* regulation between strains, organ, and environment effects, with combinations denoting their interactions (e.g., β_{CO} models the organ-specific *cis* effect). This is also compatible with any of the three previously described *trans* hypotheses (log-additive, dominant, and free).

Classifying genes based on statistical test results

The results of the two tests for single sample experiments or experiments with replicates are used to categorize genes as outlined in Table 6.1.

Single sample: Binomial test Replicates: $\beta_C = 0$ (null: <i>no cis</i>)	Single sample: Binomial ratio test Replicates: $\beta_T = 0$ (null: <i>no trans</i>)	Classification
Fail to reject	Fail to reject	conserved
Reject	Fail to reject	<i>cis</i>
Fail to reject	Reject	<i>trans</i>
Reject	Reject	<i>cis & trans</i> (<i>cis + trans</i> or <i>cis × trans</i> by geometry)

Table 6.1: Statistical assignments of gene regulatory differences. Genes assigned to *cis*, *trans*, *cis + trans* or *cis × trans* regulation based on two statistical tests, described above for single sample or replicate experiments. If both statistical tests are rejected, regulatory assignment can be made by locating the point $(R_P - R_H, R_H)$ in the transformed space and following the geometric assignments listed in Supp. Table 1 of [48]

Note that some care must be taken when interpreting a *cis* or *trans* assignment within this hypothesis testing framework. For example, a *cis* assignment is made when "no *cis*" can be rejected but "no *trans*" cannot be rejected. Failure to reject "no *trans*" does not imply that one can accept the null hypothesis, but we utilize the assignment of *cis* in this case for simplicity for users. However, this caveat should be taken into account when interpreting the assignments of *cis*, *trans*, and *cis & trans* assignments.

Simulation Study

We also performed genome-wide simulations of parental and F1 hybrid allele-specific RNA-seq count data, generating a total of 6,000 genes per simulated dataset. Genes were randomly assigned to low, medium, or high expression classes with mean counts of 20, 100, or 500. We simulated 45 parameter scenarios, varying biological replication per group (4, 6, 8, 16, and 32), negative-binomial dispersion ($\alpha = 0.02, 0.20, 0.50$), ASE-informative read fraction (0.05, 0.20, 0.40), and absolute $\log_2$ cis and trans effect sizes (0.125, 0.25, 0.5, 1, 1.5, 2, and 3). For each of the 45 scenarios we performed a full analysis on each of the 10 datasets using XgeneR and GLM-based tests for cis and trans effects and also an analysis using parent, ASE, and parent-vs-hybrid ratio tests. Performance was evaluated using power, type I error, ROC/AUC, parameter recovery, and regulatory-class assignment accuracy. Full details of the data-generating model, parameter grid, and diagnostic results are provided in the Simulation Supplement Figures and Simulation Methods of [48].

Application of hypothesis testing

We applied the single sample tests to data from [195], which consist of bulk RNA-seq performed on hybrid and parental strains of genetically divergent *Saccharomyces cerevisiae*. Briefly, [195] generated hybrid crosses from 26 parental yeast isolates derived from diverse environmental conditions. In [195], genes were classified according to the representation shown in Fig. 1D(i). First, both R_P and R_H were tested for statistically significant differences from 0 and assigned "null" (our *conserved*) if both tests failed to reject the null hypothesis (one-sample allele-specific expression tests); then, changes between parental and hybrid allelic ratios were tested for significance (two-sample allele-specific expression tests). For genes that passed significance thresholds, regulatory assignments were made by dividing the untransformed 2D plane into cones (Fig. 2A). We reassigned genes based on our hypothesis tests as derived from the transformed coordinate system (Fig. 2C), with one test with the null hypothesis of *cis* regulation and one with the null hypothesis of *trans* regulation, thereby putting the two regulation strategies on equal footing. Compared to the original study, we found major differences in assignment (of Fig. 2B, Supp. Fig. 3 of [48]).

Interestingly, whereas [195] conclude that "the transcriptome is globally buffered at the genetic level mainly due to trans-regulatory variation in the population", we find that a considerable amount of difference in gene expression can be attributed to

cis (Supp. Fig. 3C of [48]), with the difference due to [195] deriving assignments in the untransformed coordinate system using a statistical testing procedure that treats *cis* and *trans* regulation asymmetrically. Specifically [195] report 57,253 cases where gene expression difference is due to *trans* regulation (Supp. Fig. 3A of [48]). We find 39,063 cases that can be assigned to *trans* (Supp. Fig. 3B of [48]). These numbers are similar; however, [195] report 2,804 cases assigned to *cis* (Supp. Fig. 3A of [48]), whereas we find 7,441 (Supp. Fig. 3B of [48]). Furthermore, we find 2,313 cases assigned to *cis* + *trans* (Supp. Fig. 3B of [48]) versus 1,727 in [195] (Supp. Fig. 3A of [48]). While these discrete classifications can be compared, reflecting the outcomes of statistical tests, there is a range of proportions *cis* for genes within a classification (Fig. 6.2A and C). Notable, the *trans* classifications of [195] extend up to a calculated proportion *cis* of almost 1.0 (Fig. 6.2A), while our assignments are restricted to a region closer to the *trans* axis, explaining the possible over-assignment of *trans* regulation in the original study. While it has been suggested that *trans* regulation accounts for more variation in expression within species and *cis* regulation accounts for interspecific divergence [202], our results indicate that *cis* regulation, with still fewer assigned instances than *trans*, may play a larger role than previously assessed. Overall, there is a marked difference between our results and those of [195] (Supp. Fig. 3C of [48]).

We tested our approach for multiple samples on data from [201], in which two wild-derived mouse strains and their F1 hybrids were used to study the effect of genotype and environment (temperature) on gene expression divergence in rapidly evolving strains. Two inbred lines of house mice were derived from mice from a cold environment (Saratoga Springs, New York, USA) and a warm environment (Manaus, Amazonas, Brazil) [201], and two tissues, brown adipose tissue (BAT) and liver, were evaluated in each of the strains and their F1 hybrids, also raised at the two different temperatures. The experiment was performed in six replicates each of male and female parents, as well as six replicates each of male and female F1 hybrids [201] (Fig. 3A). The GLM framework allows the joint analysis of all samples (parents and hybrids) rather than requiring that separate tests be run on separate subsets of the data (e.g., differential expression between parents, differential expression between hybrid allelic expression, and a ratio test as in [201]). It further enables explicit modeling of specific biological hypotheses of *trans* regulatory action.

Restricting our analysis to a single tissue, temperature, and sex (brown adipose tissue collected from male mice reared in the cold temperature), we fit a GLM with

the free hypothesis and classified genes based on the reduced models of no *cis* or no *trans* regulation (Fig. 3B). Consistent with our single sample analyses, we noted an increase in the number of *cis* assigned genes from 478 in the original study [201] to 877 (Fig. 3C). This trend held in different tissue and temperatures (Supp. Fig. 7 of [48]). We further fit the free model to all male samples, including environment (cold/warm) and organ (liver/BAT) specific *cis* and *trans* weights (Fig. 3D). We identified genes that exhibited significant overall *cis* and *trans* regulatory changes (Fig. 3E, top panel), as well as those that were only significantly different in one organ (Fig. 3E, middle panel) or environment (Fig. 3E, bottom panel). Among these genes were several with functions related to body fat and temperature. For example, *Acad10*, which was identified as having an overall significant *trans* regulatory difference between the strains (Fig. 3E, top panel), codes for Acyl-coA dehydrogenase family member 10. This protein plays a role in lipid metabolism and fatty acid oxidation, and deficiency has been shown to lead to abnormal weight gain and deficiency in glucose tolerance in mice [203]. *Serinc5*, which was shown to have organ specific regulatory differences between parental strains, codes for serine incorporator 5, a protein that incorporates serine into cell membranes to facilitate the synthesis of sphingolipids [204], which themselves have been implicated in thermal adaptation and response to heat across eukaryotic organisms [205, 206]. Notably, *Notch2* was found to have only environment (temperature) specific and not overall *cis* regulatory differences between parental strains, concordant with [207] reporting temperature sensitivity of the Notch signaling pathway underlying species' level differences in development and plasticity in mice. These genes are plausible candidates for differential regulation between mouse strains adapting to life in cold and warm environments. The flexibility of the GLM framework facilitates identification of genes with unique regulatory patterns, such as the tissue and organ-specific effects, a key improvement over post-hoc comparisons of multiple model fits that would be otherwise required.

Proportion *cis*

In addition to naturally revealing the appropriate hypothesis tests to conduct for attribution of gene expression difference in parents to *cis* or *trans*, the linear transformation we propose leads directly to a meaningful measure of the proportion of difference in gene expression that can be attributed to *cis* regulation (or *trans*) (Equation 6.3). To illustrate this, we re-analyzed a dataset of gene expression from human-chimpanzee cell line hybrids [208], calculating the proportion *cis* according

to Equation 6.3. In [208], the proportion *cis* was calculated using slope, as is natural to do when working in the untransformed coordinate framework, whereas the correct calculation (Equation 6.3) uses angle in the transformed coordinate system. While the absolute differences are small (Fig. 4A), with a maximum difference of 0.045, the relative difference is large when the proportion *cis* is small (Fig. 4B), and can be as high as 57%. Moreover, our computation provides a biologically interpretable measure of proportion *cis*. We found several genes in [208] exhibiting high variance in proportion *cis* (Fig. 4C), and identified interesting differences between cell types (Supp. Fig. 8 – 11 of [48]), perhaps relating to cell cycle status or developmental stage of different cell types. For example, *Mageh1* codes for a member a family of proteins that mediate apoptosis, cell growth, and cell cycle [209, 210]: the difference in proportion *cis* across the assayed cell types could indicate different regulatory strategies at different stages of cell cycle, implicating it as an interesting candidate for follow-up experimental validation to explore this question. *Dnmt1* encodes DNA methyltransferase 1, a protein that copies the methylation patterns from template DNA strands onto newly synthesized strands during cell replication [211]: its variation in proportion *cis* across cell types may be evidence of varying regulation between actively proliferating versus quiescent cells, or cell types at different stages of development. These examples show that an analysis of the variation in proportion *cis* across cell types allows the identification of patterns (and the genes displaying them) beyond considering classifications into discrete groups (e.g., *cis* or *trans*).

Discussion

The use of crosses between strains to identify the nature of differential regulation is a powerful tool for genetics studies that is particularly relevant now that single-cell RNA-seq can be used for cell type resolution. Moreover, while original studies were limited to a handful of genes, genome-wide single-cell RNA-seq assays can complement genome-wide eQTL studies.

We have shown that geometric considerations reveal the need for applying a linear transformation prior to visualization. The linear transformation highlights independent axes that lead naturally to hypothesis tests for classifying genes according to the type of regulation underlying differences in gene expression between parents. While we have focused on explaining differences between parents that fall into five categories (conserved, *cis*, *trans*, *cis* + *trans*, *cis* × *trans*), our approach can be extended to finer classifications such as in [195]. We note that in the hypothesis testing framework, referring to a gene as having differences in expression explained

by *cis* or *trans* is technically incorrect. This is because the rejection of the *cis* null hypothesis only shows that the difference in gene expression is not due solely to *cis* regulation. This does not mean that the difference in gene expression can or should be attributed solely to *trans*. The same is the case for rejecting the *trans* hypothesis. In Fig. 2, our coloring of genes as *cis* or *trans* is therefore not precise, but we have done so to facilitate comparisons to previous work.

Our single sample statistical tests depend on the assumption that read counts are binomially distributed, which is standard in the absence of biological replicates. However, when replicate samples are available, we can better assess the extent of technical variation and use generalized linear models. Although our results are shown using commonly applied estimation techniques for negative binomially distributed counts [199], we also measure the extent to which differing estimates of variation change the results (see the Supplement if [48]) [199, 200]. We encourage experimentalists to include enough samples in the experimental design such that heuristic estimations are unnecessary and the observed sample variation can be used for hypothesis testing.

Finally, we note that our framework is general and can be applied to more complex experimental designs, regulatory hypotheses, and phenotypes other than gene expression. For example, the versatility of our GLM framework allows modeling of *cis* or *trans* differences in specific conditions and in interactions between specific conditions. In terms of additional phenotypes, with single-cell RNA sequence data our approach could be used in conjunction with methods such as [212, 47] to assess the regulation mechanisms underlying differences in biophysical aspects of gene transcription, splicing and degradation. Such extensions will be particularly interesting to explore in conjunction with complementary modalities [213, 214].

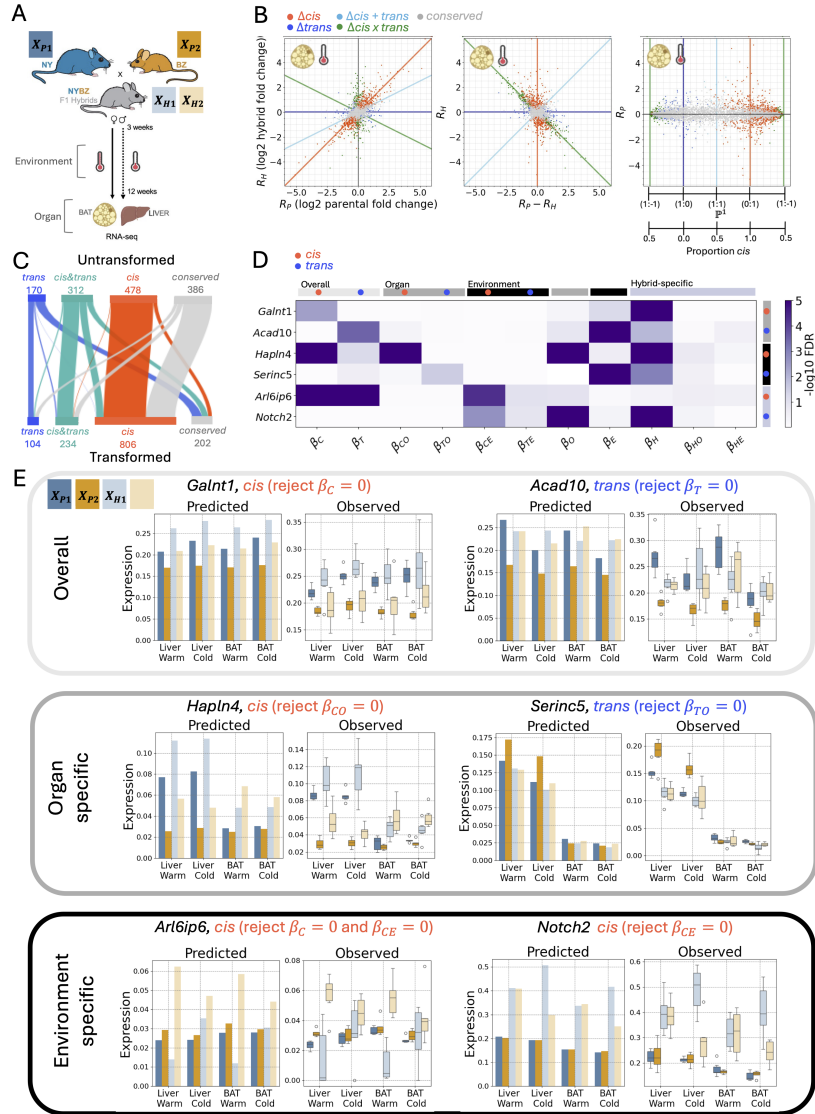


Figure 6.3: Multi-sample regulatory assignment can be determined using generalized linear models and condition-specific hypothesis tests. A) Cartoon of experimental design set up from original study (adapted from Fig. 1C of [201]). B) Model predicted ratios colored by GLM determined regulatory assignments as described in. The free model was fit to brown adipose tissue from male mice reared in the cold ($n = 6$ NY mice, $n = 6$ BZ mice, and $n = 6$ hybrid crosses, and $n = 5,970$ genes). C) Untransformed (original) regulatory assignments (reported in [201]) derived from three independent model fits and statistical tests differ from transformed regulatory assignments derived using the unified GLM framework (genes and assignments from B, excepting genes that were categorized as conserved by both the original and our testing procedure, which are not shown). D) The free model was fit to samples from all four environment/organ pairs (BAT/cold, BAT/warm, liver/cold, liver/warm, $n = 6$ NY/BZ/hybrid samples per condition) with weights for environment and organ-specific regulatory effects. The negative log10 of false discovery rates (FDR) highlights several genes that exhibit significant *cis* or *trans* behavior overall or only in an organ or environment-specific manner. E) For the same genes, GLM predicted (bar plots) and observed gene expression (box plots with counts normalized by the total count per sample for parents and per haplotype for hybrids) for parental and hybrid alleles ($n = 6$ samples per box, with the box showing values' interquartile ranges and a line at the mean and whiskers extending to min/max values).

