A biophysical approach to normalization and trajectory inference in single-cell RNA sequencing data analysis

Thesis by
Meichen Fang

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Biological Engineering

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 28, 2025

Meichen Fang
ORCID: 0000-0002-8217-0710

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Lior Pachter, for his continuous support throughout my PhD. His integrity and compassion, along with a sense of humor and an endless supply of anecdotes, make him a perfect advisor to have. I am grateful for his willingness to listen to my ideas and provide thoughtful feedback, no matter how random the topic. I never cease to be amazed by his ability to see through things and identify the fundamental part of the question.

I want to thank my committee members, Matt Thomson, Justin Bois, and Shasha Chong, for their guidance and exceptional support. They have always been available whenever I wanted to discuss my research, generously offering their help and genuinely caring about my future.

I want to thank the Pachter laboratory for providing an inclusive and supportive environment. I am especially grateful to my mentor and collaborator, Gennady Gorin, who introduced me to the world of biophysical modeling of single-cell genomics data and has since continued to help shape my research. Also to Tara Chari, who is always willing to provide help to me and others in the lab. I am grateful to my collaborator, Catherine Felce, whose kindness and sharpness make working together enjoyable. Thanks to Delaney Sullivan for his invaluable help with kallisto. Thanks to everyone else in the biophysics subgroup: Maria Carilli, Kayla Jackson, Conrad Oakes, Joseph Rich, who have provided engaging discussions. Thanks to our lab manager Charlene Kim, who is thoughtful and responsible. Thanks to current and past member in Pachter lab: Vera Beilinson, Anne Yeokyoung Kil, Rebekah Loving, Nikhila Swarna, Laura Luebbert, Taleen Dilanyan, Lambda Moses, Ángel Merchán Gálvez, Kristján Eldjarn Hjörleifsson, and Sina Booeshaghi.

Outside of the Pachter lab, I want to thank my collaborator John Vastola, a soft-spoken man with keen insight. I also want to thank my previous mentor, Fangzhou Xiao, who is always fun to talk to. Thanks to my rotation mentors and friends, Jan Gregrowicz and Jialong Jiang.

I want to thank my friends at Caltech. I want to thank Jia Yao, who made my PhD experience more colorful through our fun conversations and activities. I also want to thank my fellow students in the BBE division, Yutian Li and Yujing Yang, who have helped me both in research and daily life. I want to thank my high school and college friends, whose friendship has stood the test of time.

Finally, I want to thank my parents for their unconditional love.

ABSTRACT

Single-cell genomics assays, particularly single-cell RNA sequencing that enables genome-wide profiling of gene expression, have been driven forward by a combination of technological and computational advances. While producing extraordinary large amounts of data for biological discovery, methods for mining results currently rely heavily on heuristics and lack of modeling has resulted in limited mechanistic biological insight. This thesis presents two models for normalization and trajectory inference in single-cell RNA sequencing analysis to demonstrate how biophysical modeling, when combined with principled statistical inference, can yield interpretable insights grounded in rigorous theoretical frameworks.

We begin by explaining the two cultures in single-cell RNA sequencing analysis. Next, we present the chemical master equation, which forms the theoretical foundation for biophysically informed stochastic models of gene expression, and explore an existing gap in developing uniform approximations over time under the large-volume limit. Returning to scRNA-seq data analysis, we introduce two mechanistic models for normalization and trajectory inference, which are essential components of scRNA-seq analysis.

PUBLISHED CONTENT AND CONTRIBUTIONS

Fang, Meichen, Gennady Gorin, and Lior Pachter (2025). "Trajectory inference from single-cell genomics data with a process time model". In: *PLoS Comput. Biol.* 21.1, e1012752. DOI: 10.1371/journal.pcbi.1012752.
Meichen Fang participated in the model formulation, performed analysis, and participated in writing the manuscript.

Fang, Meichen and Lior Pachter (2025). "Extrinsic biological stochasticity and technical noise normalization of single-cell RNA sequencing data". In: *bioRxiv*, p. 2025.05.11.653373. DOI: 10.1101/2025.05.11.653373.
Meichen Fang participated in the conception of the project, performed analysis, and participated in writing the manuscript.

Gorin, Gennady, Meichen Fang, Tara Chari, and Lior Pachter (2022). "RNA velocity unraveled". In: *PLoS Comput. Biol.* 18.9, e1010492. DOI: 10.1371/journal.pcbi.1010492.
Meichen Fang participated in analysis and editing the manuscript.

Gorin, Gennady, John J Vastola, Meichen Fang, and Lior Pachter (2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nat. Commun.* 13.1, p. 7620. DOI: 10.1038/s41467-022-34857-7.
Meichen Fang participated in analysis and editing the manuscript.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

In his seminal 2001 essay on statistical modeling, Leo Breiman identified two distinct modeling cultures: data models, which posit a stochastic mechanism for data generation and emphasize interpretability, and algorithmic models, which prioritize predictive performance without necessarily modeling the underlying process (Breiman, 2001). At the time, data models dominated the statistical modeling field, and Breiman advocated for broader adoption of algorithmic approaches. A parallel to this can be seen in early studies of (single) gene expression at single-cell resolution, where the focus was on the development of models for the data, with the goal of providing mechanistic insight into transcriptional dynamics.

Two decades later, the landscape of single-cell biology has changed dramatically. Rapid expansion of single-cell genomic data, driven by advances in sequencing technologies, has led to a dominant reliance on algorithmic models . In this thesis, we advocate for greater use of data models, which provide a more principled and insightful framework for uncovering meaningful biological insights from single-cell data. Rather than replacing algorithmic approaches, data models offer essential complementary strengths, particularly in terms of interpretability and mechanistic understanding (Gorin and Lior Pachter, 2024).

In this introduction, we begin by reviewing early stochastic models of gene expression, developed from the data model perspective, which lay the theoretical foundation for interpreting single-cell sequencing data. We then introduce the development and limitations of single-cell RNA sequencing (scRNA-seq), and current practices for scRNA-seq analysis. Finally, we explain the two modeling cultures in contemporary scRNA-seq analysis and argue for more mechanistic models in this rapidly evolving field.

1.1 The study of stochastic gene expression

Gene expression, encompassing transcription (RNA synthesis) and translation (protein synthesis), is a central pillar of molecular biology research. As it inherently consists of a series of biochemical reactions such as transcriptional activation, mRNA synthesis, splicing, and degradation, one approach to studying gene expression is to

quantitatively model gene expression processes as biochemical reaction networks (or chemical reaction networks). Two major frameworks exist for such modeling: 1) deterministic models, based on the law of mass action, describe reactions using continuous concentrations and ordinary differential equations (ODEs) and focus on the graphical and algebraic structures of reaction networks (Feinberg, 2019); 2) stochastic models, grounded in the chemical master equation (CME), explicitly account for randomness in molecular interactions and focus on the properties of probability distributions of the systems (Van Kampen, 2007; C. Gardiner, 2009).

Since gene expression often involves molecules with low copy numbers (e.g., DNA in transcription initiation), stochastic models are frequently necessary to accurately describe gene expression dynamics. Stochastic fluctuations in gene expression contribute substantially to cellular heterogeneity in genetically homogeneous populations (Ko, Nakauchi, and Takahashi, 1990; Elowitz et al., 2002). As a result, the distribution of gene expression outcomes becomes critical, and stochastic chemical reaction networks, along with the chemical master equation, have become the standard framework for gene expression models. It describes the time evolution of the probability distribution over the discrete states of a biochemical system with a transition rate matrix. Furthermore, as common biochemical reactions only depend on the current states in a time-independent manner, biochemical reaction networks are modeled as homogeneous continuous time discrete state Markov chains, which have a time-independent transition rate matrix, i.e., homogeneous, and the time derivative of the probability distribution only depends on the current probability distribution, i.e., Markovian. A detailed discussion on the stochastic model, along with their deterministic counterparts as the large-volume limit, will be presented in Chapter 2.

The biochemical reactions incorporated in the gene expression models are based on the mechanistic characterization of gene expression processes. Canonical messenger RNA (mRNA) maturation involves an ordered sequence of regulated steps: (1) transcription initiation and elongation, (2) co-transcriptional splicing, (3) nuclear export, (4) cytoplasmic translation, and (5) active degradation. Among these steps, transcription initiation has been the most widely characterized and typically serves as the primary regulatory node in gene expression models. The subsequent steps are predominantly modeled as single first-order reactions with constant rates.

Let us begin with the simplest case: a constitutive zero-order transcription combined with first-order mRNA degradation. In this scenario, the stationary distribution of

mRNA levels follows a Poisson distribution. This model, which we will refer to as the constitutive model throughout this thesis, serves as the basic model for gene expression. If we extend it to include models with constitutive production and first-order degradation for mRNA which still lead to a Poisson distribution at steady states, then despite its apparent simplicity, some recent experimental evidence supports its applicability to cytoplasmic mRNA counts (Battich, Stoeger, and Pelkmans, 2015). Notably, in such cases, nuclear export—rather than transcription—behaves as a constitutive zero-order process.

Beyond the constitutive model, more sophisticated models have been inspired by experimental advances. Early electron micrograph results revealed there were active and inactive period of transcription (Miller and McKnight, 1979). Additional evidence came from studies of inducible gene expression. In 1990, Ko et al. performed single-cell quantification of a glucocorticoid-inducible reporter gene, revealing heterogeneous expression that led to the formulation of the telegraph model (Ko, Nakauchi, and Takahashi, 1990; Ko, 1991; Ko, 1992). This model introduces two fundamental states of DNA: “on” state where active promoter permits RNA polymerase binding and transcription initiation; “off” state where inactive promoter halts RNA synthesis. The promoter switches between these two states with defined activation and inactivation rates, and transcription occurs at a transcription initiation rate only in the on state. The stationary distribution of the chemical master equation of this telegraph model has been solved analytically using generating function methods (Peccoud and Ycart, 1995), with a Fano factor (variance-to-mean ratio) greater than one, in contrast to the Poissonian statistics (Fano factor = 1) expected from constitutive model. The telegraph model has also been extended to incorporate more promoter states (Ham et al., 2020), as well as to include other modalities such as protein expression (Shahrezaei and Swain, 2008; Bokes, 2022).

Building on the telegraph model, further critical insights into transcriptional dynamics have been gained through advanced imaging technologies that enable the direct observation of mRNA production at the single-molecule level. Using an MS2-GFP fusion protein to tag mRNA transcribed from inducible promoters, Golding et al. demonstrated that transcription occurs in bursts in living E. coli cells (Golding et al., 2005). This behavior reflects a limiting case of the telegraph model, characterized by long "off" periods followed by brief, intense "on" periods during which multiple mRNAs are produced in rapid succession, a regime we will refer to as the bursty model. In this regime, mRNA counts follow a geometric distribution, with each

burst corresponding to a stochastic event that generates a variable number of transcripts. A similar bursting regime has been observed in mammalian cells when fitting with the telegraph model, where the promoter activation rate is an order of magnitude lower than the inactivation rate, and the transcription initiation rate is two orders of magnitude higher than the inactivation rate (Raj et al., 2006). Under these conditions, the telegraph model can be accurately approximated by the bursty model, in which only the ratio of the transcription initiation rate to the inactivation rate is identifiable. This ratio defines the size of the burst, effectively reducing the number of model parameters by one and simplifying the description of the system. Therefore, in conjunction with the telegraph model, the bursty model has since been routinely employed in stochastic modeling of gene expression (Singh and Bokes, 2012), as it offers a reasonable approximation to the complex biological reality (Jiao et al., 2024).

However, these studies have primarily focused on a limited number of genes. A central question that remains is whether all genes follow the bursty model, or more broadly, how gene expression patterns are distributed across different models. To address this question, the bursty model has been applied to a broader set of genes beyond the scope of previous studies, providing a genome-wide portrait of transcriptional dynamics (Taniguchi et al., 2010; Suter et al., 2011; Dar et al., 2012). For example, it has been fit to fluorescence data driven by the HIV-1 LTR promoter integrated into more than 8,000 individual human genomic loci, and revealed that the bursty model rather than the constitutive model is the predominant mode of gene expression (Dar et al., 2012). However, *in vivo* mRNA and/or protein data across the entire genome without relying on artificial integration of reporter constructs would be valuable for gaining a more comprehensive understanding of transcriptional kinetics.

The limited throughput and perturbation are related to the fluorescent imaging technologies predominantly used in earlier studies on stochastic gene expression. Excitingly, recent advances in single cell sequencing have enabled us to measure high-throughput *in vivo* gene expression.

1.2 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) enables the isolation and high-throughput sequencing of mRNA transcripts from individual cells, providing transcriptome-wide resolution at the single-cell level. Compared to imaging-based approaches,

sequencing technologies are more amenable to high-throughput analysis, allowing researchers to measure gene expression across tens of thousands of genes in individual cells. Since the first study of single cell RNA sequencing (scRNA-seq) was published in 2009 (F. Tang et al., 2009), the field has rapidly evolved, with numerous methodological innovations expanding its applications (Hashimshony et al., 2012; Ramsköld, Luo, et al., 2012; Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). Initially focused solely on transcriptomic profiling, scRNA-seq technologies now facilitate multimodal measurements, enabling simultaneous detection of RNA, chromatin accessibility (ATAC-seq), and proteins in the same cell (Mimitou et al., 2021).

Despite providing unprecedented single-cell resolution of gene expression, scRNA-seq also presents significant challenges and limitations. First, these measurements are notoriously noisy, presenting new challenges for model fitting. This noise arises from factors such as low capture efficiency, dropout events, amplification bias, and variability in transcript detection. Many studies have sought to characterize the technical noise in scRNA-seq data (Brennecke et al., 2013; Grün, Kester, and Oudenaarden, 2014; Kim, Kolodziejczyk, et al., 2015); nevertheless, there remains ongoing debate about the most appropriate statistical models for handling unique molecular identifier (UMI) counts (Svensson, 2020; Sarkar and Stephens, 2021). These challenges underscore the necessity of incorporating a well-calibrated technical noise model to accurately infer transcription kinetics (Gorin and Lior Pachter, 2023). Currently, the most common measurement models for scRNA-seq are Poisson and Bernoulli sampling with cell-wise capture rates (Sarkar and Stephens, 2021; W. Tang et al., 2023; Gorin and Lior Pachter, 2022b).

Another key limitation of scRNA-seq is that these methods provide only static snapshots of gene expression, as they require cell lysis and therefore cannot directly capture temporal dynamics. Metabolic labeling of newly synthesized mRNA can provide partial insight into past transcriptional events; however, the measurement still represents the distribution at a single time point (Erhard et al., 2022). However, scRNA-seq data often capture cells at different stages along underlying biological processes, which motivates the development of trajectory inference methods.

1.3 Current practices for scRNA-seq analysis

In summary, scRNA-seq and fluorescent imaging technologies represent two complementary approaches: scRNA-seq captures the expression of thousands of genes

at a single time point, while fluorescent imaging enables dynamic tracking of a limited number of genes over time. Perhaps not surprisingly, the analysis of scRNA-seq data marks a clear departure from the mechanistic strategies that have traditionally guided the study of gene expression. Contemporary approaches are largely rooted in the algorithmic model culture, frequently relying on heuristics. To see this, let us look at steps in common scRNA-seq analysis workflow.

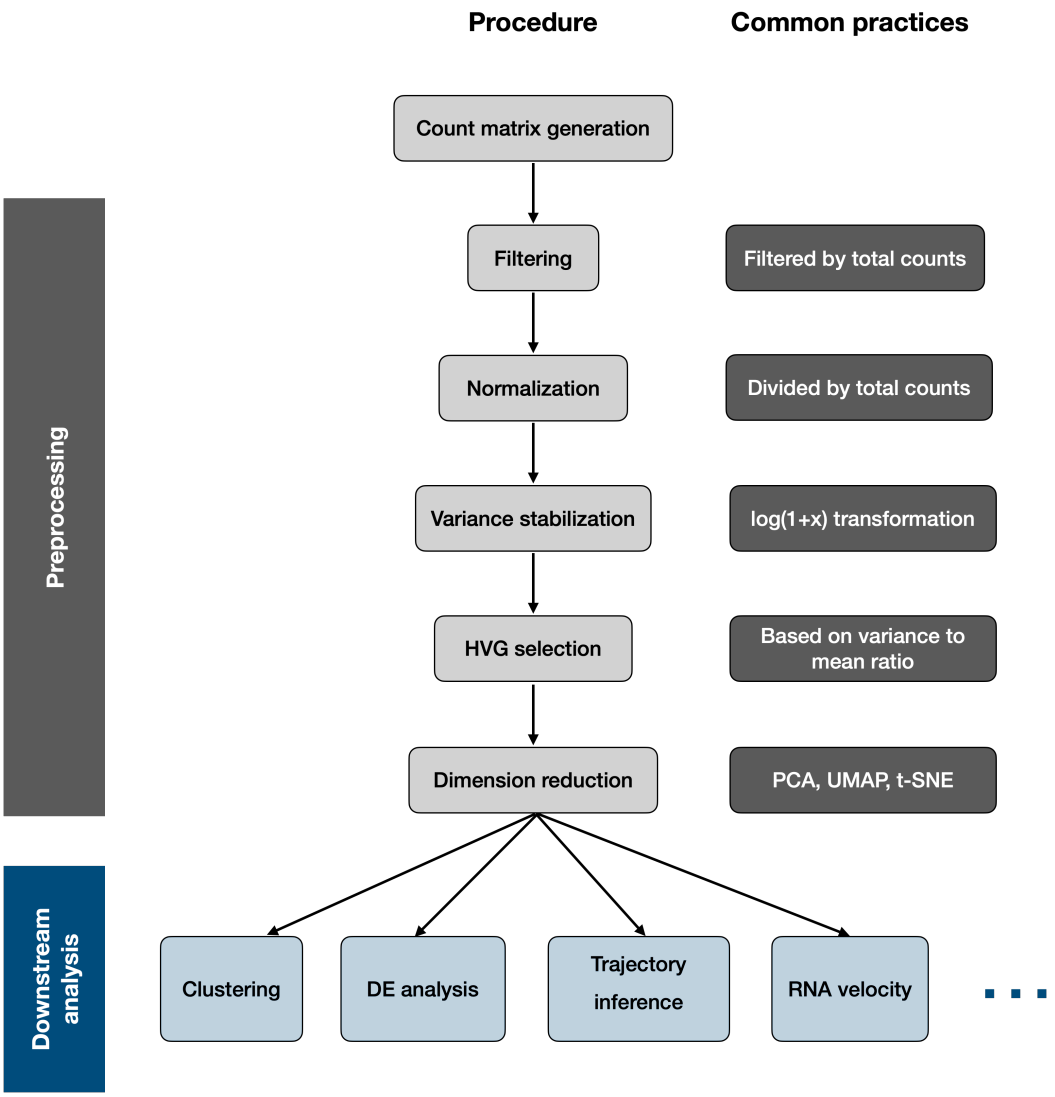

Figure 1.1: Common scRNA-seq analysis workflow.

The direct output of scRNA-seq experiments is a collection of sequencing reads, which are processed to generate count matrices. Then count matrices undergo an analysis workflow (Figure 1.1). A standard workflow can be broadly divided into two phases. The first phase involves data preprocessing, including quality control

(e.g., cell filtering), normalization, variance stabilization, feature selection, and dimensionality reduction. The second phase focuses on downstream analysis, such as clustering, differential expression (DE) analysis, trajectory inference, and RNA velocity.

At the beginning of analysis, cells are filtered based on their total transcript counts. A knee plot is used to identify high-quality cells by ranking barcodes based on total UMI or gene counts. The plot typically shows a sharp bend, i.e., knee, separating barcodes likely to represent real cells (above the knee) from low-quality or empty droplets (below). This helps determine a threshold for filtering valid cells during quality control. Genes are also filtered by retaining those expressed in a minimum number of cells.

Normalization in scRNA-seq aims to correct for technical variability (e.g., differences in sequencing depth, capture efficiency). Typically, a global-scaling normalization method is used, which calculates a single normalization factor per cell (cell size factor) using the sum of total counts. Then raw counts are scaled by cell size factors.

In scRNA-seq data, genes often exhibit a strong mean–variance relationship, where genes with higher mean expression levels also display greater variance. This heteroskedasticity poses challenges for downstream analyses such as clustering and differential expression. A common approach to mitigate this issue is to apply a variance-stabilizing transformation. One widely used method is the log-transformation, typically $log(1+x)$, which effectively reduces the dependence of variance on the mean. Despite its simplicity, this transformation performs surprisingly well in practice (Ahlmann-Eltze and Huber, 2023).

Both highly variable gene (HVG) selection and dimensionality reduction address the high dimensionality of scRNA-seq data, which captures expression levels for tens of thousands of genes in each cell. Typically, genes with high dispersion (variance-to-mean ratio) are selected as HVGs. For dimensionality reduction, methods such as PCA, t-SNE, and UMAP are commonly used, although they can introduce great distortions of the data (Chari and Pachter, 2021).

Here, we provide only a brief summary of the preprocessing steps. There are reviews offering more comprehensive discussions on scRNA-seq analysis (Luecken and Theis, 2019), and articles providing thorough benchmark of certain steps in the workflow (Booeshaghi and Lior Pachter, 2021; Booeshaghi, Hallgrímsdóttir, et al.,

2022; Ahlmann-Eltze and Huber, 2023; Rich et al., 2024).

1.4 The two cultures in scRNA-seq analysis

We can clearly see the pattern that the methods in standard workflow do not have underlying models but are defined by the algorithms.

The prominence of algorithmic models can be partially explained by both the strengths and limitations of scRNA-seq technology. Although it allows high-throughput profiling of thousands of genes across large numbers of individual cells, it is also plagued by significant technical noise, reducing its reliability as a direct quantitative readout. Furthermore, the high dimensionality of the data makes it difficult to construct mechanistic models that fully capture the complexities of gene regulation.

However, we argue that data models should not be ignored in analyzing scRNA-seq data. A balanced approach to scRNA-seq analysis requires integrating the strengths of both algorithmic and data modeling cultures. Algorithmic models excel at extracting patterns from high-dimensional data and scaling to large datasets, enabling powerful exploratory analyses. However, they often lack interpretability and mechanistic grounding. In contrast, data models, particularly mechanistic models grounded in biophysical principles, offer interpretable parameters and insights into the underlying biological processes. By combining the predictive power of algorithmic approaches with the interpretability and rigor of mechanistic modeling, we can gain a deeper and more principled understanding of single-cell gene expression. Given the prevalence of algorithmic models, there is a strong case for renewed investment in mechanistic modeling.

Several studies have successfully integrated traditional gene expression models with scRNA-seq data. For example, the telegraph and bursty models have been applied to scRNA-seq data to infer transcriptional bursting kinetics in various biological contexts (Kim and Marioni, 2013; Ramsköld, Hendriks, et al., 2024; Larsson et al., 2019; Gorin and Lior Pachter, 2022b; Tara Chari, Gorin, and Lior Pachter, 2024b). These efforts reflect a mechanistic perspective that aims to explain the observed data by explicitly modeling the underlying biological processes.

However, a mechanistic approach to scRNA-seq analysis goes beyond applying bursting models, since scRNA-seq has been used in a variety of ways. What we advocate for is a data model culture that emphasizes the formulation of biophysically inspired models and rigorous inference.

1.5 Outline

This thesis summarizes my work in advancing this approach. Chapter 2 provides a review of the chemical master equation for stochastic chemical reaction systems, along with a theoretical study of the large-volume limit for infinite time. Chapter 3 introduces an extrinsic noise model for normalization. Chapters 4 presents a process time model for trajectory inference. The thesis concludes with a discussion of future directions in Chapter 5.

Chapter 2

STOCHASTIC CHEMICAL REACTION SYSTEMS AND APPROXIMATIONS

2.1 Introduction

As discussed in Chapter 1.1, the chemical master equation (CME) offers a probabilistic framework for modeling the time evolution of a system's state, capturing the intrinsic stochastic fluctuations present in small, discrete systems. However, obtaining analytical solutions to the CME is generally intractable (Schnoerr, Sanguinetti, and Grima, 2017), and qualitative understanding is often achieved through various approximations.

A natural direction is to consider the limit of large system volumes, with the expectation that as the number of molecules becomes sufficiently large, the system's stochastic behavior will converge to either deterministic dynamics or a simplified form of stochastic dynamics. The convergence of the CME to the reaction rate equations and to the Chemical Langevin Equation (CLE) over finite time intervals has been well established (Kurtz, 1972; Gillespie, 2000). By applying the system-size expansion (also known as the linear noise approximation) and retaining the first two orders, one obtains a Fokker–Planck equation that approximates the evolution of the probability distribution (Van Kampen, 2007). A recent study has also established the validity of the linear noise approximation for stationary distributions under certain restricted conditions (Grunberg and Del Vecchio, 2023).

However, approximations that remain uniform in time in the large volume limit are generally lacking. In fact, it is believed that the limits of large volume and infinite time are not interchangeable (Hanggi et al., 1984; Baras, Mansour, and Pearson, 1996; Srivastava et al., 2002; Vellela and Qian, 2007; Vellela and Qian, 2009; Assaf and Meerson, 2017). In this chapter, we review various approximations to the CME and discuss when the large volume and infinite time limits can, or cannot, be interchanged. We argue that the (non)interchangeability of these limits arises from the specific asymptotic approximation method employed. Notably, a time-uniform large volume approximation may be attainable through the Wentzel–Kramers–Brillouin (WKB) approximation and Poisson representation, which offers a different perspective on the system's stochastic dynamics.

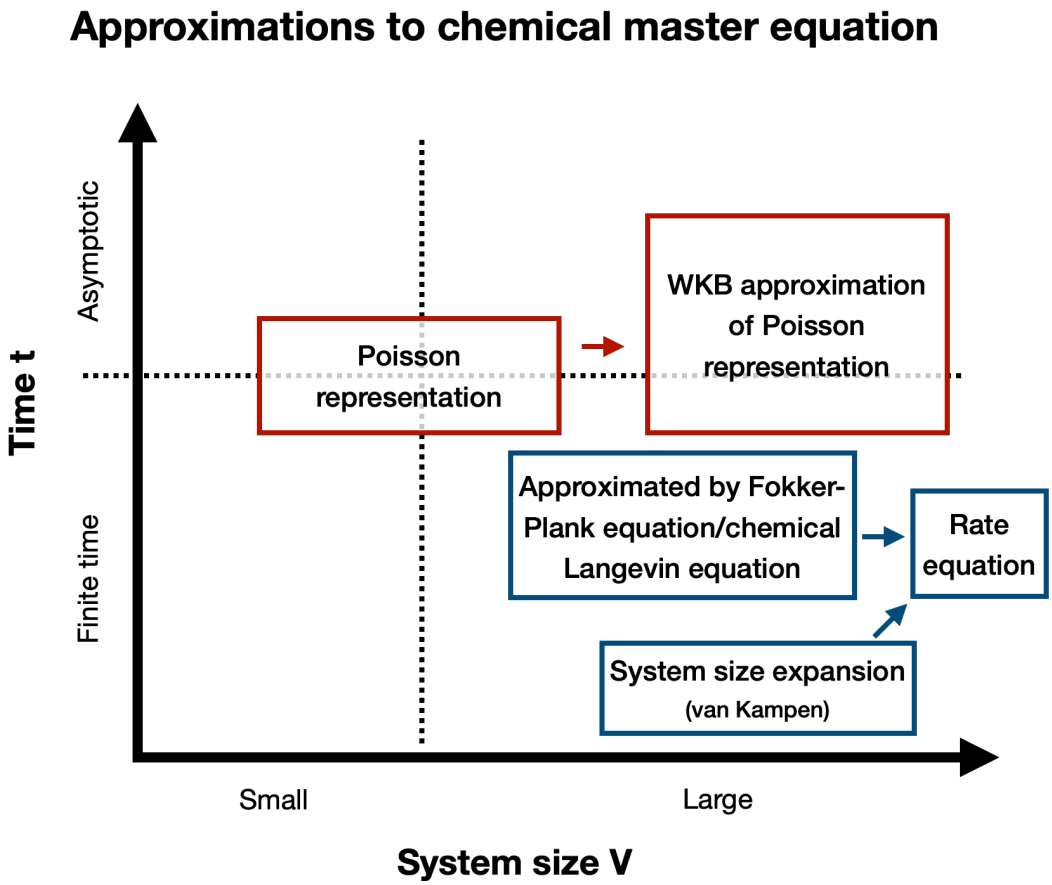

Figure 2.1: Approximations to CME

CME

Generally, a master equation for the probability distribution function $P(\mathbf{x}, t)$ can be expressed as follows:

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = \sum_{\mathbf{x}' \neq \mathbf{x}} \left[W(\mathbf{x}' \to \mathbf{x}) P(\mathbf{x}', t) - W(\mathbf{x} \to \mathbf{x}') P(\mathbf{x}, t) \right],$$

where $P(\mathbf{x}, t)$ is the probability of being in state $\mathbf{x}$ at time t and $W(\mathbf{x}' \to \mathbf{x})$ is the transition rate from state $\mathbf{x}'$ to $\mathbf{x}$.

Now consider specifically for chemically reacting system. Let a mixture of n molecular species $S_1, ..., S_n$ chemically interact through m chemical reactions $R_1, ..., R_m$:

$$\sum_{i=1}^{n} s_{ij} S_i \xrightarrow{k_j} \sum_{i=1}^{n} r_{ij} S_i, \quad j = 1, ..., m, \tag{2.1}$$

where s_{ij}, r_{ij} are stoichiometric constant and k_j is the reaction rate constant.

Let $X(t) = (X_1(t), ..., X_n(t))$ denote the numbers of molecules of species in the system at time t and P the probability. The chemical master equation assumes the system is a continuous time Markov chain and the transition rate matrix is completely determined by the reactions. Then we have that the probability distribution $P(x,t) := P(X(t) = x|X(t_0) = x_0)$ follows (Gillespie, 2000)

$$\frac{d}{dt}P(x,t) = \sum_{j=1}^{m} k_j \left(\Pi_{i=1}^{n} \frac{(x_i - (r_{ij} - s_{ij}))!}{(x_i - r_{ij})!} P(x - (r_{\cdot j} - s_{\cdot j}), t) - \Pi_{i=1}^{n} \frac{x_i!}{(x_i - s_{ij})!} P(x,t) \right). \tag{2.2}$$

Let V be the volume of the system and z be the concentration $z_i = \frac{x_i}{V}$. The corresponding reaction rate equation for Equation 2.1 is

$$\frac{dz_i}{dt} = \sum_j (r_{ij} - s_{ij}) \kappa_j \Pi_l {z_l}^{s_{lj}}, \tag{2.3}$$

where $\kappa_j = k_j V^{\sum_i s_{ij} - 1}$ is the macroscopic reaction rate.

2.2 Large-volume approximations to the chemical master equation

Kramers–Moyal expansion/Chemical Langevin equation

The first type of approximation is the chemical Langevin equation:

$$\begin{aligned} \partial_t P(y,t) = &- \sum_{j=1}^{m} k_j \sum_{i=1}^{n} \left(r_{ij} - s_{ij}\right) \cdot \partial_{y_i} \left(\prod_{l=1}^{n} \frac{y_l!}{(y_l - s_{lj})!} P(y,t) \right) \\ &+ \frac{1}{2} \sum_{j=1}^{m} k_j \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \left(r_{i_1 j} - s_{i_1 j}\right) \left(r_{i_2 j} - s_{i_2 j}\right) \partial_{y_{i_1}} \partial_{y_{i_2}} \left(\prod_{l=1}^{n} \frac{y_l!}{(y_l - s_{lj})!} P(y,t) \right). \end{aligned} \tag{2.4}$$

The first derivation to this form, which is straightforward but not mathematically rigorous, is based on the Kramers–Moyal expansion. Kramers–Moyal expansion refers to a Taylor series expansion of the chemical master equation (Equation 2.2):

$$\frac{d}{dt}P(x,t) = \sum_{j=1}^{m} k_j \sum_{k=1}^{\infty} \frac{(s_{\cdot j} - r_{\cdot j})^k}{k!} \partial_x^k \left(\Pi_{i=1}^{n} \frac{x_i!}{(x_i - s_{ij})!} P(x,t) \right). \tag{2.5}$$

By only keeping only the first two terms of the series, we derive Equation 2.4.

However, this truncation is not rigorously justified and is primarily adopted for analytical convenience. A rigorous foundation was established by Kurtz in the

1970s, albeit under strong assumptions on the reaction rate functions (Kurtz, 1978). He proved a central limit theorem for the random variable governed by the chemical master equation (CME). Later, Gillespie proposed a more intuitive derivation in 2000 (Gillespie, 2000). His approach relies on the existence of a macroscopically infinitesimal time scale during which the propensity functions remain effectively constant, while a large number of reactions occur. Both derivations follow the underlying stochastic process directly, rather than analyzing the associated CME, and are valid only over finite time intervals.

System size expansion/Linear noise approximation

A more rigorous approach to analyzing the CME than the Kramers–Moyal expansion is the system size expansion developed by Van Kampen (Van Kampen, 2007). This approach is analogous to the small noise approximation used in the analysis of the Fokker–Planck equation (C. Gardiner, 2009), which employs boundary-layer theory to handle regions near the solution of ODE (Bender and Orszag, 2010).

Let $z(t)$ be the solution to the rate equation (Equation 2.3). Then, the probability distribution $P(x,t)$ can be approximated in terms of the rescaled deviation y as

$$P(y,t) = P\left(z(t) + \frac{y}{V}, t\right),$$

which describes fluctuations around the deterministic trajectory $z(t)$. The evolution of $P(y,t)$ is governed by the following Fokker–Planck equation:

$$\begin{aligned}\partial_t P(y,t) = &- \sum_{j=1}^{m} \kappa_j \sum_{i=1}^{n} \left(r_{ij} - s_{ij}\right) \partial_{y_i} \left(\prod_{l=1}^{n} \frac{y_l!}{(y_l - s_{lj})!}\right) \partial_{y_i} \left(y P(y,t)\right) \\ &+ \frac{1}{2} \sum_{j=1}^{m} \kappa_j \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \left(r_{i_1 j} - s_{i_1 j}\right) \left(r_{i_2 j} - s_{i_2 j}\right) \left(\prod_{l=1}^{n} \frac{y_l!}{(y_l - s_{lj})!}\right) \partial_{y_{i_1}} \partial_{y_{i_2}} P(y,t).\end{aligned} \tag{2.6}$$

WKB approximation

Although not commonly presented in the literature, one can also apply the WKB (Wentzel–Kramers–Brillouin) approximation directly to Equation 2.2. Let $P(x,t) = \exp(V\phi(x,t))$ and plug it into Equation 2.5. Keeping only the $O(V)$ terms gives

$$\begin{aligned}\partial_t \phi &= \sum_{j=1}^{m} k_j \sum_{k=1}^{\infty} \frac{(s_{\cdot j} - r_{\cdot j})^k}{k!} (\partial_x \phi)^k \left(\Pi_{i=1}^{n} \frac{x_i!}{(x_i - s_{ij})!}\right) \\ &= \sum_{j=1}^{m} k_j \Pi_{i=1}^{n} \frac{x_i!}{(x_i - s_{ij})!} \exp\left(\left(s_{\cdot j} - r_{\cdot j}\right) \cdot \partial_x \phi\right).\end{aligned}$$

This can be solved numerically but does not provide an intuitive understanding of the dynamics of the systems.

Non-interchangeability of the limits of infinite system size and infinite time

It has been pointed out in several studies that the limits of infinite system size and infinite time do not, in general, commute (Hanggi et al., 1984; Baras, Mansour, and Pearson, 1996; Srivastava et al., 2002; Vellela and Qian, 2007; Vellela and Qian, 2009; Assaf and Meerson, 2017). The underlying argument can be summarized as follows: (1) both the linear noise approximation and the deterministic rate equations fail to capture multistability, and (2) the Fokker–Planck equation yields asymptotically incorrect predictions for switching times and the relative stability of states in the stationary distribution. These shortcomings arise because the linear noise approximation is only valid in the vicinity of a fixed point, while the Fokker–Planck approximation is formally justified only for finite time horizons.

Therefore, the issue is not strictly that the infinite-time and infinite-system-size limits are fundamentally non-interchangeable, but rather that the system-size expansion and Fokker–Planck approximation is not valid uniformly in time.

2.3 Large-volume approximations to the chemical master equation through Poisson representation

Below we describe a preliminary and incomplete attempt to construct a large-volume approximation to the chemical master equation that is uniform in time through Poisson representation.

The chemical master equation for bimolecular reactions is translated into a Fokker-Planck equation in the complex domain using the positive Poisson representation developed by Drummond and Gardiner (Drummond and C. W. Gardiner, 1980). We note that the large volume limit and the infinity time limit are interchangeable for Fokker-Planck equation since it is a linear PDE, i.e., the WKB method provides an uniform approximation in time. Graham and Tel has shown that the stationary distribution of Fokker-Planck equation in the weak-noise limit is associated to the ODE by the drift term (Graham and Tél, 1985). Therefore, the long-term behavior of the chemical master equation in the large volume limit is determined by the fixed points of the corresponding reaction rate equation in the complex domain. We show that a stable fixed point cannot exist outside the real axis.

The Poisson representation of $P(x,t)$ is $P(x,t) = \int_D \frac{\lambda^x}{x!} e^{-\lambda} f(\lambda,t) d\lambda$. Assume the

surface terms of the domain of the integration vanish. Then

$$\partial_t f(\lambda, t) = \sum_j \left[\left(\Pi_i (1 - \partial_{\lambda_i})^{r_{ij}} - \Pi_i (1 - \partial_{\lambda_i})^{s_{ij}} \right) k_j \Pi_i \lambda_i^{s_{ij}} \right] f(\lambda, t).$$

Let $\alpha = \frac{\lambda}{V}$ and $g(\alpha, t) := V^n f(\alpha V, t)$. Recall that $\kappa = k_j V^{\sum_i s_{ij} - 1}$. We have

$$\partial_t g(\alpha, t) = \sum_j \left[\left(\Pi_i (1 - \frac{1}{V}\partial_{\alpha_i})^{r_{ij}} - \Pi_i (1 - \frac{1}{V}\partial_{\alpha_i})^{s_{ij}} \right) V \kappa_j \Pi_i \alpha_i^{s_{ij}} \right] g(\alpha, t).$$

We only consider bimolecular reactions where $\sum_i s_{ij} \leq 2$ and $\sum_i r_{ij} \leq 2$ for all j, which includes the majority of elementary reactions.

