STORM: INCORPORATING TRANSIENT STOCHASTIC DYNAMICS TO INFER THE RNA VELOCITY WITH METABOLIC LABELING INFORMATION

A bioRχiv PREPRINT

Qiangwei Peng
LMAM and School of Mathematical Sciences
Peking University
Beijing 100871, China

Xiaojie Qiu✉
Whitehead Institute
Massachusetts Institute of Technology
Cambridge, USA
xqiu@wi.mit.edu

Tiejun Li✉
LMAM and School of Mathematical Sciences, Peking University
Center for Machine Learning Research, Peking University
Beijing 100871, China
tieli@pku.edu.cn

June 12, 2023

ABSTRACT

The time-resolved scRNA-seq (tscRNA-seq) provides the possibility to infer physically meaningful kinetic parameters, e.g., the transcription, splicing or RNA degradation rate constants with correct magnitudes, and RNA velocities by incorporating temporal information. Previous approaches utilizing the deterministic dynamics and steady-state assumption on gene expression states are insufficient to achieve favorable results for the data involving transient process. We present a dynamical approach, Storm (Stochastic models of RNA metabolic-labeling), to overcome these limitations by solving stochastic differential equations of gene expression dynamics. The derivation reveals that the new mRNA sequencing data obeys different types of cell-specific Poisson distributions when jointly considering both biological and cell-specific technical noise. Storm deals with measured counts data directly and extends the RNA velocity methodology based on metabolic labeling scRNA-seq data to transient stochastic systems. Furthermore, we relax the constant parameter assumption over genes/cells to obtain gene-cell-specific transcription/splicing rates and gene-specific degradation rates, thus revealing time-dependent and cell-state specific transcriptional regulations. Storm will facilitate the study of the statistical properties of tscRNA-seq data, eventually advancing our understanding of the dynamic transcription regulation during development and disease.

Keywords tscRNA-seq · Metabolic labeling enabled scRNA-seq · RNA velocity · Cell-specific Poisson model

Background

Cells are dynamic identities that are subject to intricate transcriptional and post-transcriptional regulations. Understanding the tight regulation of the RNA life cycle will shed light on not only the regulatory mechanism of RNA biogenesis, but also cell fate transitions. Based on the observation that most scRNA-seq approaches capture both premature unspliced mRNA and mature spliced mRNA information, La Manno et al. La Manno et al. (2018) pioneered the concept of RNA velocity or the time derivative of spliced RNA to reveal the local fate of each individual and designed a RNA kinetic parameter inference method called velocyto based on the steady state assumption. In a later work, scVelo Bergen et al. (2020) relaxed the steady-state assumption and proposed a dynamic RNA velocity model to infer gene-specific reaction rates of transcription, splicing and degradation as well as cell-specific hidden time using the expectation-maximization (EM) algorithm. Li et al. Li et al. (2021) derived a stochastic model of RNA

velocity based on the chemical master equation (CME) satisfied by the probabilistic mass function (PMF) rather than the deterministic ordinary differential equation (ODE) satisfied by the mean, and presented a mathematical analysis framework of RNA velocity. MultiVelo Li et al. (2022) extends the dynamic RNA velocity model by incorporating epigenome data that can be jointly measured with emerging multi-omics approaches. Protaccel Gorin et al. (2020) extends the concept of RNA velocity to protein. UniTVelo Gao et al. (2022) uses a top-down design for more flexible estimation of the RNA velocity, as opposed to the usual bottom-up strategy. DeepVelo Cui et al. (2022) uses graph convolutional neural networks to infer cell-specific parameters to extend RNA velocity to cell populations containing time-dependent dynamics and multiple lineages which were proven to be challenging in previous methods Bergen et al. (2021). Other deep learning-based approaches include VeloVI Gayoso et al. (2022), VeloVAE Gu et al. (2022), LatentVelo Farrell et al. (2022), cellDancer Li et al. (2023), and so on. However, due to the absence of physical time information, the above methods usually suffer the issue of scale invariance, that is, amplifying the parameters by an arbitrary amount will not change the solution if the time shrinks with the same amount, e.g., the exact physical time remains undetermined. This issue makes the inferred parameters and the RNA velocity have physical significance only up to a multiplicative constant Li et al. (2021). In addition, the missing time information enters the model as hidden variables, which makes the parameter inference difficult.

Technological innovations in scRNA-seq now enable us to directly measure the amount of newly synthesized mRNA molecules over a short period of time, either through chemically introduced mutations in the sequencing reads or direct biotin pull-down of RNA analogs such as 4sU metabolically labeled RNA molecules, which subtly introduces physical time information. These time-resolved metabolic labelingaugmented scRNA-seq (tscRNA-seq) include scSLAM-seq Erhard et al. (2019), scNT-seq Qiu et al. (2020), sci-fate Cao et al. (2020), NASC-seq Hendriks et al. (2019) and scEU-seq Battich et al. (2020). Qiu et al. Qiu et al. (2022) recently developed Dynamo to reconstruct analytical vector fields from discrete RNA velocity vectors by taking advantage of tscRNA-seq data to infer more robust and time-resolved RNA velocity, however, they only used the deterministic model and largely relied on the steady-state assumption.

To overcome the shortcomings of Dynamo and fully explore the potential of tscRNA-seq data, we present the Storm approach (Stochastic models of RNA metabolic-labeling) to improve the estimation of RNA kinetic parameters and the inference of the RNA velocity of the metabolic labeling scRNA-seq data by incorporating the transient stochastic dynamics of gene expressions. Importantly, we focus on modeling the kinetics/pulse metabolic labeling data as it follows the RNA synthesis across multiple short time periods and is thus ideal to capture temporal RNA kinetics. In order to properly model both biological noise and cell-specific technical noise (due to the variations in sequencing depth across individual cells and dropout resulting from imperfect RNA capture in scRNA-seq), we implemented in Storm three stochastic models of new mRNA (or new unspliced and spliced mRNA). Depending on the biological processes considered, Storm indicates that new mRNA sequencing data obeys different types of cell-specific Poisson (CSP) distributions. On this basis, Storm also includes hypothesis testing, parameter inference and goodness of fit evaluation methods for CSP-type distribution. In addition, we analyze the similarities and differences of the model considering RNA splicing or not. For one-shot data, we introduce the steady-state assumption to make the parameter inference possible. We verified the effectiveness of Storm in the cell cycle data set of kinetic experiments from the scEU-seq study Battich et al. (2020) and several one-shot datasets, including scSLAM-seq, scNT-seq and sci-fate. Storm is incorporated in Dynamo Qiu et al. (2022) of the Aristotle ecosystem that facilitates rich downstream analytical vector field modeling.

Results

Overall description of Storm

We established three stochastic gene expression models for new mRNA (or new unspliced and spliced mRNA) (**Fig. 1A**) for the inference of the RNA kinetic parameters and thus the RNA velocity in the Storm approach. In Model 1, only transcription and mRNA degradation were considered. In Model 2, we considered transcription, splicing, and spliced mRNA degradation. And in Model 3, we considered the switching of gene expression states, transcription in the active state, and mRNA degradation.

The complete workflow of Storm is demonstrated in **Fig. 1B**. We first analytically solve the new RNA (or new unspliced and spliced mRNA) stochastic dynamics corresponding to the above three models, which are Poisson distribution, independent Poisson distribution and zero-inflated Poisson distribution, respectively. In addition, we model the technical noise as the cell-specific binomial distribution. By integrating the biological noise and the technical noise together, we obtain the distribution for the measured number of new/labeled mRNA molecules (or new unspliced and spliced mRNA molecules), which are cell-specific Poisson distribution, independent cell-specific Poisson distribution

and cell-specific zero-inflated Poisson distribution, respectively. Maximum likelihood estimation (MLE) is used to fit the data and make inferences for the parameters shown in the corresponding models.

To ensure the general applicability of Storm in common nascent RNA labeling schemes, such as one-shot or kinetics/pulse experiments (See Figure 2 of Qiu, et. al Qiu et al. (2022)), we designed specific estimation strategies for each labeling scheme. For the one-shot labeling experiments, since there is only one labeling duration, the steady-state assumption under the stochastic dynamics framework is reinvoked to infer parameters. For kinetics/pulse-labeling experiments with multiple labeling durations, the transient stochastic dynamics framework is used without the steady-state assumption. Furthermore, the goodness-of-fit index $R_D^2 = 1 - D/D_0$ based on deviance commonly used in generalized linear models is utilized to quantify the goodness of fit of our models in kinetics/pulse datasets. The index is then used to select genes that are more consistent with model assumptions for later downstream analysis, such as the enrichment analysis of different gene-specific parameters. Furthermore, we relaxed the previous assumption of constant parameters in genes or cells and assumed that only degradation rates (γ_t in Models 1 and 3; γ_s in Model 2) are constant while the other parameters (α in three models; β in Model 2; p_{off} in Model 3) are cell specific and depend on the state of gene expression in each cell. This relaxation would be useful for modeling lineage-specific kinetics resulted from hierarchical lineage bifurcation, which is common in cell developments. Finally, in order to calculate and visualize the RNA velocity, we reduced the considered stochastic models to derive the deterministic equation for the mean gene expression. The inferred parameters, after filtering with the goodness-of-fit index are then used in RNA velocity analysis and visualization. Notably, to demonstrate Storm's performance, we conducted systematic comparison with the state-of-the-art method Dynamo Qiu et al. (2022) for processing metabolic labeling scRNA-seq experiment datasets.

In the continued subsections we will present the details of each step in the Storm workflow, starting from the introduction of our mathematical models.

CSP modeling of counts data with metabolic labeling information

We proposed and analytically solved three aforementioned stochastic gene expression models for the dynamics of new mRNAs (or new unspliced and spliced mRNAs).

For simplicity of modeling, we followed La Manno et al. (2018); Bergen et al. (2020) to assume that the genes are independent. In the stochastic gene expression model, the generation of new/labeled mRNA $\tilde{l}(t)$ (or new unspliced and spliced mRNA $(\tilde{u}_l(t), \tilde{s}_l(t))$) is a stochastic process, and we are interested in the evolution of its PMF over time,which is denoted by

$$\begin{aligned} \tilde{P}_n(t) &:= \text{Prob}\left(\tilde{l}(t) = n\right),\ n \in \mathbb{N} \\ \tilde{P}_{mn}(t) &:= \text{Prob}\left((\tilde{u}_l(t), \tilde{s}_l(t)) = (m, n)\right),\ (m, n) \in \mathbb{N}^2. \end{aligned} \tag{1}$$

In Model 1 and Model 2, since the initial value of $\tilde{l}(t)$ (or $(\tilde{u}_l(t), \tilde{s}_l(t))$) is 0, we obtained the following closed-form solution (see "Methods" section).

$$\begin{aligned} &\text{Model 1:} \quad \tilde{P}_n(t) = \frac{a^n(t)}{n!} e^{-a(t)}, \quad n \in \mathbb{N}, \\ &\text{Model 2:} \quad \tilde{P}_{mn}(t) = \frac{b^m(t) c^n(t)}{m! n!} e^{-b(t)-c(t)}, \quad (m, n) \in \mathbb{N}^2, \end{aligned} \tag{2}$$

where

$$\begin{aligned} a(t) &= \frac{\alpha}{\gamma_t}(1 - e^{-\gamma_t t}), \\ b(t) &= \frac{\alpha}{\beta}(1 - e^{-\beta t}), \\ c(t) &= \begin{cases} \frac{\alpha}{\gamma_s}(1 - e^{-\gamma_s t}) + \frac{\alpha}{\gamma_s - \beta}(e^{-\gamma_s t} - e^{-\beta t}), & \beta \neq \gamma_s, \\ \frac{\alpha}{\beta}(1 - e^{-\beta t}) - \alpha t e^{-\beta t}, & \beta = \gamma_s, \end{cases} \end{aligned} \tag{3}$$

which means that $\tilde{l}(t)$ obeys the Poisson distribution with mean $a(t)$ in Model 1, and $(\tilde{u}_l(t), \tilde{s}_l(t))$ obey independent Poisson distributions with mean $b(t)$ and $c(t)$ in Model 2. Here α, β are the transcription and splicing rates, and γ_s, γ_t are the spliced and total mRNA degradation rates, respectively.