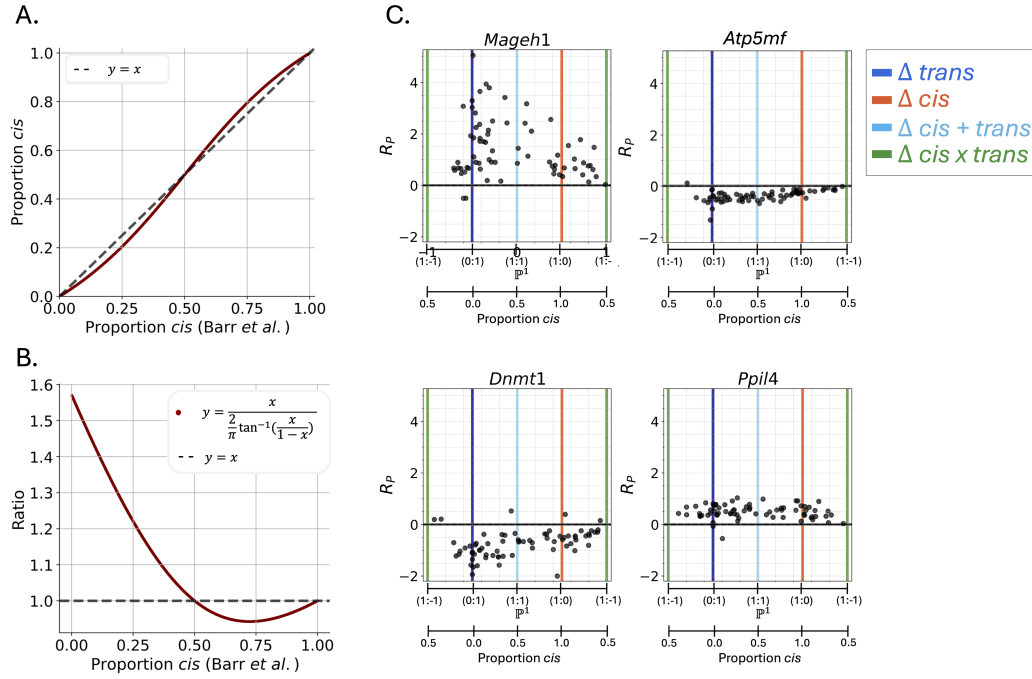


Figure 6.4: In addition to broad regulatory categories, genes can display subtle differences in their proportion of *cis* regulation. A) Comparison of slope (as previously defined [208], x-axis) and angle (y-axis) for determining proportion *cis*. B) Ratio of slope and angle determination of proportion *cis* (dark red line, dashed line displays equality). Note that at small values of proportion *cis*, the slope can be over 50% the determined angle. C) Gene expression data from [208] (human-chimpanzee crossed cell lines) was used to determine cell type specific proportion *cis*. These four genes display high variance in proportion *cis* across cell types (black points are determined proportion *cis* in $n = 72$ cell types).

Chapter 7

REGULATION OF BIOPHYSICAL PROCESSES IN FOUNDER MOUSE STRAINS

In Chapter II, we showed that differences in gene expression *variance* between two groups is as revealing differences in gene expression means. In Chapters III, we showed how the *variation* (in fact, the whole distribution!) of scRNA-seq counts is appropriately modeled with a mathematical framework motivated by what we know about the data generation process, giving insight into the underlying biophysics of the system that produced the them. In Chapters IV and V, we presented strategies for scaling these models to inference loads required for genetic association tests. In Chapter VI, we developed a strategy for determining the contribution of two orthogonal gene regulatory strategies given genetic *variation* in homozygous strains. In this Chapter, we use the theory and tools of the preceding chapters to examine differences in regulation of *biophysical mechanism* between eight founder mouse strains. We show that distinct regulatory patterns underlie transcriptional, splicing, and export rates of RNA, and that expression changes between strains can be attributed to changes in the processing of RNA.¹

7.1 Adapting Modeling Strategies

In this section, we describe how strategies from the previous chapters (biophysical models including technical capture and GLM hypothesis testing framework) as they had been developed for one particular technology and setting had to be adapted for this new study. This demonstrates that analyses of transcriptomic data cannot be "plug and play," but require careful consideration of the experimental technology and values being compared.

The Collaborative Cross Founder Mice

Complex human traits and diseases arise from the interaction of many naturally occurring variants in DNA and environmental factors. To facilitate the study of system genetics, or how variants interact and influence genes and biochemical pathways across contexts, the Collaborative Cross (CC) mouse strains were designed

¹It could be argued that this a more coherent uses of the terms *cis* and *trans* regulation: what does "regulation of a gene" mean but the regulation of the processes that it undergoes?

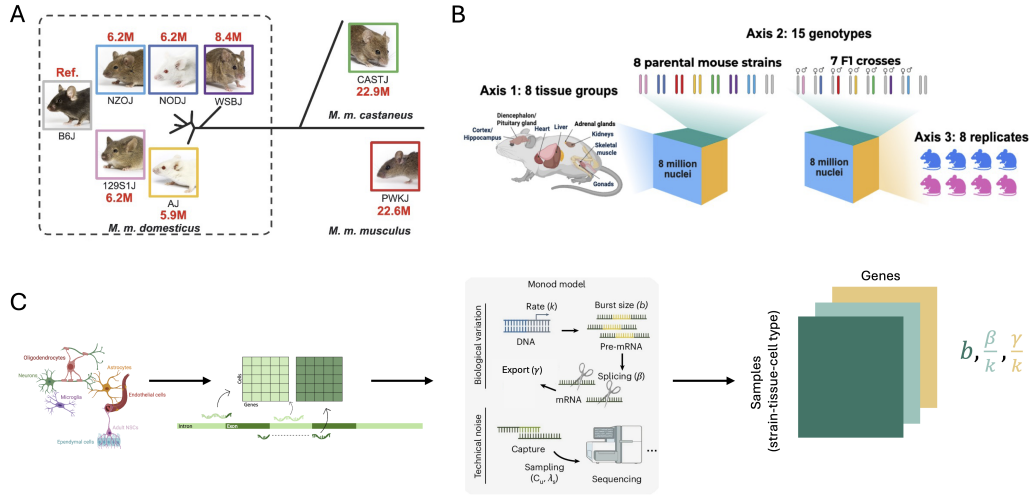


Figure 7.1: An overview of the dataset and fitting procedure. A. The eight genetically diverse founder mice being compared. B. Eight tissues from eight individuals from each of the eight strains and 7 crosses (B6 with all others) were collected and snRNA-seq performed. C. Collected samples (cells) are aligned to a joint reference genome to produce counts for unspliced and spliced molecules, then each unique strain-individual-tissue-cell type was used to fit the a biophysical model to obtain the parameters burst size, relative splicing rate, and relative degradation rate per gene. Images in A. and B. were adapted from [215], and the sequencing step in C. from [37].

[216, 217]. The mouse is the primary model organism for human disease, with high protein-coding sequence similarity to humans and shared mammalian physiology, but even within the species, different mouse strains display different behaviors, phenotypes, and disease predispositions. The CC strains were designed to "optimally" investigate the relationship between genetic variation and complex traits by selectively interbreeding eight founder mouse strains [216, 217]. The eight founder mouse strains were chosen to capture over 90% of the genetic diversity in *Mus musculus* and consist of five classical inbred/lab strains: C57BL/6J (B6J, which is the most commonly used lab strain and the mouse whose genome serves as reference), A/J (AJ), NOD/LtJ (NODJ), 129S1/SvImJ (129S1J), and NZO/H1LtJ (NZOJ); and three wild-derived strains: WSB/EiJ (WSBJ), PWK/PhJ (PWKJ), CAST/EiJ (CASTJ). [216, 217]. These strains range from about 5.9 million (AJ) to 22.9 million SNPs compared to the B6J reference with at least one in 60-90% of genes [215].

To comprehensively characterize the variation in gene expression in these parental mouse strains, a recent publication of a single nucleus RNA-seq dataset (snRNA-seq) assayed 5.2 million nuclei from eight core tissues (heart, gonads, gastrocnemius,

adrenal gland, kidney, diencephalon and pituitary gland, liver, and cortex and hippocampus) from at least eight individuals from each of the eight strains [218]. This revealed extensive transcriptomic variation across the strains at cell type resolution, with positive correlation between transcriptomic variation and strain genetic divergence [218]. The observed gene expression variation in the parents that is due to genetic divergence can be explained by either *cis* or *trans* variants, as discussed in Chapter II, with *cis* regulation attributable to sequence alteration near or within the target gene and *trans* regulation to variants in distal regulatory regions (like enhancers or *cis* variants that affect transcription factors).

To identify if the observed differences in CC founder strain gene expression could be attributed to *cis* or *trans* regulation, seven heterozygous F1 crosses were made with a dam of B6 and each of the other seven founder strains as sires, and snRNA-seq data was collected from the same eight tissues were assayed in eight individuals from each cross (and eight additional B6J samples for technical comparison with the initial study) [215]. The F1 dataset comprises 5,346,886 nuclei (after quality control) from eight tissues classified into 92 distinct cell types [215]. We applied the GLM framework developed in Chapter VI to the total counts of genes summed over unique cell types ("pseudobulk" expression per individual sample) to identify 25,777 genes (91% of detected genes) as being regulated (non-conserved) in at least one of the 92 cell types in one or more crosses [48, 215]. It was found, in agreement with previous work [201, 219], that while *cis* regulation seems to be the primary driver of strain gene expression differences, *trans* regulatory differences displays more cell type and tissue specificity (present in a given cell type or tissue and not in others and changes with context) [215].

While this comprehensive analysis of gene regulatory strategies over the founder mice strain will be of great utility for interpreting expression patterns in the collaborative cross lines derived from them, we can go a step deeper. Pseudobulking and comparing the mean of single-nucleus data obscures the biological signal in the distribution of counts. By applying biophysical models to the snRNA-seq data, we can link regulatory differences in cellular processes.

Aligning to joint reference genomes for allele specific quantification. To obtain allele specific counts of unspliced and spliced molecules, we built a strain specific reference for each of the non-B6J founder strains by incorporating variants from each of the founder mouse https://ftp.ebi.ac.uk/pub/databases/mousegenomes/REL-2112-v8-SNPs_Indels/mgp_REL2021_snps.vcf.gz strain Variant Call Format (VCF)

file [220] into the GRCmm39 (B6J) reference genome using `g2gtoolsv 0.2.0`. We then generated a joint alignment index per non-B6 founder strain by concatenating the strain-specific reference and the unedited GCRmm39 reference and aligned reads using `kb-python 0.28.0` [25]. This process treats genes from B6 and the other founder strains as different "genes." We note that this does not give access to all genes, but only those containing SNPs in regions that reads overlap. If a read is ambiguously assigned (compatible with genes from both strains), it is discarded, and thus not all allele specific expression can be resolved² We pseudo-aligned all data (founder and F1) to the seven joint references using the `--nac` option, which generates nascent, mature, and ambiguous counts per gene [25]. For each cross, we retained reads that mapped to the "correct" allele for the homozygous founder strains (e.g. reads we know came from a B6 mice that mapped to the GRCmm39 copy of the gene, reads we know came from a CASTJ mouse that aligned to the CASTJ-altered reference), and allele specific reads that were consistent with parentage for the hybrids (e.g. for each B6-CASTJ cross F1, we had now two "cells," one set with counts from genes aligning to GRCmm39 and the other with counts aligning to the CASTJ-altered reference). We treated the nascent counts as spliced, and summed the counts belonging mature and ambiguous reads for spliced counts.

Length biased capture of spliced molecules. The CME technical noise model discussed and fit in Chapter III includes a technical capture rate dependent on the length of gene (length-biased capture) for nascent/unspliced molecules [37, 181]. That is, in the Poisson process of capture, each nascent molecule has a rate $\lambda_{Ng} = C_N \times l_g$ of being captured and reported, with l_g being the gene length and C_N a constant:

$$\mathcal{N}'_g \rightarrow \mathcal{N}'_z + \mathcal{N}_z \text{ at rate } \lambda_{Ng} = C_N \times l_g, \quad (7.1)$$

were $\mathcal{N}'_g$ are the unspliced/nascent RNA molecules for a given gene in the cell and $\mathcal{M}_g$ are the observed or captured unspliced/nascent RNA molecules for that gene.

This was motivated by the observation that in 10x Genomics scRNA-seq technologies, which capture mRNA using poly(T) probes, probes sometimes hybridize to internal poly(A) tracts within the gene body. There is thus positive correlation between the number of unspliced molecules and the gene length, as longer genes contain more poly(A) tracts [181]. Inferring parameters for these counts without correcting for this bias leads to estimates of burst sizes that *increased* with gene

²Long-read RNA-seq data would allow better quantification of allele specific expression.

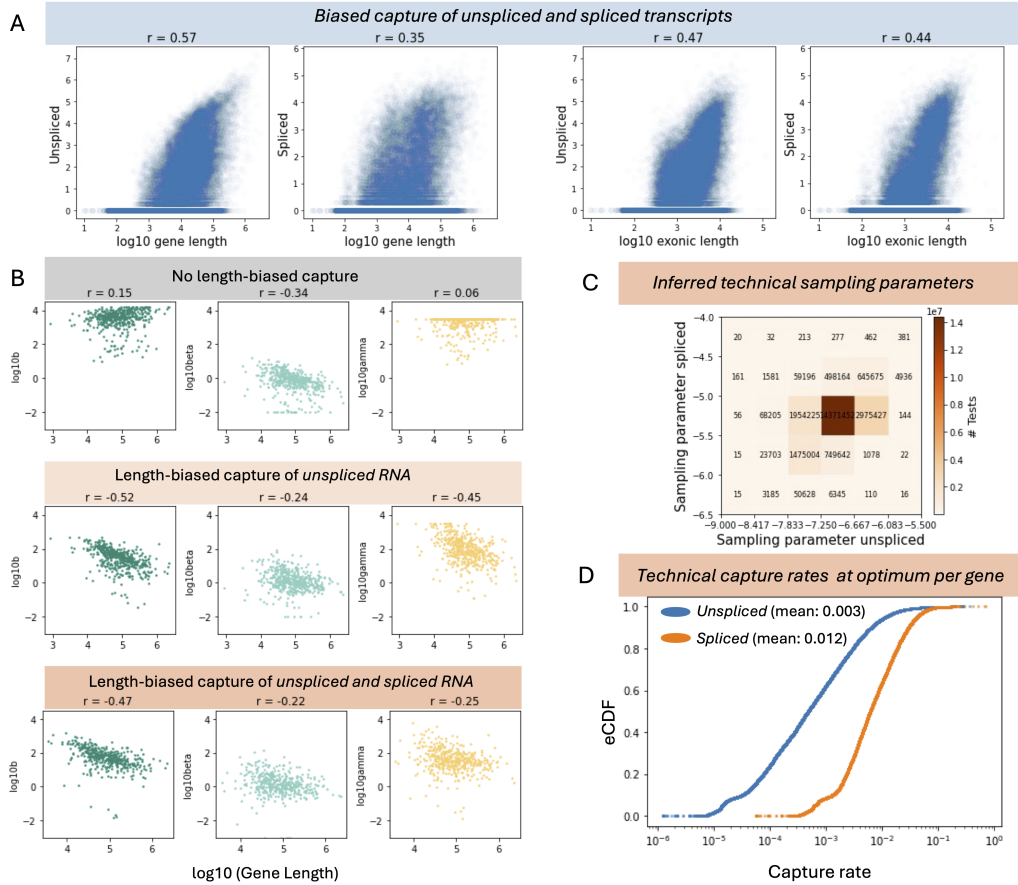


Figure 7.2: A. As an example of length bias in Parse snRNA-seq data, summing over all cells in the adrenal gland, unspliced RNA molecule total counts sum positively with gene length; spliced molecules also showed correlation with exonic length (union of all exons in a gene). B. For fits with no length-biased capture, length-biased capture of unspliced molecules, and length-biased capture of both spliced and unspliced molecules, the correlation of inferred burst sizes ("b"), relative splicing rates ("beta"), and relative export rates ("gamma") for sample "024 B6J 10F 08 B6J" in zona glomerulosa cells in the adrenal gland. C. Over the grid of fit sampling parameters, there was one clear optimum. D. Multiplying the optimum sampling parameter ($\log_{10}-6.9, \log_{10}-5.25$) by gene lengths recovers the inferred per gene capture rate.

length, while previous work suggests the opposite [181, 221]. When length-biased capture is including in the technical noise model in the CME, the positive correlation is removed for more consistent estimations [181, 221].

The snRNA-seq data for the founder and F1 datasets were generated using Parse Biosciences' split-pool barcoding technology (Evercode WT MEga Kits v2), in which cells are put into wells on a plate, pooled and mixed and then re-plated iteratively to generate unique barcodes per cell. This technology includes *two*

primers per capture well, in contrast to 10x capture, one set of poly(T) probes for mRNA poly(A) hybridization and another set of random hexamers that can bind to random (internal or non-coding) locations along RNA molecules. We noted that this introduces a length bias for spliced molecules as well: spliced molecules from genes with longer total *exonic* lengths had longer tracts for probe hybridization, and thus are captured at a higher frequency than shorter molecules 7.2A. To account for this, we introduce a spliced/mature capture rate that depends on the total *exonic* length of the gene (not distinguishing between isoforms, but summing over the union of all exonic content l_{ge}):

$$\mathcal{M}'_g \rightarrow \mathcal{M}'_g + \mathcal{M}_g \text{ at rate } \lambda_{Mg} = C_M \times l_{ge}. \quad (7.2)$$

Here $\mathcal{M}'_g$ are the spliced/mature RNA molecules in the cell and $\mathcal{M}_g$ are the observed or captured spliced/mature RNA molecules.

Examples of parameters inferred without a length-biased capture rate in the model, length-biased capture of only unspliced molecules, and length-biased capture of both spliced and unspliced molecules are shown in Fig. 7.2B. Without length-bias, the inferred parameters were unphysically large (for example, burst sizes around 10^4), or at the bounds of what we set for plausible physicality (7.2B). Including only unspliced gene length-bias correctly induced negative correlations between the inferred burst sizes and gene length ([181, 221]), but also negative correlation between the other two parameters and gene length (Fig. 7.2B). By including both spliced and unspliced length-biased capture, the inferred parameters remained within physically reasonable limits and displayed trends in line with expectations from previous work [181, 221, 222]

Monod fits. We fit *Monod* to all genes in all cell types per individuals in the founder and F1 dataset if the gene had at minimum two spliced and two unspliced counts and an unspliced and spliced mean over that individual's cell type of at least 0.001. We fit over a 6 by 5 grid of technical capture rates to find the one at which genes had highest likelihood.