Then

$$\partial_t g(\alpha, t) = -\sum_i \partial_{\alpha_i} \left(A_i(\alpha) g \right) + \frac{\varepsilon}{2} \sum_{i_1, i_2} \partial_{\alpha_{i_1}} \partial_{\alpha_{i_2}} \left(B_{i_1, i_2}(\alpha) g \right),$$

where $\varepsilon = \frac{1}{V}$, $A_i(\alpha) = \sum_j (r_{ij} - s_{ij}) k_j \Pi_i \alpha_i^{s_{ij}}$ and $B_{i_1,i_2}(\alpha) = \sum_j (r_{i_1 j} r_{i_2 j} - s_{i_1 j} s_{i_2 j} - \delta_{i_1, i_2} r_{ij}) k_j \Pi_i \alpha_i^{s_{ij}}$.

Note that the diffusion term is no longer positive semidefinite, and the corresponding SDE can have imaginary noise. To resolve that, Gardiner proposed the positive Poisson representation (C. Gardiner, 2009), where α is a complex variable $\alpha = x + iy$ and $d\mu(\alpha) = dxdy$. Write the drift and diffusion terms also explicitly with real and imaginary parts $A(x, y) = A_x(x, y) + iA_y(x, y)$ and $B(x, y) = C(x, y)C(x, y)^T$ where $C = C_x(x, y) + iC_y(x, y)$. By doubling the dimension, it becomes a Fokker-Planck equation with positive semidefinite:

$$\partial_t h(\mathbf{x}, \mathbf{y}, t) = -\sum_i \partial_i \left(\mathcal{A}_i h \right) + \frac{\varepsilon}{2} \sum_{i_1, i_2} \partial_{i_1} \partial_{i_2} \left(\mathcal{B}_{i_1, i_2} h \right), \tag{2.7}$$

where $\mathcal{A} = [A_x(x, y), A_y(x, y)]^T$ and $\mathcal{B} = \begin{bmatrix} C_x C_x^T & C_x C_y^T \\ C_y C_x^T & C_y C_y^T \end{bmatrix}$.

Note that Poisson representation (Equation 2.7) is valid for both finite and infinite time and for all volume.

Apply WKB approximation and assume $h(\mathbf{x}, \mathbf{y}, t) = \exp(-\frac{\sum_{n=0}^{\infty} \varepsilon^n \varphi_n(\mathbf{x}, \mathbf{y}, t)}{\varepsilon})$. To the leading order in ε, Equation 2.7 can be approximated as

$$\partial_t \varphi_0 = -\sum_i \mathcal{A}_i \partial_i \varphi_0 + \frac{1}{2} \sum_{i_1,i_2} \mathcal{B}_{i_1,i_2} \partial_{i_1} \partial_{i_2} \varphi_0. \tag{2.8}$$

Let $\phi_0 = \varphi_0(\mathbf{x}, \mathbf{y}, \infty)$, then

$$0 = -\sum_i \mathcal{A}_i \partial_i \phi_0 + \frac{1}{2} \sum_{i_1,i_2} \mathcal{B}_{i_1,i_2} \partial_{i_1} \partial_{i_2} \phi_0. \tag{2.9}$$

Graham and Tél has showed that ϕ_0, interpreted as the non-equilibrium potential, is continuous, even though its derivatives may have infinitely many discontinuities (Graham and Tél, 1985). Importantly, given that $\mathcal{B}$ is positive semidefinite, the quasi-potential φ_0 inherits the asymptotic structure of the underlying deterministic system: φ_0 attains local minima at the attractors, local maxima at the repellors, and saddle points at the deterministic saddles, provided that $\mathcal{B}$ is nonzero at those fixed points. If $\mathcal{B}$ is zero at a fixed point, then the point is absorbing and thus dynamically stable, which explains the Keizer's paradox discussed in (Vellela and Qian, 2007).

Therefore, the systems in the large volume limit is determined by the attractors of the deterministic system expanded in the complex domain:

$$\frac{d\alpha}{dt} = \mathcal{A}(\alpha). \tag{2.10}$$

We focus on systems where attractors consist only of fixed points. As the ODE (eq 2.10) is in the complex domain, all fixed points in the real domain remain fixed points, and generally, additional fixed points exist outside the real domain. However, we will show that the fixed points with non-zero imaginary parts cannot be stable.

Recall that we only consider bimolecular reactions. Therefore, we can write

$$A_i = \sum_{j=1}^{N} \sum_{k=1}^{N} a_{ijk} \alpha_j \alpha_k + \sum_{j=1}^{N} b_j \alpha_j + c_i, \tag{2.11}$$

where $a_{ijk} = a_{ikj}$. Consequently,

$$A_{x;i} = \sum_{j=1}^{N} \sum_{k=1}^{N} a_{ijk}(x_j x_k - y_j y_k) + \sum_{j=1}^{N} b_j x_j + c_i$$
$$A_{y;i} = \sum_{j=1}^{N} \sum_{k=1}^{N} a_{ijk}(x_j y_k + y_j x_k) + \sum_{j=1}^{N} b_j y_j.$$

Then the Jacobi matrix is

$$J = \begin{bmatrix} \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} \\ \frac{\partial A_y}{\partial x} & \frac{\partial A_y}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} \\ -\frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial x} \end{bmatrix}.$$

The second equality follows from Cauchy–Riemann equations and

$$\frac{\partial A_{x;i}}{\partial x_j} = \sum_{k=1}^{N} 2a_{ijk}x_k + \sum_{j=1}^{N} b_j.$$

Now suppose $\alpha^* = (x^*, y^*)^T$ is a fixed point satisfying $\mathcal{A}(\alpha) = [A_x(x, y), A_y(x, y)]^T = 0$ and denote the Jacobi matrix evaluated at α^* by J^*. Then

$$\left(\frac{\partial A_x}{\partial x} y^{*T}\right)_i = \sum_{k=1}^{N} 2a_{ijk}x_k^* y_j^* + \sum_{j=1}^{N} b_j y_j^* = \sum_{k=1}^{N} a_{ijk}(x_k^* y_j^* + x_j^* y_k^*) + \sum_{j=1}^{N} b_j y_j^* = 0,$$

and

$$(y^*, y^*)J^*(y^*, y^*)^T = [y^*, y^*] \begin{bmatrix} \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} \\ -\frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial x} \end{bmatrix} \begin{bmatrix} y^* \\ y^* \end{bmatrix} = 0.$$

Therefore, unless $y^* = 0$, J^* cannot be negative definite, which means that the fixed points with non-zero imaginary parts cannot be stable.

Chapter 3

AN EXTRINSIC NOISE MODEL FOR NORMALIZATION

Fang, Meichen and Lior Pachter (2025). "Extrinsic biological stochasticity and technical noise normalization of single-cell RNA sequencing data". In: *bioRxiv*, p. 2025.05.11.653373. DOI: 10.1101/2025.05.11.653373.

3.1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables genome-wide expression profiling at unprecedented scale, but current data are notoriously noisy, partially due to variability in sequencing depth per cell due to random sampling of libraries during sequencing. This issue becomes particularly acute when the measurement rate is low, and can cause biological signal to be overwhelmed by noise. Therefore, a standard and critical step at the beginning of scRNA-seq analysis is normalization, which is intended to mitigate the effects of technical noise before downstream analysis (Luecken and Theis, 2019). Typically, a global-scaling normalization method is used, which calculates a single normalization factor per cell (cell size factor) using the sum of total counts to adjust for variability in sequencing depth and technical artifacts. This approach is implemented in common packages for scRNA-seq analysis (Wolf, Angerer, and Theis, 2018; Hao et al., 2024; Booeshaghi, Hallgrímsdóttir, et al., 2022). Beyond this, more sophisticated approaches have been developed, including some popular methods that calculate the cell size factor by pooling cells (scran) (Lun, Bach, and Marioni, 2016) or using homogeneously expressed genes (Linnorma) (Yip et al., 2017), introducing multiple cell size factors for different groups of genes (SCnorm) (Bacher et al., 2017), and utilizing negative binomial regression (sctransform) (Hafemeister and Satija, 2019). Regardless of the methods used, the common goal of normalization techniques is to use one or more scaling factors to account for technical variation and try to remove it through methods such as scaling and regression.

We argue that common practices for normalization inadvertently remove extrinsic noise. The concept of extrinsic noise emerged in the study of biological stochasticity in gene expression, and refers to fluctuations in the cellular environment that affect all genes (Elowitz et al., 2002). Applying this concept to scRNA-seq suggests that normalization, particularly scaling, can eliminate biological variance present

in extrinsic noise. This may be critical as extrinsic noise of biological origin may carry meaningful signals relevant to specific biological questions. As a result, current normalization procedure tend to diminish biological variation (Gorin and Lior Pachter, 2022a).

To fully extend the concept of extrinsic noise to scRNA-seq, both biological and technical sources of extrinsic noise must be modeled. In fact, both biological and technical extrinsic noise are prominent and have been well characterized. In the context of biological noise (i.e., stochasticity), gene expression variability has been classified into extrinsic and intrinsic components based on their underlying mechanisms (Elowitz et al., 2002). Since extrinsic noise affects all genes within a single cell, the normalized covariance between genes has been identified as an effective measure of extrinsic noise in dual-reporter studies (Swain, Elowitz, and Siggia, 2002; Hilfinger and Paulsson, 2011; Fu and Lior Pachter, 2016). Furthermore, the impact of biological extrinsic noise on transcriptome-wide inference has been explored using a telegraph model, highlighting the importance of accounting for biological extrinsic noise itself, which can also be estimated using normalized covariance (Grima and Esmenjaud, 2024).

On the other hand, technical noise in scRNA-seq experiments has been widely studied since the development of scRNA-seq assays. For example, technical noise has been assessed experimentally using ERCC spike-ins to assess technical variance (Brennecke et al., 2013; Grün, Kester, and Oudenaarden, 2014; Kim, Kolodziejczyk, et al., 2015). ERCC-derived technical noise has been linked to global tube-to-tube variations in sequencing efficiency and a correspondence between technical noise and the observed constant coefficient of variation (CV) for highly expressed transcripts has been noted (Grün, Kester, and Oudenaarden, 2014). Currently, UMI counts are typically modeled using binomial or Poisson distributions, corresponding to Bernoulli or Poisson sampling, respectively, with cell-specific capture rates incorporated to account for detection efficiency variability (Wang et al., 2018; Sarkar and Stephens, 2021; W. Tang et al., 2023; Öcal, 2023). Notably, in the regime of low capture rates, the Poisson distribution approximates the binomial distribution.

Currently, models for scRNA-seq that account for both biological and technical extrinsic noise are typically based on specific gene expression frameworks and often assume that biological extrinsic noise influences particular kinetic parameters (W. Tang et al., 2023; Öcal, 2023). However, a more general model that accounts for extrinsic noise without assuming a specific gene expression form could be valuable.

Such a model would allow for flexible characterization and validation of biological intrinsic noise, while generating specific and potentially insightful predictions.

We develop such an extrinsic noise model for scRNA-seq data that combines the results of previous studies to account for both biological and technical sources of noise. We derive a general relationship between observed and intrinsic moments (covariance/variance) under a Bernoulli technical noise model and a scaling assumption for *in vivo* gene expression. In the specific case where genes are independent and exhibit Poisson intrinsic noise, we show that the extrinsic noise is equal to both the normalized covariance and overdispersion. This extends and unifies two previous approaches: the estimation of extrinsic noise using normalized covariance (Elowitz et al., 2002; Swain, Elowitz, and Siggia, 2002; Hilfinger and Paulsson, 2011; Fu and Lior Pachter, 2016), originally applied to biological noise, and the interpretation of technical variability as a baseline overdispersion observed in pseudocell data (Grün, Kester, and Oudenaarden, 2014), generalizing both to total extrinsic noise in scRNA-seq datasets. We test this equality on RNA solution datasets where counts are intrinsically Poisson-distributed, thereby validating the technical model as well as identifying any abnormalities. Second, when applied to single-cell datasets, this equality enables us to quantify biological and technical extrinsic noise, predict the overdispersion of intrinsically Poisson genes, and identify Poisson genes, whose total expression provides a principled approach for estimating cell size factors. Overall, we demonstrate how a mechanistic and detailed model of extrinsic noise clarifies the normalization step in scRNA-seq analysis and offers new insights into the observed variability in scRNA-seq data.

3.2 Results

A single-cell RNA-seq extrinsic noise model

Modeling extrinsic noise in scRNA-seq requires both a biological model of gene expression and a technical model of scRNA-seq measurement so as to jointly account for biological stochasticity and technical noise (Figure 3.1b). As genes can have very different expression mechanisms and resultant distributions, we do not assume any specific distribution for *in vivo* counts at first. Instead, we only assume that the means of genes in each cell are proportional to a cellular random variable c^{bio}, which represents the cell-wise size factor that summarize the biological extrinsic noise. The value of c^{bio} could be influenced by many factors such as the cell volume and the cell cycle phase. As there could be many unknown sources of cell-to-cell variability, c^{bio} is a phenomenological parameter that captures the combined effects

of various extrinsic factors. We denote the *in vivo* amount of gene j in cell i by Y_i^j and assume $\mathrm{E}\left[c_i^{bio}\right] = 1$ without loss of generality, this model for biological extrinsic noise means

$$\mathrm{E}\left[Y_i^j | c_i^{bio}\right] = c_i^{bio}\ \mathrm{E}\left[Y^j\right], \tag{3.1}$$

where $\mathrm{E}\left[Y^j\right] = \mathbb{E}_{c_i^{bio}}\left[\mathrm{E}\left[Y_i^j | c_i^{bio}\right]\right]$ is the mean of gene j across cells.

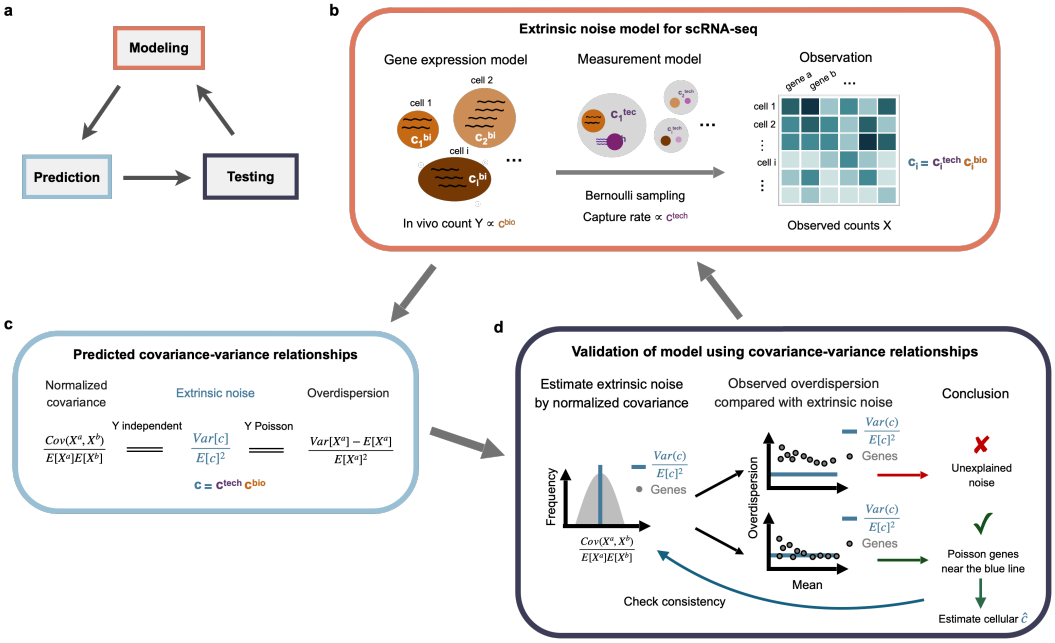

Figure 3.1: **Modeling extrinsic noise in scRNA-seq data**. **a)** Model-based closed-loop paradigm. The process begins with the formulation of mechanistic models, followed by rigorous mathematical analysis to generate testable predictions. These predictions are then tested on data, allowing models to be refined or rejected. The cycle repeats with updated models, creating an iterative loop of modeling. **b)** Schematic of the extrinsic noise model. **c)** Predicted relationships among normalized covariance, extrinsic noise, and overdispersion. **d)** Procedure for validating these relationships.

For the technical measurement model, we make much stronger assumptions: we assume a Bernoulli sampling of each transcript in single-cell experiment; this is based on previous studies (Klein et al., 2015) and the assumptions leads to a binomial distribution of observed counts given *in vivo* counts. This is also the low detection approximation of Poisson sampling (Section 3.4). Similarly, we introduce a cellular

random variable c^{tech} to summarize the relative success probability in the binomial distribution. This c^{tech} can be interpreted as affecting relative read depth during sequencing and is independent of the biological model and the *in vivo* counts. To account for differences in capture efficiency between genes, we introduce a constant capture rate, λ, as an unknown constant for each molecular species, which cancels out in normalized quantities. Finally, we denote the observed counts after single cell sequencing by X, this measurement model yields

$$X_i^j \sim binomial(n = Y_i^j, p = \lambda^j c_i^{tech}). \tag{3.2}$$

The two assumptions (Equation 3.1 and Equation 3.2) lead to simple expressions that relate the intrinsic normalized (co)variance to the observed normalized (co)variance (Section 3.4):

$$\begin{aligned}
\frac{\mathrm{Cov}\left(X^a, X^b\right)}{\mathrm{E}\left[X^a\right]\mathrm{E}\left[X^b\right]} &= \frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} + \left(1 + \frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2}\right)\frac{\mathrm{E}\left[\mathrm{Cov}\left(Y^a, Y^b | c^{bio}\right)\right]}{\mathrm{E}\left[Y^a\right]\mathrm{E}\left[Y^b\right]} \\
\frac{\mathrm{Var}\left[X^a\right] - \mathrm{E}\left[X^a\right]}{\mathrm{E}\left[X^a\right]^2} &= \frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} + \left(1 + \frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2}\right)\frac{\mathrm{E}\left[\mathrm{Var}\left[Y^a | c^{bio}\right]\right] - \mathrm{E}\left[Y^a\right]}{\mathrm{E}\left[Y^a\right]^2},
\end{aligned} \tag{3.3}$$

where a, b are gene indices and $c = c^{tech}c^{bio}$ is the overall size factor for each cell. We denote $\frac{\mathrm{Cov}\left(X^a, X^b\right)}{\mathrm{E}\left[X^a\right]\mathrm{E}\left[X^b\right]}$ as normalized covariance following previous literature (Hilfinger and Paulsson, 2011), and $\frac{\mathrm{Var}\left[X^a\right] - \mathrm{E}\left[X^a\right]}{\mathrm{E}\left[X^a\right]^2}$ as normalized variance for convenience, since it directly indicates the extent of over-dispersion. The first term on the right hand side ($\frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2}$) represents the extrinsic noise, representing a combination of both biological and technical extrinsic noise:

$$\frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} = \frac{\mathrm{Var}\left[c^{bio}\right]}{\mathrm{E}\left[c^{bio}\right]^2} + \frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2} + \frac{\mathrm{Var}\left[c^{bio}\right]}{\mathrm{E}\left[c^{bio}\right]^2}\frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2}. \tag{3.4}$$

The second terms on the right hand side of Equation 3.3 represent the contribution of intrinsic noise. For example, for two independent and intrinsically Poisson-distributed genes, the intrinsic normalized covariance ($\frac{\mathrm{E}\left[\mathrm{Cov}\left(Y^a, Y^b | c^{bio}\right)\right]}{\mathrm{E}\left[Y^a\right]\mathrm{E}\left[Y^b\right]}$) and normalized variance ($\frac{\mathrm{E}\left[\mathrm{Var}\left[Y^a | c^{bio}\right]\right] - \mathrm{E}\left[Y^a\right]}{\mathrm{E}\left[Y^a\right]^2}$) would both be 0. Therefore, the second

terms denote the effect of intrinsic normalized (co)variance convoluted with technical extrinsic noise.

The Equation 3.3 leads to two important observations. First, if we assume that genes are intrinsically uncorrelated, the normalized covariance can be used to estimate the extrinsic noise (Figure 3.1c), which is the canonical approach in previous studies on biological extrinsic noise (Swain, Elowitz, and Siggia, 2002; Hilfinger and Paulsson, 2011; Grima and Esmenjaud, 2024; Fu and Lior Pachter, 2016). Although the exact distribution of normalized covariance between uncorrelated gene pairs depends on the distribution of c, it should nevertheless center around the value of extrinsic noise, and we can use the mean/mode of the distribution to estimate the extrinsic noise. If not all but most genes are uncorrelated, the normalized covariance can still provide a reasonable estimate of the extrinsic noise using the mode of the distribution of normalized covariance. The distance between the mode and the mean provides some insight into the correlation between genes since the mode should coincide with the mean if all genes are uncorrelated. Furthermore, if we have an empirical distribution of c, we can verify whether the distribution of the normalized covariance aligns with the model (Figure 3.1d).

Second, if the genes are intrinsically Poisson distributed, then the normalized variance also equals the extrinsic noise (Figure 3.1c). Therefore, after estimating extrinsic noise using normalized covariance between genes, we can test whether each gene is Poisson distributed using this expected equality (Figure 3.1d). Note that the extrinsic noise term in normalized variance contributes to the observed over-dispersion in scRNA-seq data, as it introduces a constant offset visible when plotting the coefficient of variation (CV) against the mean. Therefore, genes are intrinsically less over-dispersed than observed counts might suggest, and it is possible for some genes to not be over-dispersed after taking into account the extrinsic noise.

In summary, assuming genes a and b are intrinsically uncorrelated and the superscript $Pois$ denotes intrinsically Poisson distributed genes, these two observations can be expressed as follows:

$$s := \frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} = \frac{\mathrm{Cov}\left(X^a, X^b\right)}{\mathrm{E}\left[X^a\right]\mathrm{E}\left[X^b\right]} = \frac{\mathrm{Var}\left[X^{Pois}\right] - \mathrm{E}\left[X^{Pois}\right]}{\mathrm{E}\left[X^{Pois}\right]^2}. \tag{3.5}$$

This moment relationship can be validated by estimating extrinsic noise and testing Poisson distributed genes (Section 3.4). Notably, the Poisson distribution property is particularly useful for cell size estimation, as the maximum likelihood estimate

(MLE) of the cell size for Poisson genes is simply their sum (Section 3.4). In practice, we filter genes based on a mean expression threshold (>0.1), as the normalized (co)variance of low-expression counts tends to be noisy. We then calculate the normalized covariance between the filtered genes to determine the mean and mode (Section 3.4). To determine whether the normalized variance equals the extrinsic noise (average normalized variance), we calculate its bootstrap confidence intervals and the equality holds for a gene if its 95% confidence interval contains the estimated value of extrinsic noise (Section 3.4). We call those genes "Poisson".

In the above derivation, we assume that c^{bio} and c^{tech} are the same for all genes or species within a single cell. However, this may not be the case and we need to validate it on data. Nevertheless, the model can be extended to cases where multiple c^{bio} and c^{tech} values exist for different groups of species and genes. In such cases, we simply need to introduce different c^{bio} and c^{tech} for each group and all equations still hold.

Validating the technical noise model with homogeneous RNA solutions

We first validated our technical noise model and the covariance-variance relationships (Equation 3.3) on scRNA-seq data of homogeneous RNA solution from K562 cells with ERCC using inDrop (Klein et al., 2015). As the RNA solution was homogeneous, the *in vivo* count for each gene in every droplet followed the same Poisson distribution and was mutually independent (Figure 3.2a). Therefore, there was no biological extrinsic stochasticity, only technical extrinsic noise. The mode of the normalized covariance was expected to be close to the mean and all genes were expected to be Poisson.

We calculated the distribution of normalized covariance within ERCC, mature mRNA and nascent mRNA respectively to see if they shared the same c^{tech} and extrinsic noise (Figure 3.3a). We found that ERCC and endogenous mRNA seemed to have slightly different c^{tech}, as the estimated extrinsic noise value of ERCC was slightly smaller that that of mRNA. Within the endogenous mRNA, the extrinsic noise appeared to be the same. This suggests that the capture mechanism of ERCC in scRNA-seq might be different from endogenous mRNA, though the difference is small.

Next, we tested whether the covariance-variance relationships held for the ERCC, mature mRNA and nascent mRNA respectively. Given the genes are independent, the covariance-variance relationship holds if the Bernoulli sampling model holds.

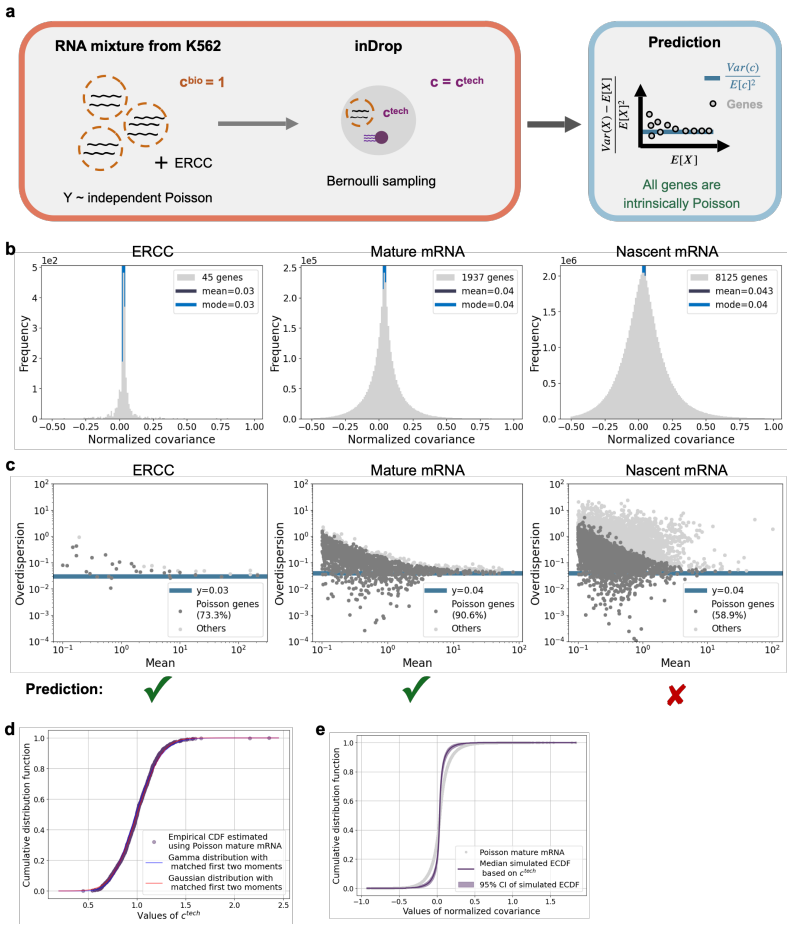

Figure 3.2: **Results for homogeneous RNA solution**. **a)** Schematic of the experiment and model predictions. **b)** Distribution of normalized covariance between gene pairs with mean expression greater than 0.1, shown separately for ERCC, mature mRNA, and nascent mRNA counts. **c)** Overdispersion-mean relationship for genes with mean expression greater than 0.1, for ERCC, mature mRNA, and nascent mRNA counts, respectively. **d)** Cumulative distribution function of c^{tech}. Gray dots represent the empirical CDF of estimated c^{tech} using selected Poisson mature mRNA counts. The blue line shows a Gamma distribution fitted by matching the first two moments (mean and variance), and the red line shows a Gaussian distribution with the same mean and variance. **e)** Cumulative distribution function of normalized covariance. Gray dots represent the empirical CDF of normalized covariance for mature mRNA counts shown in **b)**. Using the estimated c^{tech} values and the mean expression levels of the selected genes, 1,000 bootstrap samples were generated. The purple line indicates the median empirical CDF across bootstrap replicates, and the light purple band represents the 95% confidence interval.

We plotted the normalized variance against mean, and colored Poisson genes on which the relationships held among all genes (Figure 3.2b). We also calculated

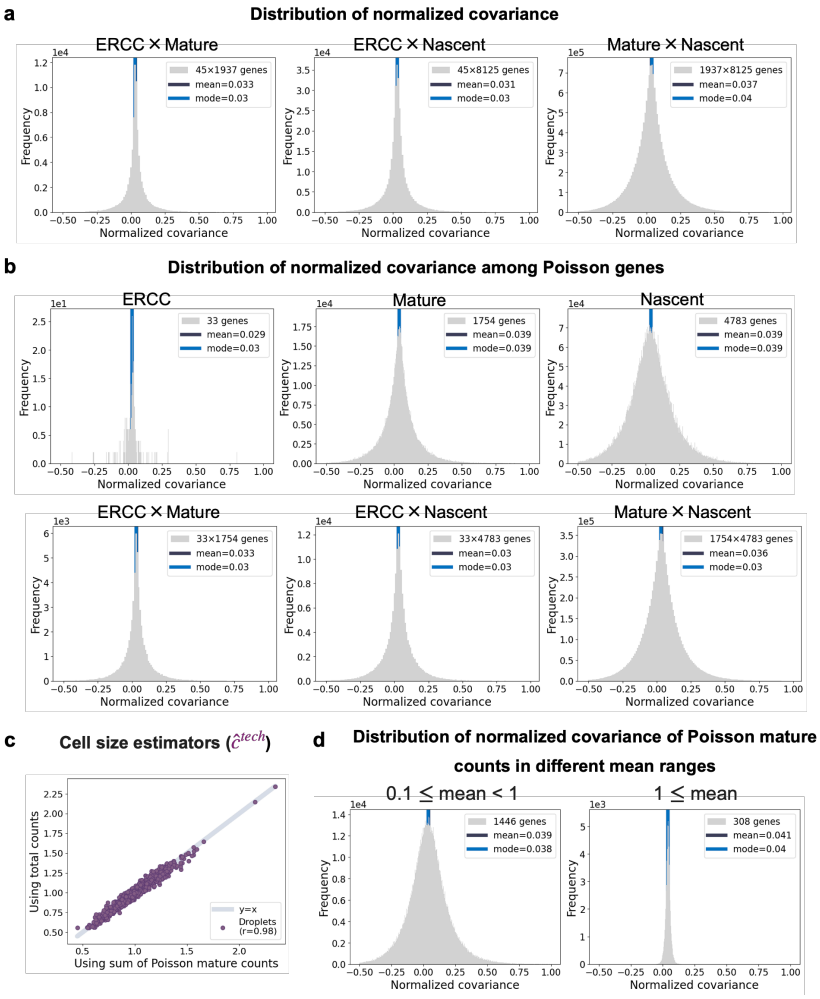

Figure 3.3: **Supplementary figures for homogeneous RNA solution a)** Distribution of normalized covariance between different species. **b)** Distribution of normalized covariance between Poisson genes of different species. **c)** Comparison of cell size estimators using the sum of total counts and Poisson mature counts. **d)** Distribution of normalized covariance within Poisson mature counts across different mean expression ranges.

the normalized covariance among selected Poisson genes and found that they were consistent (Figure 3.3b). The Bernoulli sampling model seemed to work well for the ERCC and mature mRNA but not for the nascent mRNA: the normalized variance of almost half of the nascent counts was much noisier than predicted, while most of the mature mRNA (90%) was within the 95% confidence intervals (Figure 3.2b). This suggests that the nascent mRNA requires a different measurement model than Bernoulli sampling.

We then sought to characterize c $(= c^{tech})$, which represents the cell size factors

commonly used in scRNA-seq analysis (Luecken and Theis, 2019). We estimated c^{tech} as the sum of Poisson-distributed mature RNA counts and found that it correlated well with the total count sum, which was expected since there were no differentially expressed genes (Figure 3.3c). As the negative binomial distribution is commonly used to model mRNA counts in both pseudo and real cells, which implies that cell size follows a gamma distribution given a Poisson distribution of *in vivo* counts, we asked whether c^{tech} indeed followed a gamma distribution. We plotted and compared the empirical cumulative distribution function (CDF) of estimated c^{tech}, computed from the sum of Poisson-distributed mature RNA counts, with the CDFs of gamma and Gaussian distributions that shared the same first two moments as c^{tech}. We found that the distribution of c^{tech} followed a gamma distribution reasonably well, but also fit a Gaussian distribution equally well, if not better (Figure 3.2d), which is consistent with the fact that the gamma distribution approaches the Gaussian distribution when the shape parameter is large.

To assess whether a single c^{tech} was shared across all genes, we compared the empirical CDF of the normalized covariance of mature mRNAs with simulations generated from a Poisson distribution coupled with the empirical distribution of c^{tech} (Figure 3.2e). The empirical CDF was less sharp than the empirical CDF of simulations, which suggested that a single c^{tech} could not fully explain the variability, even for mature mRNA. The single c^{tech} per cell is the average of the distribution of capture rates for different mature mRNAs in a cell. The capture rates did not seem to relate to the expression levels (Figure 3.3d). Nevertheless, based on the consistency between the mean and mode of the normalized covariance among all genes and the selected Poisson genes, we concluded that a single c^{tech}, while not entirely accurate, provided a useful approximation.

In summary, homogeneous RNA solution data revealed that ERCCs, mature mRNAs, and nascent mRNAs are captured through distinct mechanisms, which leads to varying levels of extrinsic noise within ERCCs and endogenous mRNAs and a significantly higher variance in nascent mRNA than would be expected under a simple Bernoulli sampling model. We therefore advocate for more control experiments of this kind to validate technical noise models prior to large-scale data generation.

Validating the "biological" extrinsic noise of heterogeneous RNA solution

To validate our interpretation of extrinsic noise using a heterogeneous RNA solution we examined CEL-seq2 data that contained ERCC spike-ins alongside varying

amounts of endogenous RNA (Tian et al., 2019). The experiment comprised eight distinct RNA mixtures extracted from three human lung cancer cell lines, each present at four RNA amounts within different wells (Figure 3.4a). Here, the eight RNA mixtures represented eight cell types and the four RNA amounts. Since cell counts were derived from wells containing RNA in solution, we assumed that they were independent and Poisson distributed. However, the different RNA amounts could give rise to biological extrinsic noise and the cell type specific mean parameters could lead to intrinsic variance and covariance. Assuming that the three human lung cancer cell lines have similar concentrations for most genes, we expected the biological extrinsic noise to arise mostly from the variation of RNA amounts. Specifically, the biological extrinsic noise $\frac{\mathrm{Var}\left[c^{bio}\right]}{\mathrm{E}\left[c^{bio}\right]^2}$ equals the CV^2 of RNA amounts, which could be calculated to be approximately or slightly above 0.33 based on the experimental design (Section 3.4). Furthermore, those genes that did not vary across the three human lung cancer cell lines were intrinsically Poisson, meaning they had a normalized variance equal to the extrinsic noise, similar to a homogeneous mRNA solution. In contrast, genes that did vary displayed greater dispersion than intrinsically Poisson genes, resulting in a normalized variance that exceeded the extrinsic noise (Prediction in Figure 3.4a).

We estimated the technical extrinsic noise using the ERCC spike-ins, and also estimate the total extrinsic noise using mature mRNA and nascent mRNA respectively, to see if they were the same. The total extrinsic noise differed within mature and nascent mRNA: the estimated extrinsic noise was slightly higher for mature mRNA (0.46) than nascent mRNA (0.42) (Figure 3.4b). The distribution of normalized covariance between mature and nascent mRNA had a similar mode to that of nascent mRNA (Figure 3.5a), indicating that the extrinsic noise of mature mRNA included both components shared with nascent mRNA and components unique to mature mRNA. Most ERCC (92%) and mature mRNA (87%) fell within the 95% confidence intervals and satisfied the Poisson criteria, whereas nascent mRNA counts had a smaller percentage (78%) and were noisier with larger normalized variance (Figure 3.4b). As a consistency check, the normalized covariance between Poisson genes showed similar values, with differences within 0.01 (Figure 3.5b). Based on these observations, we speculate that the capture of mature and nascent mRNA shared similar mechanisms as well as distinct differences, which led to the small difference in extrinsic noise. Importantly, nascent mRNA is likely to require a slightly nosier technical model than Bernoulli sampling.

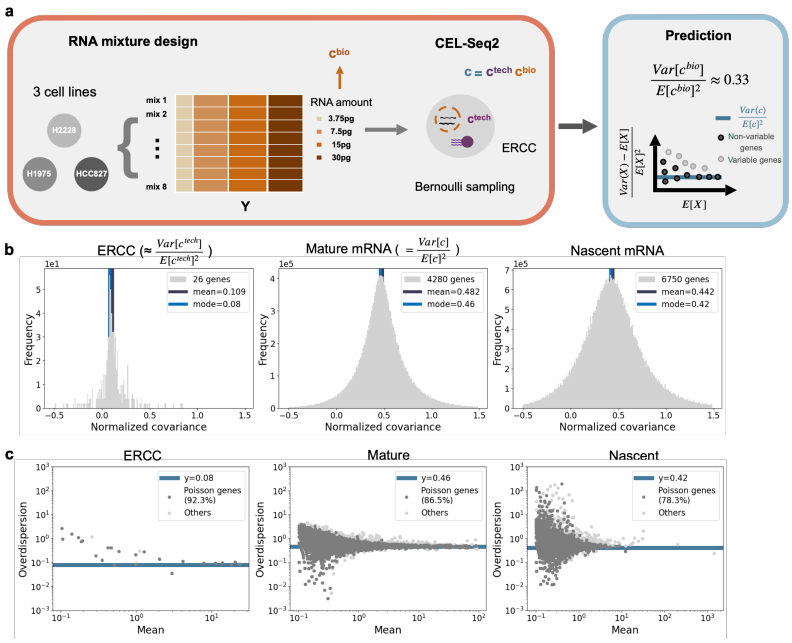

Figure 3.4: **Results for heterogeneous RNA solution**. **a)** Schematic of the experiment and model predictions. **b)** Distribution of normalized covariance between gene pairs with mean expression greater than 0.1, shown separately for ERCC, mature mRNA, and nascent mRNA counts. **c)** Overdispersion-mean relationship for genes with mean expression greater than 0.1, for ERCC, mature mRNA, and nascent mRNA counts, respectively.

Therefore, we used normalized covariance of mature mRNA for calculating the total extrinsic noise and ERCC for the technical extrinsic noise to estimate the biological extrinsic noise based on Equation 3.4. The estimated biological extrinsic noise was 0.35, which was reasonably closed to the expectation (0.33). Then we calculated the total and technical cell size (c_i and c_i^{tech}) using the sums of Poisson mature mRNA and ERCC respectively, which were similar to those using total counts (Figure 3.5c). We estimated biological cell size (c_i^{bio}) by taking the ratio of c_i and c_i^{tech}, and compared the distribution of c_i^{bio} to the expected distribution from the experimental design (Section 3.4). Cells were centered around the expected values but the variance seemed to be large (Figure 3.5d and e).