In Model 3, following Chong et al. (2014), we assumed that switching rates k_{on} and k_{off} are significantly smaller than α and γ_t, which implies that the gene expression is either always on or always off during transcription/degradation

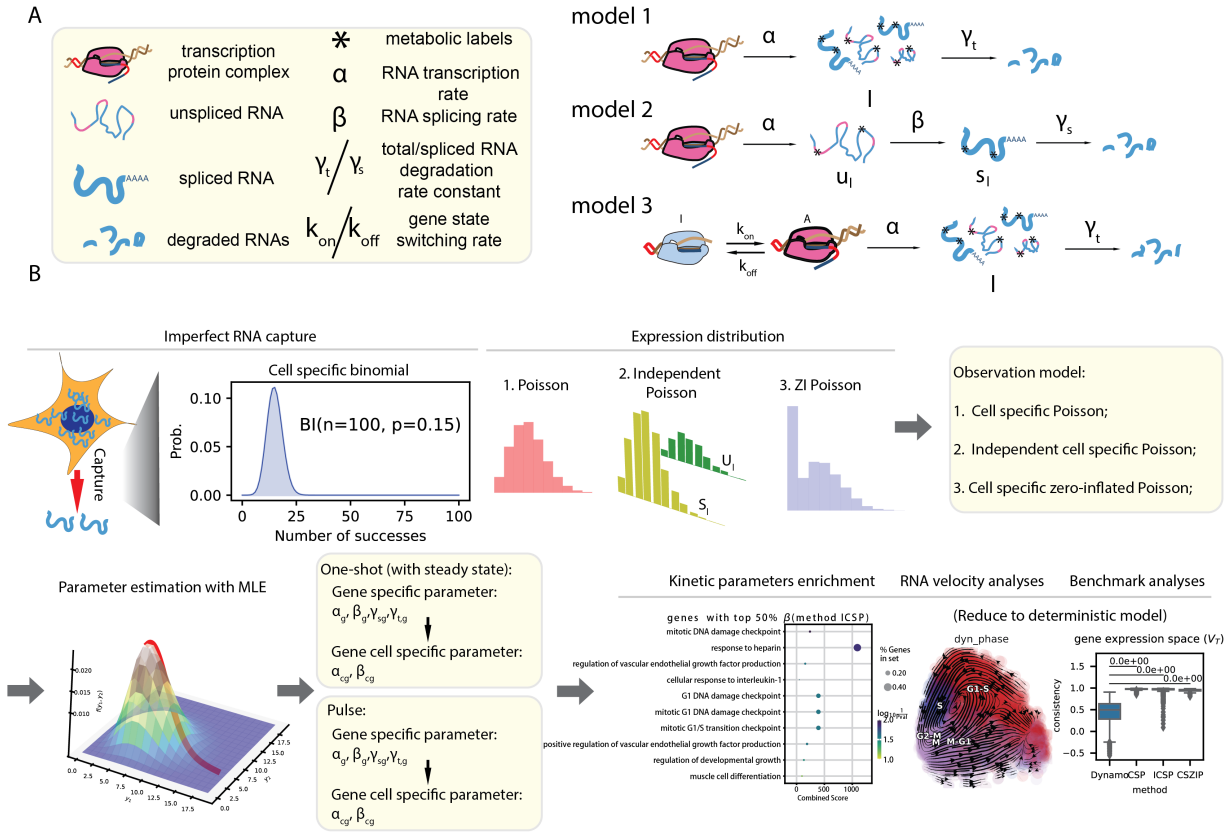

Figure 1: **Schematic overview of Storm. A.** Three models of RNA life cycle considering different biological processes: **Model 1**: Reaction dynamics model for new RNA $l(t)$ ignoring the splicing process, where α is the transcription rate and γ_t is the total mRNA degradation rate. **Model 2**: Reaction dynamics model of new unspliced and spliced mRNA $(u_l(t), s_l(t))$ considering the splicing process, where β is the splicing rate, γ_s is the spliced mRNA degradation rate, and α is the same as **Model 1**. Reaction dynamics model of new RNA $l(t)$ considering gene state switching, where α and γ_t are the same as in **Model 1**, k_{on} is the rate at which the gene switches from the inactive state to the active state, k_{off} is the opposite. **B.** Complete workflow diagram for parameter inference and downstream analysis based on stochastic dynamics of new mRNA considering technical noise.

period. Therefore, $\tilde{l}(t)$ obeys a zero-inflated Poisson (ZIP) distribution, then we have

Model 3:
$$\begin{aligned} \tilde{P}_0(t) &= (1 - p_{\text{off}})e^{-a(t)} + p_{\text{off}}, \\ \tilde{P}_n(t) &= (1 - p_{\text{off}})\frac{a(t)^n}{n!}e^{-a(t)}, n \geq 1, \end{aligned} \tag{4}$$

where $p_{\text{off}} = k_{\text{off}}/(k_{\text{on}} + k_{\text{off}})$ is the probability that gene expression is in the off state, i.e., the extra proportion of zeros in the ZIP distribution (see "Methods" section).

We also specifically modeled technical noise of the measured number of new RNA (or new unspliced and spliced mRNA) molecules in scRNA-seq experiments. Such noises often lead to dropout of RNA measurements during the sequencing process and generally result in variations in sequencing depth across cells. To account the noise, in Storm we modeled the dropout process of sequencing technology as cell-specific binomial distributions. Adopting the common practice in many preprocessing pipelines through a size factor to normalize the data La Manno et al. (2018); Bergen et al. (2020); Cui et al. (2022); Gayoso et al. (2022); Qiu et al. (2022), we assumed that the total numbers of mRNA molecules across all genes in different cells are close. Probabilistically, this assumption implies that

$$p_j \propto n_j,$$

where p_j is the probability of mRNA molecules being captured in cell j and n_j is the total number of mRNA molecules across all genes in cell j in scRNA-seq experiments.

Combining the stochastic models for biological and technical noise, we can obtain different formalism of the distribution for the measured number of new/labeled mRNA molecules $l(t)$ (or new unspliced and spliced mRNA molecules $(u_l(t), s_l(t))$) in the scRNA-seq experiments (see "Methods" section) for each model. Specifically, in Model 1, $l(t)$ obeys the cell-specific Poisson (CSP) distribution, that is,

$$P_{n,j}(t) = \frac{(p_j a(t))^n}{n!} e^{-p_j a(t)}, \tag{5}$$

where $P_{n,j}(t)$ is the PMF for the measured number of new mRNA molecules in cell j. In Model 2, $(u_l(t), s_l(t))$ obeys the independent cell-specific Poisson (ICSP) distribution, that is,

$$P_{mn,j}(t) = \frac{(p_j b(t))^m}{m!} e^{-p_j b(t)} \frac{(p_j c(t))^n}{n!} e^{-p_j c(t)}, \tag{6}$$

where $P_{mn,j}(t)$ is the joint PMF for the measure number of new unspliced and spliced mRNA molecules in cell j. In Model 3, $l(t)$ obeys the cell-specific zero-inflated Poisson (CSZIP) distribution, that is,

$$\begin{aligned} P_{0,j}(t) &= (1 - p_{\text{off}}) e^{-p_j a(t)} + p_{\text{off}}, \\ P_{n,j}(t) &= (1 - p_{\text{off}}) \frac{(p_j a(t))^n}{n!} e^{-p_j a(t)}, n \geq 1. \end{aligned} \tag{7}$$

We call the above distributions as *cell-specific* because different cells obey the distributions with different parameters.

Note that Grün et al. also modeled the scRNA-seq data by integrating biological noise and technical noise Grün et al. (2014). Our work is different from them in the following aspects: (1) Our work models the transient dynamics of new mRNA and solves their distribution for the proposed stochastic models analytically. However, in Grün et al. (2014), they instead modeled the total mRNA and derived that the biological noise follows a negative binomial distribution as the steady state of the transcriptional bursting model. (2) Our work accurately models the technical noise as a cell-specific binomial distribution, while they approximated the cell-sepcific binomial distribution with a Poisson distribution and modeled the capture probability as a random variable subject to the Gamma distribution, which finally leads to a negative binomial distribution (Poisson-Gamma mixture distribution) of the technical noise.

As one-shot labeling experiments are much more convenient than pulse experiments in practice, in the following, we will first demonstrate how Storm can be applied to the one-shot case. We will then extensively show Storm's power in analyzing the pulse datasets.

Stochastic models combined with steady-state assumptions for one-shot data

Since one-shot data has only one labeling duration, we designed the corresponding parameter inference method which invokes the steady-state assumption under the stochastic model, focusing specifically on Model 1 (see "Methods" section). Similar steady-state methods of the stochastic model can also be designed for both Model 2 and Model 3 as well, although they are not the focus of this paper.

We validated our method in several one-shot datasets (**Fig. 2**, S1). We first analyzed a dataset from the sci-fate study Cao et al. (2020) in which cell cycle progression and glucocorticoid receptor (GR) activation were explored. Similar to Dynamo, the RNA velocity flow from our method also revealed a sequential transition of cells following the DEX (dexamethasone) treatment times in the first two principal components (PCs) (**Fig. 2A Left**). In the second two PCs, we observed an orthogonal circular progression of the cell cycle (**Fig. 2A Middle**). From the first two UMAP dimensions projected further from the four PCs, we observed a combined dynamics of GR responses and cell cycle progression (**Fig. 2A Right**). Next, we analyzed the neuronal activity dataset from the scNT-seq study Qiu et al. (2020) to investigate cellular polarization dynamics after KCl treatment (**Fig. 2B**). Dynamo and Storm both revealed a coherent transition that nicely follows the temporal progression from time point 0 to 15, 30, 60 and finally 120 minutes. We analyzed the murine intestinal organoid system dataset from scEU-seq Battich et al. (2020). Dynamo observed a bifurcation (Fig. S1B, top row) from intestinal stem cells into the secretory lineage (left) and the enterocyte lineage (right), and Storm also observed similar results, although with some defects in the secretory lineage (Fig. S1B, bottom row). We also analyzed mouse fibroblast cells dataset from scSLAM-seq Erhard et al. (2019). We observed that both Dynamo and Storm inferred velocities further discriminated infected from non-infected cells (**Fig. S1C**).

To demonstrate the precision and robustness of the Storm method in estimating the one-shot dataset, we benchmarked the estimated kinetic parameters of different subsets of the cell cycle pulse-labeling dataset Battich et al. (2020), each with a different duration of labeling. On the 15-minute labeling sub-dataset, Storm recovers a transition that matches well with the cell-cycle progression, while the transition recovered by Dynamo is problematic near the M/M-G phase (**Fig. 2C Left**). On the 30-minute labeling sub-dataset, both methods recover the cell cycle progression correctly, but the streamlines of our method are considerably smoother compared to those of Dynamo (**Fig. 2C**

Table 1: The proposed sample-specific hypothesis test results on whether the number of new mRNA molecules in the Cell Cycle dataset obeys the CSP and CSZIP distributions. UTD means that it is unable to determine because there are too few groupings resulting in zero degrees of freedom, when it is always a perfect fit. The significance level is 0.05.

Labeling duration		15mins	30mins	45mins	60mins	120mins	180mins
CSP	Accept	0.116	0.067	0.049	0.062	0.064	0.065
	Reject	0.278	0.568	0.655	0.652	0.695	0.725
	UTD	0.606	0.365	0.296	0.286	0.241	0.210
CSZIP	Accept	**0.351**	**0.467**	**0.472**	**0.476**	**0.459**	**0.459**
	Reject	0.055	0.189	0.266	0.274	0.327	0.344
	UTD	0.594	0.344	0.262	0.250	0.214	0.197

Right). In addition, we compared the consistency of degradation rates γ_t inferred by the two methods between two sub-datasets with different labeling durations (**Fig. 2D**). The results show that our method is more consistent compared with Dynamo. Notably, although Storm shows higher consistency than Dynamo, it is still not satisfactory, perhaps due to the experimental noises from different labeling durations and the violation of the steady-state assumption. Therefore, it is crucial to integrate data of different durations of labeling when a kinetic experiment is available. Furthermore, it is equally important to design methods that do not rely on the steady-state assumption for parameter inference.