Overall, we performed 22,890,370 *Monod* gene parameter inferences (each being a unique gene / individual mouse / cell type / tissue combination). As this is data collected from individual nuclei (not cells), the estimated parameters are average burst size b , splicing rate relative to transcriptional initiation rate $\beta = \hat{\beta}/\hat{k}$, and *export rate* relative to transcription rate $\gamma = \hat{\gamma}/\hat{k}$, where hats indicate the underlying rates.

By far the most frequent technical sampling parameter optimum was $\log_{10}(C_N = -6.9, C_M = -5.25)$ (see 7.2C). Multiplying these by gene lengths to recover per-gene technical capture rates, we found that on average the rate of capture for unspliced molecules was 0.003 (or, about .3% of unspliced molecules are captured) and 0.012 for spliced molecules (or, about 1.2% of spliced molecules are captured). This also corresponds with the experimental technology. Since Parse capture includes both poly(T) and random hexamer priming, the spliced molecules with poly(A) tails have higher rates of capture than the unspliced molecules.

GLM testing for *cis* vs. *trans* regulation of biophysical parameters. The inferred parameters are best represented on a $\log_{10}$ scale, and per gene distribution over samples of the $\log_{10}$ scaled parameters appeared approximately Gaussian. We thus adapted the framework discussed in Chapter VI [48] to fit a general linear model with an identity link.

The parameters inferred by fitting on counts estimated in the F1 hybrids are estimates of transcription from a single allele, while parental counts are estimates having transcription from two alleles. If we assume that the rate governing the Poisson process of transcriptional initiation is $\hat{k}$ from each allele in the parents and initiation occurs independently, the overall rate of transcriptional initiation is $2\hat{k}$, and our steady-state inference of splicing and export rates for parents would be half of what we would infer given counts from only one allele. This makes our null hypothesis of no regulatory differences between the relative splicing and export rates representable by the following design matrix for, as one example, the model of dominant *trans* regulation:

$$\begin{bmatrix} \log_{10} \beta_{P1} \\ \log_{10} \beta_{P2} \\ \log_{10} \beta_{H1} \\ \log_{10} \beta_{H2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ \log_{10}(2) & 0 & 1 \\ \log_{10}(2) & 1 & 1 \end{bmatrix} \begin{bmatrix} W \\ W_C \\ W_T \end{bmatrix}, \quad (7.3)$$

where one could replace $\log_{10} \beta$ with $\log_{10} \gamma$. We have changed the notation for regression weights from the previous chapter to W, W_C, W_T to avoid clashing with our notation for the relative splicing rate. However, to account for sample specific differences, we first perform orthogonal distance regression over all genes comparing each sample to one reference sample (as in [37]): this regression accounts for the factor of 2 before we fit *cis* vs. *trans* GLMs, and we thus ran the same free, dominant, and log-additive GLMs described in Chapter VI [48] per parameter, cell type, and

strain. We report results for the free model, as of the three it produced the highest data likelihoods.

7.2 Large-scale Trends and Summarization

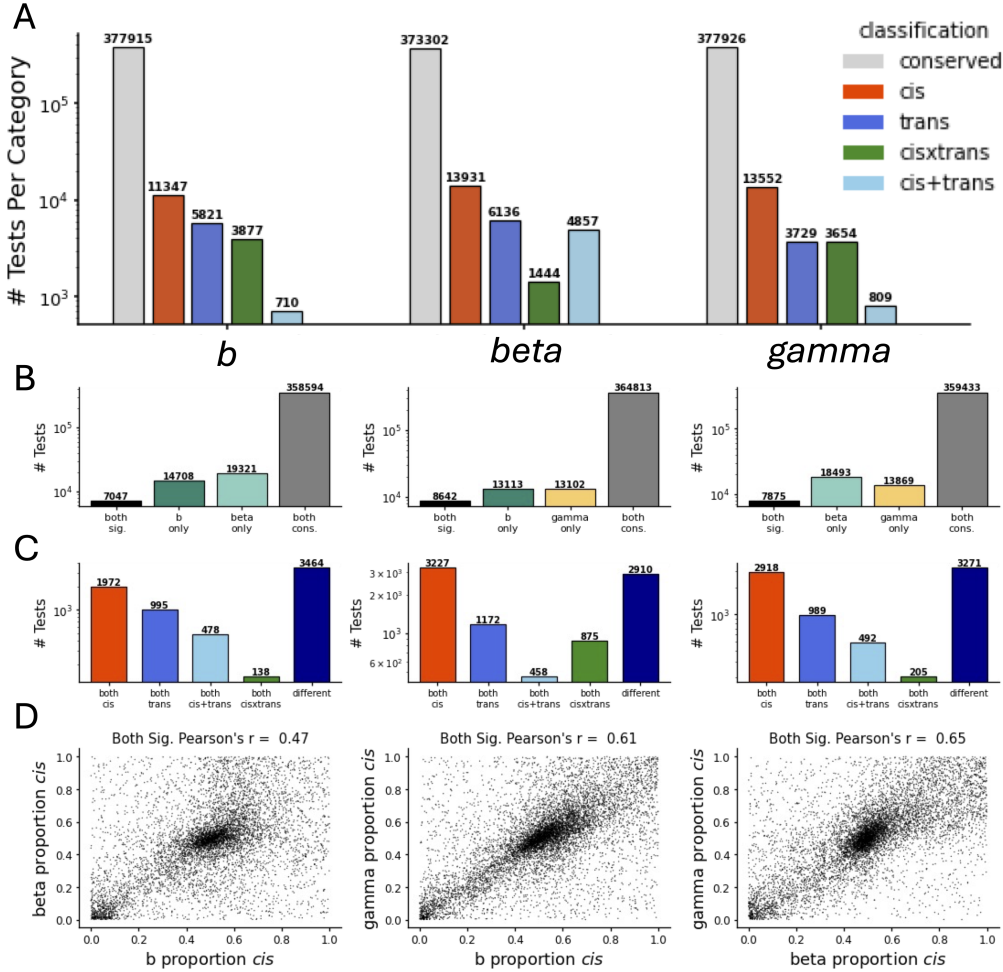


Figure 7.3: A. The number of overall regulatory classifications for burst size ("*b*"), relative splicing rate ("*beta*"), and relative export rate ("*gamma*"). B. The number of instances in which pairwise combinations of parameters (burst size and relative splicing rate, burst size and relative export rate, and relative splicing and relative export rate, from left to right) were found to be both non-conserved/significant, only one or the other significant, or both conserved. C) For the case in which both pairwise combinations were non-conserved, the number of instances in which they were classified as the same regulatory strategy, or different regulatory strategies. D) Of the times the pairwise comparisons of parameters were both non-conserved, the correlation between their inferred proportion *cis*.

In the following section, we refer to each unique parameter-gene-cell type regulatory assignment as a "test." Across all seven strain crosses, eight tissues, 92 cell types and 30,763 genes tested, we found the most instances of regulated relative splicing

(26,368) and a similar number of regulated burst sizes (21,755) and relative export rate (21,744) Fig. 7.1A. Consistent with regulation of total counts, the most frequent category of regulatory strategy for non-conserved tests was *cis*, making up 52.2% of all regulatory assignments for burst size, 52.8% for relative splicing rate, and a greater percent, 62%, for relative export rate. For all parameters, *trans* regulation was the next most common, followed by *cisxtrans* and a much smaller number of *cis+trans* for burst size and relative export rate. The hierarchy of regulatory assignments for burst size and relative export rate are the same as what we observed for total counts [215]. Notably, for relative splicing rate, there were over three times more instances of *cis+trans* regulation (which was very rare when testing total counts [215]). While burst size and degradation rate may be constrained by compensatory *cis* and *trans* activity to buffer overall levels of RNA, the rate of *splicing* is one example of how transcripts can be modified without changing overall quantity. Examining this for individual isoforms or introns could be even more revealing.

Comparing overall patterns of pairwise parameter regulatory assignments for the same gene in a strain-cell type-tissue test, there was little overlap (Fig. 7.1B). For burst size and relative splicing rate, in 14,708 instances burst size was significantly regulated while relative splicing rate was not, in 19,321 instances relative splicing rate was regulated while burst size was not, and only in 7,047 were both significantly assigned. Among those 7,047 instances, in 49.1% they were assigned different regulatory classifications, in 28.0% both were classified as *cis*, 14.1% *trans*, 6.8% *cisxtrans* and 2.0% *cis+trans*. For burst size and export rate, there were similarly more instances of only one or the other having significant regulatory differences between parental strains, although of those a greater percent (66.3%) were the same rather than different. Comparing relative splicing and export rates, there was less of an overlap in assignments than when both were non-conserved (about 58.5%) than burst size and relative export rate (66.3%), but more than burst size and relative splicing rate (50.8%). This suggests that the regulation driving differences in strains' regulatory processes of a gene in a particular context is not singly categorizable, and in fact calls into question the very terminology of "*cis* vs. *trans* regulation of a gene". To say that a gene's *expression* in a cell type is different between strains due *cis* or *trans* regulation is to elide particular regulation of points in the gene expression process.

We observed that the number of significant tests generally increased with the genetic

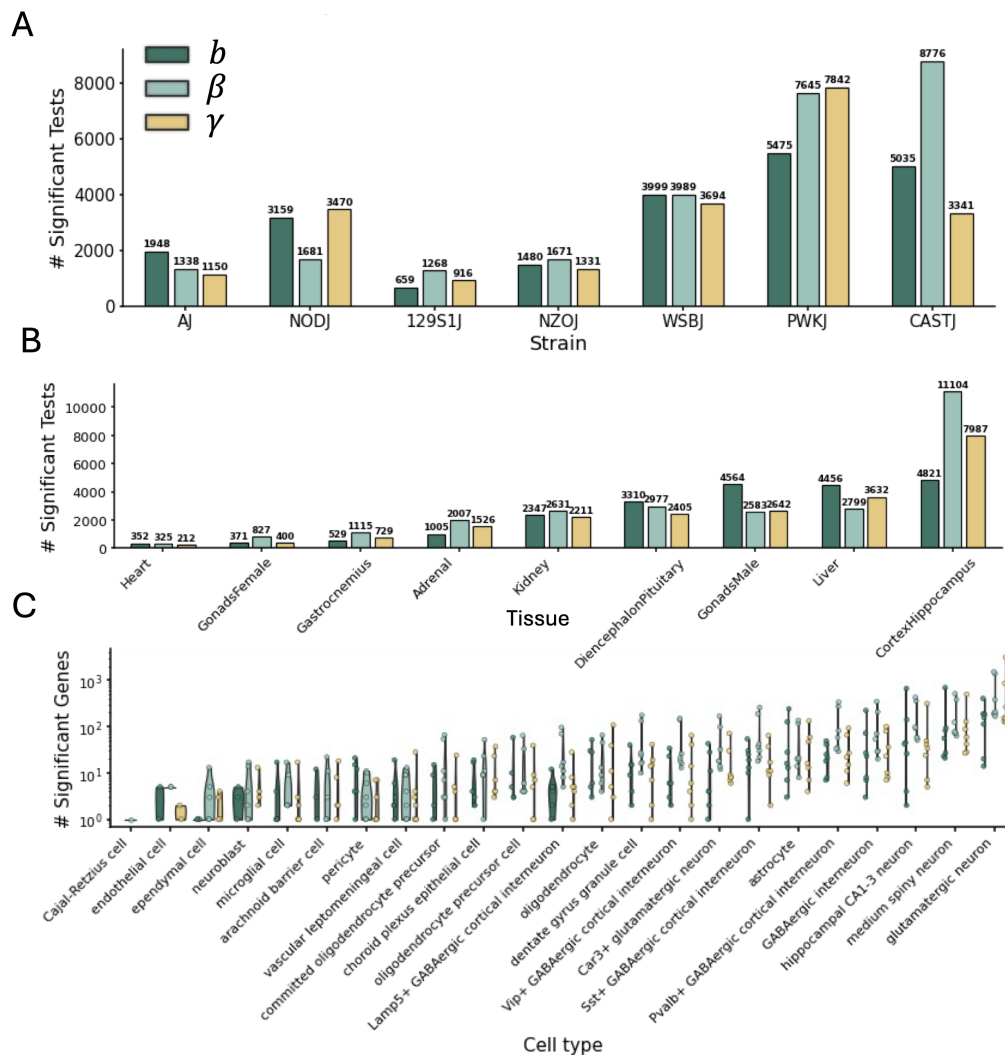


Figure 7.4: The number of overall significant tests for burst size (" b "), relative splicing rate (" β "), and relative export rate (" γ ") for A. different strain hybrids (ordered by increasing genetic diversity), B. different tissues (ordered by the sum of significant tests over all parameters), and C. cell types within the Cortex/Hippocampus, where each dot is a different strain comparison.

diversity of the strain (summing over parameters and tissues), ranging from 2,682 significant tests for B6-AJ crosses (the most genetically similar with 5.9 million SNPs from B6), to 20,962 for B6-PWKJ (22.6 million SNPs from B6) and 17,152 for B6-CASTJ (22.9 million SNPs from B6) (Fig. 7.4A). Among tissues (summing over parameters and strains), we also classified the fewest significant instances of regulatory differences in heart (572) and the greatest in the Cortex/Hippocampus (23,912) (Fig. 7.4B). Within one tissue, we can further break down regulation at the cell type level: in the Cortex/Hippocampus, we found that glutamatergic neurons in

particular to have more non-conserved regulatory differences (Fig. 7.4C).

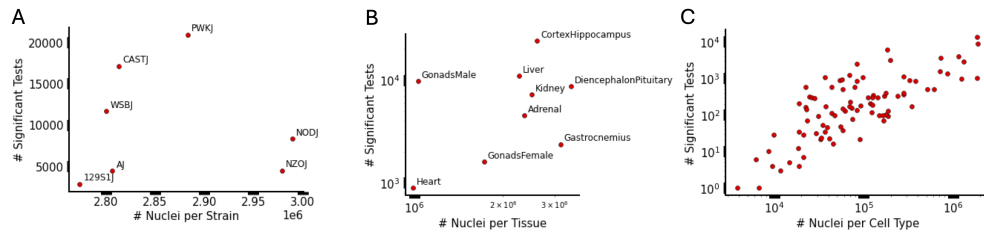


Figure 7.5: The number of overall significant tests (summed over parameters) versus the number of nuclei for A. strains, B. tissues, and C. unique cell types.

We caution in interpreting the biological meaning of the number of significant tests, as "the absence of evidence is not the evidence of absence," and, for the 92 cell types tested in all strains, there was a positive correlation between the number of nuclei in that cell type and the number of significant calls (Fig. 7.5C). Interestingly, the dependence of the number of significant tests on the number of nuclei was not as strong for strain and tissue (Fig. 7.5A and B).

7.3 Exploring the Processes Driving Allele Specificity

Regulation of processes for genes with PWKJ allele specific expression in F1 hepatocytes.

PWKJ mice are prone to liver disorders and are often used as a model to study non-alcoholic steatohepatitis (NASH), a non-alcoholic fatty liver disease [224]. As an example of a full pipeline of discovery for gene processes that may behave differently in different strains in specific contexts, we used the normalized (depth and $\log_{10}$) total counts (summing over spliced and unspliced) per allele in all 7 haplotypes of the F1 crosses to determine those with allelic expression specific to PWKJ alleles in hepatocytes (an analysis a researcher might perform to investigate this question without biophysical models) using an entropy-based metric [223]. These genes indeed showed highest mean expression for the PWKJ allele in F1 hybrids (five are displayed in Fig. 7.6A).

For five of these genes, we explored the regulatory differences between B6 and PWK that could explain the allele specificity (Fig. 7.6B). For example, *Mup2*, which codes for a member of the major urinary proteins whose knockout leads to increased hepatic lipid accumulation [225], had significant regulatory differences between parental strains of all three parameters: burst sizes, regulated in *cisxtrans*, was higher for PWK parents than for B6 parents, as well as for the PWK allele

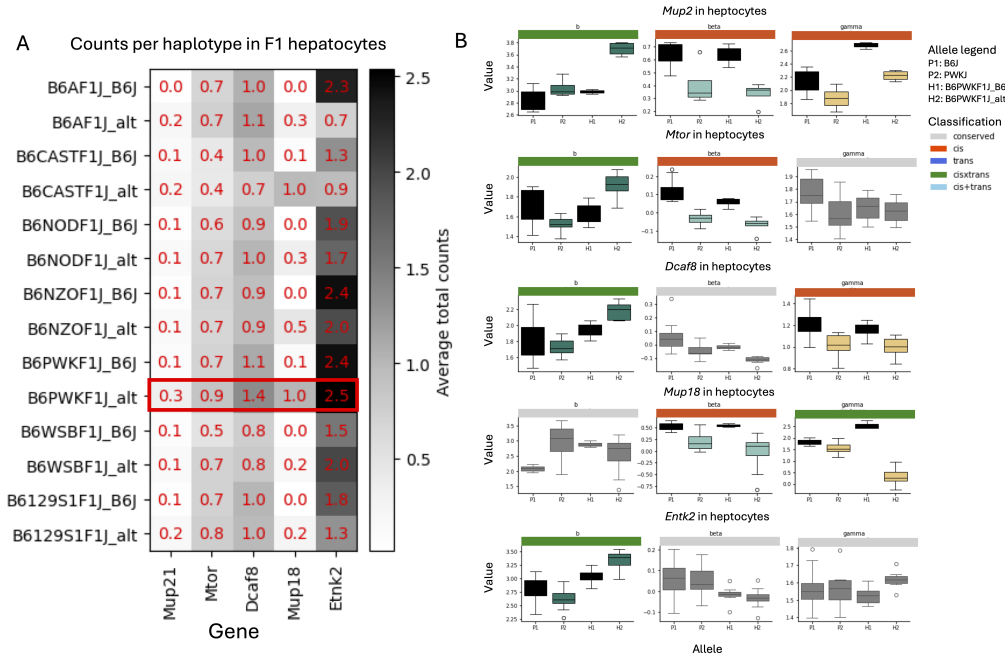


Figure 7.6: Exploring the biophysical basis of allele specific expression of genes in the F1 data in hepatocytes from liver samples. A. Genes whose expression was found to be highly specific to the PWK allele in B6-PWK F1 samples in hepatocytes using normalized total counts show higher mean expression (the average over total counts summed over spliced and unspliced, depth and log₁₀ normalized) [223]. B. The allele specificity can be traced to regulatory differences in the parent strains (B6 and PWK). Boxplots show the mean and quartiles over log₁₀ of burst size ("b"), relative splicing rate ("beta") and relative export rate ("gamma"). Bars above each subplot indicate by color the regulatory classification.

than for the B6 allele in B6-PWK F1 hybrids. Relative splicing rate and relative export rate were smaller for PWK alleles: as they are relative to transcription rate, it may be that the transcription rate (denominator) increased. An increased burst size and/or transcription rate might explain why the *Mup2* transcript shows allele specific expression (Fig. 7.6B). By contrast, *Entk2*, the human orthologue transcript of which was upregulated in gastric cancer patients who showed hepatic metastasis [226], only exhibits significant regulatory differences between B6 and PWK in burst size, and interestingly, the compensatory behavior (*cis+trans*) changed from smaller in PWK compared to B6 to larger for the PWK allele than for the B6 allele in F1 (Fig. 7.6B). This suggests that there was some *trans* difference buffering the *cis* difference between the parents; in the shared F1 environment, the un-buffered *cis* difference led to increased expression of the PWK F1 allele.