Decomposing biological and technical extrinsic noise using species-mixing experiments

Based on the results obtained with RNA in solution, we decided to focus on the mature counts. However, in this context of experiments with individual cells, the logic is reversed. Unlike in RNA solution, where genes can be assumed to follow

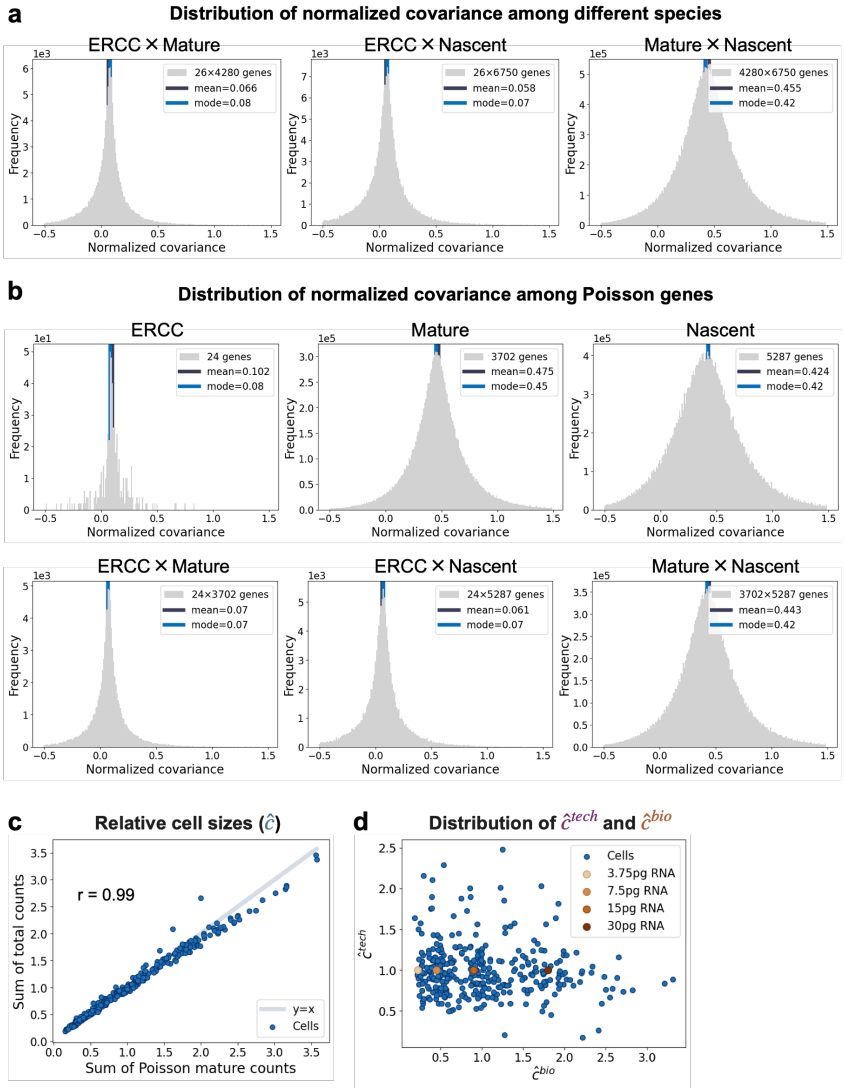

Figure 3.5: **Supplementary figures for heterogeneous RNA solution a)** Distribution of normalized covariance between different species. **b)** Distribution of normalized covariance between Poisson genes of different species. **c)** Comparison of cell size estimators using the sum of total counts and Poisson mature counts. **d)** Distribution of c^{tech} and c^{bio}. The values of c^{tech} and c are estimated from the total Poisson ERCC counts and mature mRNA counts, respectively. Based on these estimates, c^{bio} is computed. The brown dots are the theoretical c^{bio}.

a Poisson distribution in pseudocells, these assumptions do not inherently hold *in vivo*. Instead, by testing the relationship between extrinsic noise and overdispersion across genes, we identified those genes for which the assumptions of the Poisson distribution hold, at least approximately. This enables genome-scale understanding of gene expression noise and provided an additional piece of evidence in the context

of previously inconsistent findings regarding biological variability (Dar et al., 2012; Battich, Stoeger, and Pelkmans, 2015).

Therefore, we used an iterative approach to estimate extrinsic noise and to identify "Poisson" genes (Figure 3.6). Starting with all genes, we calculated the normalized covariance and estimated the extrinsic noise, which was then used to identify genes whose normalized variances were close to the extrinsic noise. Next, we re-estimated the extrinsic noise using the normalized covariance between selected genes and compared it to the previous value. This process was repeated iteratively until the extrinsic noise estimate stabilized (with differences within 10%), though typically, at most one iteration was needed. Therefore, we assumed that these selected genes were independent (on average) and had intrinsic variance similar to a Poisson distribution, so we referred to them as "Poisson". Although their exact distributions may deviate from a Poisson model, these genes were likely to exhibit low variability across cells, rendering them appropriate for estimating cell size.

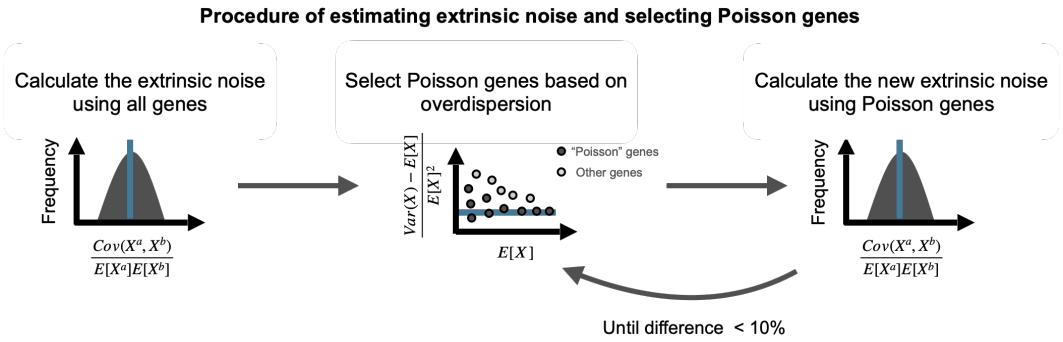

Figure 3.6: **Procedure on single cell datasets.**

We first sought to characterize the biological and technical contribution of extrinsic noise. To measure technical extrinsic noise, we needed some control RNA in the same cell. Usually, ERCC spike-ins are used as external controls to quantify technical variance. However, here we utilized the ambient mRNA as the control mRNA by leveraging species-mixing experiments, which are commonly used to assess doublet rates. In these scRNA-seq experiments, human and mouse cells are typically mixed, resulting in ambient mRNA from both species in droplets containing cells from only one species. Then the ambient mRNA from the other species serves as an external RNA control for technical extrinsic noise (Figure 3.7a).

In light of this, we calculated the biological and technical extrinsic noise of three 10x human-mouse mixture datasets. We used mature mRNA of the corresponding

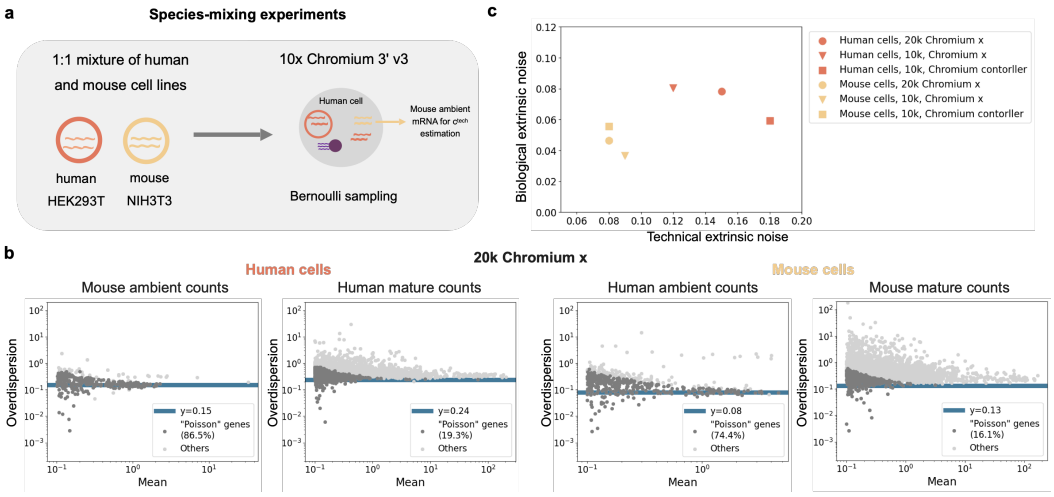

Figure 3.7: **Extrinsic noise in species-mixing experiments**. **a)** Schematic of the species-mixing experiment. **b)** Overdispersion-mean relationships for human and mouse genes in both human and mouse cells in the 20k Chromium X dataset. **c)** Biological and technical extrinsic noise in three species-mixing experiments.

species to estimate total extrinsic noise, and total ambient mRNA from the other species to estimate technical extrinsic noise. We did not distinguish between nascent and mature and used the total counts when calculating normalized covariance because their counts are low. For example, in droplets containing only human cells, the technical extrinsic noise estimated from mouse ambient mRNA is 0.15, and, similar to RNA solution, the genes were expected to follow a Poisson distribution. In contrast, the total extrinsic noise estimated from human mRNA is 0.24, with fewer than 20% of human transcripts falling within the Poisson range (Figure 3.7b). The differing percentages between ambient and cellular mRNA highlight that only a small fraction of *in vivo* transcript counts potentially follow a Poisson distribution. Based on the total and technical extrinsic noise, we estimate the biological extrinsic noise based on Equation 3.4, which leads to 0.15 for human cells (Figure 3.7b). Both the biological and technical contributions to extrinsic noise are substantial (Figure 3.7c). The estimated biological extrinsic noise are relatively robust across three datasets, and the mouse cells (NIH3T3) seem to be slightly more homogeneous than human cells (HEK293T).

Characterizing extrinsic noise and cell size factors on scRNA-seq data

Given the non-negligible contribution of biological extrinsic noise even in homogeneous cell lines, we argue that even when biological extrinsic noise cannot be explicitly distinguished due to the absence of control mRNA, a substantial portion

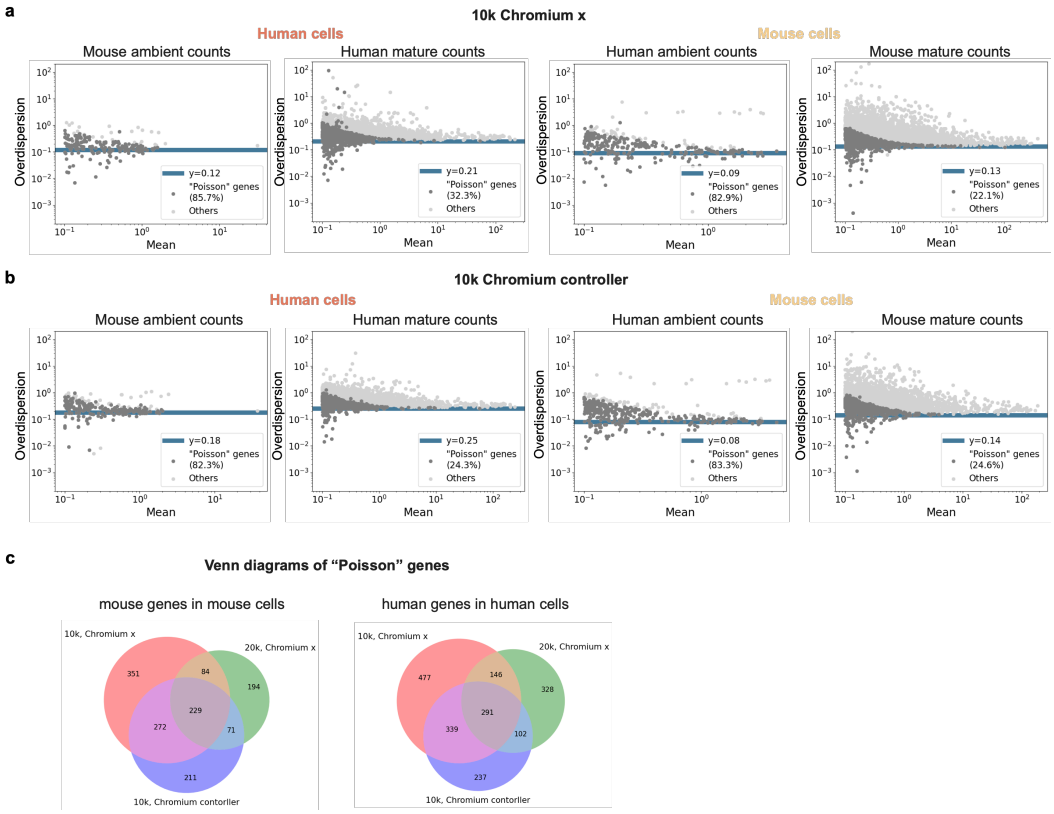

Figure 3.8: **Supplementary figures for species-mixing experiments**. **a)** Overdispersion-mean relationships for human and mouse genes in both human and mouse cells in the 10k Chromium X dataset. **b)** Overdispersion-mean relationships for human and mouse genes in both human and mouse cells in the 10k Chromium controller dataset. **c)** Venn diagram of selected Poisson genes across the three datasets.

of the total extrinsic noise likely reflects underlying biological variation and should not be disregarded. Therefore, we applied our procedure to several scRNA-seq datasets and characterized both extrinsic noise and cell size factors.

We first investigated whether extrinsic noise is related to the average abundance of genes, a topic that has been debated in previous studies (Hafemeister and Satija, 2019; Lause, Berens, and Kobak, 2021). For our analysis, we selected the 10x Flex K562 datasets because the probes covering exon junctions yield more abundant and accurate mature mRNA counts. We then plotted the distribution of normalized covariance across genes with varying mean expression levels and found that the modes of the distributions were identical (0.23), with comparable means (Figure 3.9a). The percentage of "Poisson" genes is also comparable to those observed in the species-mixing data (Figure 3.9b). The distributions of normalized covariance

across “Poisson” genes with varying mean expression levels also show no difference with the same modes (Figure 3.10a). We concluded that extrinsic noise and the resulting baseline overdispersion are not related to the average abundance of genes.

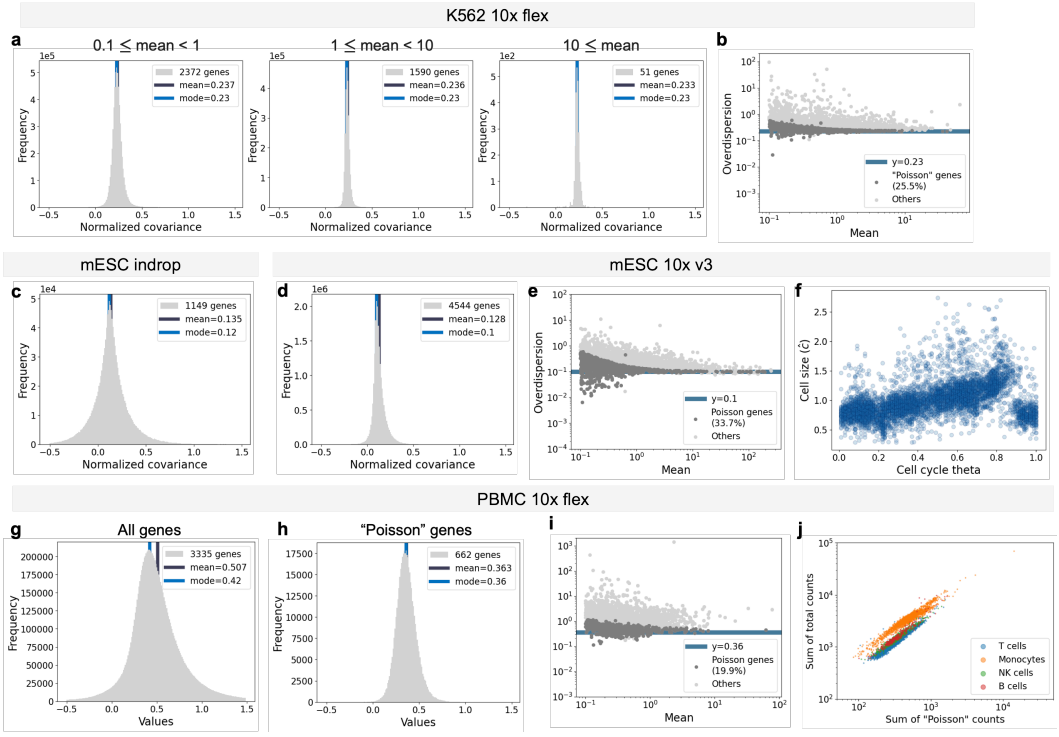

Figure 3.9: **Extrinsic noise in single cell datasets**. **a)** Distribution of normalized covariance within genes across different mean expression ranges for the K562 10x Flex dataset. **b)** Overdispersion-mean relationship for the K562 10x Flex dataset. **c)** Distribution of normalized covariance for the mESC inDrop dataset. **d)** Distribution of normalized covariance for the mESC 10x 3’ v3 dataset. **e)** Overdispersion-mean relationship for the mESC 10x 3’ v3 dataset. **f)** Cell size along cell cycle. Cell sizes are estimated using Poisson genes from panel e). Cell cycle progression is denoted by cell cycle theta, as reported by Riba et al. (2022). **g)** Distribution of normalized covariance between gene pairs with mean expression greater than 0.1 for the PBMC dataset. **h)** Distribution of normalized covariance between selected Poisson genes for the PBMC dataset. **i)** Overdispersion-mean relationship for the PBMC dataset. **ii)** Sum of total counts against sum of Poisson counts, colored by cell types.

We then investigated whether extrinsic noise is associated with cell cycle progression. To do this, we used mouse embryonic stem cells (mESC) data with inferred cell cycle stages (Riba et al., 2022). We estimated the extrinsic noise (Figure 3.9c) and also compared it to that of mESC inDrop data (Klein et al., 2015), finding that the estimates were similar (Figure 3.9c). This indicates the robustness of the extrinsic noise across different datasets. We selected Poisson genes (Figure 3.9e),

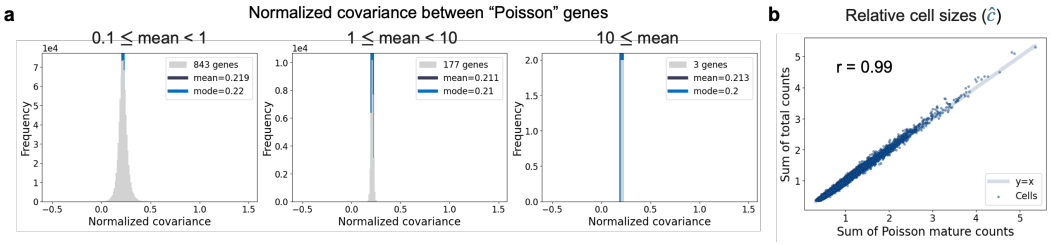

Figure 3.10: **Supplementary figures for K562 10x flex dataset**. **a)** Distribution of normalized covariance within Poisson mature counts across different mean expression ranges. **b)** Comparison of cell size estimators using the sum of total counts and Poisson mature counts. r denotes the Pearson correlation coefficient.

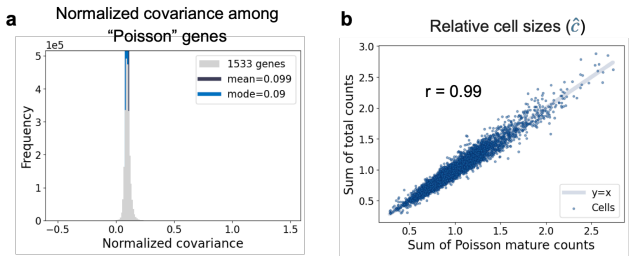

Figure 3.11: **Supplementary figures for mESC 10x dataset**. **a)** Distribution of normalized covariance within Poisson mature counts. **b)** Comparison of cell size estimators using the sum of total counts and Poisson mature counts. r denotes the Pearson correlation coefficient.

and estimate $\hat{c}$ as the sum of "Poisson" mature RNA counts, which correlates well with the total count sum (Figure 3.11c). We plotted the estimated cell size factors ($\hat{c}$) along the inferred transcriptional phase (cell cycle θ) from Riba et al. and observed a clear pattern of cell size variation across the cell cycle (Figure 3.9f), which confirms that the cell cycle contributes to extrinsic noise.

Up to this point, the sum of Poisson counts had shown a perfect correlation with the total counts. However, this may not be the case for heterogeneous cells, i.e., datasets consisting of different cell types. To investigate this, we applied our approach to the peripheral blood mononuclear cells (PBMC) dataset generated using 10x flex technology (10x Genomics, 2024). We found that the extrinsic noise in PBMCs is significantly higher than that observed in homogeneous cell types such as mESC and K562. The distribution of normalized covariance across all genes is right-skewed (Figure 3.9g), suggesting that many genes exhibit positive correlations. On the other hand, the distribution across Poisson genes is more symmetrical, with the mode and mean closely aligned (Figure 3.9h). The percentage of "Poisson" genes remains

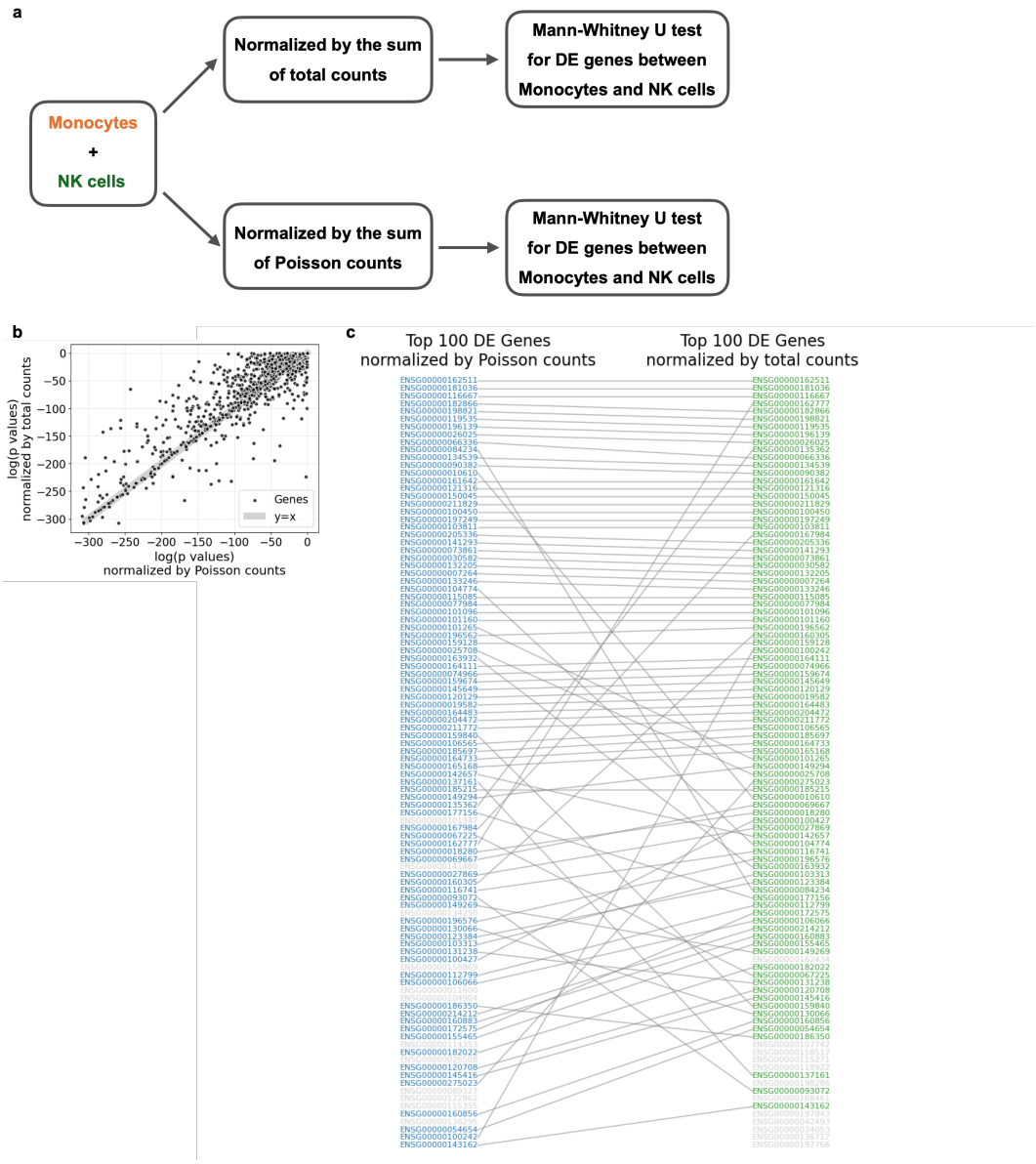

Figure 3.12: **Supplementary figures for PBMC dataset**. **a)** Procedure for comparing differential expression (DE) analysis results using different cell sizes. **b)** P-values from the Mann–Whitney U test comparing gene expression between monocytes and NK cells, computed after normalizing the data using different cell size estimates. **c)** Correspondence between top 100 genes after two normalizations.

similar (Figure 3.9i). However, the estimated $\hat{c}$ as the sum of "Poisson" mature RNA counts no longer aligns well with the total count sum (Figure 3.9j). We used the Leiden algorithm to cluster cells into T cells, Natural killer (NK) cells, B cells, and Monocytes based on marker genes. Different cell types within PBMC appear to have varying ratios of "Poisson" to total counts (Figure 3.9j), likely reflecting the

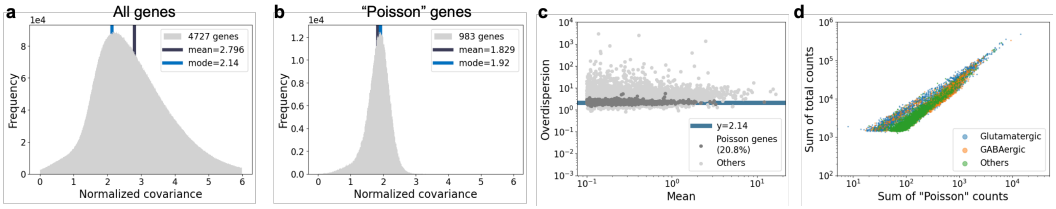

Figure 3.13: **Results for mouse forebrain dataset**. **a)** Distribution of normalized covariance between gene pairs with mean expression greater than 0.1 for the PBMC dataset. **b)** Distribution of normalized covariance between selected Poisson genes for the PBMC dataset. **c)** Overdispersion-mean relationship for the PBMC dataset. **d)** Sum of total counts against sum of Poisson counts, colored by cell types.

presence of highly differentially expressed genes specific to monocytes. Therefore, using "Poisson" counts or total counts will result in different cell size factors. To demonstrate the impact on downstream analysis, we normalized the raw counts using both the total UMI count and the sum of Poisson gene counts, respectively, and performed a Mann–Whitney U test to identify differentially expressed genes between monocytes and natural killer cells. We compared the resulting p-values (Figure 3.12a), and listed the top 100 differentially expressed genes identified under each normalization method for comparison (Figure 3.12b). We found that this phenomenon is dataset-specific and depends on the underlying cellular composition, as demonstrated by the 10x mouse forebrain data (10x Genomics, 2023), where the sum of 'Poisson' mature RNA counts aligns better with the total count sum (Figure 3.13d).

3.3 Discussion

In this work, we clarify the underlying assumptions of extrinsic noise in scRNA-seq normalization and describe an extrinsic noise model that has only been implicitly recognized in previous studies. This model establishes a direct relationship among normalized covariance, extrinsic noise, and the overdispersion observed in intrinsically Poisson genes. This relationship enables us to validate the model using RNA solution data and to identify genes whose expression variance is consistent with a Poisson distribution. By providing a baseline for overdispersion, extrinsic noise reveals that much of the observed overdispersion in scRNA-seq data can still be explained by genes that are intrinsically Poisson.

Importantly, we have shown how a mechanistic model can lead to testable predictions, and how validating these predictions can either support the model or prompt

the development of alternative explanations (Phillips, 2015). Specifically, we found that Bernoulli sampling is applicable only to mature RNA counts, likely due to differences in capture mechanisms between mature and nascent mRNA. Even for mature counts, using a single cell size factor remains a coarse approximation. Furthermore, we observed that the overdispersion in some datasets cannot be fully explained by extrinsic noise, as seen in the case of STORM-seq datasets of K562 cells (Johnson et al., 2022). Despite exhibiting similar levels of extrinsic noise to the 10x Flex dataset in Figure 3.9a (Figure 3.14a), the mean-overdispersion relationship and the behavior of Poisson genes suggest that our model does not apply in this case (Figure 3.14b). Given that the technical noise model may vary across species and technologies, we advocate for more careful assessment in future experimental designs, recommending that the technical noise model be characterized prior to large-scale data generation.

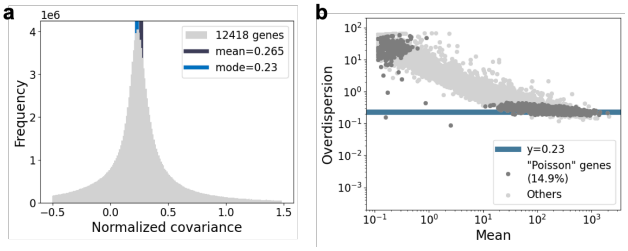

Figure 3.14: **Results for K562 STORM-seq dataset**. **a)** Distribution of normalized covariance between gene pairs with mean expression greater than 0.1. **b)** Overdispersion-mean relationship.

Beyond quantifying biological and technical extrinsic noise, a key motivation for modeling extrinsic noise and cell size is to enhance the accuracy of downstream data analysis. Rather than simply normalizing total counts by cell size, we advocate for explicitly incorporating the cell size factor when modeling variable genes with more complex gene expression models, such as those beyond constitutive expression and Poisson distributions. The cell size estimators derived from Poisson genes can be treated as constants and provided as inputs to the inference process, thereby simplifying the modeling of other variable genes. Because these estimators are based on two orthogonal groups of genes, namely those used for cell size estimation and those being modeled, this approach effectively avoids the issue of "double-dipping" and ensures a more robust and reliable analysis.

We have provided only preliminary insights into extrinsic noise, as our analysis is based on modeling single genes and does not specify a detailed gene expression

model. As a result, we cannot determine the exact forms of variance and covariance beyond what is expected from a Poisson distribution. For genes that follow Poisson statistics, all biological extrinsic noise arises from variation in their mean expression levels. In such cases, it is sufficient to decompose extrinsic noise into biological and technical components. However, for more variable genes that exhibit super-Poissonian variance, more sophisticated models such as bursty transcription are needed to accurately capture their expression dynamics (Golding et al., 2005). These models introduce additional parameters, such as burst frequency and burst size, to account for the excess variability. While most studies assume that only burst size scales with cell size (Grima and Esmenjaud, 2024; W. Tang et al., 2023; Öcal, 2023), we show that biological extrinsic noise can, in fact, be further decomposed. This allows for a more detailed, quantitative dissection of how each parameter contributes to extrinsic noise (see Section 3.4). We advocate for future studies to adopt more comprehensive modeling approaches in order to deepen our understanding of the sources and mechanisms underlying gene expression variability.

3.4 Methods

Extrinsic noise model

Extrinsic noise is global to a cell and contains both biological and technical components. Therefore, we introduce two random variables to represent biological and technical cell factors. Specifically, for the biological model, we assume that each cell has a random variable c^{bio} and the mean of every gene is proportional to this value. c_{bio} could result from the cell volume, cell cycle and factors that effect all genes. Without loss of generality, we assume $\mathrm{E}\left[c_{bio}\right] = 1$. Denote the *in vivo* number of gene j in cell i by Y_i^j, and this model for biological extrinsic noise means

$$\mathrm{E}\left[Y_i^j | c_i^{bio}\right] = c_i^{bio}\ \mathrm{E}\left[Y^j\right], \tag{3.6}$$

where $\mathrm{E}\left[Y^j\right] = \mathbb{E}_{c_i^{bio}}\left[\mathrm{E}\left[Y_i^j | c_i^{bio}\right]\right]$ is the mean of gene j across cells.

As a consequence of law of total variance, for covariance and variance of gene a and b across cells, we have

$$
\begin{aligned}
\operatorname{Cov}\left(Y^a, Y^b\right) &= \mathbb{E}_{c_i^{bio}}\left[\operatorname{Cov}\left(Y_i^a, Y_i^b | c_i^{bio}\right)\right] + \operatorname{Cov}\left(\mathrm{E}\left[Y^a | c^{bio}\right], \mathrm{E}\left[Y^b | c^{bio}\right]\right) \\
&= \mathbb{E}_{c_i^{bio}}\left[\operatorname{Cov}\left(Y_i^a, Y_i^b | c^{bio}\right)\right] + \operatorname{Var}\left[c^{bio}\right] \mathrm{E}\left[Y^a\right] \mathrm{E}\left[Y^b\right] \\
\operatorname{Var}\left[Y^a\right] &= \mathbb{E}_{c_i^{bio}}\left[\operatorname{Var}\left[Y_i^a | c^{bio}\right]\right] + \operatorname{Var}\left[\mathrm{E}\left[Y^a | c^{bio}\right]\right] \\
&= \mathbb{E}_{c_i^{bio}}\left[\operatorname{Var}\left[Y_i^a | c^{bio}\right]\right] + \operatorname{Var}\left[c^{bio}\right] \mathrm{E}\left[Y^a\right]^2 .
\end{aligned}
\tag{3.7}
$$

The first terms on the right-hand side of both equations describe the intrinsic covariance and variance, which depend on and also reflect the gene expression mechanism. The second terms describe the extrinsic noise introduced by c^{bio}. This separation of intrinsic and extrinsic terms has been addressed in previous studies (Elowitz et al., 2002; Swain, Elowitz, and Siggia, 2002; Hilfinger and Paulsson, 2011).

For a technical noise model, we assume Bernoulli sampling of transcripts in single-cell sequencing experiments, which leads to binomial distribution of observed counts given *in vivo* counts. Similarly we assume each cell has a random variable c_{tech} and the success probability in binomial distribution is proportional to this value. This c_{tech} could be interpreted as relative read depth during sequencing and is independent of the biological model and *in vivo* counts. We introduce a constant capture rate λ for each species of molecule so that we can again assume $\mathrm{E}\left[c_{tech}\right] = 1$. Denote the observed counts after single cell sequencing by X, this technical model means

$$
X_i^j \sim binomial(n = Y_i^j, p = \lambda^j c_i^{tech}), \tag{3.8}
$$

$$
\mathrm{E}\left[X_i^j | c_i^{tech}, Y_i^j\right] = c_i^{tech} \lambda^j Y_i^j, \tag{3.9}
$$

$$
\operatorname{Var}\left[X_i^j | c_i^{tech}, Y_i^j\right] = c_i^{tech} \lambda^j (1 - c_i^{tech} \lambda^j) Y_i^j . \tag{3.10}
$$

Using law of total variance (covariance) gives

$$\begin{aligned}
\operatorname{Cov}\left(X^a, X^b\right) &= \mathbb{E}\big[\operatorname{Cov}\left(X_i^a, X_i^b | c_i^{tech}, Y_i^a, Y_i^b\right)\big] \\
&\quad + \operatorname{Cov}\left(\operatorname{E}\left[X^a | c^{tech}, Y_i^a, Y_i^b\right], \operatorname{E}\left[X^b | c^{tech}, Y_i^a, Y_i^b\right]\right) \\
&= \operatorname{Cov}\left(c^{tech}\lambda^a Y_i^a, c^{tech}\lambda^b Y_i^b\right) \\
&= \mathbb{E}\big[\operatorname{Cov}\left(c^{tech}\lambda^a Y_i^a, c^{tech}\lambda^b Y_i^b | Y_i^a, Y_i^b\right)\big] \\
&\quad + \operatorname{Cov}\left(\operatorname{E}\left[c^{tech}\lambda^a Y_i^a | Y_i^a, Y_i^b\right], \operatorname{E}\left[c^{tech}\lambda^b Y_i^b | Y_i^a, Y_i^b\right]\right) \\
&= \lambda^a\lambda^b \operatorname{E}\left[Y_i^a Y_i^b\right] \operatorname{Var}\left[c^{tech}\right] + \operatorname{E}\left[c^{tech}\right]^2 \lambda^a\lambda^b \operatorname{Cov}\left(Y_i^a, Y_i^b\right) \\
&= (\operatorname{Var}\left[c^{tech}\right] + \operatorname{E}\left[c^{tech}\right]^2)\lambda^a\lambda^b \operatorname{Cov}\left(Y_i^a, Y_i^b\right) \\
&\quad + \operatorname{Var}\left[c^{tech}\right] \lambda^a\lambda^b \operatorname{E}\left[Y^a\right] \operatorname{E}\left[Y^b\right]
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Var}\left[X^a\right] &= \mathbb{E}\big[\operatorname{Var}\left[X^a | c^{tech}, Y^a\right]\big] + \operatorname{Var}\left[\operatorname{E}\left[X^a | c^{tech}, Y_i^a\right]\right] \\
&= \mathbb{E}\big[c_i^{tech}\lambda^j(1 - c_i^{tech}\lambda^j)Y^a\big] + \operatorname{Var}\left[c^{tech}\lambda^a Y^a\right] \\
&= \mathbb{E}\big[c_i^{tech}\lambda^j(1 - c_i^{tech}\lambda^j)\big] \operatorname{E}\left[Y^a\right] + \operatorname{E}\left[\operatorname{Var}\left[c^{tech}\lambda^a Y^a | Y^a\right]\right] \\
&\quad + \operatorname{Var}\left[\operatorname{E}\left[c^{tech}\lambda^a Y^a | Y^a\right]\right] \\
&= \lambda^a \operatorname{E}\left[c^{tech}\right] \operatorname{E}\left[Y^a\right] - (\lambda^a)^2 \operatorname{E}\left[(c^{tech})^2\right] \operatorname{E}\left[Y^a\right] \\
&\quad + \operatorname{E}\left[\operatorname{Var}\left[c^{tech}\right] (\lambda^a Y^a)^2\right] + \operatorname{Var}\left[\operatorname{E}\left[c^{tech}\right] \lambda^a Y^a\right] \\
&= \lambda^a \operatorname{E}\left[c^{tech}\right] \operatorname{E}\left[Y^a\right] - (\lambda^a)^2 \operatorname{E}\left[(c^{tech})^2\right] \operatorname{E}\left[Y^a\right] \\
&\quad + \operatorname{Var}\left[c^{tech}\right] (\lambda^a)^2 \operatorname{E}\left[(Y^a)^2\right] + \operatorname{E}\left[c^{tech}\right]^2 (\lambda^a)^2 \operatorname{Var}\left[Y^a\right] \\
&= \lambda^a \operatorname{E}\left[c^{tech}\right] \operatorname{E}\left[Y^a\right] - (\lambda^a)^2 \operatorname{E}\left[(c^{tech})^2\right] \operatorname{E}\left[Y^a\right] \\
&\quad + (\operatorname{Var}\left[c^{tech}\right] + \operatorname{E}\left[c^{tech}\right]^2)(\lambda^a)^2 \operatorname{Var}\left[Y_i^a\right] + \operatorname{Var}\left[c^{tech}\right] (\lambda^a \operatorname{E}\left[Y^a\right])^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\operatorname{Cov}\left(X^a, X^b\right)}{\operatorname{E}\left[X^a\right] \operatorname{E}\left[X^b\right]} &= \frac{\operatorname{Var}\left[c^{tech}\right]}{\operatorname{E}\left[c^{tech}\right]^2} + \left(1 + \frac{\operatorname{Var}\left[c^{tech}\right]}{\operatorname{E}\left[c^{tech}\right]^2}\right) \frac{\operatorname{Cov}\left(Y^a, Y^b\right)}{\operatorname{E}\left[Y^a\right] \operatorname{E}\left[Y^b\right]} \\
\frac{\operatorname{Var}\left[X^a\right]}{\operatorname{E}\left[X^a\right]^2} &= \frac{1}{\operatorname{E}\left[X^a\right]} + \frac{\operatorname{Var}\left[c^{tech}\right]}{\operatorname{E}\left[c^{tech}\right]^2} + \left(1 + \frac{\operatorname{Var}\left[c^{tech}\right]}{\operatorname{E}\left[c^{tech}\right]^2}\right) \frac{\operatorname{Var}\left[Y^a\right] - \operatorname{E}\left[Y^a\right]}{\operatorname{E}\left[Y^a\right]^2}.
\end{aligned} \tag{3.11}$$

Plugging in Equation 3.7, we arrive at the expression that connects intrinsic and observed noise under our extrinsic noise model,

$$\frac{\mathrm{Cov}\left(X^a, X^b\right)}{\mathrm{E}\left[X^a\right]\mathrm{E}\left[X^b\right]} = \frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} + \left(1 + \frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2}\right)\frac{\mathrm{E}\left[\mathrm{Cov}\left(Y^a, Y^b | c^{bio}\right)\right]}{\mathrm{E}\left[Y^a\right]\mathrm{E}\left[Y^b\right]}$$
$$\frac{\mathrm{Var}\left[X^a\right] - \mathrm{E}\left[X^a\right]}{\mathrm{E}\left[X^a\right]^2} = \frac{\mathrm{Var}\left[c\right]}{\mathrm{E}\left[c\right]^2} + \left(1 + \frac{\mathrm{Var}\left[c^{tech}\right]}{\mathrm{E}\left[c^{tech}\right]^2}\right)\frac{\mathrm{E}\left[\mathrm{Var}\left[Y^a | c^{bio}\right]\right] - \mathrm{E}\left[Y^a\right]}{\mathrm{E}\left[Y^a\right]^2}, \tag{3.12}$$

where $c = c^{tech}c^{bio}$ is the overall cell factor. This expression is rather general and follows directly from Equation 3.1 and Equation 3.2. The shared constant factors ($s := \frac{\mathrm{Var}[c]}{\mathrm{E}[c]^2}$) denote the extrinsic noise, and potentially explain the constant offset observed in the plot of the Fano factor against mean for genes. We denote $\frac{\mathrm{Cov}\left(Y^a, Y^b | c^{bio}\right)}{\mathrm{E}[Y^a]\,\mathrm{E}\left[Y^b\right]}$ and $\frac{\mathrm{Var}\left[Y^a | c^{bio}\right] - \mathrm{E}[Y^a]}{\mathrm{E}[Y^a]^2}$ by intrinsic covariance and variance.