Finally, we quantitatively compared the degradation rates γ_t inferred by the two methods. The two methods are close on the other datasets (**Fig. S1A,D**) except on 15-minute labeling cell cycle sub-dataset where Dynamo is unreasonably large (**Fig. S1A**, **third column**). Thus, our method has similar or even better performance compared to Dynamo on the one-shot dataset.

Statistical analysis of cell cycle dataset based on Storm's stochastic model

Next we first performed a goodness-of-fit test of the stochastic model proposed in Storm to a cell cycle dataset from scEU-seq Battich et al. (2020) with mutliple labeling time points to validate our proposals.

When the fixed labeling duration is t_{fixed}, $a(t_{\text{fixed}})$, $b(t_{\text{fixed}})$ and $c(t_{\text{fixed}})$ are all fixed constants. We can test whether the number of new mRNA molecules in tscRNA-seq within a fixed labeling duration matches the distribution obtained based on the stochastic models (Eqs. (5), (6) and (7)), respectively. A common method of testing whether a dataset obeys a given distribution is the chi-square (χ^2) goodness-of-fit test Pearson (1900). However, the usual χ^2 test is not directly applicable because in our case different cells obey different distributions with different parameters. By inspecting the mathematical analysis procedure of the χ^2 test Benhamou and Melot (2018), we constructed a new asymptotic χ^2 statistics and proposed a modified χ^2 test for our cell-specific distributions (see "Methods" section).

We used the proposed cell-specific χ^2 test in the cell cycle dataset from the scEU-seq study Battich et al. (2020), in which cells were labeled for 15, 30, 45, 60, 120 or 180 minutes. Because the labeled unspliced mRNA counts $u_l(t)$ were too small to be grouped/binned to create a distribution, hypothesis tests were performed only for CSP and CSZIP distributions and not for ICSP. The results are shown in **Table 1**. We found that some genes were not well determined (especially for cases with a short duration of labeling) in the sense that these genes had too few new mRNA molecules in the tscRNA-seq experiments, which results in very few groupings and perfect fittings. With so few mRNA counts for these genes, we were unable to determine whether they obeyed our proposed distribution or not. Moreover, our results revealed that the CSZIP distribution exhibited a better fit with the data than the CSP distribution when focusing on a fixed time point alone, suggesting that the data are indeed zero-inflated.

We next showed the high goodness-of-fit of the CSP and CSZIP model on two characteristic genes, namely *RPL41* and *IL22RA1* with an overall low and high gene expression respectively (**Fig. 3A**). Qualitatively, we found that the expected counts of both the CSP and CSZIP models matched well with the observed counts for the gene *RPL41*. Quantitatively, the results of the cell-specific chi-square test also showed that the distribution of CSP or CSZIP was well satisfies in most labeling durations (**Fig. 3A**, **first row**). Similar results were observed for the gene *IL22RA1* with significantly higher expression (**Fig. 3A**, **second row**). Therefore, we demonstrated CSP and CSZIP distribution accurately describes these two genes and is thus suitable for modeling the tscRNA-seq datasets.

Finally, we found that, for most genes, the number of total mRNA molecules shares the same distribution across different labeling durations. In **Fig. 3B**, we showed the number of total mRNA molecules of four example genes *TSPOAP1*, *GPRC5A*, *ADAMTS6* and *APEX1* is nearly identical across different labeling durations. Quantitatively, we performed a global chi-square independence test on the number of total mRNAs (as distinct from the new mRNAs) with different durations of labeling in all genes and found that, interestingly, there are 72.3% of the genes passed the

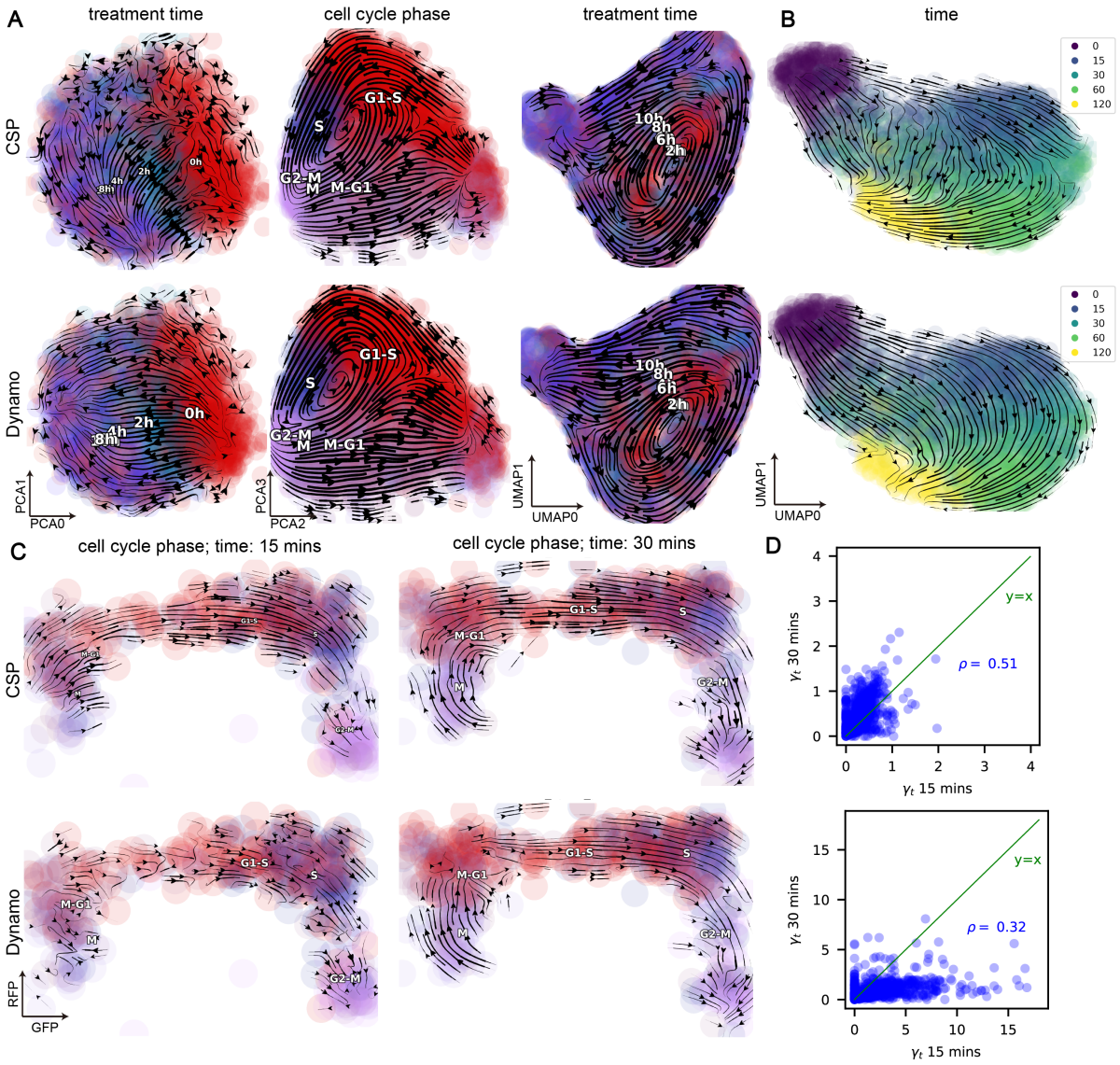

Figure 2: **Stochastic model combined with steady-state assumptions for one-shot experiments. A.** Streamline plots of the sci-fate dataset Cao et al. (2020) reveal two orthogonal processes of GR response and cell-cycle progression. From left to right: streamline plot on the first two PCs, the second two PCs, and the first two UMAP components that are reduced from the four PCs, respectively. The first row is the result of CSP and the second row is the result of Dynamo. The same applies to panels **B**, **C**, and **D**. **B.** Streamline projected in the UMAP space plots of neuronal activity under KCl polarization datasets from scNT seq Qiu et al. (2020). **C.** Streamline projected in the RFP_GFP space plots of cell cycle dataset from scEU-seq Battich et al. (2020). On the left is the result of taking only the data labelled with 15 minutes, and on the right is the data labelled with 30 minutes. **D.** Comparison of degradation rates γ_t in cell cycle datasets with labeling duration of 15 and 30 minutes.

test at a significance level of 0.05 (**Fig. 3C**). This indicates that a considerable proportion of the number of genes' total mRNA molecules obeyed the same distribution, consistent with what we observed for the four example genes.

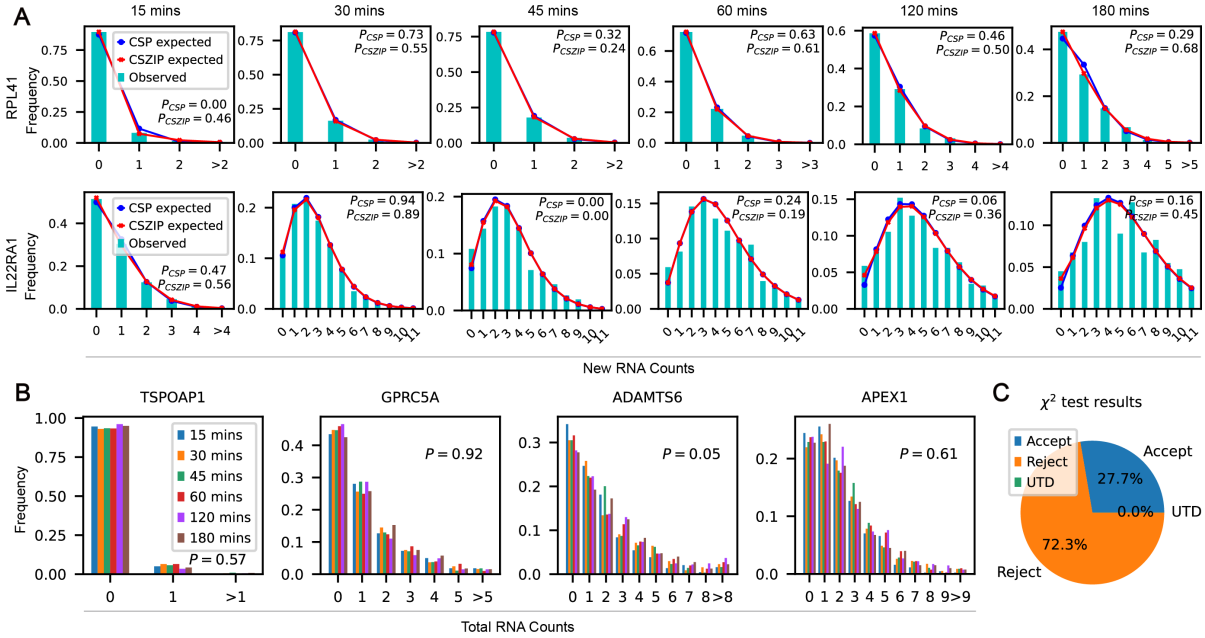

Figure 3: **Statistical analysis of cell cycle dataset. A.** Observed counts, expected counts of CSP, and expected counts of CSZIP of new mRNA molecules of the two example genes *RPL41* and *IL22RA1*. The first row: Fitting results of the *RPL41* gene with a small number of mRNA molecules; The second row: Fitting results of the *IL22RA1* gene with a higher number of new mRNA molecules (truncated to 11 for better visualization). **B.** Comparison of the total mRNA counts with different labeling durations of the four example genes *TSPOAP1*, *GPRC5A*, *ADAMTS6* and *APEX1*. **C.** Results of chi-square independence test for total RNA counts (significance level 0.05).