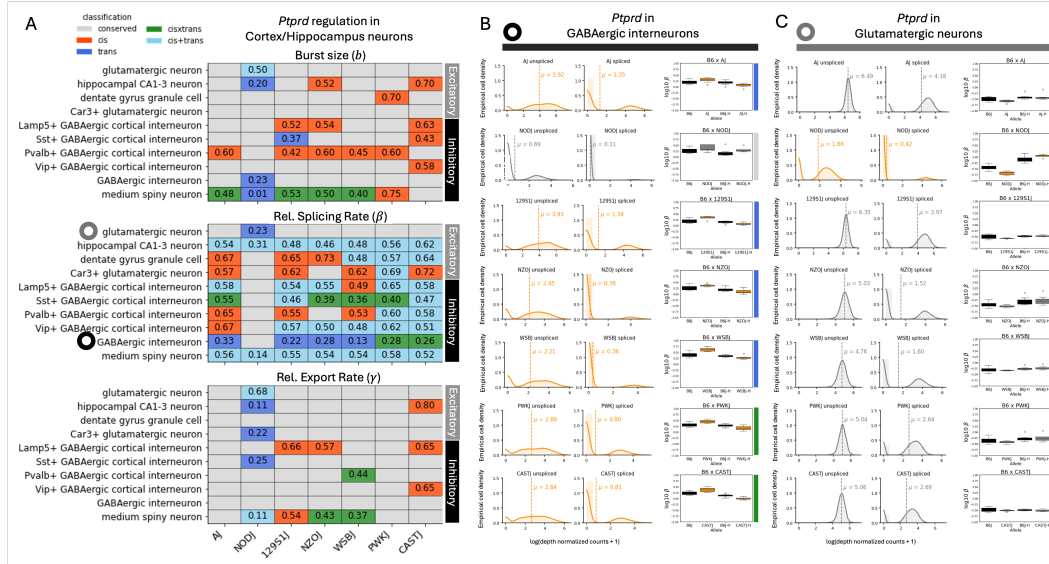


Figure 7.7: *Ptpd* regulation in the Cortex/Hippocampus across the eight founder mouse strains reveals distinct differences in regulation in GABAergic and glutamatergic neurons in the NODJ. A. Classifications of regulatory differences between parental strains (B6 vs. the indicated strain) for burst size, relative splicing rate, and relative export rate show extensive splicing regulation, with NODJ showing distinct patterns. B. Distributions of normalized (depth and log1p normalization) spliced and unspliced counts in GABAergic and glutamatergic show are similar for all six non-B6 founder strains except for NODJ. These differences can be further explored by comparing inferred relative splicing rates for alleles in the parents and hybrid F1s of each strain.

Ptpd regulatory differences in neurons across strains.

Focusing our analysis on the cortex/hippocampus, we found the transcript *Ptpd* had the most significant calls (summing over strains and cell types). *Ptpd* is a receptor type protein tyrosine phosphatase which contains an extracellular domain, a single transmembrane segment, and intracellular catalytic domains, with roles in brain development and synapse formation [227]. In humans, variants at the *PTPRD* locus have been associated restless leg syndrome, Alzheimer's disease, obsessive-compulsive disorder, intellectual disabilities, and attention-deficit/hyperactivity disorder (ADHD) [227, 228]. *Ptpd* in mice and *PTPRD* in humans are known to be regulated by extensive, cell type specific alternative splicing [229, 230], and multiple studies have shown that knock-down and knock-out of *Ptpd* in B6 background mice led to an increase in excitatory neurons in the cortex and decrease of sociability [231, 229].

Consistent with extensive regulation by splicing, we found that in all 10 neuron cell types in the brain, across strains, there was significant regulatory differences in relative splicing rate for *Ptpd* in 55 out of the 70 tested cell type-strains (Fig. 7.7A).

Particularly interesting was the regulation in excitatory glutamatergic neurons. Although glutamatergic neurons overall had the greatest number of significant tests (Fig. 7.4C), all six founder strains besides NODJ showed no significant regulatory differences compared to B6 for *any* parameter (Fig. 7.7A). NODJ, however, displayed significant regulatory differences for *all* parameters (Fig. 7.7A). Although *Ptprd* expression has not been studied in NODJ mice, the NODJ strain was shown to display low sociability in behavioral tests, as opposed to B6 mice which showed significantly higher levels of preference for sociability [232] (the other six founder strains were not included in the study). This finding could indicate NODJ and the *Ptprd* gene in glutamatergic neurons as good candidates to consider as mouse model, target gene and context for studies on behavioral (ADHD, OCD, autism-spectrum) disorders in humans.

Discussion and Limitations

We have shown that genome-wide inference of the regulation of RNA processes between genetically diverse mouse strains can be achieved using high-throughput transcriptomic data. We recapitulated known results (splicing of *Ptprd*) as well as suggested particular processes of genes as being regulated between B6 and PWK founder strains in hepatocytes in the liver. While we focused on two examples, we expect a deeper analyses of the thousands of fit genes, cell types, strains, and tissues (Fig. 7.1 and Fig. 7.4) would yield harvests of rich insight; and we are excited to provide these inferences as a resource for the mouse biologist.

This analysis is, of course, subject to a number of limitations. Firstly, considering the low capture rates per gene (Fig. 7.2C), certainly not every RNA molecule has been accounted for; but, as the same technical capture rate was inferred for parental and hybrid alleles in the F1s, we are at least reasonably confident in the assumption that the capture technology is not biased towards one or the other allele. Secondly, alignment of reads to obtain allele specific counts using short-read snRNA-seq requires that 1) there be enough SNPs between strains to distinguish between alleles, and 2) the sections of the transcript that were captured align to the region of genes that contain those SNPs. More accurate would be fluorescent studies that can label and watch real-time RNA production from two different alleles [233]. However, labeling and obtaining live-cell fluorescent images is far more intensive than performing a snRNA-seq experiment: our analysis depends on collecting enough samples from two parental strains and their hybrids to have the

statistical power to make a claim about regulation, which would be prohibitive if conducting fluorescence experiments.

Another limitation is the primitive nature of our "splicing" model. With longer reads, we could more accurately assess isoform expression, and include isoform specific or intron specific rates of splicing in the model. However, this is more an experimental than a theoretical limitation (updating the CME model for improved capture technology is straightforward). Another model limitation is that we are examining genes independently and not considering gene interactions. While interactions in and of themselves should not affect the logic we are testing, they could provide insights into which genes are likely to be *cis* vs. *trans*, as well as the possible *trans* acting factors, that could refine our current inferences.

Finally, we currently test the regulatory strategy of each biophysical parameter separately, but in future will change the design matrix to fit shared and independent effects for different parameters (this would be more appropriate, as we already know that the relative splicing and export rates share a denominator).

Despite the limitations, this study demonstrates an advancement in the challenge of understanding how genetic diversity results in regulated gene expression. The regulation of a gene occurs via regulation of specific processes, and it is both more precise conceptually and more biologically informative to test for regulatory differences in these processes rather than in the mean counts. While we have probed regulatory processes from the angle of genes, finding evidence of regulation but not the casual genetic variants themselves, increased sample sizes could be used to associate variants with the changed process in particular cell types [29, 41]. We expect that higher resolution data could more precisely pinpoint *how* variation in the genome changes molecular processes in the cell.

Chapter 8

CONCLUSION AND PERSPECTIVES

*[We were] to be redeemed by the
gift of arranging words / but must
be prepared for an earth without
grammar.*

CZESŁAW MIŁOSZ

We live in an exciting time for biology, a time of revolutions in data collection (a multitude of ‘omics) and computational advancements.

A brief summary

In Chapter I, we reviewed the journey from horticulture [1] to the sequencing of the human genome [8] and the explosions of ensuing studies [29, 18, 234, 235, 31, 10] in humankind’s ambitious drive to understand how life works. We showed in Chapter II that mathematics published over a hundred years ago still has something fresh to teach us about high-dimensional data analysis [236, 50, 51]. In Chapter III, we described some fairly simple stochastic models of transcriptional processes, and how they can be powerful tools for biological discovery when fit to scRNA-seq data [99, 37]. In Chapter IV, we presented methods for approximating solutions to these models when they become slightly more complicated [46], and in Chapter V included these approximations in a variational autoencoder framework for scalable inference of biophysical rates in single cells [47]. In Chapter VI, we used RNA-seq data from homozygous crosses to establish the presence or absence of distal and proximal gene regulatory differences. In Chapter VII, we demonstrated that the two lines of work are compatible and, having fit tens of thousands of genes for transcriptional parameters, interrogated the differences in regulation of these processes between founder mouse strains at tissue and cell type resolution.

A brief look ahead

Given the expanding number of molecular modalities assayed in single cells (including proteins [237]; chromatin accessibility [139, 165] and modifications [238])

and perturbations to these cells [239, 240], it is exciting to think ahead towards what we can learn about cellular processes and their genetic basis by analyzing such data from the biophysical perspective taken in this work. While inroads have been made for some analytical solutions [241, 242], as the complexity of proposed interactions (for example, gene networks) and the number of modalities grows, combining models and machine learning approximations trained on data from fast stochastic simulation algorithms [243] could be a fruitful way forward.

Complementary to the efforts of building predictive foundation models for molecular components of the cell, and combining them to build multi-scale and multi-model AI "virtual cells" [244], will always be efforts to use all available modeling tools and computational resources to understand what is happening in the "irl" cells still being assayed.

A brief experiential comment

A model is a representation of thought, forcing us to precisely define ideas and measure them against experimental data. Whether it is describing stochastic transcription, tracing RNA over the course of its lifecycle, or setting up a design matrix for modeling allelic expression from homozygous crosses, encoding hypotheses within mathematical frameworks formalizes our conception of biological processes. It further allows us to interrogate these assumptions, tweaking models that do not describe observations well, and conduct new experiments to test updated modes of thought. In the experience of my Ph.D., it did not take models with millions of parameters to make me question assumptions I had about a fundamental regulatory process, but a simple GLM with three or four learned weights and non-trivial implications. This is not to say that the complexity of biology can be contained in what we decide to write down or draw out about it; but the model is for the scientist, not for biology, and it is an interesting future challenge to maintain what modeling does for our thought processes while, with the advantages of deep learning, freeing them from the constraints of our limited hypotheses to expand their scope.

BIBLIOGRAPHY

- [1] Gregor Mendel. “Versuche über Pflanzen-Hybriden”. In: *Verhandlungen des naturforschenden Vereines in Brünn* 4 (1866), pp. 3–47.
- [2] Friedrich Miescher. “Ueber die chemische Zusammensetzung der Eiterzellen”. In: *Medicinisch-Chemische Untersuchungen* 4 (1871). Published in Hoppe-Seyler’s *Medicinisch-chemische Untersuchungen*, pp. 441–460.
- [3] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. “Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III”. In: *Journal of Experimental Medicine* 79.2 (1944), pp. 137–158.
- [4] Francis Crick. “Central dogma of molecular biology”. In: *Nature* 227.5258 (1970), pp. 561–563.
- [5] Walter Fiers, Roland Contreras, Fred Duerinck, Guy Haegeman, Dirk Iserebant, Jozef Merregaert, Willy Min Jou, Francis Molemans, Alex Raeymaekers, Alfons Van den Berghe, Guido Volckaert, and Marc Ysebaert. “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene”. In: *Nature* 260.5551 (1976). PMID: 1264203, pp. 500–507. DOI: 10.1038/260500a0. URL: <https://doi.org/10.1038/260500a0>.
- [6] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. “Nucleotide sequence of bacteriophage ϕ X174 DNA”. In: *Nature* 265 (1977), pp. 687–695. DOI: 10.1038/265687a0. URL: <https://www.nature.com/articles/265687a0>.
- [7] The *C. elegans* Sequencing Consortium. “Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology”. In: *Science* 282.5396 (1998), pp. 2012–2018. DOI: 10.1126/science.282.5396.2012. URL: <https://www.science.org/doi/10.1126/science.282.5396.2012>.
- [8] International Human Genome Sequencing Consortium. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921. DOI: 10.1038/35057062. URL: <https://doi.org/10.1038/35057062>.
- [9] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (2022), pp. 44–53. DOI: 10.1126/science.abj6987.

- [10] Wen-Wei Liao, Mazloomi Asri, Jana Ebler, et al. “A draft human pangenome reference”. In: *Nature* 617 (2023), pp. 312–324. DOI: 10.1038/s41586-023-05896-x.
- [11] The International HapMap Consortium. “The International HapMap Project”. In: *Nature* 426 (2003), pp. 789–796. DOI: 10.1038/nature02168.
- [12] The International HapMap Consortium. “A second generation human haplotype map of over 3.1 million SNPs”. In: *Nature* 449 (2007), pp. 851–861. DOI: 10.1038/nature06258.
- [13] Richard J. Klein, Caroline Zeiss, Emily Y. Chew, Jean Y. Tsai, Richard S. Sackler, Carol Haynes, Anna K. Henning, John P. SanGiovanni, Shrikant M. Mane, Scott T. Mayne, et al. “Complement factor H polymorphism in age-related macular degeneration”. In: *Science* 308.5720 (2005), pp. 385–389. DOI: 10.1126/science.1109557.
- [14] The 1000 Genomes Project Consortium. “A global reference for human genetic variation”. In: *Nature* 526 (2015), pp. 68–74. DOI: 10.1038/nature15393.
- [15] Majid Nikpay et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. In: *Nature Genetics* 47 (2015), pp. 1121–1130. DOI: 10.1038/ng.3396.
- [16] Gabriel E. Hoffman, Jaroslav Bendl, Georgios Voloudakis, Kelsey S. Montgomery, Laura Sloofman, Ying-Chih Wang, Hardik R. Shah, Mads E. Hauberg, Jessica S. Johnson, Kiran Girdhar, Lingyun Song, John F. Fullard, Robin Kramer, Chang-Gyu Hahn, Raquel Gur, Stefano Marengo, Barbara K. Lipska, David A. Lewis, Vahram Haroutunian, Scott Hemby, Patrick Sullivan, Shahram Akbarian, Andrew Chess, Joseph D. Buxbaum, Greg E. Crawford, Enrico Domenici, Bernie Devlin, Solveig K. Sieberts, Mette A. Peters, and Panos Roussos. “CommonMind Consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder”. In: *Scientific Data* 6 (2019), p. 180. DOI: 10.1038/sdata.2018.180.
- [17] Andrew D. Grotzinger, Josefin Werme, Wouter J. Peyrot, Oleksandr Frei, Christiaan de Leeuw, Lucy K. Bicks, Qiuyu Guo, Michael P. Margolis, Brandon J. Coombes, Anthony Batzler, Vanessa Pazdernik, Joanna M. Biernacka, Ole A. Andreassen, Verner Anttila, Anders D. Børghlum, Jerome Breen, Na Cai, Ditte Demontis, Howard J. Edenberg, Stephen V. Faraone, Barbara Franke, Michael J. Gandal, Joel Gelernter, Alexander S. Hatoum, John M. Hettima, Emma C. Johnson, Katherine G. Jonas, James A. Knowles, Karestan C. Koenen, Adam X. Maihofer, Travis T. Mallard, Manuel Mattheisen, Karen S. Mitchell, Benjamin M. Neale, Caroline M. Nievergelt, John I. Nurnberger, Kevin S. O’Connell, Roseann E. Peterson, Elise B. Robinson, Sandra S. Sanchez-Roige, Susan L. Santangelo, Jeremiah M. Scharf, Hreinn Stefansson, Kari Stefansson, Murray B. Stein, Nora I. Strom, Laura M. Thornton, Elliot M. Tucker-Drob, Brad Verhulst, Irwin D.