Procedure for estimating extrinsic noise and selecting Poisson genes

To estimate the extrinsic noise, we calculated the normalized covariance among genes with mean expression greater than 0.1, and used the mode of the resulting distribution. This was computed using histogram bins of width 0.01. The center of the bin with the highest frequency was taken as the estimated value, which was set to exactly two decimal digits by construction of the bin edges.

To select Poisson genes, we calculated the 95% bootstrap interval of overdispersion for each gene based on 1,000 bootstrap samples by default. Then we selected genes whose 95% bootstrap intervals contain the estimated extrinsic noise.

Maximum likelihood estimation of cell size

Let $X_i^j \sim \mathrm{Poisson}(c_i\mu_j)$ be the observed expression count of gene j in cell i, where c_i is the cell size (scaling factor) for cell i and μ_j is the mean of gene j.

The likelihood function for cell i given its gene expression vector $\{X_i^j\}_{j=1}^G$ is given by

$$L(c_i) = \prod_{j=1}^{G} \frac{(c_i\mu_j)^{X_i^j} e^{-c_i\mu_j}}{X_i^j!}.$$

The log-likelihood is given by

$$\log L(c_i) = \sum_{j=1}^{G} \left(X_i^j \log(c_i\mu_j) - c_i\mu_j - \log(X_i^j!)\right).$$

To find the maximum likelihood estimator (MLE) of c_i, we differentiated the log-likelihood with respect to c_i and set the derivative to zero:

$$\frac{d}{dc_i} \log L(c_i) = \sum_{j=1}^{G} \left(\frac{X_i^j}{c_i} - \mu_j \right) = 0.$$

Solving for c_i gives the MLE:

$$\hat{c}_i = \frac{\sum_{j=1}^{G} X_i^j}{\sum_{j=1}^{G} \mu_j}.$$

The expected biological extrinsic noise of the CEL-seq2 data

The RNA mixture was prepared on a 384-well plate (Supplementary Figure 1a in (Tian et al., 2019)). After processing the SRA files from GEO Series GSE117617 using kb-python (Sullivan, Min, et al., 2025) and filtering out two outlier cells, we obtained a final dataset consisting of 357 cells. The exact biological extrinsic noise is influenced by the RNA amounts of the remaining 357 cells, which remain unknown. However, assuming that the 27 removed cells each had an RNA amount of 3.75 μg, we can estimate a lower bound for the biological extrinsic noise, which is approximately 0.333.

Identifiability of parameter-specific extrinsic noise in bursty models

Here we consider the bursty model, and assume a random variable for the extrinsic noise associated with each parameter. Studying the identifiability of parameter-specific extrinsic noise is crucial, as it can help us understand how biological extrinsic noise influences gene expression. By investigating this aspect, we aim to gain insights into the mechanisms that drive variability in gene expression at the single-cell level.

We consider the following bursty model of nascent and mature mRNA:

$$\emptyset \xrightarrow{k} B \times N, \quad N \xrightarrow{\beta} M, \quad M \xrightarrow{\gamma} \emptyset, \tag{3.13}$$

where in the first reaction the number of nascent mRNA molecules synthesized in each burst (B) follows a geometric distribution on $\{0, 1, 2, ...\}$ with a mean of b, referred to as the burst size (Singh and Bokes, 2012). The distribution of nascent is well known to be negative binomial, but the joint distribution of nascent and mature counts is not analytically available. Here, following the framework in (Gorin, Vastola, and Lior Pachter, 2023), we use the generating function method

to investigate the identifiablity of extrinsic noise, which can be extended to general gene expression models.

Denote the *in vivo* nascent and mature mRNA counts by y_u and y_m. Let $G_Y(z_n, z_m, t) = \sum_{y_n}\sum_{y_m} z_n^{y_u} z_m^{y_m} P(y_n, y_m, t)$ be the generating function. Assuming $\beta \neq \gamma$, the factorial generating function $\phi_Y(u_n, u_m, \infty) := \log G_Y(u_n + 1, u_m + 1, \infty)$ is (Singh and Bokes, 2012):

$$\tilde{u_n}(s) = \frac{u_m\beta}{\beta-\gamma}e^{-\gamma s} + (u_u - \frac{u_m\beta}{\beta-\gamma})e^{-\beta s}, \tag{3.14}$$

$$\phi_Y(u_n, u_m, \infty) = k\int_0^\infty \frac{b\tilde{u_n}(s)}{1-b\tilde{u_n}(s)}ds. \tag{3.15}$$

Adding Bernoulli sampling with rate $c^{tech}\lambda$, where λ is the species-specific constant, the generating function of observed counts x is

$$\begin{aligned} G_X(\mathbf{z}, t) &= \sum_{\mathbf{x}=0}^{\infty} \mathbf{z}^{\mathbf{x}} P(\mathbf{x}, t) \\ &= \sum_{\mathbf{x}=0}^{\infty}\sum_{\mathbf{y}=0}^{\infty} \mathbf{z}^{\mathbf{x}} P(\mathbf{x}|\mathbf{y}) P(\mathbf{y}, t) \\ &= \sum_{\mathbf{x}=0}^{\infty}\sum_{\mathbf{y}=0}^{\infty} \left(c^{tech}\lambda\mathbf{z} + 1 - c^{tech}\lambda\right)^{\mathbf{y}} P(\mathbf{y}, t) \\ &= G_Y\left(c^{tech}\lambda\,(\mathbf{z}-1) + 1, t\right). \end{aligned}$$

Then, adding parameter-specific extrinsic noise to each parameter, the factorial generating function of observed counts X $\phi_X(u_n, u_m, \infty) := \log G_X(u_n{+}1, u_m{+}1, \infty)$ is

$$\tilde{v_n}(s) = \frac{\lambda_m}{\lambda_n}\frac{v_m c^\beta\beta}{c^\beta\beta - c^\gamma\gamma}e^{-c^\gamma\gamma s} + (v_u - \frac{\lambda_m}{\lambda_n}\frac{v_m c^\beta\beta}{c^\beta\beta - c^\gamma\gamma})e^{-c^\beta\beta s}, \tag{3.16}$$

$$\phi_X(v_n, v_m, \infty) = c^k k\int_0^\infty \frac{c^b b c^{tech}\lambda_n\tilde{v_n}(s)}{1 - c^b b c^{tech}\lambda_n\tilde{v_n}(s)}ds. \tag{3.17}$$

Note that all c values are shared across genes within the same cell. Given the identifiable parameters of Equation 3.14 for *in vivo* counts Y are b, $\frac{\beta}{k}$, $\frac{\gamma}{k}$ (Singh and Bokes, 2012), the identifiable parameters in Equation 3.16 from observed counts X

include $b\lambda_n$, $\frac{\beta}{k}$, $\frac{\gamma}{k}$, and $\frac{\lambda_m}{\lambda_n}$, as well as the relative values of $\frac{c^\beta}{c^k}$, $\frac{c^\gamma}{c^k}$, and $c^b c^{\text{tech}}$, under the assumption that

$$\mathbb{E}\left[\frac{c^\beta}{c^k}\right] = \mathbb{E}\left[\frac{c^\gamma}{c^k}\right] = \mathbb{E}\left[c^b c^{\text{tech}}\right] = 1.$$

With the use of external RNA controls, it becomes possible to further disentangle c^b from c^{tech}.

Data and code availability

All datasets used in this study are publicly available. Raw FASTQ files were downloaded for each dataset and processed using kb-python version 0.29.1 (Bray et al., 2016; Melsted et al., 2021; Sullivan, Min, et al., 2025), with the nac workflow (Sullivan, Hjörleifsson, et al., 2025). The links to FASTQ files are in Supplementary Table 3.1.

All code used to generate the results and figures in the paper is available at `https://github.com/pachterlab/FP_2025`.

Dataset	FASTQs	Reference
Homogeneous RNA solution (Indrops v1)	GSM1599501	Klein et al., 2015
Heterogeneous RNA solution (CEL-seq2)	GSM3305230	Tian et al., 2019
Species-mixing, 20k Chromium X (10x 3' v3)	https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-exp/6.1.0/20k_hgmm_3p_HT_nextgem_Chromium_X/20k_hgmm_3p_HT_nextgem_Chromium_X_fastqs.tar	10x Genomics, 2021c
Species-mixing, 10k Chromium X (10x 3' v3)	https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-exp/6.1.0/10k_hgmm_3p_nextgem_Chromium_X/10k_hgmm_3p_nextgem_Chromium_X_fastqs.tar	10x Genomics, 2021b
Species-mixing, 10k Chromium controller (10x 3' v3)	https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-exp/6.1.0/10k_hgmm_3p_nextgem_Chromium_Controller/10k_hgmm_3p_nextgem_Chromium_Controller_fastqs.tar	10x Genomics, 2021a
K562 (10x flex)	https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-exp/7.0.0/10k_K562_singleplex_Multiplex/10k_K562_singleplex_Multiplex_fastqs.tar	10x Genomics, 2022
mESC (10x 3' v3)	GSM5111566	Riba et al., 2022
mESC (Indrops v1)	GSM1599494	Klein et al., 2015
PBMC (10x flex)	https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-exp/8.0.0/10k_Human_PBMC_TotalSeqB_singleplex_Multiplex/10k_Human_PBMC_TotalSeqB_singleplex_Multiplex_fastqs.tar	10x Genomics, 2024
Mouse forebrain (10x flex)	https://cf.10xgenomics.com/samples/cell-exp/7.1.0/10k_mouse_forebrain_scFFPE_singleplex_Multiplex/10k_mouse_forebrain_scFFPE_singleplex_Multiplex_fastqs.tar	10x Genomics, 2023
K562 (STORM-seq)	GSE181544	Johnson et al., 2022

Table 3.1: **Datasets metadata.** Datasets used for all analyses, with their technology in parentheses, FASTQ files accession links, and references.

Chapter 4

A PROCESS TIME MODEL FOR TRAJECTORY INFERENCE AND RNA VELOCITY

Fang, Meichen, Gennady Gorin, and Lior Pachter (2025). "Trajectory inference from single-cell genomics data with a process time model". In: *PLoS Comput. Biol.* 21.1, e1012752. DOI: 10.1371/journal.pcbi.1012752.

4.1 Introduction

Single-cell RNA sequencing (scRNA-seq) has provided unprecedented insights into biological dynamical processes in which cells display a continuous spectrum of states that go beyond the confines of discrete cell types (Griffiths, Scialdone, and Marioni, 2018). Cells appear to be inherently desynchronized in cellular processes and scRNA-seq can potentially capture cells at different positions over the process even if samples are collected at only one time point. The concept of pseudotime has been developed to describe the position of a cell along the underlying process (Trapnell et al., 2014), and trajectory inference (or pseudotemporal ordering) methods aim to solve the inverse problem of inferring the latent pseudotime variable from scRNA-seq data. In light of this this concept, hundreds of methods have been developed (Cannoodt, Saelens, and Saeys, 2016; Saelens et al., 2019; Deconinck et al., 2021; Cao et al., 2019; X. Qiu et al., 2017; Wolf, Hamey, et al., 2019; Street et al., 2018; Campbell and Yau, 2016; C. Lin and Bar-Joseph, 2019; Campbell and Yau, 2019; Du et al., 2024). However, with a few exceptions that explicitly model gene expression dynamics (C. Lin and Bar-Joseph, 2019; Campbell and Yau, 2019; Du et al., 2024), trajectory inference methods mostly treat pseudotime as a descriptive concept relying on more or less arbitrary distance metrics in gene expression space. Specifically, there is no well-defined, agreed-upon meaning underlying the notion of pseudotime, and its interpretation is primarily accomplished through qualitative visuals and low dimensional embeddings.

While a descriptive approach can be powerful in exploratory data analysis, the absence of a well-posed definition for a trajectory renders model interpretation and assessment challenging, even conceptually. Firstly, assessing the credibility of results is hard, as fitting can be performed on any dataset and we have limited metrics and ground truth available to gauge the fit quality. Secondly, the interpretation of

the inferred trajectory is not straightforward, and downstream analysis based on pseudotime is employed to understand the underlying gene dynamics. However, this need for following analysis to interpret results gives rise to the problem of circularity (Section 2), which becomes evident in the context of an inflated false positive rate in the problem of detecting differentially expressed (DE) genes along pseudotime. The problem of circularity is conceptually challenging and can only be effectively remedied under restrictive assumptions (Neufeld et al., 2023). To illustrate these two points, we applied the procedure of trajectory inference and DE analysis on simulations generated from four clusters and were able to "discover" superficially plausible dynamics (Figure 4.1). As naive an example as it is, it reflects the fact that we do not have a reliable way to determine the validity of trajectory inference results. Though both are false positives, there is a subtle difference between the falsely inferred trajectory (Figure 4.1b) and the inflated false positive rate in DE analysis resulting from circularity (Figure 4.1c): the first one arises when a trajectory model is inferred from cluster data without proper assessment, while circularity stems from the double use of data for fitting and testing (double dipping) (Kriegeskorte et al., 2009). Conversely, adopting a model-based approach has the potential to mitigate this problem. With a clearly defined model of gene expression along a trajectory, the interpretation of parameters and the characterization of errors becomes more straightforward. First of all, model assessment can be conducted in a more principled manner. We can effectively address the first kind of false positive using conceptually simpler approaches, such as comparing our model to cluster models to identify the correct model. In addition, the specific question of interest like finding DE genes can be incorporated directly into the formulation of the model, rendering ad hoc analysis unnecessary. For example, if we have an probabilistic model of trajectories with transcription kinetics parameters, we can directly select DE genes using inferred parameters, without the need to go through the circular process of fitting trajectories and performing DE analysis to find interesting genes. However, we emphasize that the exact p-values still cannot be easily calculated, and circularity persists if we fit trajectories and perform DE test based on the inferred time, which still falls under the issue of double dipping.

Recently, equipped with a kinetic model of RNA dynamics, RNA velocity has emerged as another powerful concept to provide complementary information about dynamic processes (La Manno et al., 2018). By distinguishing unspliced and spliced mRNA counts as derived from unique molecular identifiers (UMIs) and fitting gene-wise parameters under an on-off model of transcription, it is able to

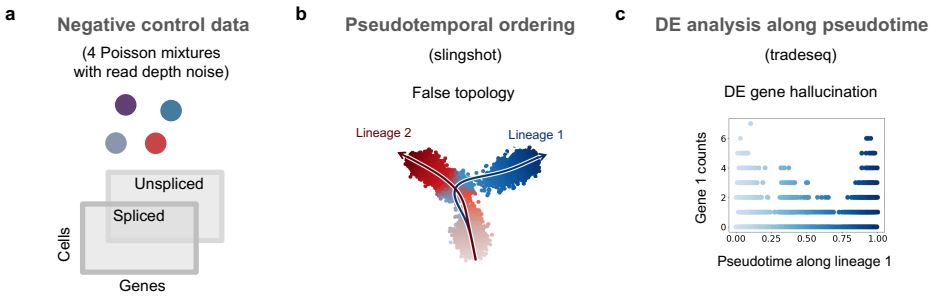

Figure 4.1: **False positive on clusters data**. **a)** Negative control data are simulated from 4 Poisson mixtures with read depth noise. **b)** As an example of false positive, specious trajectory in lower dimensional space was constructed with Slingshot (Street et al., 2018). **c)** Differential genes along pseudotime were selected with tradeSeq (Van den Berge et al., 2020), with the first gene plotted along the blue lineage.

predict the direction of future spliced counts changes. Although a time-dependent gene expression model was explicitly defined in RNA velocity, the time did not have any associated interpretation. Moreover, earlier methods often modeled genes separately with gene-wise times and fit these models after applying a series of ad hoc transformations to count data, which added excessive flexibility and hindered a clear interpretation of the time. As the velocities of different genes had non-comparable scales, they needed to be combined heuristically in a lower-dimensional space to calculate a velocity for a cell (Gorin, Fang, et al., 2022). A natural extension is to integrate the cell-wise time of trajectory inference with the mRNA dynamical model of RNA velocity, which a few methods have successfully implemented with different underlying transcription models (Aivazidis et al., 2023; Gu, Blaauw, and Welch, 2022; Li et al., 2022). Moreover, the recent *VeloCycle* developed an RNA velocity model for the cell cycle that models unspliced and spliced counts dynamics directly with harmonic functions (Lederer et al., 2024).

However, despite the implicit pseudotime modeling performed by some of the RNA velocity methods, there remain many challenges in attaching a physical meaning to pseudotime. Do the parameters of the trajectory model have underlying biophysical interpretations? How can we guarantee that our inferences align with our intended objectives? Are the assumptions of trajectory models satisfied to maintain the consistency of our inference? For instance, the application of trajectory inference or RNA velocity methods relies on the assumption of continuous dynamics in the data, which is not examined retrospectively. Though some heuristic scores

purport to distinguish between cluster-like data and trajectory-like data (Lim and P. Qiu, 2024), there is no principled approach to determine whether the data is sufficiently dynamical and whether a cluster or trajectory model is more appropriate, and the decision of applying trajectory analysis often hinges on prior knowledge and assumptions about the data.

In summary, to attach real meaning to "pseudotime" requires more than just a definition of cell-wise pseudotime. It necessitates a principled approach to statistics to ensure a meaningful inference, which is still lacking in the field of trajectory inference. Meticulous model assessment is required to ensure its relevance to the underlying biological processes and the reliability of results, which includes examining the identifiability of the model, characterizing performance to identify both ideal and failure scenarios, and establishing proper metrics for result falsification. Then pseudotime starts to have a physical meaning, which we suggest defining as "process time" to underscore its interpretation with respect to a specific cellular process.

The physical interpretation of process time is related to, but not necessarily equivalent to, physical time. Specifically, assuming that all cells share the same dynamic process, we can select a specific point along this process to serve as the starting point for all cells. At the physical sampling time, the process time denotes the relative time to that starting point, indicating how long ago in physical time the cells were at the starting point. Therefore, if the experiments establish a known starting time for when the cells enter the process, the process time should correspond to the relative physical time. On the other hand, if we could follow one cell over time, the process time would evolve in sync with physical time. In reality, where only different cells can be sampled at multiple time points, the distribution of process time should ideally evolve in parallel with physical time, provided enough cells are sampled from the same population.

Here, we build such a model and infer "process time" in a principled way with Chronocell. To strike a balance between expressiveness and identifiability, we proposed a trajectory model built on cell states (Gorin, Fang, et al., 2022). On the one hand, we incorporated different cell states so that our trajectory model is expressive enough to capture the observation that cells are generally assumed to transition through various states during development. On the other hand, we assume a constant transcription rate for each state to keep the model as simple as possible. By introducing simplifying assumptions in transcription and sequencing model, we ensure

model identifiability. We consider the influence of technical aspects and directly incorporate them into the distribution of counts, which eliminates the necessity for unjustified heuristic preprocessing steps that lead to unclear interpretation and biased results even in the large data and no noise limit (Gorin, Fang, et al., 2022). We undertook simulations to characterize estimation accuracy in the different parameter regimes to identify ideal and failure scenarios. Using simulations for which ground truth is known allowed us to characterize how large uncertainty and inconsistent parameter values serve as good indicators of potential unreliability in failure scenarios, which can be assessed even when ground truth is unavailable. Finally, we applied Chronocell to biological datasets. We assessed its appropriateness on different datasets and identified unsuitable ones. For suitable datasets, Chronocell revealed distinct cellular distributions over process time and yielded mRNA degradation rate estimates congruous with those obtained from mRNA metabolic labeling.

4.2 Challenges with the pseudotime concept

Trajectory methods overview

Single-cell genomics trajectory inference methods have mostly relied on similarity metrics: distance based methods reconstruct the trajectory based on some distance metrics in gene expression space under the assumption that cells that are more similar in gene expression space are also closer in pseudotime (Haghverdi et al., 2016; Wolf, Hamey, et al., 2019; Trapnell et al., 2014). Manifold-learning based methods draw the trajectories in a reduced dimension space based on connectivity, i.e., similarity (Campbell and Yau, 2016; Street et al., 2018). Probability/Markov chain based methods also calculate transition probabilities based on distances (Setty et al., 2019). However, pseudotime based on similarity/distance is inherently descriptive and unable to be extended to reflect physical meaning, because state spaces of dynamical processes are not isotropic. There are a few exceptions that implicitly define generative models of single-cell RNA-seq (scRNA-seq) data with pseudotime, modeling dynamics of gene expression along differentiation processes in a way that can be reformulated as driven by cell states switching models (C. Lin and Bar-Joseph, 2019). These ideas have motivated our model.

One important observation is that the usage of Markov Chain model in trajectory inference, as well as other single cell analysis, can be fundamentally flawed. This is because frequently cells are samples drawn from different chains, instead of a sequence of observations of one single chain. The Markov chain formalism is therefore not applicable to single cell studies without making further assumptions.

Circularity in pseudotime-based analysis

Due to the exploratory nature of trajectory inference, all variables are used to fit the model, but not all variables are informative. Specifically, it is natural to assume sparsity and focus on a few "marker genes" in downstream analysis. At present, a multi-stage method is commonly employed for pseudotime-based analysis. In the initial stage, trajectory and pseudotime are fitted, followed by the second stage, where hypothesis testing is utilized to select genes that are variable along trajectories.

Ideally, with a predefined and well-parameterized model, we can construct confidence intervals for parameters using Bayesian methods or the bootstrap. However, interpretation can be difficult even for PCA loadings (Cadima and Jolliffe, 1995). As heuristic methods become more popular, the variable selection problem is highly entangled with model construction, and the question of how to perform valid inference is not straightforward. Current methods usually first perform trajectory inference, and then test whether genes expression have dependency with pseudotime using the same dataset. It is well-known that such tests are not valid and can lead to inflated false positive rates (Campbell and Yau, 2016; Lähnemann et al., 2020; Z. Ji and H. Ji, 2016; Tritschler et al., 2019). The same issue for clustering has also been well discussed in several recent studies (Zhang, Kamath, and Tse, 2019; Gao, Bien, and D. Witten, 2020; Chen and D. M. Witten, 2022). Here we briefly summarize the issue in the context of trajectory inference.

First, fitting and testing using the same dataset means that one has cherry-picked the most significant association and results are consequently biased upward, which is known as post-selection inference (Taylor and Tibshirani, 2015; Kuchibhotla, Kolassa, and Kuffner, 2022). Furthermore, even if one uses separate datasets for fitting and testing, there is an inherent circularity in the hypothesis testing. Specifically, during trajectory inference, one selects a transformation (that defines a trajectory) $\hat{f}$ that maps a cell to a pseudotime based on its gene expression. Then, by testing whether genes expression associates with pseudotime, one is asking whether x associates with $\hat{f}(x)$, which just echoes the model fitted and does not perform the hypothesis testing validly.

PCA as an example

To see this circularity, consider a single component model where trajectory is replace by the first component of PCA. Denote the data by X, and let Y be the normalized data matrix, i.e. $Y_{ij} = X_{ij} - \bar{X}_j$ for covariance PCA and $Y_{ij} = \frac{X_{ij}-\bar{X}_j}{\sqrt{\sum_i (X_{ij}-\bar{X}_j)^2}}$ for

correlation PCA. Write $S = \frac{1}{n}Y^TY = V\Lambda V^T$ and denote the first eigenvector in PCA by v (first column of V) and first eigenvalue by λ. Then for first principal component scores, which are the latent variable z we want, we have $z = Yv$ and $\frac{1}{n}y_j^Tz = \frac{1}{n}y_j^TYv = \frac{1}{n}(Y^TYv)_j = \lambda v_j$. If we directly perform linear regression of z on the expression of mean-centered gene y_j, the slope is v_j.

Even after data splitting, if we follow the normal linear regression procedure and test the null hypothesis that $\beta = 0$, we will derive a t-statistic that is still biased. The correct way is to account for the projection $z = Yv$ and test $\beta = v_j$.

Current solutions

In practice, there are only a few papers that have taken this circularity into account. One possible solution is count splitting if counts number are high enough (**Neufeld2023-lh**). Another possible solution is data splitting, where we split the dataset into two parts, select our model using the first part and do the inference using the second. Specifically, to perform rigorous hypothesis testing and get some valid p-value, we can perform permutation test. However, it means that we need to generate sets of permuted data and perform the whole procedure of trajectory inference and DE analysis on each set. This approach does not seem to have been explored or adopted in any currently used tools.

More crucially, this kind of pseudotime-based analysis only answers data analytic questions, i.e., data summary and analysis. They are typically not concerned with goodness of fit/model selection, thus providing no information about the correctness of the fitted model.

4.3 Results

A trajectory model generalizing cellular states

We begin by defining the trajectory as a dynamical process underlying all cells, with potentially different lineages/branches within this process. Cells are then assumed to be sampled from various, unobserved, time points along this process. Thus, the latent variable $z = (t, l)$ is introduced to account for such heterogeneity due to process time (t) and lineage (l), and z follows a sampling distribution determined by the specific biological system and experimental conditions. Consequently, the probability distribution of the data we obtain is a mixture of cells over time and lineage,

$$P(\mathbf{x}|\theta) = \int_{\mathbf{z}} P(\mathbf{x}|\mathbf{z}, \theta)P(\mathbf{z})d\mathbf{z}, \tag{4.1}$$

where $\mathbf{x}$ is the data and θ is the set of parameters that define the trajectory. This is the common framework of trajectory inference, and developing a trajectory model requires defining the gene dynamics along the lineage and the process time $P(\mathbf{x}|\mathbf{z}, \theta)$, as well as the sampling distribution of cells over the process $P(\mathbf{z})$.

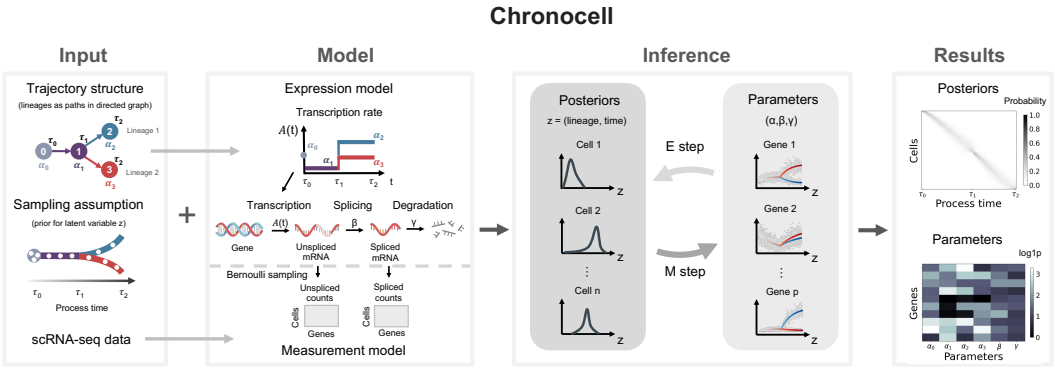

Figure 4.2: **Chronocell overview**. The **input** of Chronocell comprises three components: 1) the *trajectory structure*, which outlines the states each lineage traverses as paths on a directed graph; 2) the *sampling assumption*, which defines the prior distribution of latent variables, namely lineages and process time, with a default uniform distribution over both; and 3) the *scRNA-seq data*, consisting of unspliced and spliced count matrices. The Chronocell **model** consists of a *expression model* with piecewise-constant transcription rates, and a Bernoulli *measurement model*. Each state s is associated with a transcription rate α_s for each gene, as well as an exit time τ_k denoting the switching time to the next state, where k is the index for the time segment. The EM algorithm is used for **inference**, with each iteration alternating between E-steps and M-steps. The **results** of Chronocell primarily include the estimated parameters and posterior distributions over latent variables for each cell.

To define the dynamical process, we first state our transcription model. We consider only transcription, splicing, and degradation reactions in cells, and assume only transcription rates are time-dependent (Figure 4.2 Gene expression):

$$\emptyset \xrightarrow{A_l(t)} U,\ U \xrightarrow{\beta} S,\ S \xrightarrow{\gamma} \emptyset, \tag{4.2}$$

where $A_l(t)$ is the transcription rate function for lineage l at time t, and β and γ are the splicing, and degradation rates. The chemical master equation describing the evolving distribution of the above biochemical reaction network has an analytical solution (Jahnke and Huisinga, 2007). If we assume the initial distribution to be Poisson, the solution remains Poisson with the means (λ_u and λ_s) of U and S evolving according to the following ordinary differential equations (ODEs) of RNA velocity:

$$\begin{aligned} \frac{d\lambda_u(t)}{dt} &= A_l(t) - \beta\lambda_u(t), \\ \frac{d\lambda_s(t)}{dt} &= \beta\lambda_u(t) - \gamma\lambda_s(t). \end{aligned} \tag{4.3}$$

The modeling of transcription rate $A_l(t)$ is motivated by the common abstraction of cellular differentiation as cell state transitions: each lineage is abstracted as a series of switches in cellular states over time. The series of switches are specified by the given trajectory structure, which includes a directed graph of cell states where each lineage corresponds to a path (Figure 4.2 Input Structure). We introduce one transcription rate per gene for each cellular state (α). Switching is assumed to be instantaneous and occurs at an unknown but fixed time (τ) with the first switch leaving initial state 0 to occur at τ_0. Without loss of generality, we consider the entire process to start at time 0 (τ_0=0) and have a time length of 1 (e.g. τ_2=1 in Figure 4.2). Consequently, the transcription rate function $A_l(t)$ of lineage l is simplified as piecewise constant functions of the process time over $[0, 1]$ (Figure 4.2 Model). This piecewise constant function is defined in the limiting regime where transcriptional state switching (such as expression of master regulatory factors and changes of chromatin state) precedes gene expression and has a much faster time scale. Thus, the piecewise constant function serves as a reasonable approximation when the time scale of transcription rate changes is comparable to or larger than the mRNA half-life. It also directly reduces to discrete cell clusters in the fast dynamic limit, i.e., when dynamical timescale ($\frac{1}{\beta}$ and $\frac{1}{\gamma}$) is much smaller than sampling intervals, for example, the total time length divided by cell number, n, under a uniform sampling distribution. This connection to cluster models enables us to interpolate between discrete cell states and continuous dynamics.

The simple form of the transcription rate function lead to a tractable model and facilitates inference and analysis. In fact, it affords explicit solutions for the distribution and for its derivatives with respect to parameters. Ideally, gene regulatory networks involved in cell differentiation would be modeled, but with current (transcriptomic) data types it is difficult to include gene interactions and to model transcription rates as (protein-mediated) functions of other genes with accuracy. Thus, we assume that the dynamics of different genes are independent and that all correlations are absorbed into the shared latent process time. Additionally, for simplicity, we can assume all genes are fully synchronized, as in the synchronized model, where τ is the same for all genes. However, the desynchronized model, which allows for

different τ values for each gene, is also available. In summary, our trajectory model is suitable for capturing the coordinated global gene expression changes instead of the detailed gene dynamics.

After deriving an explicit distribution of *in vivo* counts, we turn to the measurement model. We assume simple binomial sampling of each molecule, and that the average binomial sampling probability, i.e., read depth, varies between cells but remains the same for all molecules in one cell. Then, in vitro counts remain Poisson but with means adjusted by read depth. Instead of using normalized data, we estimated read depth using the total UMI counts of near-Poissonian genes that are not used for the inference, and incorporated it into the count distribution. Therefore, we arrive at an analytical form of the conditional probabilistic distributions $P(\mathbf{x}|\mathbf{z}, \theta)$ of counts from a dynamic process by specifying the trajectory structure, gene expression model, transcription rate functions, and scRNA-seq measurement model.

The remaining part of specifying the sampling distribution $P(\mathbf{z})$ is crucial, because it breaks the scale invariance of parameters and ensures the identifiability of the model: multiplying the transcription, splicing and degradation rates by the same constant leads to the same marginal likelihood if the sampling distribution can be changed. Ideally, the specification of the sampling distribution depends on our knowledge of the studied biological system and the experimental design. For example, for stem cells that constantly divide and differentiate, we may assume a uniform sampling distribution over $(0, 1]$ with a point mass on time 0, where the point mass represents the fraction of time that cells spend in the initial proliferative state. On the other hand, for time series data, we may assume the sampling distribution of cells were centered around their captured time points with some variances. However, in practice, since there is no obvious principled way to determine such distributions, we just assume a uniform sampling distribution over process times ($t > 0$) by default, but the weight on time point 0 and different lineages can be identifiable and be updated in the inference.

A trajectory is thus defined with the above dynamical process and sampling distributions. With the parameterized form of the probabilistic distribution of counts, we can employ the Expectation-Maximization (EM) algorithm to estimate model parameters and posterior distributions of process times and cell lineages by maximizing Evidence Lower Bound (ELBO) with either warm start or random initialization (Section Inference). The synchronized model was used by default, where all genes share the same τ. Since fitting the desynchronized model from scratch is more

challenging, it is recommended to begin the fitting process using the results from the synchronized model. For desynchronized models, we also introduce a penalty term proportional to the squared difference between gene-wise τ_k and the global τ_k to encourage synchronization, and the coefficient for this penalty term is set as a parameter, with a default value of 0. We tested the EM algorithm and inference on simulations generated from our trajectory model (Section Simulations). Both the synchronized and desynchronized models are identifiable, and the parameters can be recovered accurately under reasonable conditions (Figures 4.3 and 4.4). We will elaborate on these conditions in Section 3.4 "Identifying failure scenarios reveals the fragility of inference."

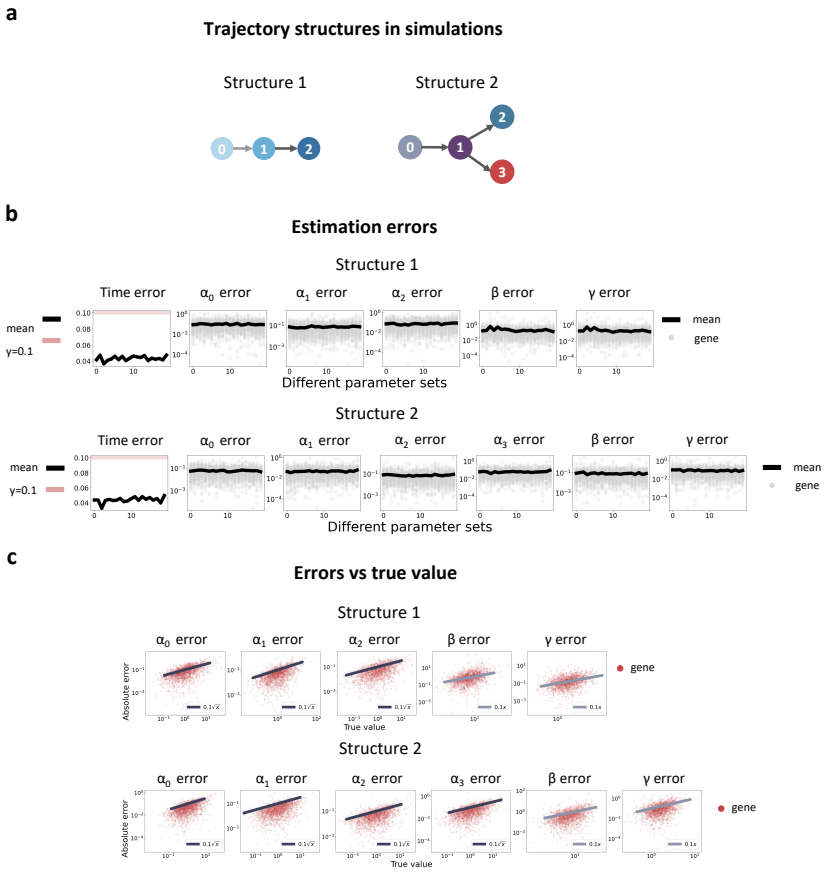

Figure 4.3: **Inference accuracy of synchronized model**. **a**) The two trajectory structures used in simulations. **b**) Estimation errors of different parameter sets. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in Section 4.4. **c**) Absolute errors with respect to the true values of parameters.