Storm accurately infers kinetic parameters that leads to rich insights of cell cycle via enrichment analysis

In the kinetic experiments, we relied on three stochastic models without the steady-state assumption to infer different set of kinetic parameters using maximum likelihood estimation (see "Methods section), namely α and γ_t for Model 1, α, β and γ_s for Model 2, and α, γ_t and p_{off} for Model 3. In addition, we defined the goodness-of-fit of each of the three models by utilizing the concept of deviance R^2 commonly used in generalized linear models Menard (2000) (see "Methods" section). According to the goodness-of-fit index, we selected genes that were more consistent with the model assumptions for downstream tasks, such as the enrichment analysis and RNA velocity analysis, etc.

Compared with Dynamo Qiu et al. (2022), the state-of-the-art method for processing tscRNA-seq datasets, our advantages are mainly in the following aspects: (1) Our method does not require steady-state assumptions on the kinetics experiments while Dynamo heavily relies on the steady-state assumptions; (2) Our stochastic model-based approach is more consistent with real biological process, while Dynamo only utilizes the deterministic model of mean value; (3) Our model takes into account all cells in the inference, while the approach based on steady-state assumptions in Dynamo only considers a small number of cells with high expression. In addition, we revealed the difference between the total mRNA degradation rate γ_t and spliced mRNA degradation rate γ_s based on their different physical roles, distinguished them in different models, and finally gave the relationship between these two (see "Methods" section). We noted that in Dynamo, to infer β, γ_t was first inferred when the splicing was ignored, then $\tilde{\gamma} := \gamma_s/\beta$ was inferred using the method based on the steady-state assumption in scVelo Bergen et al. (2020), and finally $\gamma_t/\tilde{\gamma}$ was taken as the inference of β upon assuming $\gamma_t = \gamma_s$. However, $\gamma_t/\tilde{\gamma} = \beta\gamma_t/\gamma_s$, while γ_t and γ_s are generally not equal. This point was overlooked in Dynamo, which causes a inaccurate estimate of β. In fact, under the steady-state assumption, β can be directly estimated by using only $u_l(t)$ through the formula $u_l(t) = (1 - e^{-\beta t})\alpha/\beta$, similar to the two-step method used in Dynamo to estimate γ_t through $l(t) = (1 - e^{-\gamma_t t})\alpha/\gamma_t$ since they have similar form. However, we don't use this method in Storm.

With the above inference methods and insights, we studied a cell cycle dataset from the scEU-seq study Battich et al. (2020). We compared the parameter inference results of the three models (**Fig. 4A**). When splicing was not

considered, the inference results based on CSP and CSZIP distributions were close, with high correlation coefficients, especially in genes with higher goodness of fit (**Fig. 4A Left**). However, whether or not splicing is considered significantly impacts the inference results. The inference results based on CSP and ICSP distribution were quite different, with low correlation coefficients, even in genes with higher goodness of fit (**Fig. 4A Middle**). We speculate that this is due to the assumptions of the two models are incompatible: in CSP, γ_t is assumed to be a constant; while in ICSP, γ_s is assumed to be a constant. But these two assumptions can not be held simultaneously for their different roles in the physical modeling and our analysis results (see "Methods" section). We also compared γ_t and γ_s computed by the ICSP model, and the results showed that γ_s was always greater than γ_t, and the linear correlation between the two was not high (**Fig. 4A Right**). In summary, we showed that kinetic parameters inferred from CSP and CSZIP but not CSP and ICSP, are consistent.

The inferred total mRNA degradation rates γ_t from Storm and Dynamo are close in well-fitted genes, while ICSP's inferred splicing rates β are always larger than from Dynamo. We compared the inferred results of γ_t in our CSP model with those in Dynamo (**Fig. 4B Left**). Although they were not consistent for some genes, they are quite consistent for the genes with better fitting. We also compared the inference of β in our ICSP model with those in Dynamo (**Fig. 4B Right**). The result shows that the inferred β by our approach was usually larger than those in Dynamo, even for the genes with a better fitting. A possible explanation is that the inference of Dynamo ignored the difference between γ_t and γ_s, which made the inferred β smaller. We also compared the goodness-of-fit of the three stochastic models. Overall, they are relatively close (**Fig. 4C Left**). However, when we focused on genes with higher new mRNA levels (top 10%), the ICSP model had a better fit (**Fig. 4C Right**). We speculate that this is because genes with higher expression are suitable to be fitted with more complex models.

When the parameter γ_t is small, parameter inference may not be robust enough. However, we found that the genes selected by the goodness-of-fit have robust results. We analyzed the robustness of the parameter inference in the simplest CSP model. When $\gamma_t t$ is small, $1 - e^{-\gamma_t t} \sim \gamma_t t$ holds, then

$$l(t) = \frac{\alpha(1 - e^{-\gamma_t t})}{\gamma_t} \approx \alpha t, \tag{8}$$

which implies that from the mean perspective the nonlinear fitting of α and γ_t degenerated into a linear fitting of α at this point. For a more precise analysis, let $\partial a(t)/\partial\alpha = (1 - e^{-\gamma_t t})/\gamma_t$, we have $\partial\ell(\alpha, \gamma_t)/\partial\alpha = 0$ is equivalent to

$$\alpha(\gamma_t) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{n_k} l_j(t_k)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k} p_j(t_k)\partial a(t_k)/\partial\alpha}. \tag{9}$$

But when $1 - e^{-\gamma_t t} \sim \gamma_t t$ holds, $\partial a(t)/\partial\alpha \approx t$, then we have

$$\alpha \approx \frac{\sum_{k=1}^{K}\sum_{j=1}^{n_k} l_j(t_k)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k} p_j(t_k)t_k} \tag{10}$$

is a constant, which we denoted by α_{cons}. We plotted the landscape of a typical negative log-likelihood loss function based on CSP model for gene *WWTR1* (**Fig. 4D Left**), with the black line corresponding to $\partial\ell(\alpha, \gamma_t)/\partial\alpha = 0$ (i.e. Eq. (9)) and blue line corresponding to $\alpha = \alpha_{\text{cons}}$ (i.e., Eq. (10)). The landscape of the loss function shows a fairly flat area around $\partial\ell/\partial\alpha = 0$, and the two lines almost coincide when γ_t is small, which is consistent with our previous argument. In addition, to quantitatively measure the robustness of inference on γ_t, since the optimal parameter is always located where the gradient is zero, we defined the l_1-norm of the derivative of the loss function with respect to γ_t restricted to $\partial\ell/\partial\alpha = 0$ (i.e. black line),

$$\left\| \frac{d\ell}{d\gamma_t}\Big|_{\frac{\partial\ell}{\partial\alpha}=0}(\gamma_t) \right\|_{l_1} = \int_0^{\gamma_{t,\max}} \left| \frac{d\ell}{d\gamma_t} \right| d\gamma_t, \tag{11}$$

as a measure of robustness. Since the half-life of the total mRNA molecules is usually not less than half an hour, we took $\gamma_{t,\max} = 1.5$. We analyzed the relationship between the robustness measure and the goodness-of-fit $\bar{R}_D^2$ (**Fig. 4D Right**). We found that parameter robustness was positively correlated with the goodness of fit and the correlation coefficient was as high as 0.69. Though the reason for this high correlation is not clearly understood in theory, we can utilize this fact to select the genes with high goodness of fit for downstream analysis, which also ensures the results are relatively robust.

We selected the well-fitted genes (top 40% $\bar{R}_D^2$) and performed enrichment analysis on this fraction according to the magnitude of gene-wise parameters γ_t, β, α and p_{off} (**Fig. 4E**, **Fig. S2**). The results of the enrichment analysis showed that these genes were highly correlated with the cell cycle progression.

The assumption of constant coefficients is often violated because of the time-dependent kinetics and multiple lineages Bergen et al. (2021). Many works relaxed the constant coefficient assumption and inferred cell-specific parameters to overcome this issue Cui et al. (2022); Qiu et al. (2022); Gayoso et al. (2022); Li et al. (2023) . In our proposal, we take a post-processing step to get the cell-specific parameters after inferring all parameters through previous procedures. We relaxed the constant coefficient assumption and proposed a method to infer cell-specific parameters except the constant degradation rate γ_t or γ_s, i.e., we inferred cell-specific α in Model 1, cell-specific $\alpha \times p_{\text{on}}$ in Model 3, and the cell-specific α and β in Model 2 (see "Methods" section). This partial constant coefficient assumption had support from the study in Battich et al. (2020), which showed that the degradation rate of most genes was independent of time. Finally, We plotted heat maps of the cell-wise α (based on CSP model), $\alpha \times p_{\text{on}}$ (based on CSZIP model) and β (based on ICSP model) for the well-fitted genes (**Fig. 4F**). The results show that cells in the same cell cycle phase usually have closer kinetic parameters.

Storm improves the robustness and accuracy of time-resolved RNA velocity analysis

Our three stochastic models described the evolution of the PMF (or joint PMF) of the number of new mRNA (or new unspliced and spliced mRNA) molecules over time for different settings. To estimate RNA velocity of single cells, only the evolution of the mean value over time will be considered, which requires us to reduce the stochastic models to the corresponding deterministic models (see "Methods" section).

Based on the deterministic model derived for the mean corresponding to the three stochastic models, we inferred the relevant parameters for computing different types of RNA velocity for different models. In Models 1 and 3, we computed the total RNA velocity $d\langle\tilde{r}(t)\rangle/dt$ because the splicing process was ignored. In Model 2, we calculated both total RNA velocity $d\langle\tilde{r}(t)\rangle/dt$ and spliced RNA velocity $d\langle\tilde{s}(t)\rangle/dt$ (see "Methods" section). Note that because the new RNA velocity mostly reflects the metabolic labeling process of RNA and does not reveal RNA biogenesis, it is thus not used. In addition, a derived relationship between γ_t and γ_s suggests that the total RNA velocity can be computed based on either $d\langle\tilde{r}(t)\rangle/dt = \alpha - \gamma_s\langle\tilde{s}(t)\rangle$ or $d\langle\tilde{r}(t)\rangle/dt = \alpha - \gamma_t\langle\tilde{r}(t)\rangle$. In practice, we used the former approach by default.

We compared the streamlines of the total RNA velocity of our three models with that of Dynamo on the cell cycle scEU-seq dataset (**Fig. 5A**). Almost all streamlines from our models correctly reflect the cell cycle progression, except that part of them from the ICSP model had a minor flaw in the M phase and CSZIP in the S phase. In addition, we found both ICSP and Dynamo's spliced RNA velocity (**Fig. 5B**) did not get entirely correct streamline results. The streamlines of our ICSP model were problematic in the M-G1 phase, while the streamlines of Dynamo were problematic in the S phase. We speculate that this is probably due to the fact that new unspliced mRNAs have rather low expression levels, frustrated with many dropouts and very sparse data, resulting in unreliable inferences of the parameter β and inaccurate RNA velocities.

We also quantitatively benchmarked the average correctness and consistency of the velocities in different methods in the original gene expression space and low-dimensional space (here the RFP_GFP space is used which corresponds to the Geminin-GFP and Cdt1-RFP-corrected signals of RPE1-FUCCI cells)(**Fig. 5C**,**D**; **Fig. S3A**,**B**). The definition of correctness and consistency of velocity is given in the "Methods" section. In the gene expression space, the average correctness and consistency of the total RNA velocity of CSP, ICSP, and CSZIP are significantly better than that of Dynamo (**Fig. 5C**, **D Left**), while the spliced RNA velocity of ICSP has slightly lower consistency than that of Dynamo (**Fig. 5C**, **D Right**). In the RFP_GFP space, the average correctness of total RNA velocity of all methods are significantly higher compared to that in the gene expression space, and simpler methods tend to improve more. The average correctness of CSP is highest at this time (**Fig. S3A Left**). However, the average correctness of the ICSP's spliced RNA velocity still perform slightly worse than Dynamo's (**Fig. S3A Right**). In contrast, the total RNA velocity consistency of CSP and ICSP is significantly better than that of Dynamo (**Fig. S3B Left**) and the spliced RNA velocity consistency of ICSP is also significantly better than that of Dynamo (**Fig. S3B Right**). Overall, the CSP-based total RNA velocity had the highest average correctness and consistency, significantly outperforms Dynamo, while the ICSP-based spliced RNA velocity was close to Dynamo quantitatively.