- Waldman, G. Bragi Walters, Naomi R. Wray, Dongmei Yu, Phil H. Lee, Kenneth S. Kendler, Jordan W. Smoller, et al. “Mapping the genetic landscape across 14 psychiatric disorders”. In: *Nature* 649 (2026), pp. 406–415.
- [18] Amit Sud, Ben Kinnnersley, and Richard S. Houlston. “Genome-wide association studies of cancer: current insights and future perspectives”. In: *Nature Reviews Cancer* 17 (2017), pp. 692–704. doi: 10.1038/nrc.2017.82.
- [19] Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Mariaelisa Graff, Anna U. Eliassen, Yuning Jiang, Shailaja Raghavan, et al. “A saturated map of common genetic variants associated with human height from 5.4 million individuals of diverse ancestries”. In: *Nature* 610 (2022), pp. 704–712. doi: 10.1038/s41586-022-05275-y.
- [20] Mengzhen Liu, Yuning Jiang, Robbee Wedow, Yue Li, David M. Brazel, Fang Chen, Gaurav Datta, Juan Davila-Velderrain, David McGuire, Chen Tian, et al. “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use”. In: *Nature Genetics* 51 (2019), pp. 237–244. doi: 10.1038/s41588-018-0307-5.
- [21] The ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shores, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, et al. “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In: *Nature* 583 (2020), pp. 699–710. doi: 10.1038/s41586-020-2493-4.
- [22] Mark Schena, Darryl Shalon, Ronald W. Davis, and Patrick O. Brown. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. In: *Science* 270.5235 (Oct. 1995), pp. 467–470.
- [23] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7 (July 2008), pp. 621–628. doi: 10.1038/nmeth.1226.
- [24] 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”. In: *Nature* 560 (2018), pp. 494–498. doi: 10.1038/s41586-018-0414-6.
- [25] Delaney K. Sullivan, Kristján Eldjárn Hjörleifsson, Nikhila P. Swarna, Conrad Oakes, Guillaume Holley, Páll Melsted, and Lior Pachter. “Accurate quantification of nascent and mature RNAs from single-cell and single-nucleus RNA-seq”. In: *Nucleic Acids Research* 53.1 (2025), gkae1137. doi: 10.1093/nar/gkae1137.

- [26] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. In: *Cell* 161.5 (May 2015), pp. 1187–1201. doi: 10.1016/j.cell.2015.04.044.
- [27] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets”. In: *Cell* 161.5 (May 2015), pp. 1202–1214. doi: 10.1016/j.cell.2015.05.002.
- [28] Jennifer E. Rood, Samantha Wynne, Lucia Robson, Anna Hupalowska, John Randell, Sarah A. Teichmann, and Aviv Regev. “The Human Cell Atlas from a cell census to a unified foundation model”. In: *Nature* 637 (2025), pp. 1065–1071. doi: 10.1038/s41586-024-08392-9.
- [29] The GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509 (2020), pp. 1318–1330. doi: 10.1126/science.aaz1776.
- [30] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. “Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease”. In: *Science* 376.6589 (2022). doi: 10.1126/science.abf3041.
- [31] The All of Us Research Program Genomics Investigators. “Genomic data in the All of Us Research Program”. In: *Nature* 627 (2024), pp. 340–346. doi: 10.1038/s41586-023-06957-x.
- [32] Rebecca L. Walker, Gokul Ramaswami, Christopher Hartl, Nicholas Mancuso, Michael J. Gandal, Luis de la Torre-Ubieta, Bogdan Pasaniuc, Jason L. Stein, and Daniel H. Geschwind. “Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms”. In: *Cell* 179.3 (Oct. 2019), 750–771.e22. doi: 10.1016/j.cell.2019.09.021.
- [33] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. “Stochastic mRNA Synthesis in Mammalian Cells”. en. In: *PLoS Biology* 4.10 (Sept. 2006). Ed. by Ueli Schibler, e309. issn: 1545-7885. doi: 10.1371/journal.pbio.0040309. url: <https://dx.plos.org/10.1371/journal.pbio.0040309> (visited on 08/07/2019).

- [34] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. “Real-Time Kinetics of Gene Activity in Individual Bacteria”. en. In: *Cell* 123.6 (Dec. 2005), pp. 1025–1036. ISSN: 00928674. DOI: 10.1016/j.cell.2005.09.031. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0092867405010378> (visited on 08/23/2019).
- [35] 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”. en. In: *Proceedings of the National Academy of Sciences* 109.43 (Oct. 2012), pp. 17454–17459. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1213530109. URL: <http://www.pnas.org/cgi/doi/10.1073/pnas.1213530109> (visited on 09/03/2019).
- [36] Gennady Gorin, John J. Vastola, and Lior Pachter. “Studying stochastic systems biology of the cell with single-cell genomics data”. en. In: *Cell Systems* 14 (Sept. 2023), pp. 1–22. ISSN: 24054712. DOI: 10.1016/j.cels.2023.08.004. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2405471223002442> (visited on 10/03/2023).
- [37] Gennady Gorin, Tara Chari, Maria Carilli, John J. Vastola, and Lior Pachter. “Monod: model-based discovery and integration through fitting stochastic transcriptional dynamics to single-cell sequencing data”. In: *Nature Methods* 22 (Nov. 7, 2025), pp. 2286–2300. DOI: 10.1038/s41592-025-02832-x. URL: <https://doi.org/10.1038/s41592-025-02832-x>.
- [38] Tara Chari, Gennady Gorin, and Lior Pachter. *Biophysically Interpretable Inference of Cell Types from Multimodal Sequencing Data*. en. Preprint. bioRxiv: 2023.09.17.558131, Sept. 2023. URL: <http://biorxiv.org/lookup/doi/10.1101/2023.09.17.558131> (visited on 09/22/2023).
- [39] Tara Chari and Lior Pachter. “The Specious Art of Single-Cell Genomics”. en. In: *PLOS Computational Biology* 19.8 (Aug. 2023), e1011288. DOI: 10.1371/journal.pcbi.1011288. URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288> (visited on 11/25/2023).
- [40] Rory Stark, Marta Grzelak, and James Hadfield. “RNA sequencing: the teenage years”. In: *Nature Reviews Genetics* 20 (2019), pp. 631–656. DOI: 10.1038/s41576-019-0150-2.
- [41] Anna S. E. Cuomo, Giulio Alvari, Caren B. Azodi, et al. “Optimizing expression quantitative trait locus mapping workflows for single-cell studies”. In: *Genome Biology* 22 (2021), p. 188. DOI: 10.1186/s13059-021-02407-x.
- [42] Leo Breiman. “Statistical Modeling: The Two Cultures”. In: *Statistical Science* 16.3 (2001), pp. 199–231. DOI: 10.1214/ss/1009213726.

- [43] Abhyudai Singh and Pavol Bokes. “Consequences of mRNA Transport on Stochastic Variability in Protein Levels”. en. In: *Biophysical Journal* 103.5 (Sept. 2012), pp. 1087–1096. ISSN: 00063495. DOI: 10.1016/j.bpj.2012.07.015. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0006349512007904> (visited on 08/12/2019).
- [44] Maria Carilli, Kayla Jackson, and Lior Pachter. *The Rayleigh Quotient and Contrastive Principal Component Analysis I*. Nov. 19, 2025. DOI: 10.1101/2025.11.19.689125. bioRxiv: 2025.11.19.689125. URL: <https://doi.org/10.1101/2025.11.19.689125>.
- [45] Maria Carilli, Kayla Jackson, and Lior Pachter. *The Rayleigh Quotient and Contrastive Principal Component Analysis II*. Apr. 8, 2026. DOI: 10.64898/2026.04.08.717236. bioRxiv: 2026.04.08.717236. URL: <https://doi.org/10.64898/2026.04.08.717236>.
- [46] Maria Carilli, Gennady Gorin, Tara Chari, and Lior Pachter. “Spectral neural approximations for models of transcriptional dynamics”. English. In: *Biophysical Journal* (May 2024). Publisher: Elsevier. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2024.04.034. URL: [https://www.cell.com/biophysj/abstract/S0006-3495\(24\)00314-X](https://www.cell.com/biophysj/abstract/S0006-3495(24)00314-X) (visited on 06/26/2024).
- [47] Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. “Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data”. In: *Nature Methods* 21.8 (2024), pp. 1466–1469.
- [48] Ingileif B. Hallgrímsdóttir, Maria Carilli, and Lior Pachter. *Estimating cis and trans contributions to differences in gene regulation*. Feb. 18, 2026. DOI: 10.1101/2024.07.13.603403. bioRxiv: 2024.07.13.603403. URL: <https://doi.org/10.1101/2024.07.13.603403>.
- [49] Qiqing Fu, Chenyu Dong, Yunhe Liu, Xiaoqiong Xia, Gang Liu, Fan Zhong, and Lei Liu. “A comparison of scRNA-seq annotation methods based on experimentally labeled immune cell subtype dataset”. In: *Briefings in Bioinformatics* 25.5 (Sept. 2024), bbae392. DOI: 10.1093/bib/bbae392.
- [50] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.
- [51] Harold Hotelling. “Analysis of a complex of statistical variables into principal components.” In: *Journal of educational psychology* 24.6 (1933), p. 417.
- [52] Abubakar Abid, Vivek Zhang, Vibhor Bagaria, and James Zou. “Contrastive principal component analysis”. In: *Advances in Neural Information Processing Systems* 31 (2018).
- [53] Philippe Boileau, Nima S Hejazi, and Sandrine Dudoit. “Exploring high-dimensional biological data with sparse contrastive principal component analysis”. In: *Bioinformatics* 36.11 (2020), pp. 3422–3430.

- [54] Didong Li, Andrew Jones, and Barbara Engelhardt. “Probabilistic contrastive dimension reduction for case-control study data”. In: *The Annals of Applied Statistics* 18.3 (2024), pp. 2207–2229.
- [55] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. 2nd. Cambridge, UK: Cambridge University Press, 2012. ISBN: 978-0-521-54823-6.
- [56] Ronald A. Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of Eugenics* 7.2 (1936), pp. 179–188.
- [57] Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert Müller. “Fisher discriminant analysis with kernels”. In: *Neural Networks for Signal Processing IX*. IEEE. 1999, pp. 41–48.
- [58] Elisabeth Rebboah, Ryan Weber, Elnaz Abdollahzadeh, Nikhila Swarna, Delaney K Sullivan, Diane Trout, Fairlie Reese, Heidi Yahan Liang, Ghassan Filimban, Parvin Mahdipoor, et al. “Systematic cell-type resolved transcriptomes of 8 tissues in 8 lab and wild-derived mouse strains captures global and local expression variation”. In: *bioRxiv* (2025), pp. 2025–04.
- [59] Lingyun Xiong, Jing Liu, Seung Yub Han, Kari Koppitch, Jin-Jin Guo, Megan Rommelfanger, Zhen Miao, Fan Gao, Ingileif B. Hallgrímsdóttir, Lior Pachter, Junhyong Kim, Adam L. MacLean, and Andrew P. McMahon. “Direct androgen receptor control of sexually dimorphic gene expression in the mammalian kidney”. In: *Developmental Cell* 58.21 (2023), 2338–2358.e5. DOI: 10.1016/j.devcel.2023.08.010.
- [60] Andrew Ransick, Nils O. Lindström, Jing Liu, Qin Zhu, Jin-Jin Guo, Gregory F. Alvarado, Albert D. Kim, Hannah G. Black, Junhyong Kim, and Andrew P. McMahon. “Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney”. In: *Developmental Cell* 51.3 (2019), 399–413.e7. DOI: 10.1016/j.devcel.2019.10.005.
- [61] Hirofumi Watanabe, Robert L. Paxton, Matthew R. Tolerico, Vidya K. Nagalakshmi, Shinji Tanaka, Mark D. Okusa, Shin Goto, Ichiei Narita, Seiji Watanabe, Maria Luisa S. Sequeira-López, and R. Ariel Gomez. “Expression of *Acsn2*, a kidney-specific gene, parallels the function and maturation of proximal tubular cells”. In: *American Journal of Physiology–Renal Physiology* 319.4 (2020), F569–F580. DOI: 10.1152/ajprenal.00348.2020.
- [62] Siqi Chen, Ruiyang Liu, Chia-Kuei Mo, et al. “Multi-omic and spatial analysis of mouse kidneys highlights sex-specific differences in gene regulation across the lifespan”. In: *Nature Genetics* 57 (2025), pp. 1213–1227. DOI: 10.1038/s41588-025-02161-x.
- [63] Ming-Zhi Zhang, Kensuke Sasaki, Yan Li, Zhilian Li, Yu Pan, Guan-Nan Jin, Yinqiu Wang, Aolei Niu, Suwan Wang, Xiaofeng Fan, Jian Chun Chen, Corina Borza, Haichun Yang, Ambra Pozzi, Agnes B. Fogo, and Raymond C. Harris. “The Role of the EGF Receptor in Sex Differences in Kidney

- Injury”. In: *Journal of the American Society of Nephrology* 30.11 (2019), pp. 2015–2033. DOI: 10.1681/ASN.2018121244.
- [64] Hannah Dueck, James Eberwine, and Junhyong Kim. “Variation is function: are single cell differences functionally important? Testing the hypothesis that single cell variation is required for aggregate function”. In: *Bioessays* 38.2 (2016), pp. 172–180.
 - [65] Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. “Spatial patterns underlying population differences in the background EEG”. In: *Brain topography* 2.4 (1990), pp. 275–284.
 - [66] Michael X Cohen. “A tutorial on generalized eigendecomposition for denoising, contrast enhancement, and dimension reduction in multichannel electrophysiology”. In: *Neuroimage* 247 (2022), p. 118809.
 - [67] Fabien Lotte and Cuntai Guan. “Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms”. In: *IEEE Transactions on biomedical Engineering* 58.2 (2010), pp. 355–362.
 - [68] Trevor Hastie and Robert Tibshirani. “Discriminant analysis by Gaussian mixtures”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 155–176.
 - [69] Sergey Ioffe. “Probabilistic linear discriminant analysis”. In: *European Conference on Computer Vision*. Springer. 2006, pp. 531–542.
 - [70] Kristen A Severson, Soumya Ghosh, and Kenney Ng. “Unsupervised learning with contrastive latent variable models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 4862–4869.
 - [71] Michael E Tipping and Christopher M Bishop. “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (1999), pp. 611–622.
 - [72] Sam Roweis. “EM algorithms for PCA and SPCA”. In: *Advances in neural information processing systems* 10 (1997).
 - [73] James O. Ramsay and Bernard W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer, 2002.
 - [74] James O Ramsay and Bernard W Silverman. *Functional data analysis*. Springer, 2005.
 - [75] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. “Functional Data Analysis”. In: *Annual Review of Statistics and Its Application* 3 (2016), pp. 257–295. DOI: 10.1146/annurev-statistics-041715-033624.
 - [76] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. “Functional data analysis for sparse longitudinal data”. In: *Journal of the American statistical association* 100.470 (2005), pp. 577–590.

- [77] Gareth M. James, Trevor J. Hastie, and Catherine A. Sugar. “Principal component models for sparse functional data”. In: *Biometrika* 87.3 (Sept. 2000), pp. 587–602. DOI: 10.1093/biomet/87.3.587.
- [78] Ci-Ren Jiang and Jane-Ling Wang. “Covariate adjusted functional principal components analysis for longitudinal data”. In: *The Annals of Statistics* 38.2 (Apr. 2010), pp. 1194–1226. DOI: 10.1214/09-AOS742.
- [79] Eric Zhang and Didong Li. “Contrastive Functional Principal Component Analysis”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 2025, pp. 22380–22388.
- [80] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2005. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001.
- [81] Darawan Rinchai, Sara Deola, Gabriele Zoppoli, Basirudeen Syed Ahamed Kabeer, Sara Taleb, Igor Pavlovski, Selma Maacha, Giusy Gentilcore, Mohammed Toufiq, Lisa Mathew, Li Liu, Fazlur Rehaman Vempalli, Ghada Mubarak, Stephan Lorenz, Irene Sivieri, Gabriella Cirmena, Chiara Dentone, Paola Cuccarolo, Daniele Roberto Giacobbe, Federico Baldi, Alberto Garbarino, Benedetta Cigolini, Paolo Cremonesi, Michele Bedognetti, Alberto Ballestrero, Matteo Bassetti, Boris P. Hejblum, Tracy Augustine, Nicholas Van Panhuys, Rodolphe Thiebaut, Ricardo Branco, Tracey Chew, Maryam Shojaei, Kirsty Short, Carl G. Feng, PREDICT-19 Consortium, Susu M. Zughaier, Andrea De Maria, Benjamin Tang, Ali Ait Hssain, Davide Bedognetti, Jean-Charles Grivel, and Damien Chaussabel. “High-temporal resolution profiling reveals distinct immune trajectories following the first and second doses of COVID-19 mRNA vaccines”. In: *Science Advances* 8.45 (2022). DOI: 10.1126/sciadv.abp9961.
- [82] Prabhu S. Arunachalam, Madeleine K. D. Scott, Thomas Hagan, Chunfeng Li, Yupeng Feng, Florian Wimmers, Lilit Grigoryan, Meera Trisal, Venkata Viswanadh Edara, Lilin Lai, Sarah Esther Chang, Allan Feng, Shaurya Dhingra, Mihir Shah, Alexandra S. Lee, Sharon Chinthrajah, Sayantani B. Sindher, Vamsee Mallajosyula, Fei Gao, Natalia Sigal, Sangeeta Kowli, Sheena Gupta, Kathryn Pellegrini, Gregory Tharp, Sofia Maysel-Auslender, Sydney Hamilton, Hadj Aoued, Kevin Hrusovsky, Mark Roskey, Steven E. Bosinger, Holden T. Maecker, Scott D. Boyd, Mark M. Davis, Paul J. Utz, Mehul S. Suthar, Purvesh Khatri, Kari C. Nadeau, and Bali Pulendran. “Systems vaccinology of the BNT162b2 mRNA vaccine in humans”. In: *Nature* 596 (2021), pp. 410–416. DOI: 10.1038/s41586-021-03791-x.
- [83] Matthew C. Altman, Darawan Rinchai, Nicole Baldwin, Mohammed Toufiq, Elizabeth Whalen, Mathieu Garand, Basirudeen Syed Ahamed Kabeer, Mohamed Alfaki, Scott R. Presnell, Prasong Khaenam, Aaron Ayllón-Benítez, Fleur Mougin, Patricia Thébault, Laurent Chiche, Noemie Jourde-Chiche,