By having an explicitly parameterized distribution of raw counts, we can easily

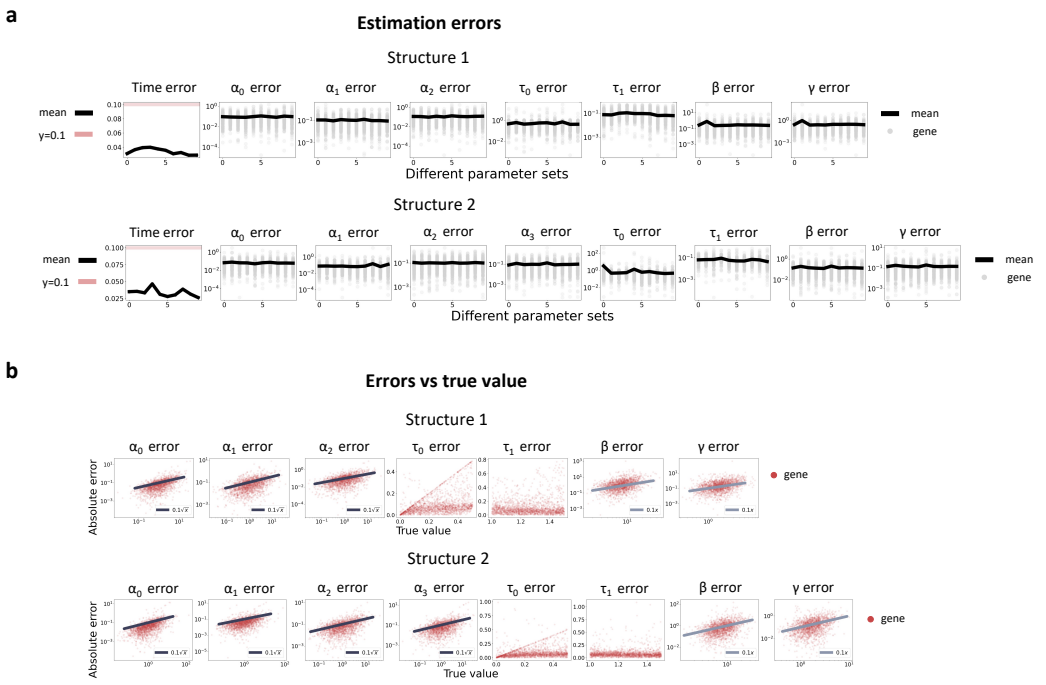

Figure 4.4: **Inference accuracy of desynchronized model**. Estimation errors of the desynchronized model tested on simulations with trajectory structures in Figure 4.3a. **a**) Errors of different parameter sets. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in Section 4.4. **b**) Absolute errors with respect to the true values of parameters.

interpret results and systematically assess the model under a more principled framework. Since parameters all have biophysical meanings, we cannot only directly interpret them but also validate their accuracy by comparing them to orthogonal experiments that measure the same parameters. Furthermore, we can directly select DE genes by fold changes in transcription rates across states, after filtering genes by goodness of fit ((see Section 4.5 "Gene selection"). The performance can be quantified through parameter errors, aiding in the identification of both confident and uncertain scenarios.

Not only are the parameters and results more interpretable, but we can also compare different models systematically. Regarding false positives from clustered data, we can evaluate when a trajectory model is no longer appropriate by comparing it to a cluster model using standard model selection methods like AIC (Figure 4.5). Importantly, the posterior distributions and multiple minima of AIC scores can hint at the lack of continuity (Figure 4.5), serving as a retrospective metric when model selection methods are compromised by unaccounted noise. We also demonstrate

model selection on a disconnected trajectory, which includes both a single cluster and a bifurcation (Figure 4.6). Our representation of the trajectory structure as paths on a directed graph naturally includes this scenario. We compare the results of the true model with those of a cluster model and a connected structure. AIC and BIC can correctly identify the true structure when compared to the two incorrect models (Figure 4.6).

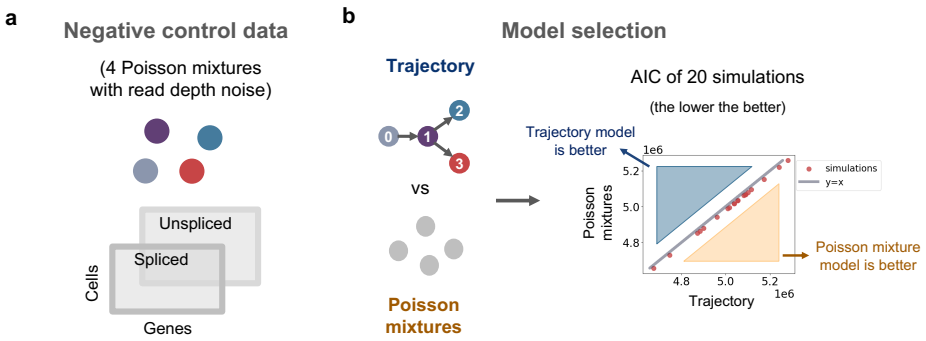

Figure 4.5: **Inference and model selection on clusters data**. **a)** The same data from 4 Poisson mixtures as in Figure 4.1. **b)** To compare the trajectory and Poisson mixture models, AIC scores of Poisson mixtures model and trajectory model are compared on 20 simulations with different random parameter sets. Dots below y=x indicate Poisson mixture model is better.

Demonstrating performance of Chronocell on ground-truth simulations

Since the true trajectory structure may differ from both our prior knowledge and the initial structure used, we first demonstrate the inference pipeline under a slightly incorrect trajectory. The inferred parameters provide insights into the true structure, and we then apply model selection to identify the correct model. Additionally, we include non-variable genes to assess the performance of the differential expression (DE) procedure.

We used a two-lineage trajectory and uniformly sampled 10,000 cells over lineages and process times with biological plausible parameters extracted from literature (Rabani et al., 2011) (Section Simulations). 100 out of 200 genes are variable (Figure 4.7a). We fit under a slightly wrong assumption of the trajectory structures and assumed all genes are variable (Figure 4.7b). For random initialization, we initialized randomly 100 times and picked the one with highest ELBO score (Figure 4.7c). We also fit with warm start by initiating the fitting process from correct cell clusters grouped by true time and lineages. Both types of initialization were able to converge to the ELBO with true parameters, and random initialization yielded a

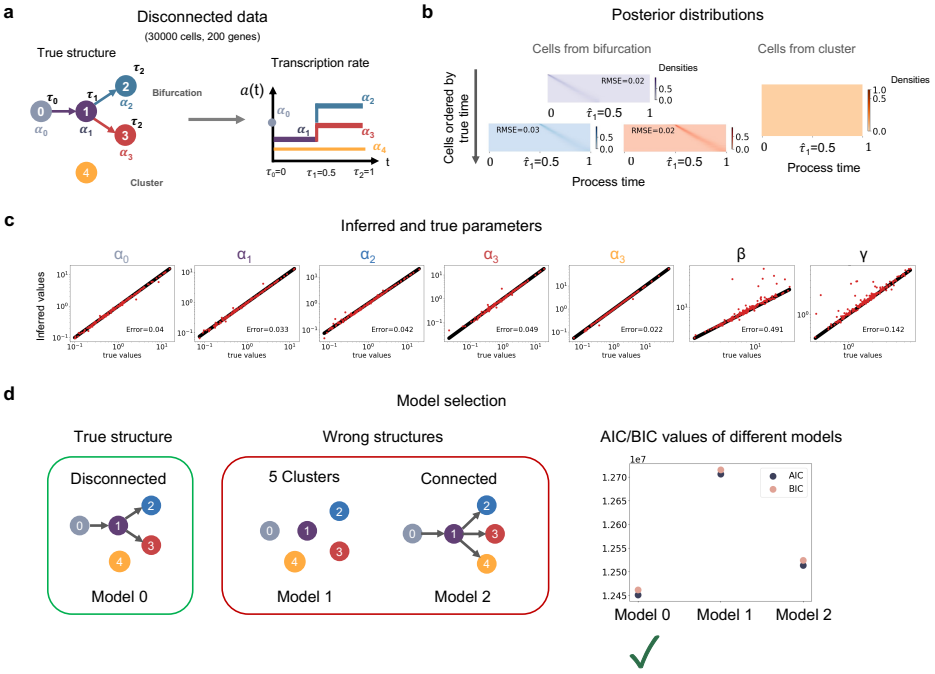

Figure 4.6: **Inference and model selection on disconnected data**. **a)** The ground truth trajectory structure. A subset of cells is from a bifurcation trajectory and the other cells are from a disjoint cluster. For the bifurcation trajectory, cells start from the state 0 and jump to the state 1 at $\tau_0 = 0$, and then bifurcate into two lineages with different ending states (2 and 3) at $\tau_1 = 0.5$. The process ends at $\tau_2 = 1$. **b)** Heatmaps of the inferred posterior distributions for cells from both the bifurcation and the cluster. x-axis is time grids, and y-axis is cells aligned by their true times and grouped by their true lineages. The intensity of color indicates the weights of posterior distributions of cells on the grids. Heatmap of cells from τ_0 to τ_1 use a purple color palette. Heatmap of cells from τ_1 to τ_2 of first lineage use blue, and those of second lineage use red. Heatmap of cells from cluster use an orange color palette. RMSE stands for root mean square errors. **c)** Inferred parameters values compared to true values. Error is mean normalized error across genes as described in Section 4.4. **d)** AIC and BIC of the true model and two wrong models.

slightly higher ELBO compared to warm start, with a negligible difference (Figure 4.7d). For the following analysis, we used the fitting results of random initialization. The posterior distributions correctly recapitulate the time and lineages of cells, with a root mean squared error (RMSE) around 0.05 for the posterior mean of process time in comparison to the true time (Figure 4.7e). This means that the error in time of a cell is around 5% of the total time length of the trajectory in average. However, the error is not completely uniform: as cells in the second interval are closer to the steady state, they have more spread posteriors and larger errors. Since the

posteriors of each cell are accurate, the posteriors averaged over cells also resemble the empirical distribution of the true time (Figure 4.7f).

The parameters are also recovered accurately (Figure 4.7g). For variable genes, all parameters are identifiable, while for non-variable genes, only α and the ratio of β to γ are identifiable. We noticed a trend that the absolute errors of α scale with the square root of true values, while for β and γ, they scale with the true values (Figure 4.8a). This trend also appeared in other simulations (Figure 4.3b and 4.3c). Therefore, to calculate the errors of parameters, we divide the absolute error of α by the square root of true values and β, γ by true values, so that errors of different genes are more comparable. We refer to them as normalized errors in the text. The parameter α tends to be estimated with higher accuracy compared to β and γ (Figure 4.7g). This aligns with intuition because, although parameters are identifiable in our trajectory model, the Evidence Lower Bound (ELBO) is nevertheless insensitive to proportional changes in β and γ, which have minimal impact on the phase portrait in the unspliced and spliced space of each gene. To confirm this, we used the true model to calculate the Fisher information matrix and the smallest eigenvalues of each gene (Figure 4.8b). The corresponding eigenvectors describe the flattest direction of the ELBO. To validate that the flattest direction primarily lies in the β and γ parameters, we add the corresponding eigenvectors to the true parameters after normalizing it by the square root of eigenvalues, and calculate the new ELBO with the modified parameters. Indeed the resultant changes in ELBO are indeed small and β and γ values were mainly varied (Figure 4.8c), which confirms that the β and γ are harder to estimate accurately (Figure 4.7g).

To select genes whose dynamics are well-fit by the model, we evaluate their goodness of fit by comparing the gene-wise likelihood with that of clustering models (Section Gene selection). We adopt this relative likelihood criterion because absolute likelihoods of different genes are not directly comparable and clusters model serves as a natural reference point for comparison to filter genes without continuous dynamics. All 91 genes selected belong to the variable class (Figure 4.7h). The similar transcription rates of states 1 and 2 of selected genes would also suggest us that those two states could be merged into a single one, if we didn't know the true structure (Figure 4.7i). Therefore, we applied model selection methods to compare the ground truth model and the assumed model. We generate another 20 parameter sets, fit under both models, and compute in-sample ELBO scores (ELBO), AIC, BIC, and out-of-sample ELBO scores (test ELBO). The ELBO is always better in

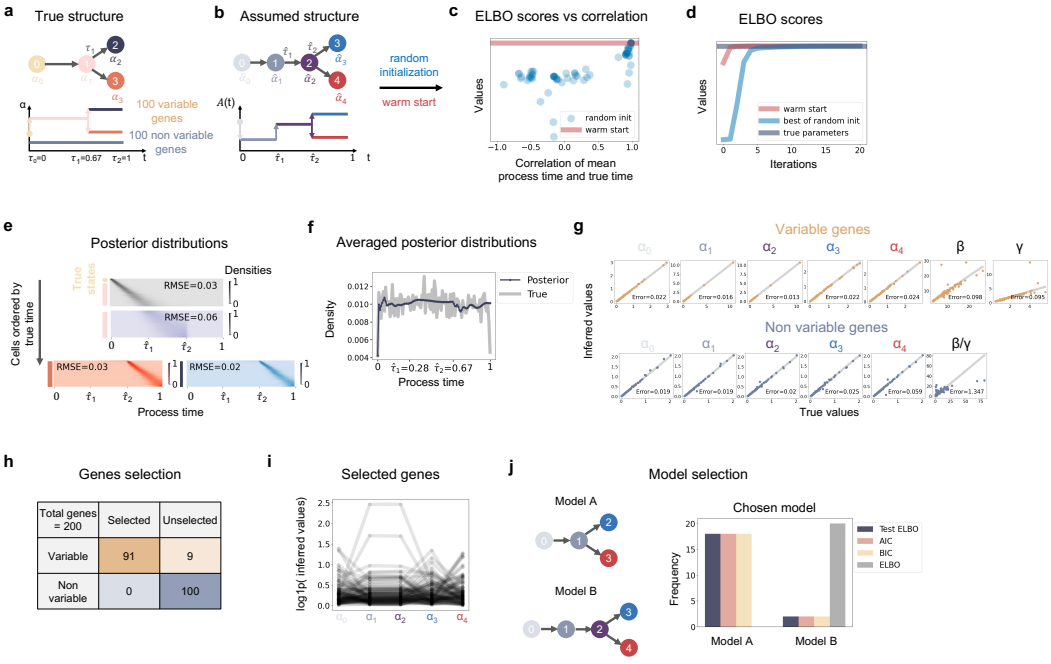

Figure 4.7: **Demonstration of inference on simulation**. **a)** The ground truth trajectory structure. Cells jump to the next state (2) from starting state (1) at $\tau_0 = 0$, and then bifurcate into two lineages with different ending states (3 and 4) at τ_1. The process ends at $\tau_2 = 1$. Out of 200 total genes, 100 genes are non variable with the same distributions along time. **b)** The falsely assumed structure that does not know the first two states are supposed to be merged into one. All genes are assumed to vary along time. **c)** The ELBO scores of 100 random initializations (blue dots) compared to those of warm start (red line). The x-axis is the Pearson's correlation between the mean process time of each random initialization and the true time. **d)** ELBO scores over fitting iterations of both warm start (red line) and the best random initialization (blue line), with the ELBO calculated with true parameters (gray line) as reference. **e)** Heatmaps of inferred posterior distributions. x-axis is time grids, and y-axis is cells aligned by their true times with true transcription states on the left. The intensity of color indicates the weights of posterior distributions of cells on the grids. Heatmap of cells from τ_0 to τ_1 use a gray color palette. Heatmap of cells from τ_1 to τ_2 use purple color palette. Heatmap of cells from τ_2 to τ_3 of first lineage use blue, and those of second lineage use red. RMSE stands for root mean square errors. **f)** The averaged posterior distributions across cells (dark blue) and true empirical distribution (gray) of process time. **g)** Inferred parameters values compared to true values. For non variable genes, only $\frac{\beta}{\gamma}$ are identifiable and compared. Error is mean normalized error as described in the text, and the mean is computed across genes. **h)** The confusion matrix for gene selection. **i)** α values of selected genes over states. **j)** Two models and the distribution of the chosen one by (train) ELBO, AIC, BIC, and test ELBO, calculated on 20 samples each with a different set of parameter.

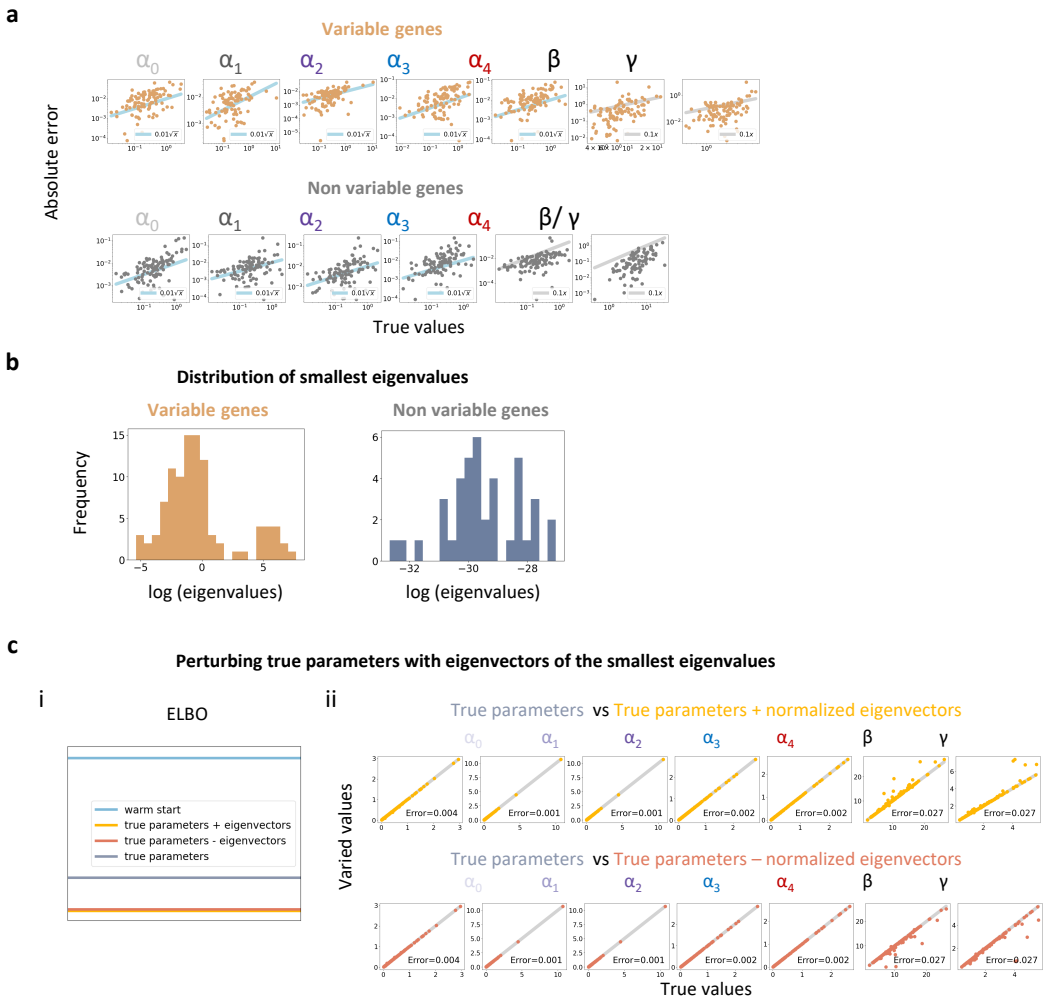

Figure 4.8: **Supplementary figures for demonstration of inference on simulation**. **a**) Absolute errors with respect to the true values of parameters. **b**) Distribution of the smallest eigenvalues of the Fisher information matrix of each gene. **c**) Marginal likelihood (ELBO) and varied parameters compared to true parameters. The difference between varied parameters and true parameters are the eigenvectors corresponding to the smallest eigenvalues of the Fisher information matrix, divided by the square root of the respective eigenvalues, specific to variable genes.

model B due to overfitting. All of the last three metrics (AIC, BIC, and test ELBO) favor the true model most of the time (90%) (Figure 4.7j). However, we want to emphasize that model selection worked because simulations were strictly generated under our trajectory model. In reality, if true transcription rates are far away from piece-wise constant functions, the resulting deterministic noise can introduce bias in AIC/BIC and cross-validation, leading them to prefer more complex models (Abu-Mostafa, Magdon-Ismail, and H.-T. Lin, 2012).

Identifying failure scenarios reveals the fragility of inference

Given the accuracy on perfect data, we sought to characterize the impact of different factors on inference accuracy and identify potential failure scenarios that could lead to unreliable results. We first study the requirements on some obvious factors like the number of cells, the number of genes, and the means of counts. Relatively small numbers of cells and genes appear to be sufficient (Figures 4.9 and 4.10). As the number of cells increases, parameter errors decrease, while time errors remain the same. On the other hand, time errors decrease with an increasing number of genes while parameter errors remain the same. Additionally, counts means must be sufficiently high to enable accurate inference, which necessitates adequate sequencing depth (Figure 4.11). We notice that when the trajectory structure becomes more complex, the requirement for count means also increases: Chronocell needs higher means for similar accuracy (Figure 4.12). Furthermore, when count means are low, increasing the number of cells can actually compromise the accuracy, underscoring the critical importance of obtaining sufficient counts.

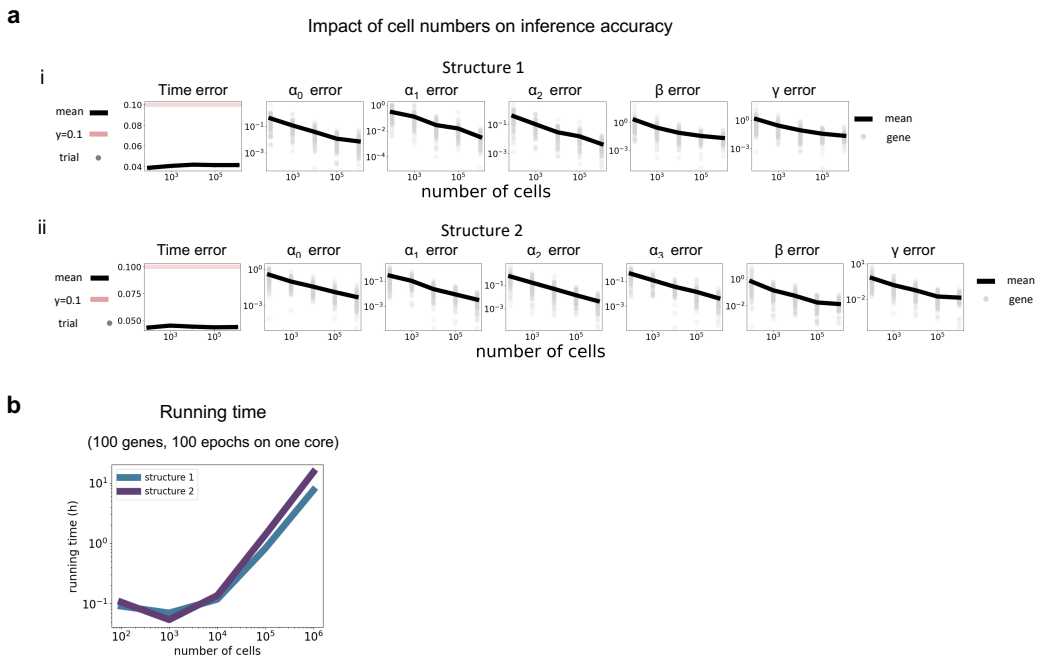

Figure 4.9: **Impact of cell numbers on inference accuracy and running time**. **a**) Estimation errors for datasets with varying cell numbers. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. Estimation errors of different cell numbers. **b**) Running time of 100 epochs on a single core on datasets with varying cell numbers.

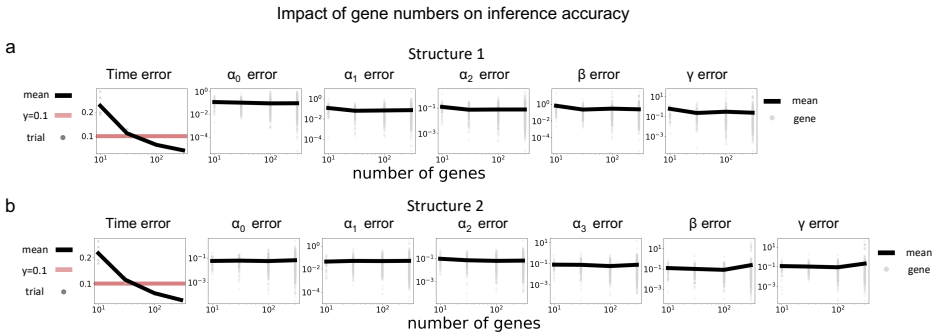

Figure 4.10: **Impact of gene numbers on inference accuracy**. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. **a**) Results for trajectory structure 1. **b**) Results for trajectory structure 2.

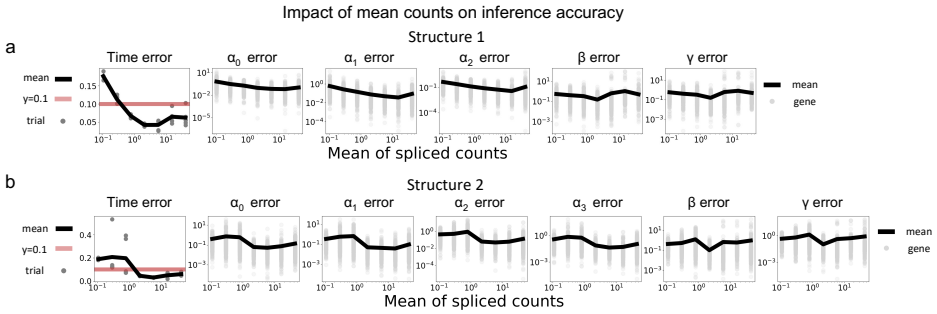

Figure 4.11: **Impact of mean counts on inference accuracy**. Simulations of different counts mean are generated by scaling the transcription rates while keeping other parameters the same. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. **a**) Results for trajectory structure 1. **b**) Results for trajectory structure 2.

As we use piecewise-constant functions to approximate transcription rates, we also test how this simplification impacts results when transcription rates are, in fact, not piecewise-constant. We generate simulations using piecewise-exponential functions for transcription rates, which ranges from almost linear to almost piecewise as the rate constants increase. We fit the model with Chronocell under the piecewise-constant assumption, and the inference accuracy remains satisfactory when the rate constants are comparable to or larger than the mRNA half-life (Figure 4.13).

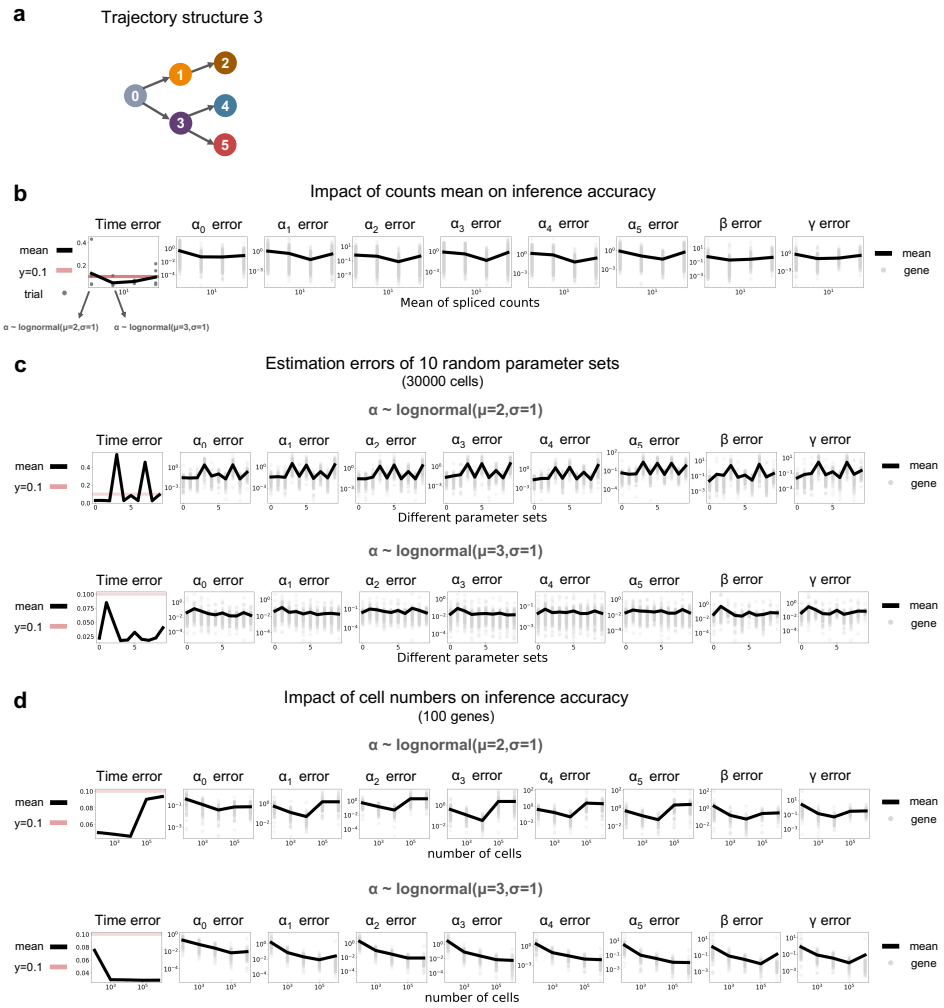

Figure 4.12: **Impact of trajectory structure complexity on inference accuracy**. **a**) The trajectory structure used in this figure. **b**) Impact of counts means on inference accuracy. Datasets with increasing counts means are generated by increasing the mean parameters μ in log-normal distributions for α. **c**) Results on 10 random parameter sets with different distributions for α. **d**) Impact of cell numbers on inference accuracy on simulations with different distributions for α.

As unaccounted noise is prevalent in scRNA-seq datasets, we then characterize the effects of noise on inference. The first type of noise is cell-wise read depth (or cell size) that influences all genes similarly. The inference is sensitive to such noise: when read depth variance is not correctly accounted for in the fitting, inference results may be highly inaccurate when its squared coefficient of variation (CV^2) exceeds 0.1 (Figure 4.14a). As the CV^2 typically observed in real datasets exceeds 0.1 (Figure 4.19), an accurate estimate of cellwise read depth is critical. Considering this fact, instead of simply using total counts as normalization factors, we use normalized

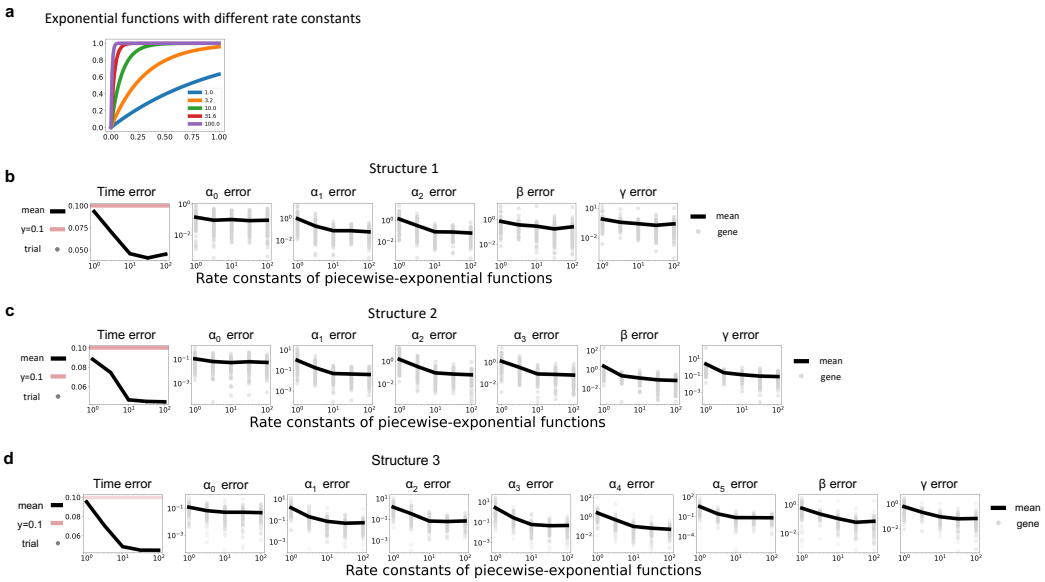

Figure 4.13: **Impact of non piecewise-constant transcription rate functions on inference accuracy**. Simulations are generated using piecewise-exponential functions for transcription rates and fitted under Chronocell's piecewise-constant assumption. The three trajectory structures have been defined in Figure 4.3a and Figure 4.12a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. **a**) One piece of the piecewise-exponential functions used for transcription rates. Different rate constants are used in the simulation to span the range from an almost linear transition to an almost step function. **b**) Results for trajectory structure 1. **c**) Results for trajectory structure 2. **d**) Results for trajectory structure 3.

covariance between genes to decompose extrinsic noise (influencing all genes) and intrinsic noise (gene specific). We estimate the read depth CV^2 across cells from the normalized covariance between genes. Subsequently based on the read depth CV^2, we subtract the extrinsic variance caused by the read depth from total variance and identify Poissonian genes whose remaining variances (intrinsic variances) are close to their means (variance $<$ 1.2 mean). We then estimate cell-wise read depth using the sum of those Poissonian genes. Different sets of genes are used for estimating the CV^2 of read depth and fitting trajectories. In reality, these read depth estimates correlate well with total counts number for most datasets, with one interesting exception (Figure 4.19).

On the other hand, gene specific noise seems to have less impact on inference. We added gene-wise gamma noise in simulation which generates negative binomial distributions in steady states and approximates bursting noise. Parameter errors

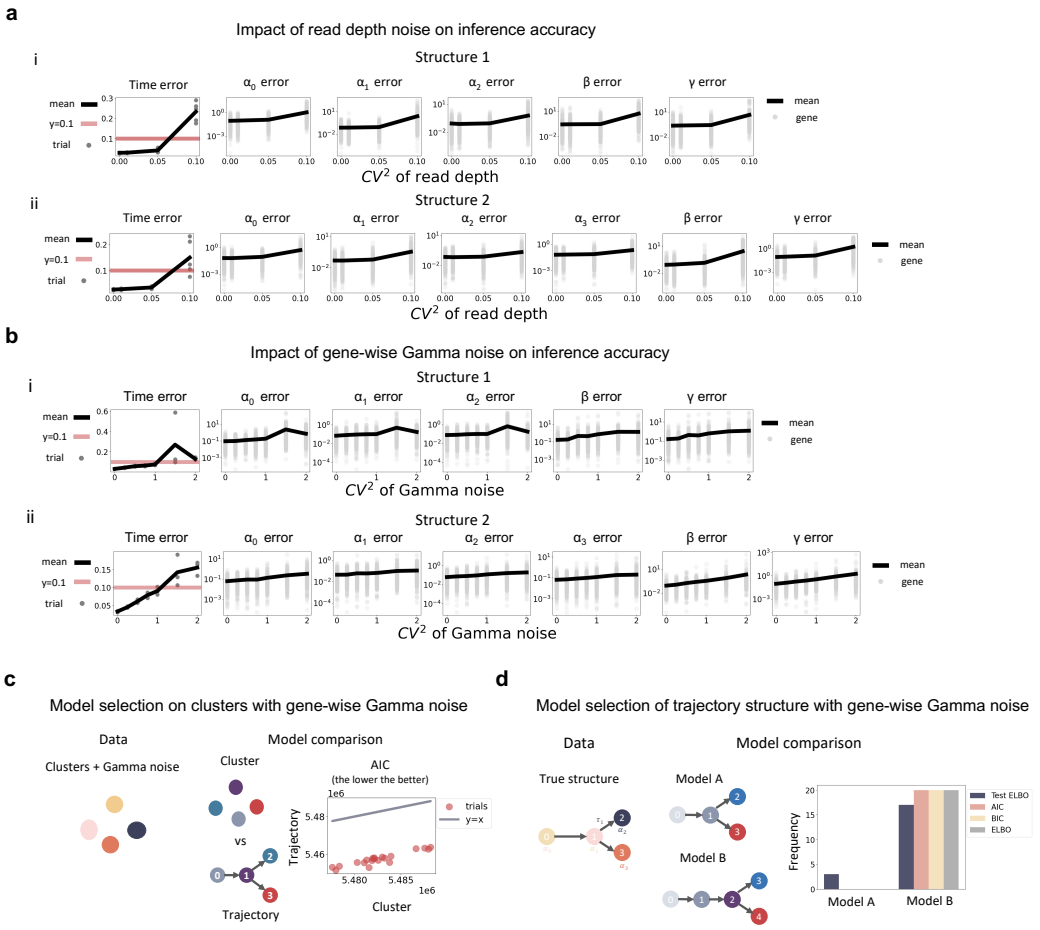

Figure 4.14: **Impact of noise on inference accuracy and model selection**. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. **a**) Estimation errors as read depth noise increases. **b**) Estimation errors as gene-wise Gamma noise increases. **c**) Impact of gene-wise Gamma noise on model selection on clusters data. Same as in Figure 2 except Gamma noise with CV^2 1 was added to Poisson mixtures to generate simulation data (Figure tion 4.4). **d**) Impact of gene-wise Gamma noise on model selection of trajectory structure. Same as in Figure 3j except Gamma noise with CV 1 was added in simulation.

gradually increase as CV^2 of Gamma noise increases but remain reasonably small even with a CV^2 of 1 (Figure 4.14b). However, while it may not completely undermine the fitting process, it can lead to failures in model selection. When repeating the model selection procedures in Figs 4.5 and 4.7 after introducing Gamma noise, the AIC/BIC and cross-validation metrics favor the wrong models that were more complex than the true one (Figure 4.14c and 4.14d). This highlights the importance of providing reasonable trajectory structure to the fitting based on

prior knowledge, and exploring alternative methods for quality control against false positives caused by clusters.