We now illustrate the advantages of our method in the estimation of kinetic parameters and the calculation of RNA velocity with two example genes: *DCBLD2* and *HIPK2*. In gene *DCBLD2*, the cells at M and M-G1 have the highest overall expression and the correct RNA velocity should be negative (**Fig 5E**). However, Dynamo returned the positive velocity, which is problematic (**Fig. 5F last column**). In contrast, CSP, CSZIP and ICSP all returned negative velocities (the **first three columns** in **Fig. 5F**). We speculated one possible explanation is that the expression of the gene *DCBLD2* has not yet reached a steady state. Consistent results were also observed from phase portraits of new-total RNA planes of *DCBLD2* (**Fig. 5G**, **Fig. S3C**). For gene *HIPK2*, similarly, cells in phase M and M-G1 have the highest expression overall and the correct velocity should be negative (**Fig. S3F**), but Dynamo and CSP both returned

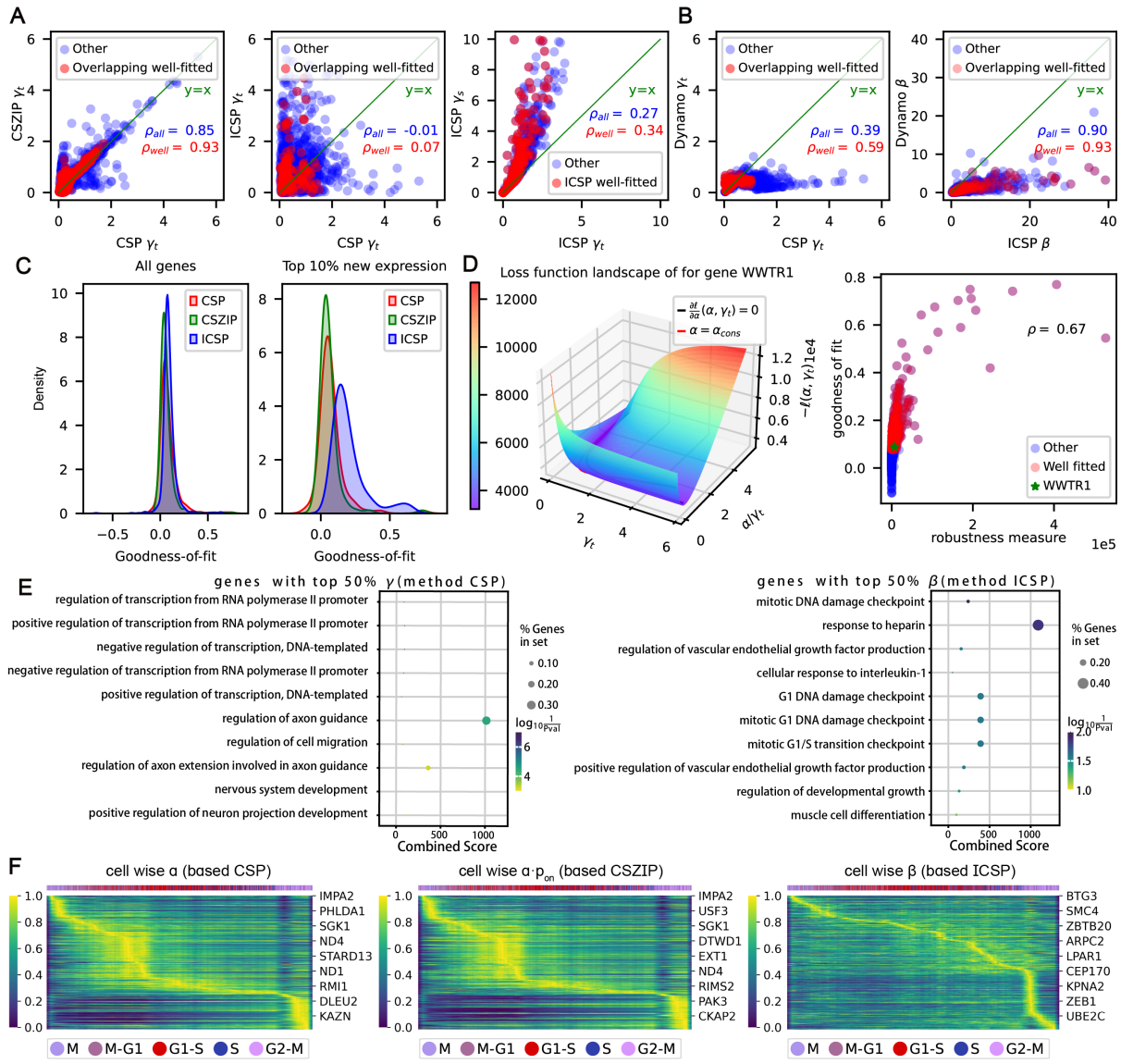

Figure 4: **Parameter inference and enrichment analysis for the cell cycle dataset. A.** Comparison of parameter inference results of our three stochastic models. From left to right are the comparison of γ_t of CSP and CSZIP, the comparison of γ_t of CSP and ICSP, the comparison of γ_t and γ_s in ICSP. The overlapping well-fitted genes were set as the overlap set of genes in the top 40% of the goodness-of-fit for both methods. **B.** Comparison of inferred parameters between our stochastic models and Dynamo's method. **Left**: the comparison of γ_t between CSP and Dynamo. **Right**: the comparison of β between ICSP and Dynamo. **C.** Comparison of the goodness-of-fit of the three stochastic models. **Left**: all highly variable genes. **Right**: genes in the top 10% of average new mRNA expression in highly variable genes. **D.** Robust analysis. **Left**: Landscape of Model 1-based loss functions for the a typical gene *WWTR1*. **Right**: Scatter plot of robustness measure and goodness of fit for parameter inference. **E.** Enrichment analysis results of genes with high γ_t, β (top 50%) in well fitted genes (top 40% of goodness of fit). **F.** Heat map of cell-wise parameters for well-fitted genes. From left to right, cell-wise α based on the CSP model, cell-wise αp_{on} based on the CSZIP model and cell-wise β based on the ICSP model, respectively. Across all three heatmaps, the X-axis is the relative cell cycle position while the order of genes in the y-axis is arranged such that the peak time of each gene increases from the top left to bottom right.

positive velocities while CSZIP got the correct results (**Fig. S3D**,**E**). We speculated one possible explanation for this is that the expression switch plays an important role in *HIPK2*.

To demonstrate the value of using gene-cell-wise parameters (except degradation rates), we visualized the streamlines of total RNA velocity based on gene-cell-wise parameters and those based only on gene-wise parameters (**Fig. S3G**). We observed that the streamlines of the CSP model and the CSZIP model in the S to G2-M phase are incorrectly reversed (**Fig. S3G Left** and **Middle**), and the streamlines of the ICSP model are also less smooth and accurate than those when gene-cell-wise parameters are used (**Fig. S3G Right**).

Finally, to demonstrate the significance of inferring time-resolved velocities with physical units, we calculated the duration time of each cell cycle phase of the human RPE1-FUCCI system based on the total RNA velocities (see "Methods" section, **Fig. 5H**). Indeed, the human RPE1-FUCCI system has a cell-cycle time of about 21 hours (about 6 hours for G1-S phase, 8 hours for S phase, 4 hours for G2-M phase, 1 hour for M phase and 2 hours for M-G1 phase) Chao et al. (2019).

Discussion

Storm utilizes three stochastic models for the dynamical description of new mRNAs and allows the estimation of the RNA velocity for kinetics experiments without the need for the steady-state assumption. It can also generally handle one-shot data when the steady-state assumption is enforced. One possible limitation of our model is that it does not fully utilize the total mRNA information in kinetics experiments. According to the results of the chi-square independence test, the number of total mRNA molecules of most genes obeys the same distribution. Noting that the old mRNA molecules with a labeling duration of zero are the total mRNA molecules, we think that it is a feasible direction to establish the stochastic dynamics of old mRNA and use the Wasserstein distance in optimal transport approach Vallender (1974); Zhang et al. (2021) to measure the differences between discrete distributions. Therefore, the optimal transport modeling of old RNAs may be integrated with Storm to obtain more robust RNA velocity inference. In addition, it is also worth exploring stochastic models that consider switching of gene expression states, transcription in the active state, splicing and spliced mRNA degradation simultaneously (i.e., integration of Model 2 and Model 3).

Some recent works, such as MultiVelo Li et al. (2022), Chromatin Velocity Tedesco et al. (2022), and protacel Gorin et al. (2020), extend RNA velocity to multi-omics. It is expected that the combination of metabolic labeling technology with other multi-omics measurements will bring new opportunities, which allows for simpler parameter inference and more accurate results.

Finally, most of the existing methods make the independent gene expression assumption, and do not consider the regulatory mechanism between genes. Deep neural network approaches are promising to solve this problem. This will be an important future direction.

Conclusions

We present Storm for estimating absolute kinetic parameters and inferring the time-resolved RNA velocity of metabolic labeling scRNA-seq data by incorporating the transient stochastic dynamics of gene expressions. Storm establishes three stochastic models of new mRNA which take into account both biological noise and cell-specific technical noise, and makes inference to the gene-specific degradation rates and other gene-cell-specific parameters without relying on the steady-state assumption in kinetics experiments. It can also handle one-shot data when the steady-state assumption is adopted. Numerical results show that Storm is able to accurately fit the kinetic cell cycle dataset and many one-shot experimental datasets. In addition, our numerical experience suggests that Model 1 (i.e., the CSP model) outperforms the other two models when splicing dynamics is not of interest, and the Model 2 (i.e., the ICSP model) is the valid choice if the data contains both labeling and splicing information and splicing dynamics is of interest. However, further applications and performance evaluations for more challenging datasets with temporal information are desired and it will be studied in the future.

Methods

Derivation of three stochastic dynamical models

Here we developed three stochastic models for the dynamical description of new mRNAs: Model 1) a stochastic dynamical model of new mRNA involving only metabolic-labeling transcription and degradation; Model 2) a stochastic dynamical model of new unspliced and spliced mRNA involving metabolic-labeling transcription, splicing and spliced mRNA degradation; and Model 3) a stochastic dynamical model of new mRNA involving gene state switching, metabolic-labeling transcription and degradation.