- J. Theodore Phillips, Goran Klintmalm, Anne O’Garra, Matthew Berry, Chloe Bloom, Robert J. Wilkinson, Christine M. Graham, Marc Lipman, Ganjana Lertmemongkolchai, and Damien Chaussabel. “Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data”. In: *Nature Communications* 12 (2021), p. 4385. DOI: 10.1038/s41467-021-24719-2.
- [84] Dejan Mesner, Ann-Kathrin Reuschl, Matthew V. X. Whelan, Taylor Bronzovich, Tafhima Haider, Lucy G. Thorne, Roberta Ragazzini, Paola Bonfanti, Greg J. Towers, and Clare Jolly. “SARS-CoV-2 evolution influences GBP and IFITM sensitivity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 120.5 (2023), e2212577120. DOI: 10.1073/pnas.2212577120.
- [85] Y. Furutani, M. Toguchi, S. Higuchi, K. Yanaka, L. Gailhouste, X. Y. Qin, et al. “Establishment of a rapid detection system for ISG20-dependent SARS-CoV-2 subreplicon RNA degradation induced by interferon-alpha”. In: *International Journal of Molecular Sciences* 22 (2021), p. 11641. DOI: 10.3390/ijms222111641.
- [86] Rashid Mir, Mohammad Fahad Ullah, Imadeldin Elfaki, Mohammad A. Alanazi, Naseh A. Algehainy, Faisal H. Altemani, Mamdoh S. Moawadh, Faris J. Tayeb, Badr A. Alsayed, Mohammad Muzaffar Mir, Jaber Alfaifi, Syed Khalid Mustafa, Jameel Barnawi, and Salma Saleh Alrdahe. “Whole-Exome Sequencing Reveals Rare Genetic Variants in Saudi COVID-19 Patients with Extreme Phenotypes”. In: *Viruses* 17.9 (2025), p. 1198. DOI: 10.3390/v17091198.
- [87] Michal Witkowski, Carla Tizian, Mariana Ferreira-Gomes, et al. “Untimely TGF β responses in COVID-19 limit antiviral functions of NK cells”. In: *Nature* 600 (2021), pp. 295–301. DOI: 10.1038/s41586-021-04142-6.
- [88] Mateusz A. Maździarz, Katarzyna Krawczyk, Ewa Lepiarczyk, et al. “Poly(A) tail dynamics, non-adenine incorporation and alternative polyadenylation shape the host transcriptome in COVID-19 pathogenesis”. In: *Scientific Reports* 15 (2025), p. 37986. DOI: 10.1038/s41598-025-21969-5.
- [89] Anru Zhang and Dong Xia. *Tensor SVD: Statistical and Computational Limits*. 2020. arXiv: 1703.02724 [math.ST].
- [90] Gennady Gorin. “Stochastic foundations for single-cell RNA sequencing”. en. PhD Dissertation. Pasadena, California: California Institute of Technology, May 2023. URL: <https://thesis.library.caltech.edu/16062/>.
- [91] Gennady Gorin, Mengyu Wang, Ido Golding, and Heng Xu. *Stochastic simulation platform for visualization and estimation of transcriptional kinetics*. en. Preprint. bioRxiv: 825869, Nov. 2019. DOI: 10.1101/825869. URL: <http://biorxiv.org/lookup/doi/10.1101/825869> (visited on 12/04/2019).

- [92] Peter Bokes, John R. King, and Martin Loose. “Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression”. In: *Journal of Mathematical Biology* 64.5 (2012), pp. 829–854. DOI: 10.1007/s00285-011-0433-5.
- [93] Gennady Gorin and Lior Pachter. “Modeling bursty transcription and splicing with the chemical master equation”. In: *Biophysical Journal* 121.6 (Feb. 2022), pp. 1056–1069. DOI: 10.1016/j.bpj.2022.02.004. URL: [https://www.cell.com/biophysj/fulltext/S0006-3495\(22\)00104-7](https://www.cell.com/biophysj/fulltext/S0006-3495(22)00104-7) (visited on 03/26/2021).
- [94] Tobias Jahnke and Wilhelm Huisinga. “Solving the chemical master equation for monomolecular reaction systems analytically”. en. In: *Journal of Mathematical Biology* 54 (Sept. 2006), pp. 1–26. ISSN: 0303-6812, 1432-1416. DOI: 10.1007/s00285-006-0034-x. URL: <http://link.springer.com/10.1007/s00285-006-0034-x> (visited on 06/07/2019).
- [95] K. Jayakumar and Davis Antony Mundassery. “On Moran’s Bivariate Gamma and Bivariate Negative Binomial Distribution”. en. In: *Calcutta Statistical Association Bulletin* 59.1-2 (Mar. 2007), pp. 15–28. ISSN: 0008-0683, 2456-6462. DOI: 10.1177/0008068320070102. URL: <http://journals.sagepub.com/doi/10.1177/0008068320070102> (visited on 01/25/2022).
- [96] Gennady Gorin and Lior Pachter. *Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production*. en. Preprint. bioRxiv: 2020.09.25.312868, Sept. 2020. DOI: 10.1101/2020.09.25.312868. URL: <http://biorxiv.org/lookup/doi/10.1101/2020.09.25.312868> (visited on 09/30/2020).
- [97] Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. “Exactly solvable models of stochastic gene expression”. en. In: *The Journal of Chemical Physics* 152.14 (Apr. 2020), p. 144106. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5143540. URL: <http://aip.scitation.org/doi/10.1063/1.5143540> (visited on 11/22/2021).
- [98] Lucy Ham, Marcel Jackson, and Michael P.H. Stumpf. *Pathway dynamics can delineate the sources of transcriptional noise in gene expression*. en. Preprint. bioRxiv: 2020.09.30.319814, Sept. 2020. DOI: 10.1101/2020.09.30.319814. URL: <http://biorxiv.org/lookup/doi/10.1101/2020.09.30.319814> (visited on 10/07/2020).
- [99] 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”. en. In: *Nature Communications* 13.1 (Dec. 2022), p. 7620. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34857-7. URL: <https://www.nature.com/articles/s41467-022-34857-7> (visited on 12/15/2022).

- [100] 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”. en. In: *Nature Communications* 12.1 (Dec. 2021), p. 2618. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22919-1. URL: <http://www.nature.com/articles/s41467-021-22919-1> (visited on 11/06/2021).
- [101] Gennady Gorin, Shawn Yoshida, and Lior Pachter. “Assessing Markovian and Delay Models for Single-Nucleus RNA Sequencing”. en. In: *Bulletin of Mathematical Biology* 85.11 (Oct. 2023), p. 114. DOI: 10.1007/s11538-023-01213-9. URL: <https://link.springer.com/article/10.1007/s11538-023-01213-9> (visited on 10/18/2023).
- [102] Hao Lu, Hua Yan, Xiaoyu Li, Yuan Xing, Yumeng Ye, Siao Jiang, Luyu Ma, Jie Ping, Hongyan Zuo, Yanhui Hao, Chao Yu, Yang Li, Gangqiao Zhou, and Yiming Lu. “Single-cell map of dynamic cellular microenvironment of radiation-induced intestinal injury”. In: *Communications Biology* 6 (2023), p. 1248. DOI: 10.1038/s42003-023-05556-5.
- [103] Ainhua Arina, Michael Beckett, Christian Fernandez, Wenxin Zheng, Sean Pitroda, Steven J Chmura, Jason J Luke, Martin Forde, Yuzhu Hou, Byron Burnette, Helena Mauceri, Israel Lowy, Tasha Sims, Nikolai Khodarev, Yang-Xin Fu, and Ralph R Weichselbaum. “Tumor-reprogrammed resident T cells resist radiation to control tumors”. en. In: *Nat. Commun.* 10.1 (Sept. 2019), p. 3959.
- [104] Abobakr K Shadad, Frank J Sullivan, Joseph D Martin, and Laurence J Egan. “Gastrointestinal radiation injury: symptoms, risk factors and mechanisms”. en. In: *World J. Gastroenterol.* 19.2 (Jan. 2013), pp. 185–198.
- [105] Ryan C Augustin, Riyue Bao, and Jason J Luke. “Targeting Cbl-b in cancer immunotherapy”. en. In: *J. Immunother. Cancer* 11.2 (Feb. 2023), e006007.
- [106] Miao Yu, Gang Guo, Lei Huang, Libin Deng, Chang-Sheng Chang, Bhagelu R Achyut, Madison Canning, Ningchun Xu, Ali S Arbab, Roni J Bollag, Paulo C Rodriguez, Andrew L Mellor, Huidong Shi, David H Munn, and Yan Cui. “CD73 on cancer-associated fibroblasts enhanced by the A2B-mediated feedforward circuit enforces an immune checkpoint”. en. In: *Nat. Commun.* 11.1 (Jan. 2020), p. 515.
- [107] Miao Yu, Gang Guo, Lei Huang, Libin Deng, Chang-Sheng Chang, Bhagelu R Achyut, Madison Canning, Ningchun Xu, Ali S Arbab, Roni J Bollag, Paulo C Rodriguez, Andrew L Mellor, Huidong Shi, David H Munn, and Yan Cui. “CD73 on cancer-associated fibroblasts enhanced by the A2B-mediated feedforward circuit enforces an immune checkpoint”. en. In: *Nat. Commun.* 11.1 (Jan. 2020), p. 515.

- [108] Linzhuo Qu, Huiying Che, Jingyu Zhao, Xin Lu, Zijun Ren, Yu Wu, Qian Liu, and Hongjian Guan. “NCAPD3 is a prognostic biomarker and is correlated with immune infiltrates in glioma”. en. In: *Histol. Histopathol.* 39.11 (Nov. 2024), pp. 1473–1484.
- [109] Ying Zhong, Xinyu Ren, Xi Cao, Yali Xu, Yu Song, Yidong Zhou, Feng Mao, Songjie Shen, Zhe Wang, and Qiang Sun. “Insulin-like growth factor 2 receptor is a key immune-related gene that is correlated with a poor prognosis in patients with triple-negative breast cancer: A bioinformatics analysis”. en. In: *Front. Oncol.* 12 (Oct. 2022), p. 871786.
- [110] Xiaotong Xu, Weyland Cheng, Shuai Zhao, Yuchun Liu, Lifeng Li, Xiaorui Song, Yaodong Zhang, and Cong Ding. “Pan-cancer analysis of the role of MPP7 in human tumors”. en. In: *Heliyon* 10.16 (Aug. 2024), e36148.
- [111] Jürgen Dittmer. “The biology of the Ets1 proto-oncogene”. en. In: *Mol. Cancer* 2.1 (Aug. 2003), p. 29.
- [112] Hyewon Phee, Marianne N Mollenauer, and Arthur Weiss. “Role of GIT2 in T cell migration and development (95.8)”. en. In: *J. Immunol.* 182.1_Supplement (Apr. 2009), pp. 95.8–95.8.
- [113] Kyle L O’Hagan, S Miller, and H Phee. “Pak2 is essential for the function of Foxp3+ regulatory T cells through maintaining a suppressive Treg phenotype”. In: *Sci. Rep.* 7 (Dec. 2017).
- [114] Germán Belenguer, Gianmarco Mastrogiovanni, Clare Pacini, Zoe Hall, Anna M Dowbaj, Robert Arnes-Benito, Aleksandra Sljukic, Nicole Prior, Sofia Kakava, Charles R Bradshaw, Susan Davies, Michele Vacca, Kourosh Saeb-Parsy, Bon-Kyoung Koo, and Meritxell Huch. “RNF43/ZNRF3 loss predisposes to hepatocellular-carcinoma by impairing liver regeneration and altering the liver lipid metabolic ground-state”. en. In: *Nat. Commun.* 13.1 (Jan. 2022), p. 334.
- [115] Fei Yue, Amy T Ku, Payton D Stevens, Megan N Michalski, Weiyu Jiang, Jianghua Tu, Zhongcheng Shi, Yongchao Dou, Yi Wang, Xin-Hua Feng, Galen Hostetter, Xiangwei Wu, Shixia Huang, Noah F Shroyer, Bing Zhang, Bart O Williams, Qingyun Liu, Xia Lin, and Yi Li. “Loss of ZNRF3/RNF43 unleashes EGFR in cancer”. en. In: *bioRxivorg* (Jan. 2024).
- [116] Fei Yue, Amy T Ku, Payton D Stevens, Megan N Michalski, Weiyu Jiang, Jianghua Tu, Zhongcheng Shi, Yongchao Dou, Yi Wang, Xin-Hua Feng, Galen Hostetter, Xiangwei Wu, Shixia Huang, Noah F Shroyer, Bing Zhang, Bart O Williams, Qingyun Liu, Xia Lin, and Yi Li. “Loss of ZNRF3/RNF43 unleashes EGFR in cancer”. en. In: *bioRxivorg* (Jan. 2024).
- [117] Jin-Li Wei, Si-Yu Wu, Yun-Song Yang, Yi Xiao, Xi Jin, Xiao-En Xu, Xin Hu, Da-Qiang Li, Yi-Zhou Jiang, and Zhi-Ming Shao. “GCH1 induces immunosuppression through metabolic reprogramming and IDO1 upregu-

- lation in triple-negative breast cancer”. en. In: *J. Immunother. Cancer* 9.7 (July 2021), e002383.
- [118] Q Zeng, L Li, Z Feng, L Luo, J Xiong, Z Jie, and Li. “LCP1 is a prognostic biomarker correlated with immune infiltrates in gastric cancer”. In: *Cancer Biomarkers* 30.1 (2021), pp. 105–125.
 - [119] Shengli Pan, Yingying Deng, Jun Fu, Yuhao Zhang, Zhijin Zhang, Xiaokun Ru, and Xianju Qin. “Decreased expression of ARHGAP15 promotes the development of colorectal cancer through PTEN/AKT/FOXO1 axis”. en. In: *Cell Death Dis.* 9.6 (June 2018), p. 673.
 - [120] Bo Dou, Gang Jiang, Wang Peng, and Chentao Liu. “OTULIN deficiency: focus on innate immune system impairment”. en. In: *Front. Immunol.* 15 (May 2024), p. 1371564.
 - [121] Wen-Yuan Wang, Ling Pan, Susan C Su, Emma J Quinn, Megumi Sasaki, Jessica C Jimenez, Ian R A Mackenzie, Eric J Huang, and Li-Huei Tsai. “Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons”. en. In: *Nat. Neurosci.* 16.10 (Oct. 2013), pp. 1383–1391.
 - [122] Danixa Martínez, Daniela Nualart, Carlos Loncoman, Juan C Opazo, Katina Zabala, Francisco J Morera, Gonzalo A Mardones, and Luis Vargas-Chacoff. “Discovery of BbX transcription factor in the patagonian blennie: Exploring expression changes following combined bacterial and thermal stress exposure”. en. In: *Dev. Comp. Immunol.* 149.105056 (Dec. 2023), p. 105056.
 - [123] Lab Mouse. “BIOART-000279. 10/07/2024. NIAID NIH BIOART Source”. In: (). bioart.niaid.nih.gov/bioart/279. URL: bioart.niaid.nih.gov/bioart/279.
 - [124] Human Anatomy. “BIOART-000519. 10/07/2024. NIAID NIH BIOART Source”. In: (). bioart.niaid.nih.gov/bioart/519. URL: bioart.niaid.nih.gov/bioart/519.
 - [125] Generic Cells. “BIOART-000172. 10/07/2024. NIAID NIH BIOART Source”. In: (). bioart.niaid.nih.gov/bioart/172. URL: bioart.niaid.nih.gov/bioart/172.
 - [126] 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 Booeslaghi, 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 Kancharla, 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”. en. In: *Nature* 598.7879 (Oct. 2021), pp. 103–110. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03500-8. URL: <https://www.nature.com/articles/s41586-021-03500-8> (visited on 11/15/2021).
- [127] Valentine Svensson. “Droplet scRNA-seq is not zero-inflated”. en. In: *Nature Biotechnology* 38.2 (Feb. 2020), pp. 147–150. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0379-5. URL: <https://www.nature.com/articles/s41587-019-0379-5> (visited on 08/29/2020).
- [128] Joseph Rodriguez and Daniel R. Larson. “Transcription in Living Cells: Molecular Mechanisms of Bursting”. In: *Annual Review of Biochemistry* 89. Volume 89, 2020 (2020), pp. 189–212. ISSN: 1545-4509. DOI: <https://doi.org/10.1146/annurev-biochem-011520-105250>. URL: <https://www.annualreviews.org/content/journals/10.1146/annurev-biochem-011520-105250>.
- [129] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. “Demystifying “drop-outs” in single-cell UMI data”. en. In: *Genome Biology* 21 (Dec. 2020), p. 196. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02096-y. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02096-y> (visited on 02/27/2021).
- [130] Crispin Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Third. Springer, 2004.
- [131] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. “Distribution shapes govern the discovery of predictive models for gene regulation”. en. In: *Proceedings of the National Academy of Sciences* 115.29 (2018), pp. 7533–7538. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1804060115. (Visited on 08/24/2019).
- [132] Xiaoming Fu, Heta P Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L Lenstra, and Ramon Grima. “Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions”. en. In: *eLife* 11 (Oct. 2022), e82493. ISSN: 2050-084X. DOI: 10.7554/eLife.82493. URL: <https://elifesciences.org/articles/82493> (visited on 01/07/2023).