Another probable factor in real datasets that can lead to suboptimal results is insufficient dynamics. Intuitively, for a dynamical model to be appropriately fitted, the data must capture a significant amount of transient dynamics to recapitulate the evolving processes over time, without which the data start to resemble discrete clusters. Such cases can occur in at least two possible scenarios: fast timescales and concentrated sampling distributions, both of which result in clusters in the extreme. Therefore, false positives caused by clusters are naturally included as a component of identifying unreliable results.

The first situation, fast timescales, arises when mRNA half-lives are significantly shorter than the timescale of biological processes, and thus cells are mostly near steady state and provide little information about the intermediate dynamics. By setting the length of the time interval to unity and increasing β and γ, which is equivalent to increasing processes timescale with unchanged mRNA half-lives, we found that ideally, the mean value of γ should not fall out of the range of 1 to 10, and the mean ratio of γ to β over genes should not be too small (Figure 4.15). Hence, if the average half-life of spliced mRNA is approximately 30 minutes (Rabani et al., 2011), snapshot data sampled from processes involving steps exceeding 10 hours are no longer suitable.

Furthermore, given an appropriate timescale, the sampling distribution still has to cover the region where the transient dynamics occur. Imagine, for example, that all cells are from one unknown time point; there is no way for a trajectory to be inferred. Instead, a cluster should be used to fit the data. Thus, it is crucial to determine the minimum level of uniformity required in the sampling distribution and verify whether this requirement is met. By gradually changing sampling distributions from a uniform distribution to a Gaussian with a random mean, we generate datasets with sampling distributions that exhibit decreasing levels of uniformity, which was quantified using entropy (Figure 4.16a). As the sampling distribution deviates from uniformity, the errors of process time and parameters quickly increase as expected (Figure 4.16b). In fact, even when the sampling distribution is not far away from a uniform one with entropy 4, the errors are big enough that the results are completely wrong (Figure 4.16b). Further, even using the true sampling distribution as a prior for a warm start with correct initialized time cannot mitigate the lack of dynamics: the errors in parameter estimations remained substantial (Figure 4.16c). Therefore,

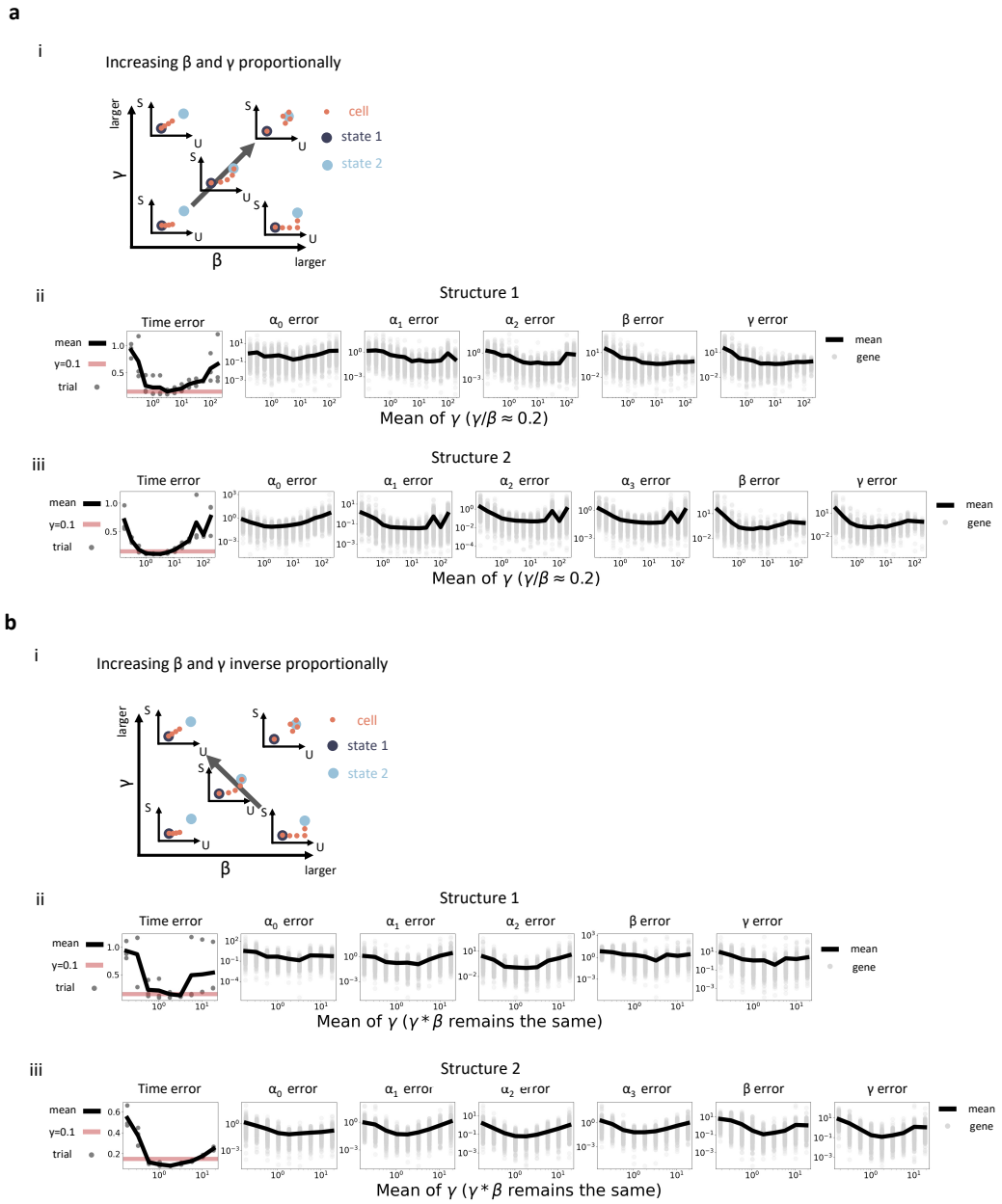

Figure 4.15: **Impact of dynamic timescale on inference accuracy**. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in the Section 4.4. **a**) **i** Schematics of phase plots with increasing timescale. **ii** Estimation errors as time scale increases. **b**) **i** Schematics of phase plots with the ratio of γ to β increasing while keeping their product constant. **ii** Estimation errors as $\frac{\gamma}{\beta}$ increases.

sufficiently transient dynamics is an inherent requirement for trajectory inference even when perfect prior information is provided.

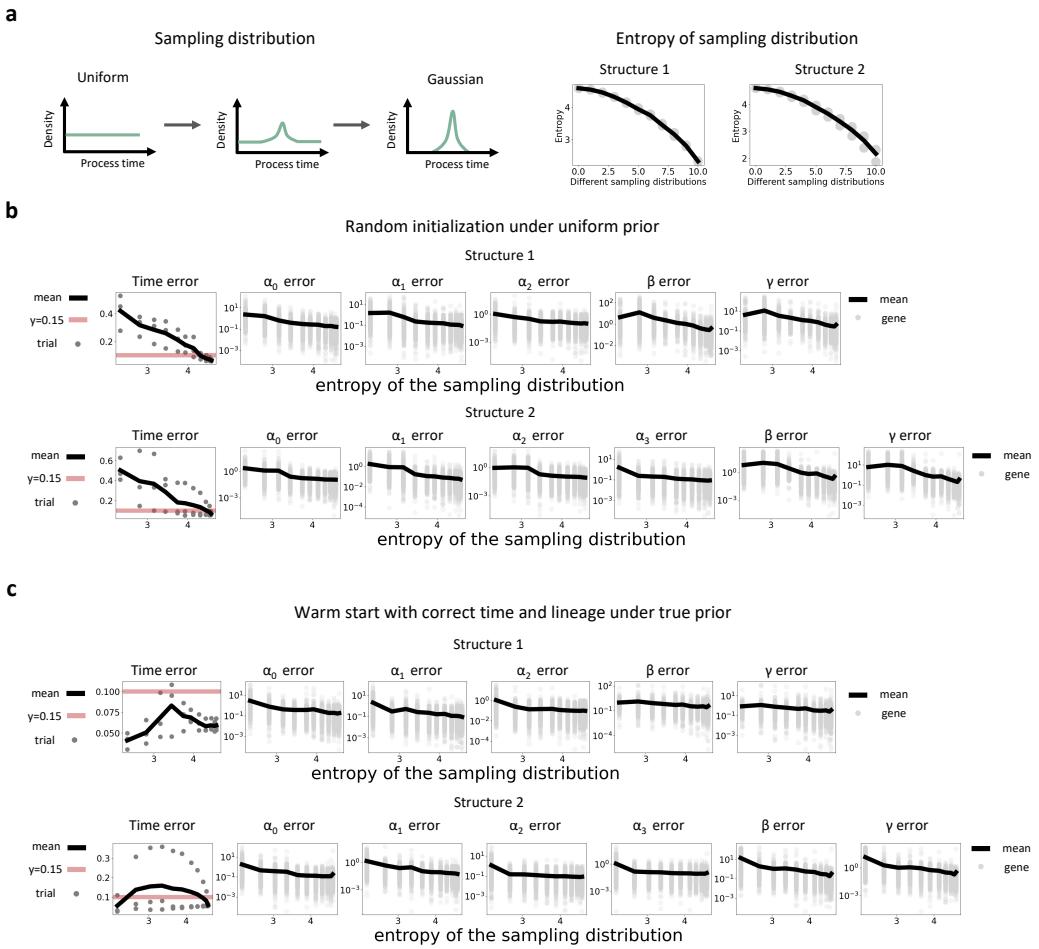

Figure 4.16: **Impact of sampling distribution uniformity on inference accuracy**. The trajectory structures are the same as in Figure 4.3a. For time, error is root mean square error. For α, β, γ, error is mean normalized error as described in Section 4.4. **a**) Schematics of sampling distributions used in simulation with decreasing uniformity. The sampling distributions were gradually changing from uniform distribution to Gaussian distribution. Right plot shows the entropy of the sampling distributions. **b**) Estimation errors as uniformity decreases under uniform prior. **c**) Estimation errors as uniformity decreases warm started with correct position under true prior. Fitting was initialized with posteriors calculated under true parameters, and empirical distribution of process time of samples were provided as prior for the sampling distribution.

In summary, a suitable dataset for fitting process time needs to contain enough dynamic information as well as limited noise. This stringent requirement for a successful trajectory inference highlights the need for suitable datasets and careful model assessment. Straightforwardly, we could make a consistency check by verifying if the remaining noise, parameter values, and uniformity of the average posterior

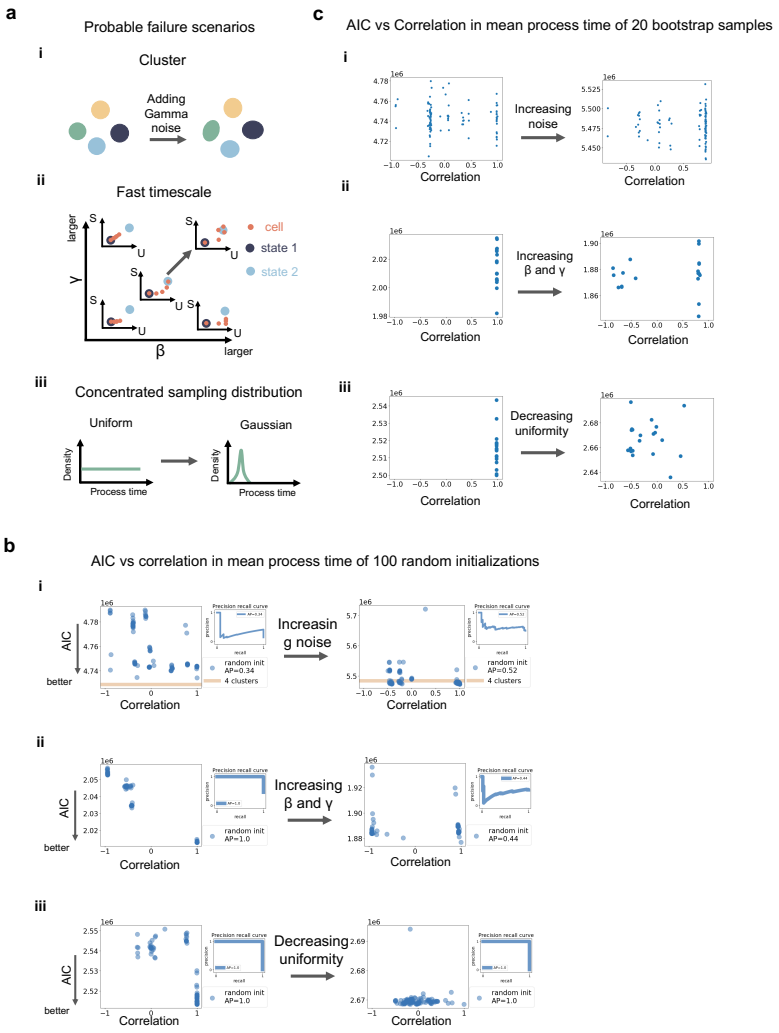

Figure 4.17: **Uncertainty and lack of robustness as an indicator of failure scenarios**. **a**) Schematics of three probable failure factors: clusters data, fast timescale, concentrated sampling distribution. Two example simulations were used for each case and the gray arrows indicate the their difference. **i** The two simulations are the cluster data in Figure 2 and noisy cluster data in Figure 4.14c respectively. **ii** The two simulations are the 5th and 13th instances of structure 1 in Figure 4.15b. **iii** The two simulations are the 1st and 11th instances of structure 2 in Figure 4.16b.**b)** AIC vs. correlation of mean process time of 100 random initializations in different scenarios. Results of two example simulations were showed. The x-axis is the correlation of mean process time between each initialization and the best one. The gray arrows correspond to those in **a**). **c**) AIC vs correlation of mean process time of 20 bootstrap samples in different scenarios. Results of the two example simulations were presented in the same position as in **b**). The x-axis is the correlation of mean process time between each bootstrap sample and the original one (which is the best one in **b**). The gray arrows correspond to those in **a**).

distribution fall into the appropriate range determined in simulations. However, when a result is incorrect, it may not necessarily exhibit large unexplained noise, high splicing/degradation rates, or a concentrated cellular distribution over process time, as it could inadvertently fit undesired patterns and output plausible results. Thus, inconsistency is a sufficient but not necessary indicator of unreliable results.

It turns out that the large uncertainty of results can be a better indicator. As both noise and limited dynamics tend to diminish or obscure the difference of scores like ELBO between correct and incorrect outcomes, they introduce multiple comparable maxima and make the fitting results unstable. The resultant large uncertainty can be measured by two different approaches. First, as each random initialization outputs a different ELBO/AIC score and cell ordering, we can inspect the distribution of scores with respect to a summary parameter of cell orderings, which describes the global landscape of the score function. Ideal scenarios usually give one distinct maximum of ELBO (or minimum of AIC) at the correct ordering of cells (correlation around one), while failure scenarios usually lead to multiple comparable maxima of ELBO (or minima of AIC) at different cell orderings beside the correct one (Figure 4.17b). We use average precision (AP) of the 100 random initializations to quantify the distinctiveness of the correct outcomes and summarize the uncertainty. One random initialization is considered correct if its resultant mean process time correlates well with that of the best one, e.g., a Pearson's correlation of 0.8 (Section Uncertainty assessment). Ideal cases lead to AP close to one and low AP indicates instability but not vice versa, which means low AP is a sufficient indicator for instability (Figure 4.17b). Second, the uncertainty can also be revealed by standard bootstrap analysis. We generated 100 sets of resampled data and computed the correlation in process time between the original and each resampled set (see Section 4.5 "Uncertainty assessment"). In failure scenarios the results of original data and resampled data often differ and correlations are scattered with large variance. On the other hand, in ideal scenarios, process time estimates of resampled results agree with the original one and correlations are centered around one (Figure 4.17c).

Uncovering distinct underlying cellular distributions in process time

We applied Chronocell to a variety of datasets with different anticipated sampling distributions over time. For those datasets, we estimated read depth as described in Section Read depth estimation (Figure 4.19), and then filtered genes for fitting based on their means, variances and unspliced to spliced ratios (see Section 4.5 "Real datasets preprocessing"). The trajectory structures are determined based on

prior knowledge. Random initialization was always performed and its uncertainty was assessed by both AP and bootstrapping. Warm start was applied as well if cell type annotations were available, which were used to initialize the fitting process. The corresponding clusters (Poisson mixtures) model was fitted with the same read depth and used for comparison. The genes were selected based on goodness of fit, in the same manner as in the simulation (see Section 4.5 "Gene selection"). Then, DE genes are selected based on the fold changes in transcription rates across states.

We first tested our method on the T cells of PBMC (Peripheral Blood Mononuclear Cells) dataset from 10x Genomics, which is typically expected to exhibit a few distinct clusters (Figure 4.20a). Indeed, though the AIC of trajectory model is lower than clusters model likely due to unaccounted noise, the scores of 100 random initializations display multiple minima: multiple different cell orderings result in similar AIC values (Figure 4.20b). The low average precision (0.28) of random initializations suggests clusters model would be more suitable for PBMC. Serving as a negative control, it confirms our ability to reject unreliable results on real datasets, even when standard model selection methods are invalidated by incorrect modeling of noise.

The second dataset contains a snapshot collection of glutamatergic neuronal lineage cells in a developing human forebrain (Figure 4.21a) (La Manno et al., 2018), which is presumed to capture cells along a continuous trajectory. However, the unstable result of random initializations indicates its unreliability, as results with reversed directions yield comparable AIC scores (Figure 4.21b), resembling simulations that lack dynamics information (Figure 4.17b). This observation is further supported by examining the cellular posterior distributions obtained through warm start with cell type annotations, where the average posterior distribution reveals that cells are concentrated around starting time τ_0, i.e., the initial state, with a low entropy (Figure 4.21c). Therefore, both the inconsistency indicated by the concentrated posterior and instability indicated by comparable peaks of AIC suggest this is not a suitable dataset for Chronocell and likely lacks enough dynamics.

The third dataset contains erythroid lineage cells during mouse gastrulation collected from multiple time points (Figure 4.22a) (Pijuan-Sala et al., 2019). Biological time availability offers a valuable means to evaluate results by comparing the posterior distributions of cells to their corresponding physical times, which is particularly useful since real snapshot datasets lack a definitive ground truth. The AIC scores of random initializations show a clear minimum at a correct direction that align

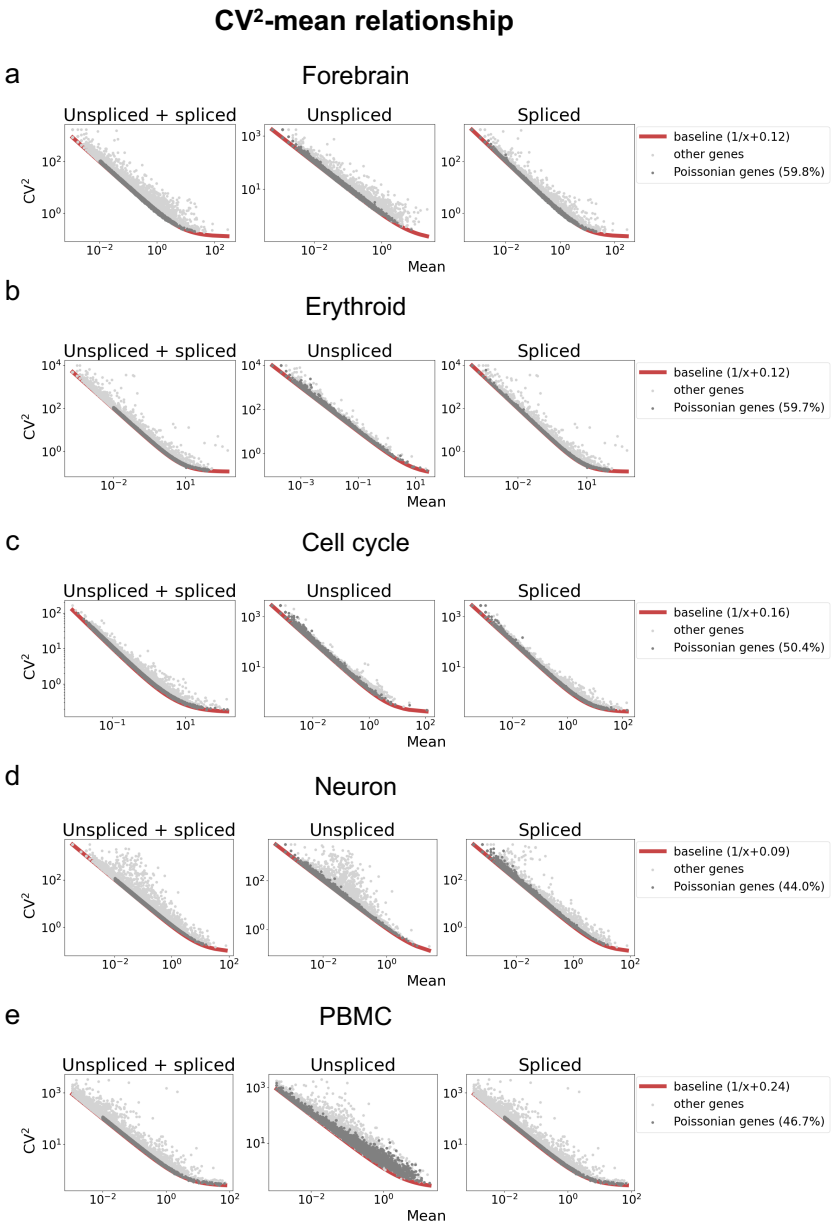

Figure 4.18: *CV*^{2}**-mean relationship of total, unspliced and spliced counts**. **a**) Forebrain data. **b**) Erythroid data. **c**) Cell cycle data. **d**) Neuron data. **e**) PBMC data.

with cell type annotations (Figure 4.23a), and the average precision of random initializations is 0.79 which is notably higher than those of PBMC and Forebrain datasets (Figures 4.20b and 4.21b). The process times of bootstrap samples are reasonably stable as well (Figure 4.23a). The fitted dynamics were able to explain most of the variance, leaving the CV^2 of unexplained noise of most genes under 1 (Figure 4.23c). Therefore, the Erythroid dataset appears to be a suitable dataset

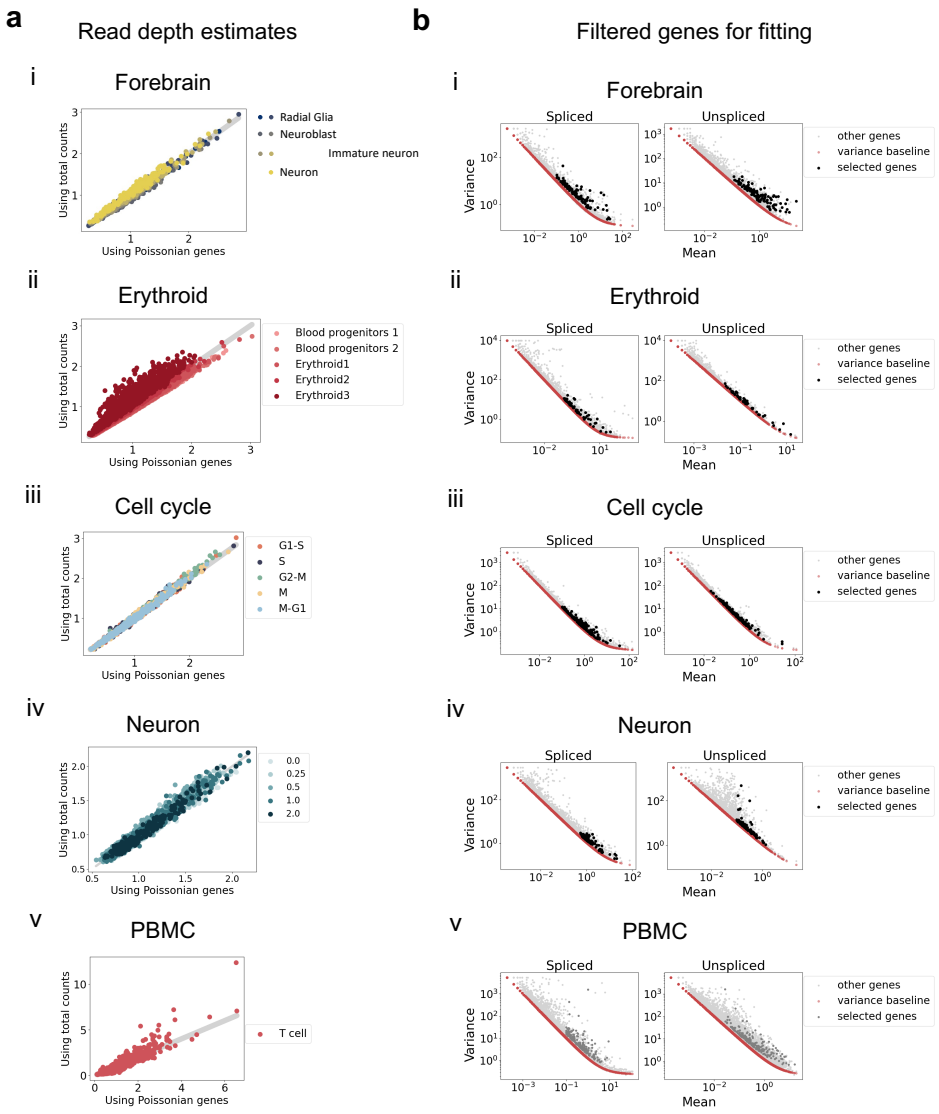

Figure 4.19: **Read depth estimation using normalized covariance for highly variable gene selection**. **a**) Read depth estimation based on total counts of Poissonion genes. Cells are colored by their cell types. **b**) Selected genes (black) for fitting plotted in the same CV^2-mean plot as in Figure 4.18).

for our trajectory model. The best result from random initialization is used for analysis (Figure 4.22b), and cellular posterior distributions confirm that cells have a broad distribution over process time (Figure 4.22c). Furthermore, the posterior distributions of cells do not differ significantly until E7.5, after which they roughly progress along the process time in sync with real-time progression (Figure 4.22d). This observation aligns with the understanding that erythroid differentiation in a mouse embryo is believed to begin around embryonic day 7.5 (E7.5) (Baron, Isern,

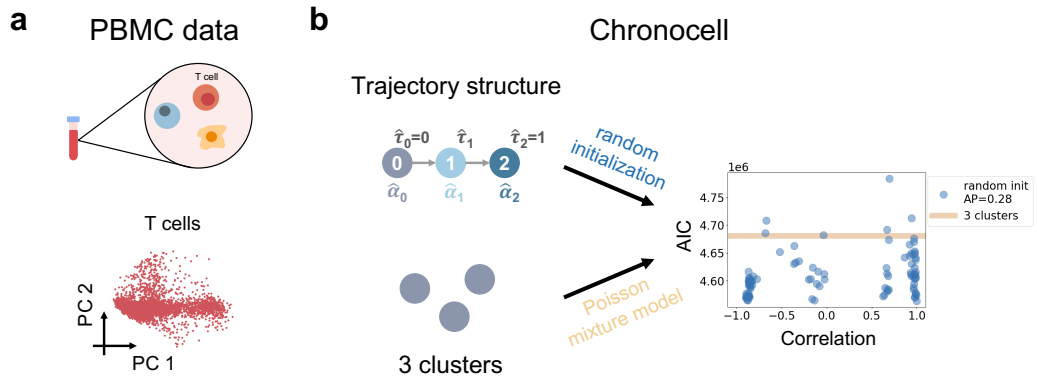

Figure 4.20: **Inference results for T cells from PBMC data**. **a**) Schematics of T cells from PBMC dataset and PCA plots. **b**) The fitted trajectory structure and AIC scores of 100 random initializations (blue dots) compared to those of three clusters (Poisson mixtures) model (yellow line). AP stands for average precision.

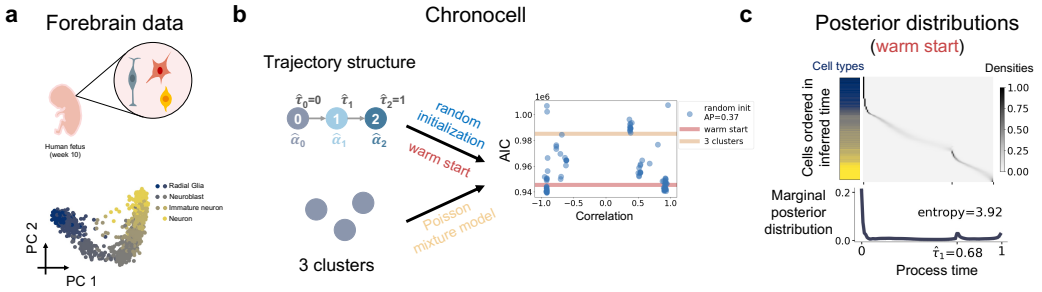

Figure 4.21: **Inference results for Forebrain data**. **a**) Schematics of Forebrain data and PCA plot of cells colored by cell type annotations. **b**) The AIC scores of the trajectory and cluster models. The x-axis is the mean process time correlations of 100 random initializations (blue dots). The AIC scores of random initializations are compared to those of warm start (red line) as well as three Poisson mixture model (yellow line). AP stands for average precision. **c**) Posterior distributions of process time of the trajectory model with warm start. Cells were ordered in y-axis by their inferred mean process time and the left bar displays their cell types using the colors in **a**). The below histogram shows average posterior distribution averaged over cells. The entropy of the average posterior distribution was calculated using its weights on the 100 discretized time grids.

and Fraser, 2012). Although the alignment with physical time is not perfect, it still suggests that our trajectory model can successfully capture the correct trend of process time, even when assuming a uniform sampling distribution for all cells.

After applying our gene selection procedure based on relative likelihood and discarding genes with extreme values, we ended up with 24 (49%) genes (Figure

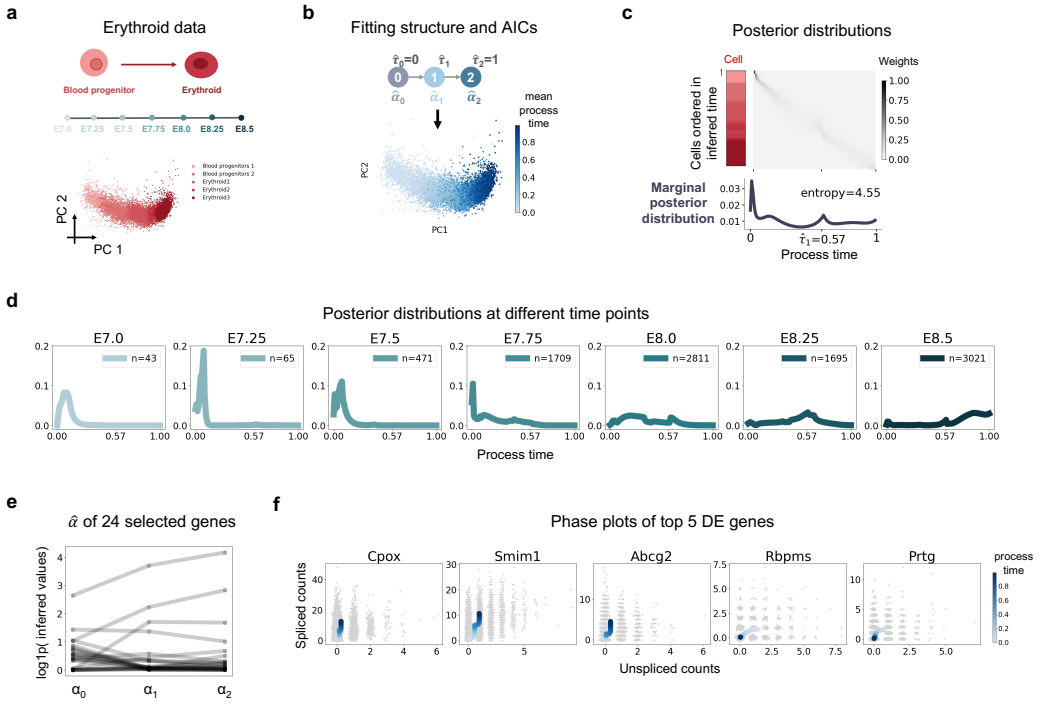

Figure 4.22: **Inference results for Erythroid data**. **a**) Schematics of Erythroid data and PCA plot of cells colored by cell type annotations. **b**) The fitted trajectory structure and inferred mean process time from random initialization indicated in blue on the same PCA plot as in **a**). **c**) Posterior distributions of process time. Cells were ordered in y-axis by their inferred mean process time and the left bar displays their cell types using the colors in **a**). The below histogram shows average posterior distribution over cells. The entropy of the average posterior distribution was calculated using its weights on the 100 discretized time grids. **d**) Averaged posterior distribution across cells from different experimental time points. n is the number of cells. **e**) α values of 24 selected genes over states. **f**) Phase plots of top five DE genes of 24 selected genes. The x-axis is the raw unspliced counts and y-axis is the raw spliced counts. The blue curve is the fitted mean of product Poisson distributions of unspliced and spliced counts over process time, and its darkness corresponds to the value of process time.

4.22e). Subsequently for demonstration, we chose the top five DE genes (*Cpox, Smim1, Abcg2, Rbpms, Prtg*) with the largest fold change of transcription rates, and plotted their phase portraits (Figure 4.22f). Interestingly, four of them (*Cpox* (Taketani, Furukawa, and Furuyama, 2001), *Smim1* (Aniweh et al., 2019), *Abcg2* (Zhou et al., 2005), *Rbpms* (Rooij et al., 2017)) were reported to be directly relevant to erythroid development, which illustrates that selecting DE genes based on inferred transcription rates is a straightforward and effective approach.

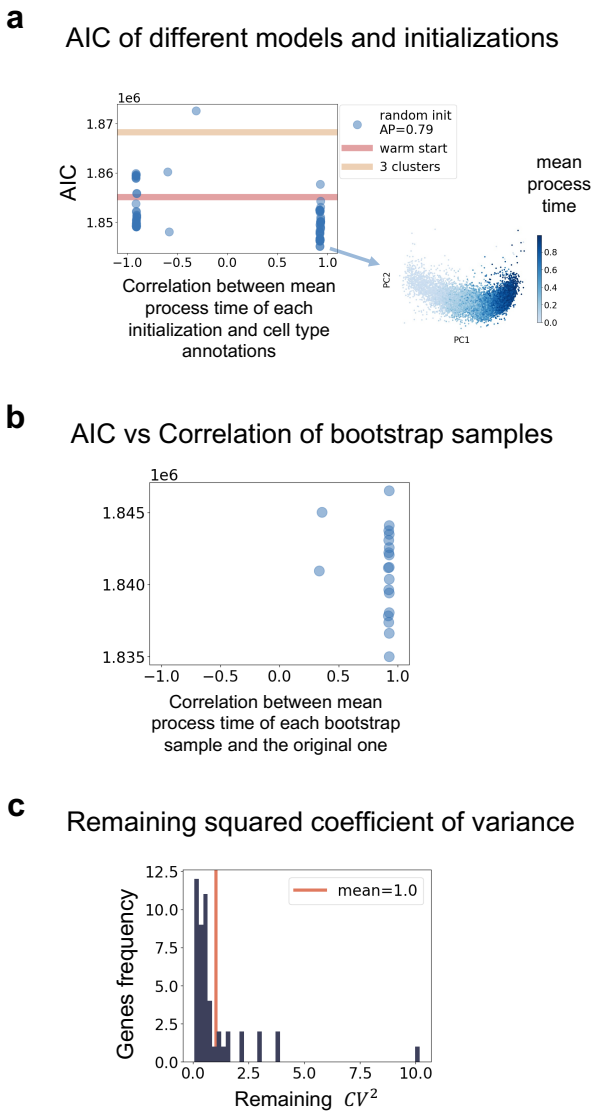

Figure 4.23: **Supplementary figures for Erythroid data**. **a**) AIC scores and mean process time correlations of 100 random initializations (blue dots) compared to those of warm start (red line) as well as three clusters (Poisson mixtures) model (yellow line). AP stands for average precision. Mean process time of the initialization with lowest AIC is indicated in blue on the same PCA plot as in **a**). **b**) AIC scores and mean process time correlations of 100 bootstrap samples. The x-axis is the Pearson's correlation between the mean process time of each bootstrap and the those of original data, i.e., the plotted one in **a**). **c**) Distribution of remaining squared coefficient of variance of 49 genes used in the fitting. Remaining squared coefficient of variance is calculated by dividing the remaining unexplained variance by mean squared.

Based on the results from datasets ranging from clusters to trajectories, it becomes evident that real datasets can display a spectrum of continuity in cellular process

time distribution. Therefore, it is critical to assess the quality of inference and verify the requirements for reliable results are indeed met.

Degradation rates estimates agree with metabolic labeling data

In addition to validating the process time, we also sought to validate the inferred parameters, which motivated us to use a metabolic labeling dataset from scEU-seq, comprising human retinal pigment epithelial (RPE1) cells undergoing cell cycles (Figure 4.24a) (Battich, Beumer, et al., 2020). Metabolic labeling of new mRNA allows for the estimation of degradation rates from cells with varying labeling times, enabling a comparison with our parameter estimations.

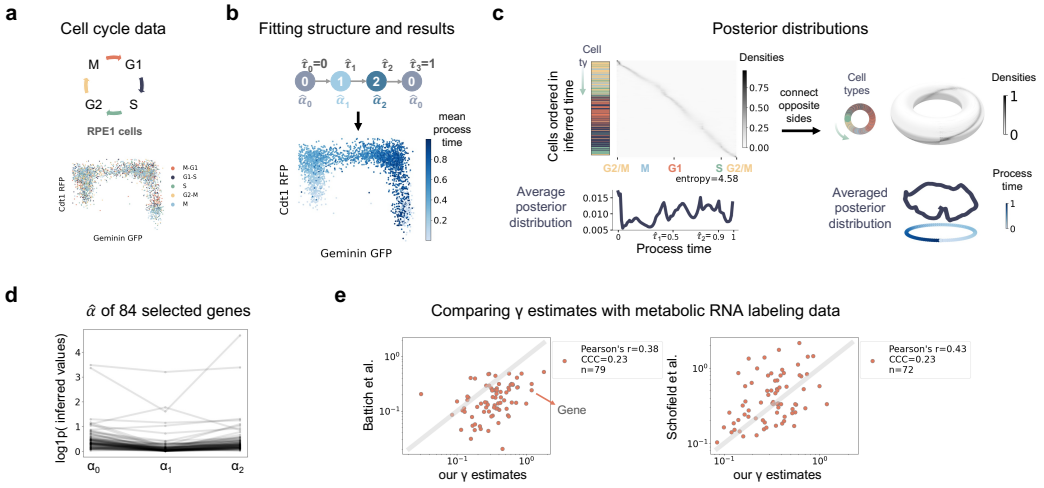

Figure 4.24: **Inference results for Cell cycle data**. **a**) Schematics of Cell cycle data and scatter plot of the Geminin-GFP and Cdt1-RFP of RPE1 cells colored by cell type annotations. **b**) The fitted trajectory structure and inferred mean process time from random initialization indicated in blue on the same scatter plot as in **a**). **c**) Posterior distributions of process time. Cells were ordered in y-axis by their inferred mean process time and the left bar displays their cell types using the colors in **a**). The below histogram shows average posterior distribution averaged over cells. The entropy of the average posterior distribution was calculated using its weights on the 100 discretized time grids. **d**) α values of 84 selected genes over states. **e**) Comparison of γ estimates for 84 selected genes with estimates derived from metabolic RNA labeling data. CCC stands for concordance correlation coefficient. n is the number of genes for which estimates are available in each respective paper.