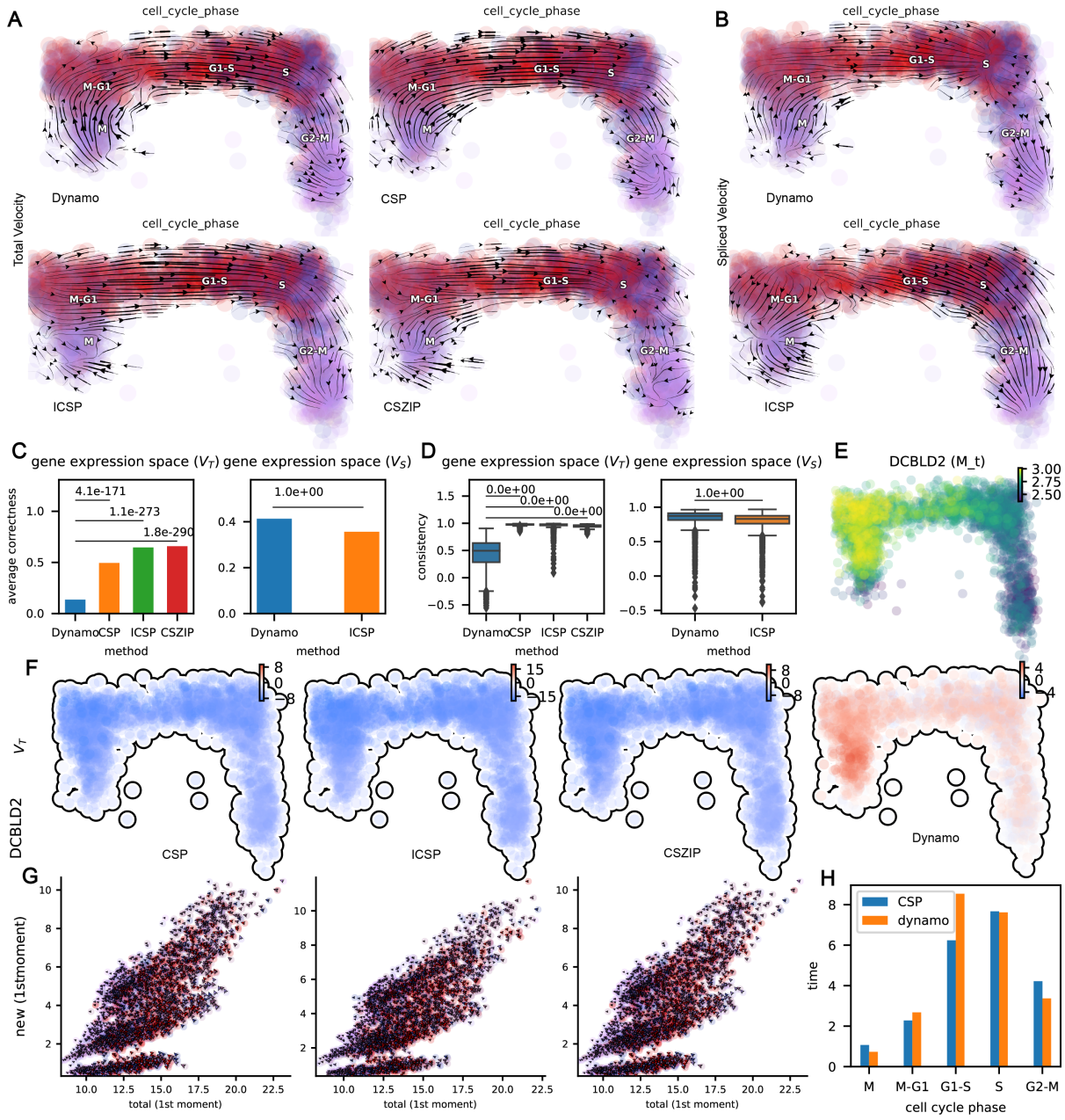

Figure 5: **RNA velocity analysis of the cell cycle dataset. A.** Comparison of total RNA velocity streamline visualizations between three stochastic methods and Dynamo. **B.** Comparison of spliced RNA velocity streamline visualizations between ICSP and Dynamo. **C.** Comparison of average correctness of velocity in gene expression space. **Left**: total RNA velocity. **Right**: spliced RNA velocity. The p-values are given by the one-sided Wilcoxon test. **D.** Similar to **C**, comparison of velocity consistency. **E.** The smoothed expression of *DCBLD2* in different cells. **F.** Comparison of total RNA velocity in *DCBLD2* between three stochastic models and Dynamo. **G.** Phase portraits of new-total RNA planes of *DCBLD2*. Quivers correspond to the total (x-component) or new (y-component) RNA velocity calculated by the different methods. **H.** The duration time (unit: hour) of each cell cycle phase of the human RPE1-FUCCI system based on Storm's CSP model and Dynamo.

Model 1: Stochastic dynamical modeling of new mRNA

Following Battich et al. (2020); Qiu et al. (2022), we made the following assumptions: (1) Genes are independent. (2) Both the transcription rate α and the degradation rate of total mRNA γ_t are constants.

The chemical master equation (CME) for the new/labeled mRNA $\tilde{l}(t)$, corresponding to the chemical reactions shown in **the first row of Fig. 1A**, is given by

$$\frac{d\tilde{P}_n}{dt} = -(\alpha + n\gamma_t)\tilde{P}_n + \alpha\tilde{P}_{n-1} + \gamma_t(n+1)\tilde{P}_{n+1}, \tag{12}$$

where $\tilde{P}_n(t) = \text{Prob}(\tilde{l}(t) = n)$. The initial value of new mRNA count is zero, i.e., $\tilde{P}_n(0) = \delta_{0n}$, where

$$\delta_{mn} = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise} \end{cases}$$

is the Kroneckers delta function. The solution of Eq. (12) is

$$\tilde{P}_n(t) = \frac{a(t)^n}{n!}e^{-a(t)}, \quad n \in \mathbb{N}, \tag{13}$$

where $a(t) = \alpha(1 - e^{-\gamma_t t})/\gamma_t$. This means that $\tilde{l}(t)$ obeys the Poisson distribution with mean $a(t)$.

The above stochastic model only describes the true expression count of new mRNA $\tilde{l}(t)$ in a cell with labeling duration t, but the measured sequencing data is different from this count due to technical noise. Denote by $l(t)$ the number of measured new mRNA molecules, and assume that $l(t)$ is associated with $\tilde{l}(t)$ through a dropout process, which we modeled as a binomial distribution:

$$\text{Prob}(l(t) = n \mid \tilde{l}(t) = N) = C_N^n p^n (1-p)^{N-n} := B_n(N, p), \tag{14}$$

where p is the capture probability of a single mRNA molecule. We further assume that the total number of mRNA molecules across all genes in different cells are close, which was commonly adopted in the preprocessing step La Manno et al. (2018); Bergen et al. (2020); Qiu et al. (2022). Denote by n_j the total number of mRNA molecules across all genes in cell j, i.e., $n_j = \sum_i r_{ij}$, where r_{ij} refers to the number of mRNA molecules in gene i of cell j in the scRNA-seq measurements. This assumption implies that the capture probability of mRNA molecules in different cells is different, and $p_j \propto n_j$. In our computation, we took $p_j = n_j/n_{\text{med}}$, where n_{med} is the median of n_j.

We denoted the PMF of new mRNA sequencing result $l_j(t)$ of cell j with labeling duration t by

$$P_{n,j}(t) := \text{Prob}\left(l_j(t) = n\right). \tag{15}$$

Then

$$P_{n,j}(t) = \sum_{N=n}^{\infty} \tilde{P}_N(t) B_n\left(N, p_j\right) = \frac{(p_j a(t))^n}{n!} e^{-p_j a(t)}, \tag{16}$$

which means that $l_j(t)$ obeys the Poisson distribution with mean $p_j a(t)$.

In summary, the former derivation shows that the number of new mRNA molecules in different cells in scRNA-seq measurements obeys Poisson distribution with cell-specific parameters, and these parameters were proportional to p_j, i.e., proportional to n_j. We call this distribution the *cell-specific Poisson distribution*.

Model 2: Stochastic dynamical modeling of new unspliced and spliced mRNAs

Compared with Model 1, we distinguished whether an mRNA molecule is spliced or not and incorporated the splicing process, which was shown in **the first row of Fig. 1A**. Again we assumed that the genes are independent. In addition, we further assumed that the transcription rate α, splicing rate β, and spliced mRNA degradation rate γ_s are all constants.

The CME for the new/labeled unspliced and spliced mRNAs $(\tilde{u}_l(t), \tilde{s}_l(t))$, corresponding to the considered chemical reactions shown in **the first row of Fig. 1A**, is given by

$$\begin{aligned} \partial_t \tilde{P}_{mn} = {} & \alpha(\tilde{P}_{m-1,n} - \tilde{P}_{mn}) + \beta\left[(m+1)\tilde{P}_{m+1,n-1} - m\tilde{P}_{mn}\right] \\ & + \gamma_s\left[(n+1)\tilde{P}_{m,n+1} - n\tilde{P}_{mn}\right], \end{aligned} \tag{17}$$

where $\tilde{P}_{mn}(t) = \text{Prob}((\tilde{u}_l(t), \tilde{s}_l(t)) = (m, n))$. The initial distribution of new unspliced and spliced mRNA is $\tilde{P}_{mn}(0) = \delta_{m0}\delta_{n0}$. The solution of Eq. (17) is

$$\tilde{P}_{mn}(t) = b(t)^m c(t)^n e^{-b(t)-c(t)}/m!n!, \quad (m, n) \in \mathbb{N}^2, \tag{18}$$

where

$$\begin{aligned} b(t) &= \alpha(1-e^{-\beta t})/\beta, \\ c(t) &= \begin{cases} \frac{\alpha}{\gamma_s}(1-e^{-\gamma_s t}) + \frac{\alpha}{\gamma_s-\beta}(e^{-\gamma_s t}-e^{-\beta t}), & \beta \neq \gamma_s, \\ \frac{\alpha}{\beta}(1-e^{-\beta t}) - \alpha t e^{-\beta t}, & \beta = \gamma_s, \end{cases} \end{aligned} \tag{19}$$

which means that $\tilde{u}_l(t)$ and $\tilde{s}_l(t)$ obey independent Poisson distributions with mean $b(t)$ and $c(t)$, respectively. We refer interested readers to Li et al. (2021) for derivation details.

Denote by $(u_l(t), s_l(t))$ the number of measured new unspliced and spliced mRNA molecules in the scRNA-seq experiments with labeling duration t. By assuming that the dropout processes for new unspliced and spliced mRNAs are independent and the capture probability is independent of whether they are spliced or not, we modeled the dropout process for $\tilde{u}_l(t)$ and $\tilde{s}_l(t)$ as independent binomial distributions with the same parameter p. So we got

$$\begin{aligned} &\text{Prob}\Big((u_l(t), s_l(t)) = (m,n) \mid (\tilde{u}_l(t), \tilde{s}_l(t)) = (M,N)\Big) \\ &= C_M^m p^m (1-p)^{M-m} C_N^n p^n (1-p)^{N-n} := B_m(M,p) B_n(N,p). \end{aligned} \tag{20}$$

For the same reason as Model 1, we take p_j proportional to n_j. And we took $p_j = n_j/n_{\text{med}}$ in the computation.

We denoted the joint PMF of new unspliced and spliced mRNA sequencing counts $(u_{l,j}(t), s_{l,j}(t))$ of cell j with labeling duration t by

$$P_{mn,j}(t) := \text{Prob}\left((u_{l,j}(t), s_{l,j}(t)) = (m,n)\right).$$

Then

$$\begin{aligned} P_{mn,j}(t) &= \sum_{M=m}^{\infty}\sum_{N=n}^{\infty} \frac{b(t)^M c(t)^N}{M!N!} e^{-b(t)-c(t)} B_M(m,p_j) B_N(n,p_j) \\ &= \sum_{M=m}^{\infty} \frac{b(t)^M}{M!} e^{-b(t)} B_M(m,p_j) \sum_{N=n}^{\infty} \frac{c(t)^N}{N!} e^{-c(t)} B_N(n,p_j) \\ &= \frac{(p_j b(t))^m}{m!} e^{-p_j b(t)} \frac{(p_j c(t))^n}{n!} e^{-p_j c(t)}, \end{aligned} \tag{21}$$

which means that $u_{l,j}(t)$ and $s_{l,j}(t)$ are independently Poisson distributed with mean $p_j b(t)$ and $p_j c(t)$, respectively.

In summary, $(u_l(t), s_l(t))$ obeys independent cell-specific Poisson distribution.