- [133] Brian Munsky, Zachary Fox, and Gregor Neuert. “Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics”. en. In: *Methods* 85 (2015), pp. 12–21. ISSN: 10462023. DOI: 10.1016/j.ymeth.2015.06.009. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1046202315002510> (visited on 06/07/2019).
- [134] 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”. en. In: *Nature Communications* 8.1 (Apr. 2017), p. 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: <http://www.nature.com/articles/ncomms14049> (visited on 08/22/2019).
- [135] Kristján Eldjárn Hjörleifsson, Delaney K. Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. *Accurate quantification of single-nucleus and single-cell RNA-seq transcripts*. en. Preprint. bioRxiv: 2022.12.02.518832, Dec. 2022. URL: <http://biorxiv.org/lookup/doi/10.1101/2022.12.02.518832> (visited on 12/12/2022).
- [136] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastrioti, 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”. en. In: *Nature* 560.7719 (Aug. 2018), pp. 494–498. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0414-6. URL: <http://www.nature.com/articles/s41586-018-0414-6> (visited on 08/22/2019).
- [137] Páll Melsted, A. Sina Boeshaghi, 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”. en. In: *Nature Biotechnology* 39.7 (July 2021), pp. 813–818. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-021-00870-2. URL: <http://www.nature.com/articles/s41587-021-00870-2> (visited on 11/15/2021).
- [138] Eleni P. Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalex, Zhengqing Ouyang, Rahul Satija, Neville E. Sanjana, Sergei B.

- Koralov, and Peter Smibert. “Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells”. en. In: *Nature Methods* 16.5 (May 2019), pp. 409–412. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0392-0. URL: <http://www.nature.com/articles/s41592-019-0392-0> (visited on 08/22/2019).
- [139] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. “Joint profiling of chromatin accessibility and gene expression in thousands of single cells”. en. In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730. URL: <http://www.sciencemag.org/lookup/doi/10.1126/science.aau0730> (visited on 06/07/2019).
- [140] Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding. “Stochastic Kinetics of Nascent RNA”. en. In: *Physical Review Letters* 117.12 (2016), p. 128101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.117.128101. URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.117.128101> (visited on 09/15/2019).
- [141] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalex, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. “Integrated analysis of multimodal single-cell data”. en. In: *Cell* 184.13 (June 2021), 3573–3587.e29. ISSN: 00928674. DOI: 10.1016/j.cell.2021.04.048. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833> (visited on 09/27/2021).
- [142] John J Vastola. “In search of a coherent theoretical framework for stochastic gene regulation”. en. PhD thesis. Vanderbilt, Mar. 2021. URL: <https://ir.vanderbilt.edu/handle/1803/16646>.
- [143] Daniel T Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. en. In: *Journal of Computational Physics* 22.4 (Dec. 1976), pp. 403–434. ISSN: 00219991. DOI: 10.1016/0021-9991(76)90041-3. URL: <https://linkinghub.elsevier.com/retrieve/pii/0021999176900413> (visited on 08/25/2019).
- [144] Brian Munsky and Mustafa Khammash. “The finite state projection algorithm for the solution of the chemical master equation”. en. In: *The Journal of Chemical Physics* 124.4 (2006), p. 044104. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2145882. (Visited on 08/25/2019).

- [145] Ankit Gupta, Jan Mikelson, and Mustafa Khammash. “A finite state projection algorithm for the stationary solution of the chemical master equation”. en. In: *The Journal of Chemical Physics* 147.15 (Oct. 2017), p. 154101. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5006484. URL: <http://aip.scitation.org/doi/10.1063/1.5006484> (visited on 07/29/2019).
- [146] Zachary Fox, Gregor Neuert, and Brian Munsky. “Finite state projection based bounds to compare chemical master equation models using single-cell data”. en. In: *The Journal of Chemical Physics* 145.7 (Aug. 2016), p. 074101. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4960505. URL: <http://aip.scitation.org/doi/10.1063/1.4960505> (visited on 07/22/2019).
- [147] 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”. en. In: *Journal of Mathematical Biology* 64.5 (Apr. 2012), pp. 829–854. ISSN: 0303-6812, 1432-1416. DOI: 10.1007/s00285-011-0433-5. URL: <http://link.springer.com/10.1007/s00285-011-0433-5> (visited on 08/09/2019).
- [148] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. “Deep generative modeling for single-cell transcriptomics”. en. In: *Nature Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: <http://www.nature.com/articles/s41592-018-0229-2> (visited on 08/10/2021).
- [149] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. “Joint probabilistic modeling of single-cell multi-omic data with totalVI”. en. In: *Nature Methods* 18.3 (Mar. 2021), pp. 272–282. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-020-01050-x. URL: <http://www.nature.com/articles/s41592-020-01050-x> (visited on 11/19/2021).
- [150] Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. “Data-driven discovery of coordinates and governing equations”. en. In: *Proceedings of the National Academy of Sciences* 116.45 (Nov. 2019), pp. 22445–22451. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1906995116. URL: <https://pnas.org/doi/full/10.1073/pnas.1906995116> (visited on 06/02/2022).
- [151] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. *Fourier Neural Operator for Parametric Partial Differential Equations*. Preprint. arXiv: 2010.08895, May 2021. URL: <http://arxiv.org/abs/2010.08895> (visited on 10/01/2022).
- [152] Long Cai, Nir Friedman, and X. Sunney Xie. “Stochastic protein expression in individual cells at the single molecule level”. en. In: *Nature* 440.7082 (Mar. 2006), pp. 358–362. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/

- nature04599. URL: <http://www.nature.com/articles/nature04599> (visited on 12/05/2019).
- [153] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulana, Christopher Cronin, Christoph Karp, Eric J. Liaw, Mina Amin, and Long Cai. “Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH”. en. In: *Cell* 174.2 (July 2018), 363–376.e16. ISSN: 00928674. DOI: 10.1016/j.cell.2018.05.035. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0092867418306470> (visited on 08/24/2019).
 - [154] Yodai Takei, Yujing Yang, Jonathan White, Jina Yun, Meera Prasad, Lincoln J Ombelets, Simone Schindler, and Long Cai. *High-resolution spatial multi-omics reveals cell-type specific nuclear compartments*. en. Tech. rep. bioRxiv, 2023. URL: <https://www.biorxiv.org/content/10.1101/2023.05.07.539762v1>.
 - [155] Shangying Wang, Sara Capponi, and Simone Bianco. “Inferring Conditional Probability Distributions of Noisy Gene Expression from Limited Observations by Deep Learning.” In: *GEN Biotechnology* 1 (6 1984), pp. 504–513.
 - [156] Ankit Gupta, Christoph Schwab, and Mustafa Khammash. “DeepCME: A deep learning framework for computing solution statistics of the chemical master equation”. en. In: *PLOS Computational Biology* 17.12 (Dec. 2021). Ed. by James R. Faeder, e1009623. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009623. URL: <https://dx.plos.org/10.1371/journal.pcbi.1009623> (visited on 03/16/2022).
 - [157] Augustinas Sukys, Kaan Öcal, and Ramon Grima. “Approximating solutions of the Chemical Master equation using neural networks”. In: *iScience* 25.9 (Aug. 2022), p. 105010. ISSN: 2589-0042. DOI: 10.1016/j.isci.2022.105010. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9474291/> (visited on 03/02/2023).
 - [158] 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, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. en. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0686-2. URL: <http://www.nature.com/articles/s41592-019-0686-2> (visited on 11/15/2021).
- [159] 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”. en. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: <https://www.nature.com/articles/s41586-020-2649-2> (visited on 06/09/2022).
- [160] 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. DOI: 10.1101/657619. URL: <http://biorxiv.org/lookup/doi/10.1101/657619> (visited on 08/24/2019).
- [161] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [162] Gennady Gorin and Lior Pachter. *Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using Monod*. Preprint.

- bioRxiv: 2022.06.11.495771, Apr. 2023. URL: <https://www.biorxiv.org/content/10.1101/2022.06.11.495771v2> (visited on 06/17/2022).
- [163] Zachary R Fox and Brian Munsky. “The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments”. en. In: *PLOS Computational Biology* 15.1 (Jan. 2019). Ed. by Oleg A Igoshin, e1006365. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006365. URL: <https://dx.plos.org/10.1371/journal.pcbi.1006365> (visited on 05/07/2021).
 - [164] Jingyao Wang, Shihe Zhang, Hongfang Lu, and Heng Xu. “Differential regulation of alternative promoters emerges from unified kinetics of enhancer-promoter interaction”. en. In: *Nature Communications* 13.1 (Dec. 2022), p. 2714. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30315-6. URL: <https://www.nature.com/articles/s41467-022-30315-6> (visited on 06/17/2022).
 - [165] Florian De Rop, Joy N Ismail, Carmen Bravo González-Blas, Gert J Hulselmans, Christopher Campbell Flerin, Jasper Janssens, Koen Theunis, Valerie M Christiaens, Jasper Wouters, Gabriele Marcassa, Joris de Wit, Suresh Poovathingal, and Stein Aerts. “HyDrop enables droplet based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads”. en. In: *eLife* 11 (Feb. 2022), e73971. ISSN: 2050-084X. DOI: 10.7554/eLife.73971. URL: <https://elifesciences.org/articles/73971> (visited on 03/18/2022).
 - [166] 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”. en. In: *Nature Methods* 14.9 (Sept. 2017), pp. 865–868. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4380. URL: <http://www.nature.com/articles/nmeth.4380> (visited on 08/22/2019).
 - [167] Gennady Gorin, Valentine Svensson, and Lior Pachter. “Protein velocity and acceleration from single-cell multiomics experiments”. en. In: *Genome Biology* 21 (Feb. 2020), p. 39. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1945-3. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1945-3> (visited on 02/24/2020).
 - [168] Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, and Joel A Klappenbach. “Multiplexed quantification of proteins and transcripts in single cells”. en. In: *Nature Biotechnology* 35.10 (Oct. 2017), pp. 936–939. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3973. URL: <http://www.nature.com/articles/nbt.3973> (visited on 08/22/2019).

- [169] Hattie Chung, Christopher N. Parkhurst, Emma M. Magee, Devan Phillips, Ehsan Habibi, Fei Chen, Bertrand Z. Yeung, Julia Waldman, David Artis, and Aviv Regev. “Joint single-cell measurements of nuclear proteins and RNA in vivo”. en. In: *Nature Methods* 18.10 (Oct. 2021), pp. 1204–1212. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-021-01278-1. URL: <https://www.nature.com/articles/s41592-021-01278-1> (visited on 10/13/2021).
- [170] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. “Exponential scaling of single-cell RNA-seq in the past decade”. en. In: *Nature Protocols* 13.4 (Apr. 2018), pp. 599–604. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2017.149. URL: <http://www.nature.com/articles/nprot.2017.149> (visited on 09/24/2019).
- [171] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. “A Python library for probabilistic analysis of single-cell omics data”. en. In: *Nature Biotechnology* 40.2 (Feb. 2022), pp. 163–166. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-021-01206-w. URL: <https://www.nature.com/articles/s41587-021-01206-w> (visited on 04/14/2023).
- [172] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. “Clustering of single-cell multi-omics data with a multimodal deep learning method”. en. In: *Nature Communications* 13.1 (Dec. 2022), p. 7705. ISSN: 2041-1723. DOI: 10.1038/s41467-022-35031-9. URL: <https://www.nature.com/articles/s41467-022-35031-9> (visited on 12/18/2022).
- [173] Tal Ashuach, Daniel A. Reidenbach, Adam Gayoso, and Nir Yosef. “PeakVI: A deep generative model for single-cell chromatin accessibility analysis”. en. In: *Cell Reports Methods* 2.3 (Mar. 2022), p. 100182. ISSN: 26672375. DOI: 10.1016/j.crmeth.2022.100182. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2667237522000376> (visited on 12/26/2022).
- [174] A. Sanchez and I. Golding. “Genetic Determinants and Cellular Constraints in Noisy Gene Expression”. en. In: *Science* 342.6163 (Dec. 2013), pp. 1188–1193. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1242975. URL: <http://www.sciencemag.org/cgi/doi/10.1126/science.1242975> (visited on 09/03/2019).
- [175] Judea Pearl. “Causal Inference in Statistics: An Overview”. In: *Statistics Surveys* 3 (2009), pp. 96–146. ISSN: 1935-7516. DOI: 10.1214/09-SS057.

- [176] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S. Baron, Marvin E. Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. “Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies”. en. In: *Science* 367.6482 (Mar. 2020), pp. 1151–1156. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax3072. URL: <https://www.science.org/doi/10.1126/science.aax3072> (visited on 11/10/2021).
- [177] XL. Kuang, XM. Zhao, HF. Xu, YY. Shi, JB. Deng, and GT. Sun. “Spatio-temporal expression of a novel neuron-derived neurotrophic factor (NDNF) in mouse brains during development”. In: *BMC Neurosci* 11 (2010).
- [178] T. K. Ulland and M. Colonna. “TREM2 — a key player in microglial biology and Alzheimer disease”. In: *Nature Reviews Neurology* 14 (2018), pp. 667–675.
- [179] Lucy Ham, Rowan D. Brackston, and Michael P. H. Stumpf. “Extrinsic Noise and Heavy-Tailed Laws in Gene Expression”. en. In: *Physical Review Letters* 124.10 (Mar. 2020), p. 108101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.124.108101. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.124.108101> (visited on 08/28/2020).
- [180] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. “Stochastic Gene Expression in a Single Cell”. en. In: *Science* 297.5584 (2002), pp. 1183–1186. DOI: 10.1126/science.1070919.
- [181] Gennady Gorin and Lior Pachter. “Length biases in single-cell RNA sequencing of pre-mRNA”. en. In: *Biophysical Reports* 3.1 (Mar. 2023), p. 100097. ISSN: 26670747. DOI: 10.1016/j.bpr.2022.100097. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2667074722000544> (visited on 01/11/2023).
- [182] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. “Interpretable factor models of single-cell RNA-seq via variational autoencoders”. en. In: *Bioinformatics* 36.11 (June 2020). Ed. by Anthony Mathelier, pp. 3418–3421. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btaa169. URL: <https://academic.oup.com/bioinformatics/article/36/11/3418/5807606> (visited on 01/09/2023).
- [183] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. “Gene expression distribution deconvolution in single-cell RNA sequencing”. en. In: *Proceedings of the National Academy of Sciences* 115.28 (July 2018), E6437–E6446. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1721085115. URL: <http://www.pnas.org/lookup/doi/10.1073/pnas.1721085115> (visited on 04/09/2020).

- [184] Allen Institute for Brain Science. *FASTQ files for Allen v3 mouse MOp samples*. Feb. 2020. URL: http://data.nemoarchive.org/biccn/grant/u19_zeng/zeng/transcriptome/scell/10x_v3/mouse/raw/MOp/.
- [185] Allen Institute for Brain Science. *Analyses for Allen v3 mouse MOp samples*. Feb. 2020. URL: http://data.nemoarchive.org/biccn/grant/u19_zeng/zeng/transcriptome/scell/10x_v3/mouse/processed/analysis/10X_cells_v3_AIBS/.
- [186] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. en. In: *Genome Biology* 19.1 (Dec. 2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0> (visited on 08/23/2019).
- [187] François Jacob and Jacques Monod. “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of molecular biology* 3.3 (1961), pp. 318–356.
- [188] Rita JS Phillips. “A cis-trans position effect at the A locus of the house mouse”. In: *Genetics* 54.2 (1966), p. 485.
- [189] John S Kovach, Antonio O Ballesteros, Marilyn Meyers, Marco Soria, and Robert F Goldberger. “A cis/trans test of the effect of the first enzyme for histidine biosynthesis on regulation of the histidine operon”. In: *Journal of Bacteriology* 114.1 (1973), pp. 351–356.
- [190] Christopher R Cowles, Joel N Hirschhorn, David Altshuler, and Eric S Lander. “Detection of regulatory variation in mouse genes”. In: *Nature genetics* 32.3 (2002), pp. 432–437.
- [191] Patricia J Wittkopp, Belinda K Haerum, and Andrew G Clark. “Evolutionary changes in cis and trans gene regulation”. In: *Nature* 430.6995 (2004), pp. 85–88.
- [192] Christian R Landry, Patricia J Wittkopp, Clifford H Taubes, Jose M Ranz, Andrew G Clark, and Daniel L Hartl. “Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*”. In: *Genetics* 171.4 (2005), pp. 1813–1822.
- [193] Angela Goncalves, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T Odom, and John C Marioni. “Extensive compensatory cis-trans regulation in the evolution of mouse gene expression”. In: *Genome research* 22.12 (2012), pp. 2376–2384.
- [194] Andreas Massouras, Sebastian M Waszak, Monica Albarca-Aguilera, Kornel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, Eric A Stone, Jeffrey D Jensen, Trudy FC Mackay, et al. “Genomic variation and its impact on gene expression in *Drosophila melanogaster*”. In: *PLoS genetics* 8.11 (2012), e1003055.

- [195] Andreas Tsouris, Gauthier Brach, Joseph Schacherer, and Jing Hou. “Non-additive genetic components contribute significantly to population-wide gene expression variation”. In: *Cell Genomics* 4.1 (2024).
- [196] C Joel McManus, Joseph D Coolon, Michael O Duff, Jodi Eipper-Mains, Brenton R Graveley, and Patricia J Wittkopp. “Regulatory divergence in *Drosophila* revealed by mRNA-seq”. In: *Genome research* 20.6 (2010), pp. 816–825.
- [197] 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, and SciPy 1.0 Contributors. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. doi: 10.1038/s41592-019-0686-2.
- [198] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.
- [199] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. issn: 1367-4803. doi: 10.1093/bioinformatics/btp616. url: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/> (visited on 02/18/2025).
- [200] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (Dec. 2014), p. 550. issn: 1474-760X. doi: 10.1186/s13059-014-0550-8. url: <https://doi.org/10.1186/s13059-014-0550-8> (visited on 02/18/2025).
- [201] Mallory A. Ballinger, Katya L. Mack, Sylvia M. Durkin, Eric A. Riddell, and Michael W. Nachman. “Environmentally robust cis-regulatory changes underlie rapid climatic adaptation”. In: *Proceedings of the National Academy of Sciences* 120.39 (2023), e2214614120. doi: 10.1073/pnas.2214614120. eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.2214614120>. url: <https://www.pnas.org/doi/abs/10.1073/pnas.2214614120>.