In the trajectory structure, we specified that the initial and final states were identical. Given our assumption that the initial state is at a steady state, capturing the cyclic nature of the cell cycle poses an additional requirement that cells must also approach a steady state in the last interval as well, which happened to be reasonably satisfied by

our results. In line with the expectation that cell cycle is a highly dynamic process, the random initialization provides a relatively stable result: the AIC exhibits a single minimum that aligns with the direction of cell cycle progression (Figure 4.25a), and bootstrap samples mostly align with the result of the original one (Figure 4.25b). Starting from the result of random initialization, a desynchronized model was fitted to enhance the accuracy of parameter estimations (Figure 4.25c). The resulting fit maintains alignment with the cell cycle progression (Figure 4.24b), and the CV^2 of the unexplained noise mostly remains below 1 for most of the genes (Figure 4.25c). The RPE1 dataset was metabolically labeled for different lengths of time but posterior distributions of RPE1 cells with different labeling times do not show significant differences, as a negative control in contrast to the erythroid data (Figure 4.25e).

The posterior distributions successfully capture the cyclic nature of the data. The cell types ordered by mean process time exhibit a cyclic pattern (Figure 4.24c); the phase plots of marker genes confirm the cycling dynamics of fitted means of unspliced and spliced counts (Figure 4.25f); the starting and ending values of fitted means of most of genes match with each other (Figure 4.25g). As both axes are cyclic, it is possible to connect the opposite edges and transform the two-dimensional cell-by-time grids posterior distributions into a torus on the surface of which the posterior distributions look like a circle (Figure 4.24c). Based on cell type annotation, total RNA counts number (Figure 4.25h), and marker genes dynamics (Figure 4.25e), we can roughly assign the three intervals to G2/M, G1/S, and S/G2 phase, and mitosis happens shortly after τ_0.

After selecting genes by relative likelihood and discarding genes with extreme values, we ended up with 84 (46%) genes (Figure 4.24d). We compared the degradation rates of 84 selected genes to those derived by scEU-seq (Battich, Beumer, et al., 2020) and TimeLapse-seq (Schofield et al., 2018). For scEU-seq, as we neglected the changes in degradation rate along cell cycle, we used the averaged degradation rates in Figure S10C of Battich et al. We observed moderate correlations between our degradation rate estimates and the respective estimates from Battich and Schofield (Figure 4.24e), and these correlations exhibited a magnitude similar to the correlation between the Battich and Schofield estimates (Figure 4.25i), which suggests that the parameters inferred by Chronocell indeed possess a meaningful biophysical interpretation.

4.4 Discussion

We have introduced Chronocell, a method constructed upon a trajectory model featuring biophysically meaningful parameters and a principled approach to fitting and analysis. Counts are directly modeled, estimation accuracy is characterized, and unreliable instances are discerned. We found several requirements regarding dynamics and noise that must be satisfied to obtain reliable results, which makes process time inference challenging and renders retrospective model assessment indispensable. We applied Chronocell to different kinds of real datasets, recognized inapplicable ones by assessing instability, and demonstrated meaningful interpretation of process time and parameters on applicable ones.

Of note, our trajectory model is simplified to balance interpretability and tractability. Building on the concept of cell state transitions, we have assumed piecewise constant transcription rates which describe genes with fast chromatin states switching in saturated regime, but may be unrealistic for many biological systems. Furthermore, we have assumed that splicing and degradation rates remain constant, which although reasonable for many genes, is not accurate for genes with peaked response (Rabani et al., 2011). We also did not incorporate transcriptional bursting due to the absence of an analytically determined temporal solution, and this leads to a under-dispersed distribution compared to what is typically observed in biological data. Nevertheless, this can be resolved by using numerical solutions (**Gorin2023-ax**). However, even with this simplified model, we have found that accurate inference imposes strict requirements on data. This question is inherently challenging and insufficiently specified due to the existence of latent variables and flexibility of the transcription rates. A more realistic model would have even more stringent requirements.

Since fitting dynamical parameters is challenging from static snapshots, physical time information could offer valuable insights when incorporated into our latent variables model. While the requirements on sampling and noise might potentially be relaxed, they would likely persist to some extent. Thus, intensely sampled time series datasets derived from well-defined cellular processes would be the ideal choice for trajectory inference. However, it remains to address the key question of how physical time should be translated into sampling distribution assumptions. The choice of sampling distribution reflects our understanding of cell heterogeneity at each time point. Should we fix the process time to physical time completely, i.e., a delta distribution? Or should we assume heterogeneity within one time point? If so, what kind of distribution should we use? There is no straightforward answer,

as it depends on our understanding of the process being studied. For instance, in a neuron dataset with cells sampled at five time points after stimulation, assuming homogeneous cell responses would suggest using a delta distribution, which aligns process time directly with physical time. However, if we account for heterogeneity within a time point, other distributions may be more appropriate. Furthermore, assuming different distributions (delta, exponential, uniform) can lead to different results (Figure 4.26). This underscores the necessity for a more comprehensive understanding and modeling of cell heterogeneity, and optimal experimental design for both the number and timing of the time points to generate more informative data.

We have not yet provided a benchmark against other methods. Each trajectory inference method assumes a different model, and comparing methods with different underlying models is often less informative without ground truth. For example, descriptive models use conceptually different approaches, so comparing them is an "apples to oranges" situation, as they cannot recover the true time of data under Chronocell trajectory model. On the other hand, they may be more suitable for exploratory analysis, with fewer data requirements and being less computation-demanding. Similarly, in model-based approaches, the model inherently reflects our understanding of the data, and discrepancies in results arise from differences between models. This underscores the importance of using models that are grounded in biophysical motivation. For demonstration purposes, we provide a comparison of Chronocell with some widely used trajectory inference methods, including Slingshot (Street et al., 2018), Monocle 3 (Cao et al., 2019), and diffusion pseudotime (Haghverdi et al., 2016), as well as recently published veloVI which also integrates trajectory inference with RNA velocity (Du et al., 2024). These comparisons are demonstrated on the simulated data used for illustration in Figure 4.7, as well as simulations generated by dyngen (Cannoodt, Saelens, Deconinck, et al., 2021) (Figure 4.27). We find that other methods cannot correctly recover true time on data generated under the Chronocell model (Figure 4.28). For dyngen simulation, Chronocell has comparable or better accuracy though all methods capture the correct trend but fail to recover true time accurately. For real data, benchmarking is more challenging due to the lack of ground truth. The closest approximation to ground truth is the experimental time in time series data, especially with short enough intervals and a clear start point. Despite the uncertainty of assumed cell heterogeneity discussed above, at least the inferred time of cells should proceed along with experimental time. Therefore, we also test other methods on the neuron data and find that only Monocle 3 captures the general trend of time progression (Figure 4.29).

In summary, our biophysically motivated model of the dynamical processes captured in single-cell data enables the inference of process times and parameters with biophysical interpretations. It presents an alternative approach to unveil continuous latent cell representations within a well-defined and rigorous framework, and highlights the limitations of what can be inferred using current snapshot single-cell genomics data.

4.5 Methods

We begin by describing our trajectory model, followed by a description of the inference procedure. Next, we explain the analysis pipeline, including our gene and model selection procedure. Then, we elucidate the simulation setup. Finally, we detail the preprocessing of real datasets.

Throughout the Methods we denote data by $\mathbf{X}$, and note that $\mathbf{X}$ corresponds to a cell by gene by species array. We use i to index cells, j to index genes, and c to index species. Parameters are denoted by θ, and latent variables by z. $p(\cdot)$ means a probability distribution.

Model

We used a simple and interpretable latent variable model for the probability distribution of the counts, with explicit biophysical meaning associated to each latent variable. We assumed cells are asynchronous, and we therefore introduced two latent variables that corresponded to the lineage and time of the cell.

Therefore, in our trajectory model, the gene expression data of each cell $\mathbf{x}$ was described as a function of latent variables $\mathbf{z} = (l, t)$ which specify the lineage l and time t of the cell. By considering a particular parametric class of gene expression dynamics that specify the distribution $p(\mathbf{y}|\mathbf{z}, \theta)$ of *in vivo* counts $\mathbf{y}$, and a sequencing noise model $p(x|y)$, we mapped latent variable z into data space and arrived at an explicit formulation of data distribution that could be trained using the expectation-maximization (EM) algorithm. In other words, we assumed the counts of each cell $\mathbf{x}$ had the following density:

$$p(\mathbf{x}|\theta) = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}, \theta) p(\mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}|\mathbf{z}, \theta) p(\mathbf{z}) d\mathbf{z},$$

where $p(z)$ is the distribution of latent variables.

In the following, we specify $p(y|z,\theta)$, $p(x|y)$ and $p(z)$, which are based on the transcription model, measurement model and sampling measure.

Transcription model

Since current transcriptomic data can be used to measure the numbers of both nascent and mature transcripts, we considered the production, processing, and degradation of individual RNA molecules in our transcription model, as in previous RNA velocity literature (La Manno et al., 2018).

The distribution of RNA counts in the reaction system (Equation4.2) is known (Jahnke and Huisinga, 2007). We assumed that the initial distribution of U and S was a product Poisson distributions, which implies that the mean parameter vector $\lambda = (\lambda_u, \lambda_s)$ of U and S evolved according to Equation(4.3).

The functional form of $A_l(t)$ reflects our assumptions of gene expression during development. However, explicitly modeling transcription rates as functions of gene expressions is hard and tends to overfit. More fundamentally, we have some prior physical intuition; this function is actually $A_l(\lambda, \mathbf{u}, t)$, where $\mathbf{u}$ is some high-dimensional vector of regulator concentrations. These data are not possible to collect using currently available technologies, although this constraint may change in the coming years. In principle, we can write down these equations, and even simulate them (using dyngen (Cannoodt, Saelens, Deconinck, et al., 2021) or the stochastic simulation algorithm), but we strived to start with an analytically tractable model, particularly to recapitulate and formalize the increasingly popular RNA velocity framework. Therefore, we used a phenomenological model and didn't consider gene interactions. Rather, the effect of transcriptional regulation on each gene during development was summarized into synchronized state switching and each state had its transcription rate. This formalized transient cell types.

To be able to account for multiple lineages, we assumed cell states $\mathcal{S} = 0, 1, ..., S$ formed a directed graph, with each lineage represented as a path of length K on this graph.. The trajectory structure described the graph, and recorded the states $s(l,k) \in \mathcal{S}, l = 1, ..., L, k = 1, ..., K$ of the L lineages during the K stages. We treated the trajectory structure as known since it can typically be obtained from our previous knowledge of the data. Specifically, for lineage l, we assumed $A_l(t)$ is a piecewise constant function: $A_l(t) = \alpha_s(l,k)$, where the cellular state index s is determined by the lineage l, and the time interval k to which t belongs, i.e., $t \in (\tau_{k-1}, \tau_k]$. The state index $s = s(l,k)$ was determined by the trajectory structure.

For example, given the trajectory structure in Figure 4.2, $s(l = 1, k = 0) = s(l = 2, k = 0) = 0$, $s(l = 1, k = 1) = s(l = 2, k = 1) = 1$, $s(l = 2, k = 2) = 2$ and $s(l = 2, k = 2) = 3$.

Then, the parametric solution of λ was given by

$$
\begin{aligned}
q &= \arg\min_k \{k | \tau_k \geq t\}, \\
\lambda_u(t) &= \sum_{k=1}^{q-1} \frac{\alpha_{s(l,k)}}{\beta} \left(e^{-\beta(t-\tau_k)} - e^{-\beta(t-\tau_{k-1})} \right) + \frac{\alpha_{s(l,q)}}{\beta} \left(1 - e^{-\beta(t-\tau_{q-1})} \right) + \lambda_u(0) e^{-\beta t}, \\
\lambda_s(t) &= \sum_{k=1}^{q-1} \frac{\beta \alpha_{s(l,k)}}{\gamma(\beta-\gamma)} \left(e^{-\gamma(t-\tau_k)} - e^{-\gamma(t-\tau_{k-1})} \right) + \frac{\beta \alpha_{s(l,q)}}{\gamma(\beta-\gamma)} \left(1 - e^{-\gamma(t-\tau_{q-1})} \right) \\
&\quad + \left(\lambda_s(0) + \frac{\beta}{\beta-\gamma} \lambda_u(0) \right) e^{-\gamma t} - \frac{\beta}{\beta-\gamma} \lambda_u(t).
\end{aligned}
$$

Assuming cells are at steady states at time 0, we have $\lambda_u(0) = \alpha_{s_{l,0}}$ and $\lambda_s(0) = \frac{\beta}{\gamma} \alpha_{s_{l,0}}$. Therefore, for each gene, the parameters are $\theta = (\alpha, \beta, \gamma, \tau)$.

Measurement model

Next, we needed to have a measurement model that connected counts number **y** in cells to the observed counts **x** in a single-cell RNA-seq experiment.

We assumed each molecule of mRNA produces $Bernoulli(q)$ number of captured RNA molecules, which is also an good approximation of a Poisson model with low mean (Gorin and Lior Pachter, 2023; Sarkar and Stephens, 2021).The capture rate q for each molecule can potentially depend on factors such as cell read depth, gene-specific, and species-specific biases in mRNA capture methods. However, in our model, we assume the capture rate only varies with cell read depth (or cell size), i.e., $q = r_i$, where r_i represents the read depth (cell size) of cell i. With i as the cell index, j as the gene index, and c as the species index, we have

$$
x_{ijc} \sim binomial(y_{ijc}, r_i), \quad y_{ijc} \sim Poisson(\lambda_{jc}),
$$

where λ_{jc} is the ODE solution for species c of gene j.

This is equivalent to multiplying the mean of Poisson distributions by a constant c_i. Since usually only the relative value of read depth c_i can be available, we absorbed the mean of c_i into α and infer their product directly:

$$x_{ijc} \sim Poisson(c_i \lambda_{jc}).$$

Sampling distribution

Now that we have defined the $p(x|z, \theta)$, the only remaining thing to complete $p(x|\theta)$ is the sampling distribution $p(z)$, which describes the prior distribution of latent variables. Given the formula of $p(y|z, \theta)$, it is easy to see that the model is not identifiable if $p(z)$ is not fixed, because we can change $p(z)$ together with β and γ easily without changing $p(x)$, for example, by scaling β and γ and transform $p(z)$ accordingly. Therefore, we assumed $t \in [0, 1]$ and fixed a uniform prior for t on $(0, 1]$. With this given prior distribution, the model was identifiable, because the parameterized form of the means of Poisson distributions is identifiable and Poisson distribution is identifiable (Teicher, 1961). If one has information about real time, one can adjust the range and prior of t to be have more physical meaning. For example, for the cell cycle dataset, if one knows that the whole cycle takes 24 hours, then one can either set the range of t to be $[0, 24]$, or scale the results by dividing both β and γ by 24, while keeping the other parameters unchanged.

Connection to cluster models

In the fast dynamic limit, as there are few cells out of steady states, transitions (edge) disappear and only states (nodes) remain, and both the weight of lineages and the length of time interval (or the weight at t=0 for state 0) determine the mixture weights of clusters. Specifically, for the state 0, its weight equals the weight at t=0, while for the state 2, its weight equals the product of the length of the second time interval $(\tau_1, \tau_2]$ and the weight of the second lineages. Thus, the Poisson mixtures model strictly belongs to the degenerate cases of our trajectory model, which connects Chronocell to biophysical cluster models like meK-Means (Tara Chari, Gorin, and Lior Pachter, 2024a), barring differences in noise models and hard/soft assignment. This does not mean that our trajectory model should be fit on cluster data, because the resulted process time is no longer meaningful. Instead, these connections allowed us to better compare the models, and determine which model was more appropriate (see Section 4.5 “Gene selection”).

Inference

Chronocell overview

Input The input of Chronocell are 1) trajectory structure, 2) sampling assumption, and 3) scRNA-seq count matrix. Trajectory structure is provided to Chronocell as a 2D array, with each lineage (path) represented as a row. Along with the structure, an initial guess of switching time is also needed as a starting point in the fitting. The sampling assumption refers to the prior distribution of the latent variables (process time and lineages) for each cell. This is represented as a 3D array with shape (n, l, m), where n is the number of cells, l is the number of lineages, and m is the number of time grids.

Model Building upon the common transcription model, we have two classes of models based on the assumption of global switch time: (1) the synchronized model, which assumes a completely synchronized switch in transcription rates across all genes; and (2) the desynchronized model, where each gene has its own switching time. The desynchronized model is more challenging to fit from scratch, so we recommend using a warm start based on the results of the synchronized model.

Inference We use the expectation–maximization algorithm to fit the trajectory model on the scRNA-seq count matrix.

Output The primary output of Chronocell consists of the parameters and posterior distribution for each cell. Other relevant information such as the Akaike Information Criterion (AIC) and the Fisher information matrix can also be calculated.

Maximum likelihood estimates of parameters by EM algorithm

We use the expectation–maximization algorithm to estimate model parameters $\hat{\theta}$. For simplicity, we discretize the latent variable t with finite regular grid points over the interval: $t = t_1, ..., t_M$, which basically approximates a continuous measure with a discrete measure. Then, the latent variable z=(l,t) describing lineage and time is discrete, and we write $\sum_{z_i}$ to denote the summation over all L lineages and M time grid points, i.e., $\sum_z = \sum_{l=1}^{L} \sum_{m=1}^{M}$. With this, the objective function becomes

$$\hat{\theta} = \arg\max_{\theta} \log p(x|\theta),$$
$$\log p(x|\theta) = \sum_{i=1}^{n} \log = \sum_{i=1}^{n} \log \int_{z_i} p(x_i, z_i|\theta) dz_i \approx \sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i|\theta),$$

where i is the cell index and $p(x_i, z_i|\theta)$ denotes the probability of observing x_i with the latent variable being z_i for cell i.

As log function is concave, we can use Jensen's inequality:

$$\begin{aligned}\sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i|\theta) &= \sum_{i=1}^{n} \log \sum_{z_i} p(z_i|x_i, \theta) \frac{p(x_i, z_i|\theta)}{p(z_i|x_i, \theta)}, \\ &\geq \sum_{i=1}^{n} \sum_{z_i} p_i(z_i|x_i, \theta) \log \frac{p(x_i, z_i|\theta)}{p_i(z_i|x_i, \theta)}.\end{aligned}$$

Since $\frac{p(x_i,z_i|\theta)}{p(z_i|x_i,\theta)} = p(x_i|\theta)$ is a constant for all z_i, the equality holds and

$$\sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i|\theta) = \sum_{i=1}^{n} \sum_{z_i} p_i(z_i|x_i, \theta) \log \frac{p(x_i|z_i, \theta)p(z_i)}{p_i(z_i|x_i, \theta)}.$$

Our trajectory model makes it possible to write out $p(x_i|z_i, \theta)$ explicitly:

$$\log p(x_i|z_i, \theta) = \sum_{j=1}^{p} \log p(x_{ij}|z_i, \theta_j),$$
$$\log p(x_{ij}|z_i, \theta_j) = \sum_{c=1}^{2} \left[\left(x_{ijc} \log \left(r_i \lambda_c(l, t_m, \theta_j) \right) - r_i \lambda_c(l, t_m, \theta_j) - \log(x_{ijc}!) \right) \right].$$

Therefore, we could use the expectation–maximization algorithm efficiently. Specifically, in the E-step of EM algorithm, we calculated the posterior distribution $p_i(z_i|x_i, \theta)$ based on $p(x_i|z_i, \theta)$,

$$\begin{aligned}p(z_i|x_i, \theta) &= \frac{p(z_i)p(x_i|z_i, \theta)}{\sum_{z_i} p(z_i)p(x_i|z_i, \theta)} \\ &= \frac{p(z_i) \exp\left(\log p(x_i|z_i, \theta)\right)}{\sum_{z_i} p(z_i) \exp\left(\log p(x_i|z_i, \theta)\right)}.\end{aligned}$$

In the M-step, based on fixed $p_i(z_i|x_i, \theta)$ and analytical form of $p(x_i|z_i, \theta)$, we optimized θ_j for each gene j separately by maximizing

$$F_j = \sum_{i=1}^{n} \sum_{z_i} p_i(z_i|x_i, \theta) \log p(x_{ij}|z_i, \theta_j),$$

Both the value and the gradient of F_j can be written out analytically and computed efficiently. Therefore, we could use off-the-shelf quasi-Newton methods for optimizing F_j with respect to θ_j, e.g., 'L-BFGS-B' method in minimize function provided by Scipy (Virtanen et al., 2020; Zhu et al., 1997).

In each step of EM algorithm, we alternated between the expectation and maximization steps. The implementation is based on defining the function that calculates $p(x_{ij}|z_i, \theta)$, so it would be easy to modify $p(x_{ij}|z_i, \theta)$ for different models in the future.

A warm start incorporating prior knowledge about data can help algorithm converge to the optimal θ^* quickly. If neither initial parameters nor $p(z|x)$ is given, we use random initializations. Multiple runs with different starting points are used to avoid local minima. By default, we tried 100 different random initializations, and run 100 steps for each initialization both for random initialization and warm start.

Poisson mixtures model

For fitting clusters, we used Poisson mixture models. Suppose there are S mixtures with transcription rates α_{js}, $s = 1, ..., S$ for each gene j, and each mixture is at steady state. Thus, the parameters of one gene are $a_j = \frac{\alpha}{\beta}$ and $\rho = \frac{\beta_j}{\gamma_j}$, since only the ratio is identifiable.

$$\begin{aligned} p(x_i|\theta) &= \sum_s p(x_i|s, a_{js}, \rho) \\ &= \sum_s \Pi_j P_{\text{Poiss}}(X = x_{ij1}; \lambda = r_i a_s) P_{\text{Poiss}}(X = x_{ij2}; \lambda = r_i a_{js}\rho). \end{aligned}$$

Similarly to the Gaussian mixture model, we could use the EM algorithm to infer parameters (including mixture weights) and posteriors.

Analysis

Fisher information

The Fisher information matrix (FIM) is defined to be

$$\mathcal{I}(\theta_0) \equiv \mathrm{E}\left[\left[\frac{\partial}{\partial\theta}\log(p(x|\theta))\right]\left[\frac{\partial}{\partial\theta}\log(p(x|\theta))\right]^{\mathsf{T}}\right]_{|\theta=\theta_0}.$$

Since

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log(p(x|\theta)) &= \frac{\partial}{\partial\theta}\log(\int p(x,z|\theta)dz) \\
&\approx \frac{\partial}{\partial\theta}\log(\sum_z p(x,z|\theta)) \\
&= \frac{1}{\sum_z p(x,z|\theta)}\sum_z\left(\frac{\partial p(x,z|\theta)}{\partial\theta}\right) \\
&= \frac{1}{\sum_z p(x,z|\theta)}\sum_z\left(p(z)\frac{\partial e^{\log p(x|z,\theta)}}{\partial\theta}\right) \\
&= \frac{1}{\sum_z p(x,z|\theta)}\sum_z\left(p(z)e^{\log p(x|z,\theta)}\frac{\partial\log p(x|z,\theta)}{\partial\theta}\right) \\
&= \sum_z\left(\frac{p(x,z|\theta)}{\sum_z p(x,z|\theta)}\frac{\partial\log p(x|z,\theta)}{\partial\theta}\right) \\
&= \sum_z\left(p(z|x,\theta)\frac{\partial\log p(x|z,\theta)}{\partial\theta}\right),
\end{aligned}$$

we could calculate FIM numerically with the explicit form of the derivative and the posterior distribution.

Gene selection

We did not use absolute likelihoods to select dynamical genes, because different genes have different scales and are not directly comparable. Instead, we noticed that the cluster model serves as a natural reference point for comparison, as genes with no dynamics can be fitted equally well, if not better, by clusters. Therefore, we decided to use the relative likelihood as a criterion for gene selection.

For trajectory model,

$$\begin{aligned}\sum_{i=1}^{n}\log\sum_{z_i}p(x_i,z_i|\theta) &= \sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log\frac{p(x_i|z_i,\theta)p(z_i)}{p_i(z_i|x_i,\theta)}\\ &= \sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log\frac{p(z_i)}{p_i(z_i|x_i,\theta)}\\ &\quad+\sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log p(x_i|z_i,\theta)\\ &= \sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log\frac{p(z_i)}{p_i(z_i|x_i,\theta)}\\ &\quad+\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{z_i}p_i(z_i|x_i,\theta)\log p(x_{ij}|z_i,\theta_j).\end{aligned}$$

Similarly for clusters model,

$$\begin{aligned}\sum_{i=1}^{n}\log\sum_{s=1}^{S}p(x_i,s|\theta) &= \sum_{i=1}^{n}\sum_{s}p_i(s|x_i,\theta)\log\frac{p(s)}{p_i(s|x_i,\theta)}\\ &\quad+\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{s}p_i(s|x_i,\theta)\log p(x_{ij}|s,\theta_j).\end{aligned}$$

We defined the gene-wise likelihood of gene j for trajectory model to be

$$\frac{1}{p}\sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log\frac{p(z_i)}{p_i(z_i|x_i,\theta)}+\sum_{i=1}^{n}\sum_{z_i}p_i(z_i|x_i,\theta)\log p(x_{ij}|z_i,\theta_j),$$

and for clusters,

$$\frac{1}{p}\sum_{i=1}^{n}\sum_{s}p_i(s|x_i,\theta)\log\frac{p(s)}{p_i(s|x_i,\theta)}+\sum_{i=1}^{n}\sum_{s}p_i(s|x_i,\theta)\log p(x_{ij}|s,\theta_j).$$

The first term is meant to represent the difference in likelihood caused by different flexibility of latent variables of two models.

For gene selection, we compared the gene likelihood of two models and selected genes whose gene likelihood of trajectory model was higher than those of the cluster model.

Uncertainty assessment

We used uncertainty/instability to falsify results. First, we evaluated the variation of different random initializations to assess the uncertainty of the inference. This was done by performing multiple (usually 100) random initializations. Specifically, we evaluated whether the process time estimation of initializations with high ELBO scores concentrated around the correct direction. To quantify this, we classified the output of each random initialization to be correct if the process time estimates had a correlation higher than 0.8 with the reference, and we used different ELBO score thresholds to compute the precision-recall curve, which were then summarized by average precision (AP). An ideal case with high ELBO scores concentrated around correlation one led to an AP close to one while multiple comparable maxima lead to low AP.

Another approach to assess uncertainty is through bootstrap resampling. The same inference procedure is applied to the resampled data, and the variation in the resultant process time serves as an indicator of instability. Specifically, we calculate the correlation of process time between the original and the resampled data and interpret instability as large variance of the correlation.

Simulations

We randomly generated parameters for 200 genes and sampled 2000 cells by default unless otherwise specified. Cells were sampled uniformly over process time and lineages. We then fitted the model with the correct trajectory structure and synchronized model if not stated otherwise.

For the simulation parameters, we assumed that the transcription parameters (α,β,γ) followed the log-normal distribution $lognormal(\mu, \sigma)$, where μ is the mean and σ is the standard deviation of the variable's natural logarithm. We attempted to derive realistic parameters for the distributions from the literature. Hence, we assumed $\beta \sim lognormal(2, 0.5), \gamma \sim lognormal(0.5, 0.5)$ so that unspliced to spliced ratio was around 0.2 and their distributions resembled those determined from metabolic labeling datasets (Rabani et al., 2011). The α were assumed to follow $lognormal(2, 1)$, so that the mean of spliced counts were similar to those in scRNA-seq data. We assumed the read depth follows Beta distribution $r \sim Beta(\mu = \frac{1}{4}, v = \frac{1}{64})$, where μ was the mean and v the variance. For simulations with Gamma noise, we multiplied the mean of Poisson distribution λ by a random variable following Gamma distribution with mean 1 and variance 0.5.

For simulations under the desynchronized model, gene-wise τ_k was sampled from a uniform distribution on $[T_k - \frac{\Delta\tau}{2}, T_k + \frac{\Delta\tau}{2}]$, where T_k corresponds to the global τ_k in the synchronized model, and $\Delta\tau$ was the smallest interval length ensuring that τ_k retains its order.

To vary the sampling distributions, we generated a Gaussian distribution with a random mean and standard deviation 0.05. We then sampled both from the Gaussian and the uniform distribution, and blended them together in different proportions. The percentages of time sampled from a Gaussian ranged from 0 to 1, increasing by 0.1 increments.

To characterize the time errors, we calculate root mean squared error (RMSE) for the posterior mean of process time in comparison to the true time. To calculate the errors of parameters, we divide the absolute error of α by the square root of true values and β, γ by true values, so that errors of different genes are more comparable. We refer to these as normalized errors throughout the text.

Real datasets preprocessing

To estimate the squared coefficient of variation of the read depth $\xi := \mathrm{E}\left[\frac{\mathrm{Cov}(X_a, X_b)}{\mathrm{E}[X]\,\mathrm{E}[Y]}\right]_{a,b}$, we calculated the covariance matrix of all genes with nonzero means, which was divided by the mean squared and averaged across gene pairs to calculate the mean normalized covariance as an estimate of ξ. We then selected Poissonian genes whose variances are close to baseline variance with reasonably large mean ($var < 1.2(\mu + \xi\mu^2), \mu > 0.01$). However, as some genes can be co-regulated and, therefore, correlated, we calculated the mean normalized covariance of Poissonian genes and repeated selected new Poissonian genes until the mean normalized covariance no longer changed. This typically occurred after two iterations. Finally, we normalized the sum of counts of the selected Poissonian genes by their mean to obtain the relative read depth estimates, which were then used as fixed parameters during the fitting process.

Based on simulations of factors of informativeness of the data, including numbers of cells and genes, the average mean of α parameters as well as β and γ ratio, we determined the procedure of filtering genes for fitting: we applied count mean thresholds (0.02 for unspliced and 0.1 for spliced), filtered out genes with a small unspliced to spliced ratio ($\frac{U}{S} < e^{-4}$), and finally, selected genes with a variance larger than 1.2 times the baseline variance, i.e., $var > 1.2(\mu + \xi\mu^2)$, where μ is the mean, var is the variance, and ξ is the read depth CV^2. For cell cycle data, we

also constrained genes to the Gene Ontology term "cell_cycle" (GO:0007049). We occasionally adjusted the variance threshold to 1.5 in order to end up with 50–200 genes. Then trajectory and Poisson mixture models were fitted on those genes.

Use of other methods

We have also used dyngen to generate simulation data (Cannoodt, Saelens, Deconinck, et al., 2021). We use a bifurcation backbone for generation of 10000 cells and 200 genes following its vignette.

For Monocle 3 and diffusion pseudotime, we always provide the correct root cells whose simulation times are 0 for simulations. For real dataset, we set to the cells from the cell type that is expected to be the progenitor. In Monocle 3, when the cells form disconnected clusters, only cells on the partition that includes the root cell has finite pseudotime. We only consider those cells and discard other cells with infinite pseudotime for comparison with simulation time. For slingshot, we manually set the true start cluster based on the simulation time for simulation and the cell type in real datasets. For veloVI, we use the mean latent time across genes for comparison with simulation time.

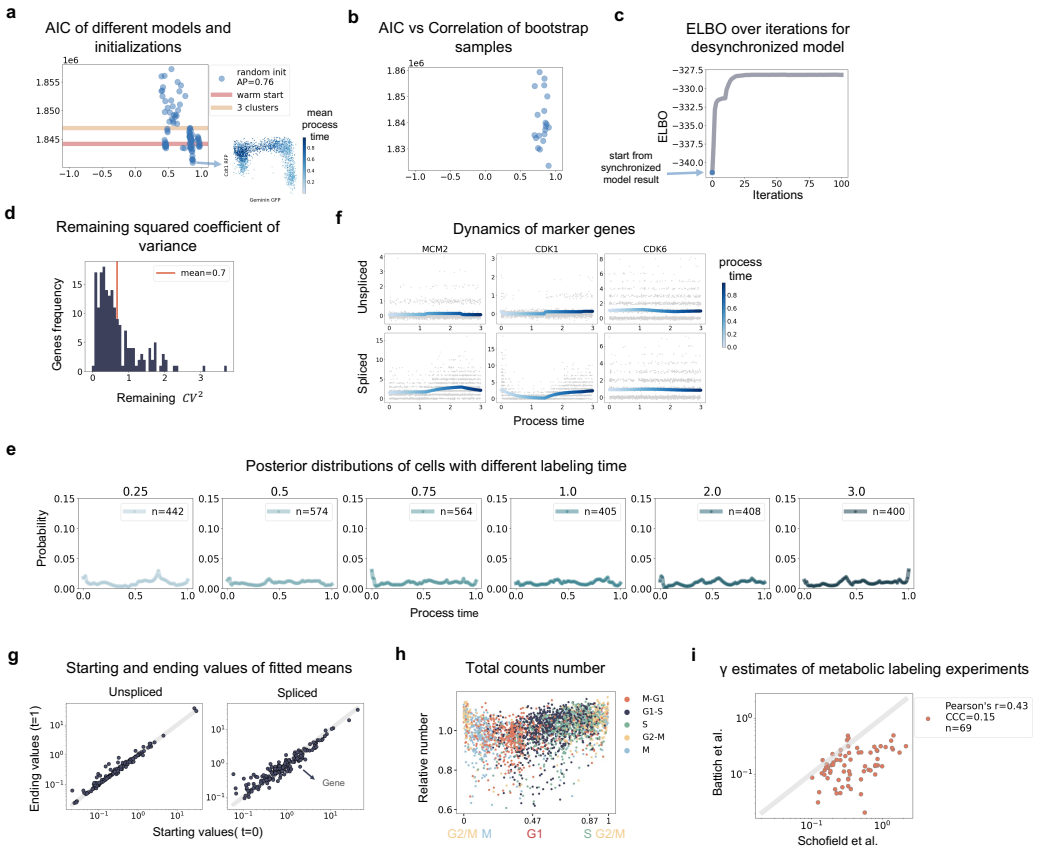

Figure 4.25: **Supplementary figures for Cell cycle data**. **a**) AIC scores and mean process time correlations of 100 random initializations (blue dots) compared to those of warm start (red line) as well as three clusters (Poisson mixtures) model (yellow line). AP stands for average precision. Mean process time of the initialization with lowest AIC is indicated in blue on the same PCA plot as in **a**). **b**) AIC scores and mean process time correlations of 100 bootstrap samples. The x-axis is the Pearson's correlation between the mean process time of each bootstrap and the those of original data, i.e., the plotted one in **a**. **c**) ELBO scores over iterations for desynchronized model. The fitting started with the best random initializations result of synchronized model. **d**) Distribution of remaining squared coefficient of variance of 182 genes used in the fitting. Remaining squared coefficient of variance is calculated by dividing the remaining unexplained variance by mean squared. **e**) Averaged posterior distribution across cells with different labeling times. n is the number of cells. **f**) Dynamics of three marker genes. The blue curve is the fitted mean of product Poisson distributions of unspliced and spliced counts over process time, and its darkness corresponds to the value of process time. Cells' raw counts (gray) are plotted against their corresponding process times. **g**) Starting and ending values of fitted mean of Poisson distributions. **h**) Total counts over process time of cells colored by cell type annotations. **i**) Comparison of γ estimates from two metabolic RNA labeling papers for 84 selected genes. Estimates of 67 genes are available in both papers. CCC stands for concordance correlation coefficient.

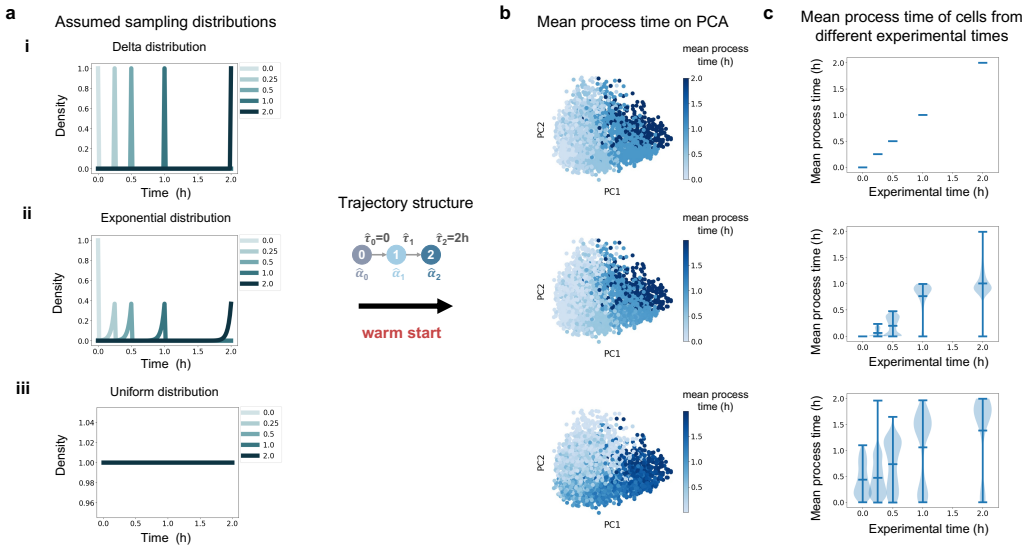

Figure 4.26: **Inference results of different sampling distribution assumption for Neuron data**. Fitting was warm started from delta distribution at physical time under different sampling distribution priors using the shown trajectory structure. **a**) The assumed sampling distribution. For **iii**, uniform distribution is assumed for cells from all time points. **b**) The fitted trajectory structure and inferred mean process time indicated in blue on the PCA plot. **c**) Violin plots of mean process time of cells with different labeling times. Three blue bars represent the mean and extremes.