Model 3: Stochastic dynamical modeling of new mRNA considering switching

In Model 3, we further considered the on/off gene state switching shown in **the first row of Fig. 1C**. We assumed that the genes are independent as well, and the transcription rate α, mRNA degradation rate γ_t, the gene on-to-off rate k_{off} and off-to-on rate k_{on} are all constants. Furthermore, we assumed that k_{on} and k_{off} are significantly smaller than α and γ_t, which implies that the gene expression is either always on or always off during the transcription/degradation period. From Eq. (12), it is known that cells in the on state obey a Poisson distribution with mean $a(t)$, while cells in the off state do not express. Define $p_{\text{off}} = k_{\text{off}}/(k_{\text{off}} + k_{\text{off}})$. Then $\tilde{l}(t)$ obeys the zero-inflated Poisson distribution

$$\begin{aligned} \tilde{P}_0(t) &= (1-p_{\text{off}})e^{-a(t)} + p_{\text{off}}, \\ \tilde{P}_n(t) &= (1-p_{\text{off}})\frac{a(t)^n}{n!}e^{-a(t)}, \;\; n \geq 1. \end{aligned} \tag{22}$$

Similarly, by taking into account the technical noise in scRNA-seq experiments, the PMF of $l_j(t)$ is

$$\begin{aligned} P_{0,j}(t) &= (1-p_{\text{off}})e^{-p_j a(t)} + p_{\text{off}}, \\ P_{n,j}(t) &= (1-p_{\text{off}})\frac{(p_j a(t))^n}{n!}e^{-p_j a(t)}, \;\; n \geq 1. \end{aligned} \tag{23}$$

In summary, different cells obey the ZIP distribution with different parameters as shown in Eq. (23), which we called cell-specific zero-inflated Poisson distribution.

Chi-square goodness-of-fit test for cell-specific distributions at a fixed time

We would construct an asymptotic χ^2 statistic for the data with common distribution type but sample-specific parameters. This goodness-of-fit test is to assess whether the null hypothesis that the considered data, at a fixed labeling duration, obeys the proposed distribution can be accepted.

We first divided the value range of the considered data into c classes. According to the range that the samples fall in, we got n independent categorically distributed random samples $X_i \in \{1, 2, \ldots, c\}$ for $i = 1, 2, \ldots, n$ with sample dependent parameter p_i, respectively. An equivalent representation for the categorical variable X_i is to denote $X_i = (X_{ij})_{j=1,\ldots,c} \in \{e_1, \ldots, e_c\}$, where $e_j = (\delta_{jk})_{k=1,\ldots,c}$ is the indicator vector for $j = 1, \ldots, c$. Correspondingly, the parameter $p_i = (p_{i1}, \ldots, p_{ic})^T$ is a c-dimensional vector with non-negative elements and sums to one, which is defined as

$$p_{ij} := \mathrm{Prob}(X_{ij} = 1) = 1 - \mathrm{Prob}(X_{ij} = 0), \quad j = 1, \ldots, c. \tag{24}$$

This implies that $\mathrm{Var}(X_{ij}) = p_{ij}(1 - p_{ij})$ and $\mathrm{Cov}(X_{ij}, X_{il}) = \mathbb{E}[X_{ij}X_{il}] - p_j p_l = -p_j p_l$ for $j \neq l$. Therefore, the covariance matrix of random vector X_i is

$$\Sigma_i = \begin{pmatrix} p_{i1}(1-p_{i1}) & -p_{i1}p_{i2} & \cdots & -p_{i1}p_{ic} \\ -p_{i1}p_{i2} & p_{i2}(1-p_{i2}) & \cdots & -p_{i2}p_{ic} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{i1}p_{ic} & -p_{i2}p_{ic} & \cdots & p_{ic}(1-p_{ic}) \end{pmatrix}. \tag{25}$$

For sample i, we defined the truncated random vector $X_i^* = (X_{i1}, \ldots, X_{i,c-1})^T$ and truncated vector $p_i^* = (p_{i1}, \ldots, p_{i,c-1})^T$, which is the first $c-1$ components of X_i and p_i, respectively. The covariance matrix of X_i^* is the submatrix consisting of the upper-left $(c-1) \times (c-1)$ block of Σ_i, denoted by Σ_i^*, which can be written as

$$\Sigma_i^* = \mathrm{diag}(p_i^*) - p_i^*(p_i^*)^T, \tag{26}$$

where $\mathrm{diag}(p_i^*)$ is the diagonal matrix formed by the components of p_i^*.

Define $\bar{X}^* := (\sum_{i=1}^n X_i^*)/n$, $\bar{p}^* := (\sum_{i=1}^n p_i^*)/n$ and $\bar{\Sigma}^* := (\sum_{i=1}^n \Sigma_i^*)/n$, and let

$$\chi^2 := n(\bar{X}^* - \bar{p}^*)^T(\bar{\Sigma}^*)^{-1}(\bar{X}^* - \bar{p}^*). \tag{27}$$

Below we would show that χ^2 is an asymptotic chi-square statistic with degrees of freedom $c-1$. First note that

$$\mathrm{E}[\bar{X}^*] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^*\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X_i^*] = \frac{1}{n}\sum_{i=1}^n p_i^* = \bar{p}^*, \tag{28}$$

then the covariance

$$\mathrm{D}[\bar{X}^*] = \mathrm{D}\left[\frac{1}{n}\sum_{i=1}^n X_i^*\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{D}[X_i^*] = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n \Sigma_i^*\right) = \frac{1}{n}\bar{\Sigma}^*. \tag{29}$$

Let $Y_n = \sqrt{n}(\Sigma^*)^{-1/2}(\bar{X}^* - \bar{p}^*)$. When n goes to infinity, Y_n converges in distribution to the normal distribution $N(0, I_{c-1})$ according to the central limit theorem for the independent sum of random variables. Thus, $\chi^2 = Y_n^T Y_n$ converges in distribution to a chi-square distribution with degrees of freedom $c-1$.

In summary, we proposed a new asymptotic χ^2 statistic for sample-specific distributions. For a fixed labeling duration t_{fixed}, $a(t_{\mathrm{fixed}})$, $b(t_{\mathrm{fixed}})$ and $c(t_{\mathrm{fixed}})$ are all constants, the proposed χ^2 statistics can be used to test whether the new mRNA sequencing data are consistent with the CSP, ICSP and CSZIP distributions based on Models 1, 2 and 3, respectively. In addition, since there are one, two and two parameters to be inferred in CSP, ICSP and CSZIP, respectively, the same number of degrees of freedom should be subtracted. Following Koehler and Larntz (1980), we ensured that the expected count $np_j \geq 0.25$ in each group when determining the group value ranges. Finally, we take p-value as 0.05 in the computation.

Parameter inference in one-shot experiments

In the one-shot experiments, we only observed new RNA $l_j(t)$ and total RNA $r_j(t)$ data for one labeling duration t. So we had to invoke the steady-state assumption for the total RNA in this case.

When the dynamics of total RNA in Model 1 is at steady state, i.e.,

$$0 = \frac{d\tilde{P}_{r,n}}{dt} = -(\alpha + n\gamma_t)\tilde{P}_{r,n} + \alpha\tilde{P}_{r,n-1} + \gamma_t(n+1)\tilde{P}_{r,n+1}, \tag{30}$$

where $\tilde{P}_{r,n} := \mathrm{Prob}(\tilde{r} = n)$ is the invariant PMF of the true expression of total RNA. From Eq. (16) we know that when technical noise is considered, the observed total RNA counts obey a similar CSP distribution

$$P_{r,n,j} = \frac{(p_j\alpha/\gamma_t)^n}{n!}e^{-p_j\alpha/\gamma_t}. \tag{31}$$

At this point, we obtained the distributions of the new RNA and total RNA observations so that parameter inference can be performed using the MLE. We want to maximize the log-likelihood function

$$\ell(\alpha, \gamma_t) = \sum_{j=1}^{n} \log\Big(\mathcal{P}(p_j a(t))|_{l_j}\Big) + \log\Big(\mathcal{P}\big(p_j \alpha/\gamma_t\big)|_{r_j}\Big), \tag{32}$$

where $\mathcal{P}(\lambda)|_n := \mathrm{Prob}(X = n) = e^{-\lambda}\lambda^n/n!$ is the probability of $X = n$ for a Poisson-distributed random variable X with mean λ. When $\partial\ell/\partial\alpha = 0$ and $\partial\ell/\partial\gamma_t = 0$, the likelihood function is maximized and it can be solved analytically

$$\gamma_t = -\frac{1}{t}\log\Big(1 - \frac{\langle l_j\rangle}{\langle r_j\rangle}\Big), \quad \alpha = \gamma_t \frac{\langle r_j\rangle}{\langle p_j\rangle}, \tag{33}$$

where $\langle\cdot\rangle$ means the population average defined by

$$\langle\cdot\rangle = \Big(\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\cdot)\Big)\Big/\Big(\sum_{k=1}^{K} n_k\Big). \tag{34}$$

Since here it is for the one-shot data set, $K = 1$. Note that Eq. (33) is similar to the formula in Dynamo Qiu et al. (2022) for estimating the parameters for one-shot data. The difference is that this formula averages the raw counts, while the method in Dynamo averages the smoothed data.

Parameter inference in kinetics experiments

In the kinetics experiments, we observed data $l_j(t_k)$ (or $(u_{l,j}(t_k), s_{l,j}(t_k))$) for new mRNA (or new unspliced and spliced mRNAs) with different labeling durations. We assumed that there are K labeling durations t_k for $k = 1, 2, \ldots, K$, and the number of cells with labeling duration t_k is n_k. We utilized the MLE to infer the unknown parameters in different models without relying on steady-state assumptions.

In Model 1, we need to maximize the log-likelihood function

$$\ell(\alpha, \gamma_t) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \log\Big(\mathcal{P}(p_j(t_k)a(t_k))|_{l_j(t_k)}\Big). \tag{35}$$

It is equivalent to minimizing the following loss function

$$L(\alpha, \gamma_t) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} -l_j(t_k)\log(p_j(t_k)a(t_k)) + p_j(t_k)a(t_k). \tag{36}$$

The optimum of the loss is achieved when the gradient equals 0. Utilizing the concrete expression of $a(t)$ (Eq. (3)) in Model 1, we got $\partial a(t)/\partial\alpha = (1 - e^{-\gamma_t t})/\gamma_t$. Then $\partial L(\alpha, \gamma_t)/\partial\alpha = 0$ has a closed form solution

$$\alpha(\gamma_t) = \frac{\langle l_j(t_k)\rangle}{\langle p_j(t_k)\partial a(t_k)/\partial\alpha\rangle}. \tag{37}$$

Another component of the Euler-Lagrange equation $\partial L/\partial\gamma_t = 0$ has no closed form solution, so we need to solve γ_t by numerical iterations. We took the initial value of γ_t as the solution from Dynamo Qiu et al. (2022) under the steady-state assumption. Denote it as $\gamma_{t,0}$, and correspondingly, we take the initial value of α as $\alpha_0 = \alpha(\gamma_{t,0})$.

In Model 2, we need to maximize the log-likelihood function

$$\ell(\alpha, \beta, \gamma_s) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \log\Big(\mathcal{P}(p_j(t_k)b(t_k))|_{u_{l,j}(t_k)} \cdot \mathcal{P}(p_j(t_k)c(t_k))|_{s_{l,j}(t_k)}\Big), \tag{38}$$

which is equivalent to minimizing the loss function

$$\begin{aligned} L(\alpha, \beta, \gamma_s) = &\sum_{k=1}^{K}\sum_{j=1}^{n_k} \Big(- u_{l,j}(t_k)\log(p_j(t_k)a(t_k)) + p_j(t_k)a(t_k)\Big) \\ &+ \Big(- s_{l,j}(t_k)\log(p_j(t_k)b(t_k)) + p_j(t_k)b(t_k)\Big). \end{aligned} \tag{39}$$

Utilizing (19), we got $\partial b(t)/\partial\alpha = (1-e^{-\beta t})/\beta$, $\partial c(t)/\partial\alpha = (1-e^{-\gamma_s t})/\gamma_s + (e^{-\gamma_s t} - e^{-\beta t})/(\gamma_s - \beta)$ when $\beta \neq \gamma_s$, and the case for $\beta = \gamma_s$ is similar. So $\partial L(\alpha, \beta_t, \gamma_s)/\partial\alpha = 0$ has a closed form solution

$$\alpha(\beta, \gamma_s) = \frac{\langle u_{l,j}(t_k) + s_{l,j}(t_k)\rangle}{\langle p_j(t_k)(\frac{\partial b}{\partial\alpha}(t_k) + \frac{\partial c}{\partial\alpha}(t_k))\rangle}. \tag{40}$$

However $\partial L/\partial\beta = 0$ and $\partial L/\partial\gamma_s = 0$ have no closed form solution, and we need to solve these equations by iterations. The choice of initial values is similar to the Model 1 case. We took the initial value of β and γ_s as the solution from Dynamo Qiu et al. (2022) under the steady-state assumption, which we denoted as $\beta_0, \gamma_{s,0}$. And then the initial value of α is taken as $\alpha_0 = \alpha(\beta_0, \gamma_{s,0})$.