- [202] Sarah A. Signor and Sergey V. Nuzhdin. “The Evolution of Gene Expression in cis and trans”. In: *Trends in Genetics* 34.7 (July 2018), pp. 532–544. DOI: 10.1016/j.tig.2018.03.007. URL: <https://doi.org/10.1016/j.tig.2018.03.007>.
- [203] Kaitlyn Bloom, Anuradha Karunanidhi, Kimimasa Tobita, Charles Hoppel, Edda Thiels, Eloise Peet, Yudong Wang, Shrabani Basu, and Jerry Vockley. “ACAD10 protein expression and Neurobehavioral assessment of Acad10-deficient mice”. In: *PLOS ONE* 15.12 (Dec. 2020), e0242445. DOI: 10.1371/journal.pone.0242445. URL: <https://doi.org/10.1371/journal.pone.0242445>.
- [204] Madoka Inuzuka, Minako Hayakawa, and Tatsuya Ingi. “Serinc, an activity-regulated protein family, incorporates serine into membrane lipid synthesis”. In: *Journal of Biological Chemistry* 280.42 (Oct. 2005), pp. 35776–35783. ISSN: 0021-9258. DOI: 10.1074/jbc.M505712200. URL: <https://doi.org/10.1074/jbc.M505712200>.
- [205] João Henrique Tadini Marilhano Fabri, Nivea Pereira de Sá, Iran Malavazi, and Maurizio Del Poeta. “The dynamics and role of sphingolipids in eukaryotic organisms upon thermal adaptation”. In: *Progress in Lipid Research* 80 (Nov. 2020), p. 101063. ISSN: 0163-7827. DOI: 10.1016/j.plipres.2020.101063. URL: <https://doi.org/10.1016/j.plipres.2020.101063>.
- [206] G. M. Jenkins. “The emerging role for sphingolipids in the eukaryotic heat shock response”. In: *Cellular and Molecular Life Sciences* 60.4 (Apr. 2003), pp. 701–710. ISSN: 1420-682X. DOI: 10.1007/s00018-003-2239-0. URL: <https://doi.org/10.1007/s00018-003-2239-0>.
- [207] Tadashi Nomura, Kohjiro Nagao, Ryo Shirai, Hitoshi Gotoh, Masato Umeda, and Katsuhiko Ono. “Temperature sensitivity of Notch signaling underlies species-specific developmental plasticity and robustness in amniote brains”. In: *Nature Communications* 13.1 (Jan. 2022), p. 96. DOI: 10.1038/s41467-021-27815-z. URL: <https://doi.org/10.1038/s41467-021-27815-z>.
- [208] Kenneth A Barr, Katherine L Rhodes, and Yoav Gilad. “The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees”. In: *Genome Biology* 24.1 (2023), p. 207.
- [209] Gyu-Tae Shin, Ji Eun Park, and Min-Jeong Lee. “MAGEH1 interacts with GADD45G and induces renal tubular cell apoptosis”. In: *PLOS ONE* 16.11 (Nov. 2021), e0260135. DOI: 10.1371/journal.pone.0260135. URL: <https://doi.org/10.1371/journal.pone.0260135>.

- [210] Anna K. Lee and Patrick Ryan Potts. “A Comprehensive Guide to the MAGE Family of Ubiquitin Ligases”. In: *Journal of Molecular Biology* 429.8 (Apr. 2017), pp. 1114–1142. DOI: 10.1016/j.jmb.2017.03.005. URL: <https://doi.org/10.1016/j.jmb.2017.03.005>.
- [211] Amika Kikuchi, Hiroki Onoda, Kosuke Yamaguchi, Satomi Kori, Shun Matsuzawa, Yoshie Chiba, Shota Tanimoto, Sae Yoshimi, Hiroki Sato, Atsushi Yamagata, Mikako Shirouzu, Naruhiko Adachi, Jafar Sharif, Haruhiko Koseki, Atsuya Nishiyama, Makoto Nakanishi, Pierre-Antoine Defossez, and Kyohei Arita. “Structural basis for activation of DNMT1”. In: *Nature Communications* 13 (Nov. 2022), p. 7130. DOI: 10.1038/s41467-022-34779-4. URL: <https://doi.org/10.1038/s41467-022-34779-4>.
- [212] Gennady Gorin, John J Vastola, and Lior Pachter. “Studying stochastic systems biology of the cell with single-cell genomics data”. In: *Cell Systems* 14.10 (2023), pp. 822–843.
- [213] Ibai Irastorza-Azcarate, Alexander Kukalev, Rieke Kempfer, Christoph J Thieme, Guido Mastrobuoni, Julia Markowski, Gesa Loof, Thomas M Sparks, Emily Brookes, Kedar Nath Natarajan, et al. “Extensive folding variability between homologous chromosomes in mammalian cells”. In: *bioRxiv* (2024), pp. 2024–05.
- [214] Zhenzhen Ma, Alexander Lorenzo Starr, David Gokhman, and Hunter Fraser. “The causes and consequences of human-specific DNA methylation”. In: *bioRxiv* (2026), pp. 2026–01.
- [215] Ryan Weber, Maria Carilli, Elisabeth Rebboah, Ghassan Filimban, Heidi Yahan Liang, Diane Trout, Margaret Duffield, Parvin Mahdipoor, Erisa Taghizadeh, Negar Fattahi, Negar Mojgani, Romina Mojaverzargar, Shimako Kawauchi, Brian A. Williams, Grant R. MacGregor, Barbara J. Wold, Lior Pachter, Ingileif B. Hallgrimsdottir, and Ali Mortazavi. “Hybrid crosses reveal a cell-type-specific landscape of mouse regulatory variation”. In: *bioRxiv* (2026). DOI: 10.64898/2026.04.02.716195.
- [216] Elissa J. Chesler, Darla R. Miller, Lisa R. Branstetter, Leslie D. Galloway, Barbara L. Jackson, Vivek M. Philip, Brynn H. Voy, Cymbeline T. Cuiat, David W. Threadgill, Robert W. Williams, Gary A. Churchill, Dabney K. Johnson, and Kenneth F. Manly. “The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics”. In: *Mammalian Genome* 19.6 (2008), pp. 382–389. DOI: 10.1007/s00335-008-9135-8.
- [217] David W. Threadgill, Darla R. Miller, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. “The Collaborative Cross: A Recombinant Inbred Mouse Population for the Systems Genetic Era”. In: *ILAR Journal* 52.1 (2011), pp. 24–31. ISSN: 1084-2020. DOI: 10.1093/ilar.52.1.24.

- [218] Elisabeth Rebboah, Ryan Weber, Elnaz Abdollahzadeh, Nikhila Swarna, Delaney K. Sullivan, Diane Trout, Fairlie Reese, Heidi Yahan Liang, Ghasan Filimban, Parvin Mahdipoor, Margaret Duffield, Romina Mojaverzargar, Erisa Taghizadeh, Negar Fattahi, Negar Mojgani, Haoran Zhang, Rebekah K. Loving, Maria Carilli, A. Sina Boeshaghi, Shimako Kawauchi, Ingileif B. Hallgrímsdóttir, Brian A. Williams, Grant R. MacGregor, Lior Pachter, Barbara J. Wold, and Ali Mortazavi. “Systematic cell-type resolved transcriptomes of 8 tissues in 8 lab and wild-derived mouse strains capture global and local expression variation”. In: *Cell Genomics* (Dec. 31, 2025), p. 101108. doi: 10.1016/j.xgen.2025.101108.
- [219] Patricia J Wittkopp, Belinda K Haerum, and Andrew G Clark. “Regulatory changes underlying expression differences within and between *Drosophila* species”. In: *Nature genetics* 40.3 (2008), pp. 346–350.
- [220] Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellåker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jérôme Nicod, Xiangchao Gan, Wei Yuan, Louise Van Der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, José Afonso Guerra-Assunção, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint, and David J. Adams. “Mouse genomic variation and its effect on phenotypes and gene regulation”. In: *Nature* 477.7364 (Sept. 2011), pp. 289–294. ISSN: 0028-0836. doi: 10.1038/nature10413.
- [221] 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”. In: *Nature* 565.7738 (Jan. 2019), pp. 251–254. ISSN: 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0836-1. URL: <http://www.nature.com/articles/s41586-018-0836-1> (visited on 06/07/2019).
- [222] J. M. Müller, K. Moos, T. Baar, K. C. Maier, K. Zumer, and A. Tresch. “Nuclear export is a limiting factor in eukaryotic mRNA metabolism”. In: *PLoS Computational Biology* 20.5 (2024), e1012059. doi: 10.1371/journal.pcbi.1012059.
- [223] Nikhila P. Swarna, A. Sina Boeshaghi, Elisabeth Rebboah, M. Grace Gordon, Pooja Kathail, Taibo Li, Marcus Alvarez, Chun Jimmie Ye, Barbara Wold, Ali Mortazavi, and Lior Pachter. “Determining gene specificity from multivariate single-cell RNA sequencing data”. In: *bioRxiv* (2025). Preprint, not peer reviewed. doi: 10.1101/2025.11.21.689845.

- [224] Giorgia Benegiamo, Giacomo V. G. von Alvensleben, Sandra Rodríguez-López, Ludger J. E. Goeminne, Alexis M. Bachmann, Jean-David Morel, Ellen Broeckx, Jing Ying Ma, Vinicius Carreira, Sameh A. Youssef, Nabil Azhar, Dermot F. Reilly, Katharine D'Aquino, Shannon Mullican, Maroun Bou-Sleiman, and Johan Auwerx. "The genetic background shapes the susceptibility to mitochondrial dysfunction and NASH progression". In: *Journal of Experimental Medicine* 220.4 (2023), e20221738. doi: 10.1084/jem.20221738.
- [225] Sarah Greve, Gisela A. Kuhn, Mara D. Saenz-de-Juano, Adhideb Ghosh, Ferdinand von Meyenn, and Katrin Giller. "The major urinary protein gene cluster knockout mouse as a novel model for translational metabolism research". In: *Scientific Reports* 12 (Aug. 2022), p. 13161. doi: 10.1038/s41598-022-17385-5.
- [226] Takashi Miwa, Mitsuro Kanda, Dai Shimizu, Shinichi Umeda, Koichi Sawaki, Haruyoshi Tanaka, Chie Tanaka, Norifumi Hattori, Masamichi Hayashi, Suguru Yamada, Goro Nakayama, Masahiko Koike, and Yasuhiro Kodera. "Hepatic metastasis of gastric cancer is associated with enhanced expression of ethanolamine kinase 2 via the p53-Bcl-2 intrinsic apoptosis pathway". In: *British Journal of Cancer* 124.8 (2021), pp. 1449–1460. doi: 10.1038/s41416-021-01271-7.
- [227] George R. Uhl and Maria J. Martinez. "PTPRD: neurobiology, genetics, and initial pharmacology of a pleiotropic contributor to brain phenotypes". In: *Annals of the New York Academy of Sciences* 1451.1 (2019), pp. 112–129. doi: 10.1111/nyas.14002.
- [228] Hideaki Tomita, Francisca Cornejo, Begoña Aranda-Pino, Cameron L. Woodard, Constanza C. Rioseco, Benjamin G. Neel, Alejandra R. Alvarez, David R. Kaplan, Freda D. Miller, and Gonzalo I. Cancino. "The Protein Tyrosine Phosphatase Receptor Delta Regulates Developmental Neurogenesis". In: *Cell Reports* 30.1 (Jan. 2020), 215–228.e5. doi: 10.1016/j.celrep.2019.12.020.
- [229] Ayako Imai, Hironori Izumi, Nagomi Ito, and Tomoyuki Yoshida. "Alternative microexon splicing code for a four-amino acid peptide of PTPRD governs behavioral development". In: *Proceedings of the National Academy of Sciences* 123.15 (Apr. 2026), e2515310123. doi: 10.1073/pnas.2515310123.
- [230] Seoyeong Kim, Jae Jin Shin, Muwon Kang, Yeji Yang, Yi Sul Cho, Hyojung Paik, Jimin Kim, Yunho Yi, Suho Lee, Hei Yeun Koo, Jinwoong Bok, Yong Chul Bae, Jin Young Kim, and Eunjoon Kim. "Alternatively spliced mini-exon B in PTP δ regulates excitatory synapses through cell-type-specific trans-synaptic PTP δ -IL1RAP interaction". In: *Nature Communications* 16 (2025), p. 4415. doi: 10.1038/s41467-025-59685-3.

- [231] Bastián I. Cortés, Rodrigo C. Meza, Carlos Ancatén-González, Nicolás M. Ardiles, María-Ingacia Aránguiz, Hideaki Tomita, David R. Kaplan, Francisca Cornejo, Alexia Nunez-Parra, Pablo R. Moya, Andrés E. Chávez, and Gonzalo I. Cancino. “Loss of protein tyrosine phosphatase receptor delta PTPRD increases the number of cortical neurons, impairs synaptic function and induces autistic-like behaviors in adult mice”. In: *Biological Research* 57 (2024), p. 40. DOI: 10.1186/s40659-024-00522-0.
- [232] Sheryl S. Moy, Jessica J. Nadler, Nancy B. Young, Randal J. Nonneman, Samantha K. Segall, Gabriela M. Andrade, Jacqueline N. Crawley, and Terry R. Magnuson. “Social Approach and Repetitive Behavior in Eleven Inbred Mouse Strains”. In: *Behavioural Brain Research* 191.1 (2008), pp. 118–129. DOI: 10.1016/j.bbr.2008.03.015.
- [233] Yichen Ge, Fuguo Wu, Mobin Cheng, Jonathan Bard, and Xiuqian Mu. “Two new genetically modified mouse alleles labeling distinct phases of retinal ganglion cell development by fluorescent proteins”. In: *Developmental Dynamics* 249.12 (2020), pp. 1514–1528. DOI: 10.1002/dvdy.233.
- [234] MV Prasad Linga Reddy, H Wang, S Liu, B Bode, JC Reed, RD Steed, SW Anderson, L Steed, D Hopkins, and J-X She. “Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population”. In: *Genes and Immunity* 12.3 (Apr. 2011), pp. 208–212. DOI: 10.1038/gene.2010.70. URL: <https://doi.org/10.1038/gene.2010.70>.
- [235] 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*. en. Preprint. bioRxiv: 2021.06.09.447782, June 2021. URL: <http://biorxiv.org/lookup/doi/10.1101/2021.06.09.447782> (visited on 06/04/2022).
- [236] John William Strutt Baron Rayleigh. *The theory of sound*. Vol. 2. Macmillan, 1896.
- [237] Eleni P. Mimitou and Peter Smibert. *Expanding the CITE-seq tool-kit: Detection of proteins, transcriptomes, clonotypes and CRISPR perturbations with multiplexing, in a single assay*. en. Mar. 2019. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126310> (visited on 08/22/2019).
- [238] Yang Wang, Jingyu Li, Andrew A. Malcolm, William Mansfield, Stephen J. Clark, Ricard Argelaguet, Laura Biggins, Richard J. Acton, Simon Andrews, Wolf Reik, Gavin Kelsey, and Peter J. Rugg-Gunn. “Combinatorial profiling of multiple histone modifications and transcriptome in single cells using scMTR-seq”. In: *Science Advances* 11.32 (2025). DOI: 10.1126/sciadv.adu3308.

- [239] Chris J. Frangieh, Johannes C. Melms, Pratiksha I. Thakore, Kathryn R. Geiger-Schuller, Patricia Ho, Adrienne M. Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S. Cuoco, Maryann Zhao, Casey R. Ager, Meri Rogava, Lila Hovey, Asaf Rotem, Chantale Bernatchez, Kai W. Wucherpennig, Bruce E. Johnson, Orit Rozenblatt-Rosen, Dirk Schaden-dorf, Aviv Regev, and Benjamin Izar. “Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion”. In: *Nature Genetics* 53 (2021), pp. 332–341. doi: 10.1038/s41588-021-00779-1.
- [240] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lan-der, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167.7 (Dec. 2016), 1853–1866.e17. doi: 10.1016/j.cell.2016.11.038.
- [241] Catherine Felce, Gennady Gorin, and Lior Pachter. *A Biophysical Model for ATAC-seq Data Analysis*. en. Tech. rep. bioRxiv: 2024.01.25.577262, Jan. 2024. URL: <http://biorxiv.org/lookup/doi/10.1101/2024.01.25.577262> (visited on 07/19/2024).
- [242] Catherine Felce, Meichen Fang, and Lior Pachter. “Joint Biophysical Mod-eling of Paired Single-Cell RNA and Protein Measurements”. In: *bioRxiv* (2025). Preprint, not peer reviewed. doi: 10.1101/2025.11.14.688548.
- [243] Francesco Mottes, Qian-Ze Zhu, and Michael P. Brenner. “Gradient-based optimization of exact stochastic kinetic models”. In: *arXiv* (2026). Preprint. arXiv: 2601.14183 [physics.comp-ph].
- [244] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. “How to build the virtual cell with artificial intelligence: Priorities and opportunities”. In: *Cell* 187.25 (Dec. 2024), pp. 7045–7063. doi: 10.1016/j.cell.2024.10.032.