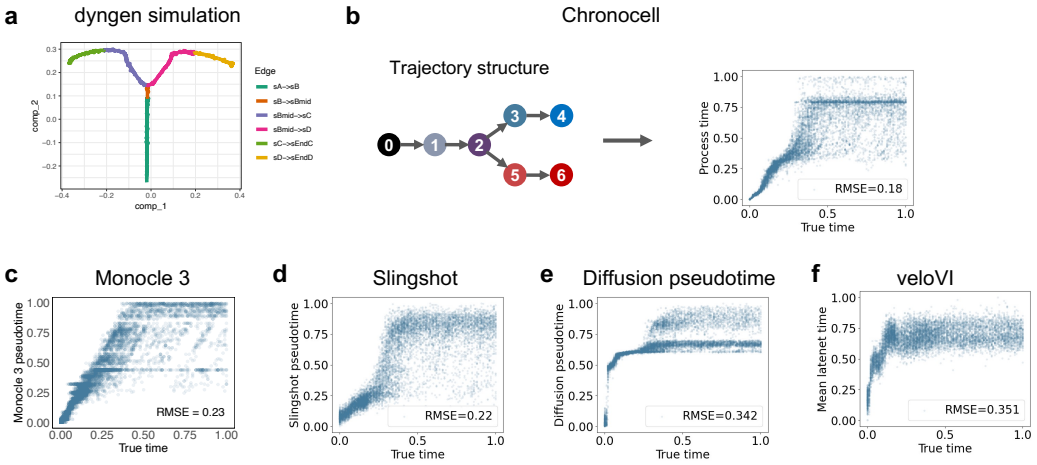

Figure 4.27: **Results of Chronocell compared to other methods on simulations generated by dyngen**. Chronocell, Monocle 3 (Cao et al., 2019), Slingshot (Street et al., 2018), diffusion pseudotime (Haghverdi et al., 2016) and veloVI (Du et al., 2024) are applied on simulation generated using dyngen (Cannoodt, Saelens, Deconinck, et al., 2021). Inferred time is plotted against true time, where x-axis is the true simulation time normalized between 0 and 1 and y-axis is corresponding inferred time normalized between 0 and 1. RMSE stands for root mean square error of inferred time. **a**) The dyngen simulation projected into the first two principal component spaces. A bifurcation backbone is used. **b**) The fit trajectory structure and results of Chronocell. **c**) The result of Monocle 3. **d**) The result of Slingshot. **e**) The result of diffusion pseudotime. **f**) The result of veloVI.

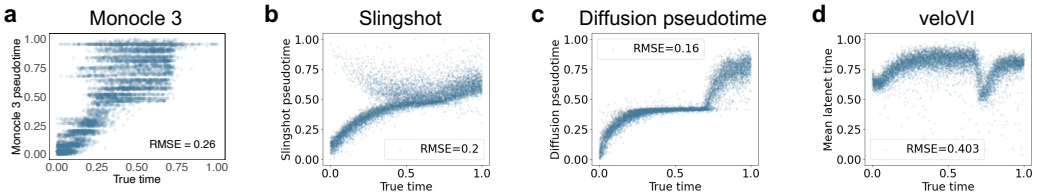

Figure 4.28: **Results of other methods on Figure 2 simulation**. Monocle 3 (Cao et al., 2019), Slingshot (Street et al., 2018), diffusion pseudotime (Haghverdi et al., 2016) and veloVI (Du et al., 2024) are applied on simulation data used in Figure 2. Inferred time is plotted against true time, where x-axis is the true simulation time and y-axis is corresponding inferred time normalized between 0 and 1. RMSE stands for root mean square error of inferred time.

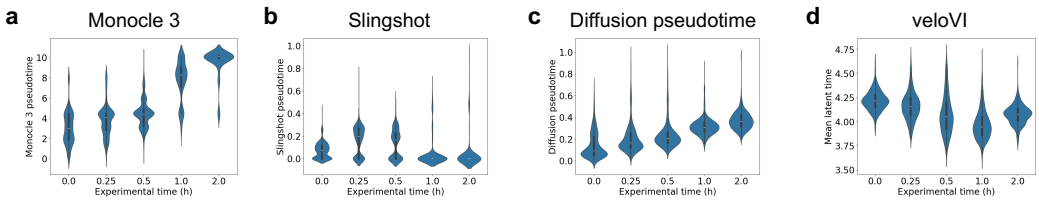

Figure 4.29: **Results of other methods on Neuron data**. Monocle 3 (Cao et al., 2019), Slingshot (Street et al., 2018), diffusion pseudotime (Haghverdi et al., 2016) and veloVI (Du et al., 2024) are applied on Neuron data used in Figure 4.26 to generate violin plots comparing inferred time to experimental time. In these plots, the x-axis represents the experimental time, while the y-axis shows the corresponding inferred time.

Chapter 5

FUTURE DIRECTIONS

Gorin, Gennady, John J Vastola, Meichen Fang, and Lior Pachter (2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nat. Commun.* 13.1, p. 7620. DOI: 10.1038/s41467-022-34857-7.

In this thesis, we advocate for a balanced coexistence of the two cultures in single-cell RNA sequencing analysis. The ideal approach is a balanced integration of data models and algorithmic models based on the questions at hand. For example, we can begin by exploring the data with algorithmic models, which provide a broad overview. Then, based on those exploratory results, we can build data models to closely examine the underlying biological mechanisms. With these mechanistic insights, we can develop more suitable and biologically informed algorithmic models. Such integration can enhance both the efficiency and depth of scRNA-seq data analysis.

In line with this perspective, we present two mechanistic models for normalization and trajectory inference. Our emphasis lies in the interpretability afforded by biophysically inspired models and the rigor of principled statistical inference, which together enable more meaningful insights into the underlying biological processes.

However, our models are admittedly naive and raise more questions than they answer. They are oversimplified and cannot account for all sources of variation. In the extrinsic noise model, a single random variable is insufficient even mature counts in pseudo-cells. In the process time model, the assumption of piecewise-constant transcription rates is unlikely to hold in realistic biological systems. We can keep listing the assumptions and simplification that we made for our models. After all, models are a compromise between tractability and the complexity of the real world. A mechanistic model must be simple enough to allow full inference from the available experimental data.

This constraint can be seen as both an advantage and a limitation, as developing mechanistic models depends more heavily on an iterative interplay between experimental data and theoretical development. The limitation is that model cannot be

improved until more informative experiments appear. For example, in the context of scRNA-seq normalization, more experiments need to be done to characterize technical noise in each scRNA-seq methods.

In an ideal scenario, scRNA-seq would achieve the same level of precision and reproducibility as physics experiments. For example, just as all free-fall experiments conducted under the same conditions yield virtually identical measurements of gravitational acceleration, each scRNA-seq experiment using the same protocol would produce data with consistent and predictable technical noise. This would allow biological variability to be interpreted with the same confidence physicists place in natural constants.

Currently, reproducibility in single-cell RNA-seq is often assessed by comparing mean expression levels across experiments. However, what truly matters is the full distribution of gene expression, not just the average. This is analogous to measuring gravity: while all objects may hit the ground, the defining feature of gravity is the consistent acceleration, not merely the fact that they fall.

The advantage of this constraint is that it helps guide rational experimental design. As we demonstrated in (Gorin, Vastola, Fang, et al., 2022), in a closed-loop framework for the rational design of transcriptomics experiments, mathematical analysis informs the design of experiments that are maximally informative for distinguishing between competing hypotheses, such as the CIR and Γ-OU models for transcription rates (Figure 5.1). Unfortunately, our work focused on the theoretical analysis and model fitting aspects of this framework, the experimental feedback loop remains to be completed. Outside of scRNA-seq, prior work in fluorescence-based transcriptomics has demonstrated the use of Fisher information criteria to optimize experimental design and detect environmental fluctuations from time-course data (Fox, Neuert, and Munsky, 2020). These studies highlight the feasibility and value of closing the loop between model development and experimental validation, which is particularly lacking in the field of scRNA-seq.

Even in the era of big data, not all data are equally informative. The value of a dataset depends not merely on its size, but on how well it is aligned with the specific scientific question being addressed. Therefore, mechanistic models remain valuable for enabling researchers to design experiments that gather the right data, not just more data, even if their interpretability and capacity for rigorous inference are set aside.

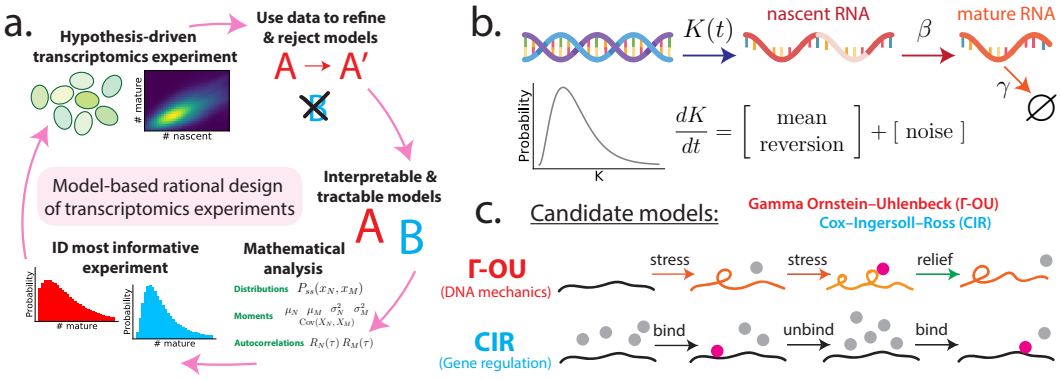

Figure 5.1: Framework for the rational design of transcriptomics experiments. **a)** Model-based closed loop paradigm. A researcher begins by representing two or more competing hypotheses as interpretable and tractable mathematical models (middle right of circle). Next, they perform a detailed mathematical analysis of each model, computing quantities (e.g., RNA count distributions and moments) that can help distinguish one hypothesis from another. Using the results of that analysis as input, they identify the experiment that best distinguishes the two models. Finally, they perform this experiment on some population of cells, use the resulting data to refine and/or reject models, and repeat the process with an updated ensemble of models. **b)** Interpretable and tractable modeling framework for transcription rate variation. We consider stochastic models of transcription involving (i) nascent/unspliced RNA, (ii) mature/spliced RNA, and (iii) a stochastic and time-varying transcription rate $K(t)$. The transcription rate is assumed to evolve in time according to a simple, one-dimensional SDE that includes a mean-reversion term (which tends to push $K(t)$ towards its mean value) and a noise term (which causes $K(t)$ to randomly fluctuate). Here, we have specifically chosen dynamics for which the long-time probability distribution of $K(t)$ is a gamma distribution (gray curve), because this assumption yields empirically plausible negative binomial-like RNA distributions. However, the framework does not require this in general. **c)** Two plausible models studied in this paper. The gamma Ornstein–Uhlenbeck (Γ-OU) model describes DNA m echanics, whereas the Cox–Ingersoll–Ross (CIR) model describes regulation by a high copy number regulator.

BIBLIOGRAPHY

10x Genomics (2021a). *10k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, 3' v3.1, Chromium Controller*. Dataset published on August 9, 2021. Licensed under CC BY 4.0. URL: https://www.10xgenomics.com/datasets/10-k-1-1-mixture-of-human-hek-293-t-and-mouse-nih-3-t-3-cells-3-v-3-1-chromium-controller-3-1-standard-6-1-0.

– (2021b). *10k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, 3' v3.1, Chromium X*. Dataset published on August 9, 2021. Licensed under CC BY 4.0. URL: https://www.10xgenomics.com/datasets/10-k-1-1-mixture-of-human-hek-293-t-and-mouse-nih-3-t-3-cells-3-v-3-1-chromium-x-3-1-standard-6-1-0.

– (2021c). *20k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, 3' HT v3.1*. Dataset published on August 9, 2021. Licensed under CC BY 4.0. URL: https://www.10xgenomics.com/datasets/20-k-1-1-mixture-of-human-hek-293-t-and-mouse-nih-3-t-3-cells-3-ht-v-3-1-3-1-high-6-1-0.

– (2022). *K562-r cells (Next GEM), Flex Gene Expression Dataset by Cell Ranger 7.0.0*. Accessed: 2025-05-11. URL: https://www.10xgenomics.com/datasets/10k-human-k562-r-cells-singleplex-sample-1-standard.

– (2023). *10k Mouse Forebrain FFPE Tissue Dissociated using gentleMACS Dissociator, Singleplex Sample (Next GEM)*. Accessed: 2025-05-11. URL: https://www.10xgenomics.com/datasets/10k-mouse-forebrain-ffpe-tissue-dissociated-using-gentlemacs-dissociator-singleplex-sample-1-standard.

– (2024). *10k Human PBMCs Stained with TotalSeq™-B Human Universal Cocktail, Singleplex Sample (Next GEM)*. Accessed: 2025-05-11. URL: https://www.10xgenomics.com/datasets/10k-human-pbmcs-stained-with-totalseq-b-human-universal-cocktail-singleplex-sample-1-standard.

Abu-Mostafa, Yaser S, Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). *Learning from data*. Vol. 4. AMLBook New York.

Ahlmann-Eltze, Constantin and Wolfgang Huber (2023). "Comparison of transformations for single-cell RNA-seq data". In: *Nat. Methods* 20, pp. 665–672. DOI: 10.1038/s41592-023-01814-1.

Aivazidis, Alexander et al. (2023). "Model-based inference of RNA velocity modules improves cell fate prediction". In: *bioRxiv*, p. 2023.08.03.551650. DOI: 10.1101/2023.08.03.551650.

Aniweh, Yaw et al. (2019). "SMIM1 at a glance; discovery, genetic basis, recent progress and perspectives". In: *Parasite Epidemiol Control* 5, e00101. DOI: 10.1016/j.parepi.2019.e00101.

Assaf, Michael and Baruch Meerson (2017). "WKB theory of large deviations in stochastic populations". In: *J. Phys. A: Math. Theor.* 50, p. 263001. DOI: 10.1088/1751-8121/aa669a.

Bacher, Rhonda et al. (2017). "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nat. Methods* 14, pp. 584–586. DOI: 10.1038/nmeth.4263.

Baras, F, M Malek Mansour, and J E Pearson (1996). "Microscopic simulation of chemical bistability in homogeneous systems". In: *J. Chem. Phys.* 105, pp. 8257–8261. DOI: 10.1063/1.472679.

Baron, Margaret H, Joan Isern, and Stuart T Fraser (2012). "The embryonic origins of erythropoiesis in mammals". In: *Blood* 119, pp. 4828–4837. DOI: 10.1182/blood-2012-01-153486.

Battich, Nico, Joep Beumer, et al. (2020). "Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies". In: *Science* 367, pp. 1151–1156. DOI: 10.1126/science.aax3072.

Battich, Nico, Thomas Stoeger, and Lucas Pelkmans (2015). "Control of Transcript Variability in Single Mammalian Cells". In: *Cell* 163, pp. 1596–1610. DOI: 10.1016/j.cell.2015.11.018.

Bender, Carl M and Steven A Orszag (2010). *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*. Springer. DOI: 10.1007/978-1-4757-3069-2.

Bokes, Pavol (2022). "Stationary and Time-Dependent Molecular Distributions in Slow-Fast Feedback Circuits". In: *SIAM J. Appl. Dyn. Syst.* 21, pp. 903–931. DOI: 10.1137/21M1404338.

Booeshaghi, A Sina, Ingileif B Hallgrímsdóttir, et al. (2022). "Depth normalization for single-cell genomics count data". In: *bioRxiv*, p. 2022.05. 06.490859. DOI: 10.1101/2022.05.06.490859.

Booeshaghi, A Sina and Lior Pachter (2021). "Normalization of single-cell RNA-seq counts by log (x+ 1) or log (1+ x)". In: *Bioinformatics* 37.15, pp. 2223–2224.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nat. Biotechnol.* 34, pp. 525–527. DOI: 10.1038/nbt.3519.

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *SSO Schweiz. Monatsschr. Zahnheilkd.* 16, pp. 199–231. DOI: 10.1214/ss/1009213726.

Brennecke, Philip et al. (2013). "Accounting for technical noise in single-cell RNA-seq experiments". In: *Nat. Methods* 10, pp. 1093–1095. DOI: 10.1038/nmeth.2645.

Cadima, Jorge and Ian T Jolliffe (1995). "Loading and correlations in the interpretation of principle compenents". In: *J. Appl. Stat.* 22, pp. 203–214. DOI: 10.1080/757584614.

Campbell, Kieran R and Christopher Yau (2016). "Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference". In: *PLoS Comput. Biol.* 12, e1005212. DOI: 10.1371/journal.pcbi.1005212.

– (2019). "A descriptive marker gene approach to single-cell pseudotime inference". In: *Bioinformatics* 35, pp. 28–35. DOI: 10.1093/bioinformatics/bty498.

Cannoodt, Robrecht, Wouter Saelens, Louise Deconinck, et al. (2021). "Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells". In: *Nat. Commun.* 12, p. 3942. DOI: 10.1038/s41467-021-24152-2.

Cannoodt, Robrecht, Wouter Saelens, and Yvan Saeys (2016). "Computational methods for trajectory inference from single-cell transcriptomics". In: *Eur. J. Immunol.* 46, pp. 2496–2506. DOI: 10.1002/eji.201646347.

Cao, Junyue et al. (2019). "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566, pp. 496–502. DOI: 10.1038/s41586-019-0969-x.

Chari, T and L Pachter (2021). "The specious art of single-cell genomics". In: *PLOS Computational Biology* 19. DOI: 10.1371/journal.pcbi.1011288.

Chari, Tara, Gennady Gorin, and Lior Pachter (2024a). "Biophysically interpretable inference of cell types from multimodal sequencing data". In: *Nat. Comput. Sci.* 4, pp. 677–689. DOI: 10.1038/s43588-024-00689-2.

– (2024b). "Stochastic modeling of biophysical responses to perturbation". In: *bioRxivorg*. DOI: 10.1101/2024.07.04.602131.

Chen, Yiqun T and Daniela M Witten (2022). "Selective inference for k-means clustering". In: *arXiv [stat.ME]*.

Dar, Roy D et al. (2012). "Transcriptional burst frequency and burst size are equally modulated across the human genome". In: *Proc. Natl. Acad. Sci. U. S. A.* 109, pp. 17454–17459. DOI: 10.1073/pnas.1213530109.

Deconinck, Louise et al. (2021). "Recent advances in trajectory inference from single-cell omics data". In: *Current Opinion in Systems Biology* 27, p. 100344. DOI: 10.1016/j.coisb.2021.05.005.

Drummond, P D and C W Gardiner (1980). "Generalised P-representations in quantum optics". In: *J. Phys. A Math. Gen.* 13, pp. 2353–2368. DOI: 10.1088/0305-4470/13/7/018.

Du, Jin-Hong et al. (2024). "Joint trajectory inference for single-cell genomics using deep learning with a mixture prior". In: *Proc. Natl. Acad. Sci. U. S. A.* 121, e2316256121. DOI: 10.1073/pnas.2316256121.

Elowitz, Michael B et al. (2002). "Stochastic gene expression in a single cell". In: *Science* 297, pp. 1183–1186. DOI: 10.1126/science.1070919.

Erhard, Florian et al. (2022). "Time-resolved single-cell RNA-seq using metabolic RNA labelling". In: *Nat. Rev. Methods Primers* 2, pp. 1–18. DOI: 10.1038/s43586-022-00157-z.

Feinberg, Martin (2019). *Foundations of chemical reaction network theory*. 1st ed. Springer Nature. DOI: 10.1007/978-3-030-03858-8.

Fox, Zachary R, Gregor Neuert, and Brian Munsky (2020). "Optimal design of single-cell experiments within temporally fluctuating environments". In: *Complexity* 2020, p. 8536365. DOI: 10.1155/2020/8536365.

Fu, Audrey Qiuyan and Lior Pachter (2016). "Estimating intrinsic and extrinsic noise from single-cell gene expression measurements". In: *Statistical applications in genetics and molecular biology* 15.6, pp. 447–471.

Gao, Lucy L, Jacob Bien, and Daniela Witten (2020). "Selective Inference for Hierarchical Clustering". In: *arXiv [stat.ME]*.

Gardiner, Crispin (2009). *Stochastic Methods*. Springer.

Gillespie, Daniel T (2000). "The chemical Langevin equation". In: *J. Chem. Phys.* 113, pp. 297–306. DOI: 10.1063/1.481811.

Golding, Ido et al. (2005). "Real-time kinetics of gene activity in individual bacteria". In: *Cell* 123, pp. 1025–1036. DOI: 10.1016/j.cell.2005.09.031.

Gorin, Gennady, Meichen Fang, et al. (2022). "RNA velocity unraveled". In: *PLoS Comput. Biol.* 18, e1010492. DOI: 10.1371/journal.pcbi.1010492.

Gorin, Gennady and Lior Pachter (2022a). "Modeling bursty transcription and splicing with the chemical master equation". In: *Biophys. J.* 121, pp. 1056–1069. DOI: 10.1016/j.bpj.2022.02.004.

– (2022b). "Monod: mechanistic analysis of single-cell RNA sequencing count data". In: *bioRxiv*, p. 2022.06.11.495771. DOI: 10.1101/2022.06.11.495771.

– (2023). "Length biases in single-cell RNA sequencing of pre-mRNA". In: *Biophys Rep (N Y)* 3, p. 100097. DOI: 10.1016/j.bpr.2022.100097.

– (2024). "New and notable: Revisiting the "two cultures" through extrinsic noise". In: *Biophys. J.* 123, pp. 1–3. DOI: 10.1016/j.bpj.2023.11.3400.

Gorin, Gennady, John J Vastola, Meichen Fang, et al. (2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". In: *Nat. Commun.* 13, p. 7620. DOI: 10.1038/s41467-022-34857-7.

Gorin, Gennady, John J Vastola, and Lior Pachter (2023). "Studying stochastic systems biology of the cell with single-cell genomics data". In: *Cell Syst.* 14, 822–843.e22. DOI: 10.1016/j.cels.2023.08.004.

Graham, R and T Tél (1985). "Weak-noise limit of Fokker-Planck models and nondifferentiable potentials for dissipative dynamical systems". In: *Phys. Rev. A Gen. Phys.* 31, pp. 1109–1122. DOI: 10.1103/physreva.31.1109.

Griffiths, Jonathan A, Antonio Scialdone, and John C Marioni (2018). "Using single-cell genomics to understand developmental processes and cell fate decisions". In: *Mol. Syst. Biol.* 14, e8046. DOI: 10.15252/msb.20178046.

Grima, Ramon and Pierre-Marie Esmenjaud (2024). "Quantifying and correcting bias in transcriptional parameter inference from single-cell data". In: *Biophys. J.* 123, pp. 4–30. DOI: 10.1016/j.bpj.2023.10.021.

Grün, Dominic, Lennart Kester, and Alexander van Oudenaarden (2014). "Validation of noise models for single-cell transcriptomics". In: *Nat. Methods* 11, pp. 637–640. DOI: 10.1038/nmeth.2930.

Grunberg, Theodore W and Domitilla Del Vecchio (2023). "A Stein's Method approach to the Linear Noise Approximation for stationary distributions of Chemical Reaction Networks". In: *arXiv [q-bio.QM]*.

Gu, Yichen, David Blaauw, and Joshua D Welch (2022). "Bayesian Inference of RNA Velocity from Multi-Lineage Single-Cell Data". In: *bioRxiv*, p. 2022.07.08.499381. DOI: 10.1101/2022.07.08.499381.

Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biol.* 20, p. 296. DOI: 10.1186/s13059-019-1874-1.

Haghverdi, Laleh et al. (2016). "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nat. Methods* 13, pp. 845–848. DOI: 10.1038/nmeth.3971.

Ham, Lucy et al. (2020). "Exactly solvable models of stochastic gene expression". In: *J. Chem. Phys.* 152, p. 144106. DOI: 10.1063/1.5143540.

Hanggi, Peter et al. (1984). "Bistable systems: Master equation versus Fokker-Planck modeling". In: *Phys. Rev. A* 29, pp. 371–378. DOI: 10.1103/PhysRevA.29.371.

Hao, Yuhan et al. (2024). "Dictionary learning for integrative, multimodal and scalable single-cell analysis". In: *Nat. Biotechnol.* 42, pp. 293–304. DOI: 10.1038/s41587-023-01767-y.

Hashimshony, Tamar et al. (2012). "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification". In: *Cell Rep.* 2, pp. 666–673. DOI: 10.1016/j.celrep.2012.08.003.

Hilfinger, Andreas and Johan Paulsson (2011). "Separating intrinsic from extrinsic fluctuations in dynamic biological systems". In: *Proc. Natl. Acad. Sci. U. S. A.* 108, pp. 12167–12172. DOI: 10.1073/pnas.1018832108.

Jahnke, Tobias and Wilhelm Huisinga (2007). "Solving the chemical master equation for monomolecular reaction systems analytically". In: *J. Math. Biol.* 54, pp. 1–26. DOI: 10.1007/s00285-006-0034-x.

Ji, Zhicheng and Hongkai Ji (2016). "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis". In: *Nucleic Acids Res.* 44, e117. DOI: 10.1093/nar/gkw430.

Jiao, Feng et al. (2024). "What can we learn when fitting a simple telegraph model to a complex gene expression model?" In: *PLoS Comput. Biol.* 20, e1012118. DOI: 10.1371/journal.pcbi.1012118.

Johnson, Benjamin K et al. (2022). "Single-cell Total RNA Miniaturized sequencing (STORM-seq) reveals differentiation trajectories of primary human fallopian tube epithelium". In: *bioRxiv*, p. 2022.03.14.484332. DOI: 10.1101/2022.03.14.484332.

Kim, Jong Kyoung, Aleksandra A Kolodziejczyk, et al. (2015). "Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression". In: *Nat. Commun.* 6, p. 8687. DOI: 10.1038/ncomms9687.

Kim, Jong Kyoung and John C Marioni (2013). "Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data". In: *Genome Biol.* 14, R7. DOI: 10.1186/gb-2013-14-1-r7.

Klein, Allon M et al. (2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". In: *Cell* 161, pp. 1187–1201. DOI: 10.1016/j.cell.2015.04.044.

Ko, M S (1991). "A stochastic model for gene induction". In: *J. Theor. Biol.* 153, pp. 181–194. DOI: 10.1016/s0022-5193(05)80421-7.

– (1992). "Induction mechanism of a single gene molecule: stochastic or deterministic?" In: *Bioessays* 14, pp. 341–346. DOI: 10.1002/bies.950140510.

Ko, M S, H Nakauchi, and N Takahashi (1990). "The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates". In: *EMBO J.* 9, pp. 2835–2842. DOI: 10.1002/j.1460-2075.1990.tb07472.x.

Kriegeskorte, Nikolaus et al. (2009). "Circular analysis in systems neuroscience: the dangers of double dipping". In: *Nat. Neurosci.* 12, pp. 535–540. DOI: 10.1038/nn.2303.

Kuchibhotla, Arun K, John E Kolassa, and Todd A Kuffner (2022). "Post-selection Inference". In: *Annu. Rev. Stat. Appl.* DOI: 10.1146/annurev-statistics-100421-044639.

Kurtz, Thomas G (1972). "The relationship between stochastic and deterministic models for chemical reactions". In: *J. Chem. Phys.* 57, pp. 2976–2978. DOI: 10.1063/1.1678692.

– (1978). "Strong approximation theorems for density dependent Markov chains". In: *Stochastic Processes and their Applications* 6, pp. 223–240. DOI: 10.1016/0304-4149(78)90020-0.

La Manno, Gioele et al. (2018). "RNA velocity of single cells". In: *Nature* 560, pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

Lähnemann, David et al. (2020). "Eleven grand challenges in single-cell data science". In: *Genome Biol.* 21, p. 31. DOI: 10.1186/s13059-020-1926-6.

Larsson, Anton J M et al. (2019). "Genomic encoding of transcriptional burst kinetics". In: *Nature* 565, pp. 251–254. DOI: 10.1038/s41586-018-0836-1.

Lause, Jan, Philipp Berens, and Dmitry Kobak (2021). "Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data". In: *Genome Biol.* 22, p. 258. DOI: 10.1186/s13059-021-02451-7.

Lederer, Alex R et al. (2024). "Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations". In: *bioRxiv*, p. 2024.01.18.576093. DOI: 10.1101/2024.01.18.576093.

Li, Chen et al. (2022). "Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction". In: *Nat. Biotechnol.*, pp. 1–12. DOI: 10.1038/s41587-022-01476-y.

Lim, Hong Seo and Peng Qiu (2024). "Quantifying the clusterness and trajectoriness of single-cell RNA-seq data". In: *PLoS Comput. Biol.* 20, e1011866. DOI: 10.1371/journal.pcbi.1011866.

Lin, Chieh and Ziv Bar-Joseph (2019). "Continuous-state HMMs for modeling time-series single-cell RNA-Seq data". In: *Bioinformatics* 35, pp. 4707–4715. DOI: 10.1093/bioinformatics/btz296.

Luecken, Malte D and Fabian J Theis (2019). "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Mol. Syst. Biol.* 15, e8746. DOI: 10.15252/msb.20188746.

Lun, Aaron T L, Karsten Bach, and John C Marioni (2016). "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome Biol.* 17, p. 75. DOI: 10.1186/s13059-016-0947-7.

Macosko, Evan Z et al. (2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161, pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002.

Melsted, Páll et al. (2021). "Modular, efficient and constant-memory single-cell RNA-seq preprocessing". In: *Nat. Biotechnol.* 39, pp. 813–818. DOI: 10.1038/s41587-021-00870-2.

Miller, O and S McKnight (1979). "Post-replicative nonribosomal transcription units in D. melanogaster embryos". In: *Cell* 17, pp. 551–563. DOI: 10.1016/0092-8674(79)90263-0.

Mimitou, Eleni P et al. (2021). "Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells". In: *Nat. Biotechnol.* 39, pp. 1246–1258. DOI: 10.1038/s41587-021-00927-2.

Neufeld, Anna et al. (2023). "Inference after latent variable estimation for single-cell RNA sequencing data". In: *Biostatistics* 25, pp. 270–287. DOI: 10.1093/biostatistics/kxac047.

Öcal, Kaan (2023). "Incorporating extrinsic noise into mechanistic modelling of single-cell transcriptomics". In: *bioRxiv*, p. 2023.09.30.560282. DOI: 10.1101/2023.09.30.560282.

Peccoud, J and B Ycart (1995). "Markovian Modeling of Gene-Product Synthesis". In: *Theor. Popul. Biol.* 48, pp. 222–234. DOI: 10.1006/tpbi.1995.1027.

Phillips, Rob (2015). "Theory in biology: Figure 1 or figure 7?" In: *Trends Cell Biol.* 25, pp. 723–729. DOI: 10.1016/j.tcb.2015.10.007.

Pijuan-Sala, Blanca et al. (2019). "A single-cell molecular map of mouse gastrulation and early organogenesis". In: *Nature* 566, pp. 490–495. DOI: 10.1038/s41586-019-0933-9.

Qiu, Xiaojie et al. (2017). "Reversed graph embedding resolves complex single-cell trajectories". In: *Nat. Methods* 14, pp. 979–982. DOI: 10.1038/nmeth.4402.

Rabani, Michal et al. (2011). "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells". In: *Nat. Biotechnol.* 29, pp. 436–442. DOI: 10.1038/nbt.1861.

Raj, Arjun et al. (2006). "Stochastic mRNA synthesis in mammalian cells". In: *PLoS biology* 4, e309.

Ramsköld, Daniel, Gert-Jan Hendriks, et al. (2024). "Single-cell new RNA sequencing reveals principles of transcription at the resolution of individual bursts". In: *Nat. Cell Biol.* 26, pp. 1725–1733. DOI: 10.1038/s41556-024-01486-9.

Ramsköld, Daniel, Shujun Luo, et al. (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nat. Biotechnol.* 30, pp. 777–782. DOI: 10.1038/nbt.2282.

Riba, Andrea et al. (2022). "Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning". In: *Nat. Commun.* 13, p. 2865. DOI: 10.1038/s41467-022-30545-8.

Rich, Joseph M et al. (2024). "The impact of package selection and versioning on single-cell RNA-seq analysis". In: *bioRxivorg*, p. 2024.04.04.588111. DOI: 10.1101/2024.04.04.588111.

Rooij, Frank J A van et al. (2017). "Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis". In: *Am. J. Hum. Genet.* 100, pp. 51–63. DOI: 10.1016/j.ajhg.2016.11.016.

Saelens, Wouter et al. (2019). "A comparison of single-cell trajectory inference methods". In: *Nat. Biotechnol.* 37, pp. 547–554. DOI: 10.1038/s41587-019-0071-9.

Sarkar, Abhishek and Matthew Stephens (2021). "Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis". In: *Nat. Genet.* 53, pp. 770–777. DOI: 10.1038/s41588-021-00873-4.

Schnoerr, David, Guido Sanguinetti, and Ramon Grima (2017). "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review". In: *J. Phys. A Math. Theor.* 50, p. 093001. DOI: 10.1088/1751-8121/aa54d9.

Schofield, Jeremy A et al. (2018). "TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding". In: *Nat. Methods* 15, pp. 221–225. DOI: 10.1038/nmeth.4582.

Setty, Manu et al. (2019). "Characterization of cell fate probabilities in single-cell data with Palantir". In: *Nat. Biotechnol.* 37, pp. 451–460. DOI: 10.1038/s41587-019-0068-4.

Shahrezaei, Vahid and Peter S Swain (2008). "Analytical distributions for stochastic gene expression". In: *Proc. Natl. Acad. Sci. U. S. A.* 105, pp. 17256–17261. DOI: 10.1073/pnas.0803850105.

Singh, Abhyudai and Pavol Bokes (2012). "Consequences of mRNA transport on stochastic variability in protein levels". In: *Biophys. J.* 103, pp. 1087–1096. DOI: 10.1016/j.bpj.2012.07.015.

Srivastava, R et al. (2002). "Stochastic vs. deterministic modeling of intracellular viral kinetics". In: *J. Theor. Biol.* 218, pp. 309–321. DOI: 10.1006/jtbi.2002.3078.

Street, Kelly et al. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19, p. 477. DOI: 10.1186/s12864-018-4772-0.

Sullivan, Delaney K, Kristján Eldjárn Hjörleifsson, et al. (2025). "Accurate quantification of nascent and mature RNAs from single-cell and single-nucleus RNA-seq". In: *Nucleic Acids Res.* 53. DOI: 10.1093/nar/gkae1137.

Sullivan, Delaney K, Kyung Hoi Joseph Min, et al. (2025). "kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus RNA-seq". In: *Nat. Protoc.* 20, pp. 587–607. DOI: 10.1038/s41596-024-01057-0.

Suter, David M et al. (2011). "Mammalian genes are transcribed with widely different bursting kinetics". In: *Science* 332, pp. 472–474. DOI: 10.1126/science.1198817.

Svensson, Valentine (2020). "Droplet scRNA-seq is not zero-inflated". In: *Nat. Biotechnol.* 38, pp. 147–150. DOI: 10.1038/s41587-019-0379-5.

Swain, Peter S, Michael B Elowitz, and Eric D Siggia (2002). "Intrinsic and extrinsic contributions to stochasticity in gene expression". In: *Proc. Natl. Acad. Sci. U. S. A.* 99, pp. 12795–12800. DOI: 10.1073/pnas.162041399.

Taketani, S, T Furukawa, and K Furuyama (2001). "Expression of coproporphyrinogen oxidase and synthesis of hemoglobin in human erythroleukemia K562 cells". In: *Eur. J. Biochem.* 268, pp. 1705–1711.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nat. Methods* 6, pp. 377–382. DOI: 10.1038/nmeth.1315.

Tang, Wenhao et al. (2023). "Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics". In: *Bioinformatics* 39, btad395. DOI: 10.1093/bioinformatics/btad395.

Taniguchi, Yuichi et al. (2010). "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells". In: *Science* 329, pp. 533–538. DOI: 10.1126/science.1188308.

Taylor, Jonathan and Robert J Tibshirani (2015). "Statistical learning and selective inference". In: *Proc. Natl. Acad. Sci. U. S. A.* 112, pp. 7629–7634. DOI: 10.1073/pnas.1507583112.

Teicher, Henry (1961). "Identifiability of Mixtures". In: *aoms* 32, pp. 244–248. DOI: 10.1214/aoms/1177705155.

Tian, Luyi et al. (2019). "Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments". In: *Nat. Methods* 16, pp. 479–487. DOI: 10.1038/s41592-019-0425-8.

Trapnell, Cole et al. (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nat. Biotechnol.* 32, pp. 381–386. DOI: 10.1038/nbt.2859.

Tritschler, Sophie et al. (2019). "Concepts and limitations for learning developmental trajectories from single cell genomics". In: *Development* 146. DOI: 10.1242/dev.170506.

Van den Berge, Koen et al. (2020). "Trajectory-based differential expression analysis for single-cell sequencing data". In: *Nat. Commun.* 11, p. 1201. DOI: 10.1038/s41467-020-14766-3.

Van Kampen, N G (2007). *Stochastic processes in physics and chemistry*. 3rd ed. North-Holland. DOI: 10.1016/b978-0-444-52965-7.x5000-4.

Vellela, Melissa and Hong Qian (2007). "A quasistationary analysis of a stochastic chemical reaction: Keizer's paradox". In: *Bull. Math. Biol.* 69, pp. 1727–1746. DOI: 10.1007/s11538-006-9188-3.

– (2009). "Stochastic dynamics and non-equilibrium thermodynamics of a bistable chemical system: the Schlögl model revisited". In: *J. R. Soc. Interface* 6, pp. 925–940. DOI: 10.1098/rsif.2008.0476.

Virtanen, Pauli et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nat. Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, Jingshu et al. (2018). "Gene expression distribution deconvolution in single-cell RNA sequencing". In: *Proc. Natl. Acad. Sci. U. S. A.* 115, E6437–E6446. DOI: 10.1073/pnas.1721085115.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biol.* 19, p. 15. DOI: 10.1186/s13059-017-1382-0.

Wolf, F Alexander, Fiona K Hamey, et al. (2019). "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells". In: *Genome Biol.* 20, p. 59. DOI: 10.1186/s13059-019-1663-x.

Yip, Shun H et al. (2017). "Linnorm: improved statistical analysis for single cell RNA-seq expression data". In: *Nucleic Acids Res.* 45, e179. DOI: 10.1093/nar/gkx828.

Zhang, Jesse M, Govinda M Kamath, and David N Tse (2019). "Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq". In: *Cell Syst* 9, 383–392.e6. DOI: 10.1016/j.cels.2019.07.012.

Zheng, Grace X Y et al. (2017). "Massively parallel digital transcriptional profiling of single cells". In: *Nat. Commun.* 8, p. 14049. DOI: 10.1038/ncomms14049.

Zhou, Sheng et al. (2005). "Increased expression of the Abcg2 transporter during erythroid maturation plays a role in decreasing cellular protoporphyrin IX levels". In: *Blood* 105, pp. 2571–2576. DOI: 10.1182/blood-2004-04-1566.

Zhu, Ciyou et al. (1997). "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". In: *ACM Trans. Math. Softw.* 23, pp. 550–560. DOI: 10.1145/279232.279236.