In Model 3, we need to maximize the log-likelihood function

$$\begin{aligned}\ell(p_{\text{off}}) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathbb{I}_{\{l_j(t_k)=0\}} \log(\text{ZIP}(p_j a(t_k), p_{\text{off}})|_0) \\ + \mathbb{I}_{\{l_j(t_k)>0\}} \log(\text{ZIP}(p_j a(t_k), p_{\text{off}})|_{l_j(t_k)}),\end{aligned} \tag{41}$$

where $\text{ZIP}(\lambda, p_{\text{off}})|_n := \text{Prob}(X = n)$ is the probability of $X = n$ for a ZIP-distributed random variable X with parameters λ and p_{off}. It is equivalent to minimizing the loss function

$$\begin{aligned}L(\alpha, \gamma_t, p_{\text{off}}) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} -\log(\text{ZIP}(p_j a(t_k), p_{\text{off}})|_0) - \\ \mathbb{I}_{\{l_j(t_k)>0\}}\Big(\log(1-p_{\text{off}}) + l_j(t_k)\log(p_j(t_k)a(t_k)) - p_j(t_k)a(t_k)\Big).\end{aligned} \tag{42}$$

Similar as before, we chose the initial value of γ_t, denoted as $\gamma_{t,0}$, based on the steady state assumption, and chose the moment estimator

$$p_{\text{off},0} = 1 - \frac{\langle l_j(t_k)\rangle^2 \langle (p_j(t_k)\frac{\partial a}{\partial\alpha}(t_k))^2\rangle}{\langle p_j(t_k)\frac{\partial a}{\partial\alpha}(t_k)\rangle^2(\langle l_j(t_k)^2\rangle - \langle l_j(t_k)\rangle)} \tag{43}$$

and

$$\alpha_0 = \frac{\langle l_j(t_k)\rangle}{(1-p_{\text{off},0})\langle p_j(t_k)\frac{\partial a}{\partial\alpha}(t_k)\rangle} \tag{44}$$

as the initial values of p_{off} and α.

According to the biological meaning of the parameters, we added the constraints $0 < \alpha < 10\alpha_0$, $0 < \beta < 10\beta_0$, $0 < \gamma_t < 10\gamma_{t,0}$, $0 < \gamma_s < 10\gamma_{s,0}$ and $0 < p_{\text{off}} < 1$, and we called the SLSQP optimizer in SciPy to solve the above optimization problem.

Goodness-of-fit test for the distribution evolution in time

In ordinary least squares (OLS) linear regression, people often use

$$R^2 := 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{45}$$

to define the goodness of fit, where y_i is the sample observation, $\hat{y}_i$ is the model prediction, and $\bar{y}_i$ is the sample mean. For the generalized linear model (GLM), the R^2 can be defined using the deviance D and null deviance D_0 Menard (2000),

$$R_D^2 := 1 - \frac{D}{D_0} = 1 - \frac{-2(\ell(\hat{\beta}) - \ell_s)}{-2(\ell_0 - \ell_s)} = 1 - \frac{\ell(\hat{\beta}) - \ell_s}{\ell_0 - \ell_s}, \tag{46}$$

where $\ell(\hat{\beta})$, ℓ_0 and ℓ_s denotes the log-likelihood function of the model with parameter $\hat{\beta}$, the null model (that is, fitted with only the intercept), and the saturated model (that is, fitted with one parameter per sample), respectively. A pictorial representation of D and D_0 is shown in Fig. 1E. R_D^2 can be seen as a generalization of R^2, which is equal to R^2 when the model is a least squares linear regression Menard (2000). Finally, to overcome the disadvantage of adding more parameters without reducing R_D^2 (similar to R^2), we used adjusted R_D^2 (denoted as $\bar{R}_D^2$) as the goodness of fit of our model, which is defined as

$$\bar{R}_D^2 := 1 - \frac{D/d_D}{D_0/d_{D_0}} = 1 - \frac{(\ell(\hat{\beta}) - \ell_s)/d_D}{(\ell_0 - \ell_s)/d_{D_0}}, \tag{47}$$

where d_D and d_{D_0} are the degrees of freedom of D and D_0, respectively.

In Model 1, ℓ_s has the closed form

$$\ell_s = \sum_{k=1}^{K}\sum_{j=1}^{n_k} l_j(t_k) \log\left(\mathcal{P}(l_j(t_k))|_{l_j(t_k)}\right). \tag{48}$$

To calculate ℓ_0, we need to maximize the log-likelihood function

$$\ell(a_0) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \log\left(\mathcal{P}(p_j(t_k)a_0)|_{l_j(t_k)}\right), \tag{49}$$

where a_0 is the intercept. The problem has a closed form solution $a_0 = \langle l_j(t_k)\rangle / \langle p_j(t_k)\rangle$. In addition, $d_D = N - 2$ and $d_{D_0} = N - 1$, where N is the number of cells.

In Model 2, ℓ_s has the closed form

$$\ell_s = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \log\left(\mathcal{P}(u_{l,j}(t_k))|_{u_{l,j}(t_k)}\right) + \log\left(\mathcal{P}(s_{l,j}(t_k))|_{s_{l,j}(t_k)}\right) \tag{50}$$

To calculate ℓ_0, we need to maximize the log-likelihood function

$$\ell(b_0, c_0) = \log\left(\mathcal{P}(p_j(t_k)b_0)|_{u_{l,j}(t_k)}\right) + \log\left(\mathcal{P}(p_j(t_k)c_0))|_{s_{l,j}(t_k)}\right) \tag{51}$$

where b_0 and c_0 are intercepts and have closed form solutions $b_0 = \langle u_{l,j}(t_k)\rangle / \langle p_j(t_k)\rangle$ and $c_0 = \langle s_{l,j}(t_k)\rangle / \langle p_j(t_k)\rangle$, respectively. In addition, $d_D = 2N - 3$ and $d_{D_0} = 2N - 2$.

In Model 3, to calculate ℓ_s, we need to maximize the log-likelihood function

$$\begin{aligned}\ell(\alpha, \gamma_t, p_{\text{off}}) = &\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathbb{I}_{\{l_j(t_k)=0\}} \log\left(\text{ZIP}(0, p_{\text{off}})|_0\right) \\ &+ \mathbb{I}_{\{l_j(t_k)>0\}} \log\left(\text{ZIP}(l_j(t_k), p_{\text{off}})|_{l_j(t_k)}\right) \\ = &\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathbb{I}_{\{l_j(t_k)>0\}} \log\left(\text{ZIP}(l_j(t_k), p_{\text{off}})|_{l_j(t_k)}\right)\end{aligned} \tag{52}$$

When p_{off} is equal to zero, Eq. (52) is maximized, and the closed form solution of ℓ_s is

$$\ell_s = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathbb{I}_{\{l_j(t_k)>0\}} \left(l_j(t_k)\log(l_j(t_k)) - l_j(t_k) - \log(l_i(t_k)!)\right). \tag{53}$$

To calculate ℓ_0, we need to maximize the log-likelihood function

$$\begin{aligned}\ell(a_0, p_{\text{off}}) = &\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathbb{I}_{\{l_j(t_k)=0\}} \log(\text{ZIP}(p_j a_0, p_{\text{off}})|_0) \\ &+ \mathbb{I}_{\{l_j(t_k)>0\}} \log(\text{ZIP}(p_j a_0, p_{\text{off}})|_{l_j(t_k)}).\end{aligned} \tag{54}$$

Similar to solving Eq. (42), $p_{\text{off},0}$ and a_0 were initialized using moment estimators with additional constraints $0 < p_{\text{off}} < 1$ and $0 < a < 10a_0$. We then called the SLSQP optimizer in SciPy to solve the problem. In addition, $d_D = N - 2$ and $d_{D_0} = N - 1$.

Post-processing for cell-specific parameters

In our cell-specific modeling of gene expression, we only assumed that γ_t (in Models 1 and 3) and γ_s (in Model 2) are constants over cells and are inferred based on the corresponding stochastic models, while the other parameters are cell-specific and continuously dependent on gene expressions. This relaxed assumption implies that only the degradation rate is common to all cells, and only cells with similar gene expressions have similar other parameters (due to continuous dependence). To realize this assumption, we first constructed the k-nearest neighbor (kNN) graph

of cells by a data preprocessing. The cell-specific parameter inference was performed by applying the inference to the kNN graph for each cell with local constant parameter assumption and already inferred degradation rates. The inference details for our three models were shown as below.

In Model 1, we have

$$l_i(t_k) \sim \text{Poisson}(p_i a_j(t_k)), \quad \forall i \in \mathcal{N}_{j,t_k}, \tag{55}$$

where $\mathcal{N}_{j,t_k}$ denotes the set of top k cells that have the most similar gene expressions as the jth cell with labeling duration t_k (including itself) and $a_j(t_k) = \alpha_j(t_k)(1-e^{-\gamma_t t_k})/\gamma_t$. Assuming that γ_t has been inferred, we can obtain a local estimator

$$\frac{\sum_{i\in\mathcal{N}_{j,t_k}} l_i(t_k)}{\sum_{i\in\mathcal{N}_{j,t_k}} p_i(t_k)} = a_j(t_k) = \frac{\alpha_j(t_k)}{\gamma_t}(1-e^{-\gamma_t t_k}) \tag{56}$$

by using the MLE. Define $\hat{l}_j(t_k) = (\sum_{i\in\mathcal{N}_{j,t_k}} l_i(t_k))/(\sum_{i\in\mathcal{N}_{j,t_k}} p_i(t_k))$. Then the cell-specific transcription rate $\alpha_j(t_k)$ has a closed form solution

$$\alpha_j(t_k) = \hat{l}_j(t_k)\gamma_t/(1-e^{-\gamma_t t_k}). \tag{57}$$

In Model 2, we have

$$(u_{l,i}(t_k), s_{l,i}(t_k)) \sim \text{independent Poisson}(p_i b_j(t_k), p_i c_j(t_k)), \quad \forall i \in \mathcal{N}_{j,t_k}. \tag{58}$$

Similarly, assuming γ_s has been inferred, and defining the local estimators

$$\hat{u}_{l,j}(t_k) = \frac{\sum_{i\in\mathcal{N}_{j,t_k}} u_{l,i}(t_k)}{\sum_{i\in\mathcal{N}_{j,t_k}} p_i(t_k)}, \qquad \hat{s}_{l,j}(t_k) = \frac{\sum_{i\in\mathcal{N}_{j,t_k}} s_{l,i}(t_k)}{\sum_{i\in\mathcal{N}_{j,t_k}} p_i(t_k)}, \tag{59}$$

we have

$$\begin{aligned} \hat{u}_{l,j}(t_k) &= b_j(t_k) = \frac{\alpha_j(t_k)}{\beta_j(t_k)}(1-e^{-\beta_j(t_k)t_k}), \\ \hat{s}_{l,j}(t_k) &= c_j(t_k) = \frac{\alpha_j(t_k)}{\gamma_s}(1-e^{-\gamma_s t_k}) + \frac{\alpha_j(t_k)}{\gamma_s-\beta_j(t_k)}(e^{-\gamma_s t_k} - e^{-\beta_j(t_k)t_k}), \end{aligned} \tag{60}$$

which is a nonlinear system. We have

$$\frac{\hat{s}_{l,j}(t_k)}{\hat{u}_{l,j}(t_k)} = \frac{\beta_j(t_k)(1-e^{-\gamma_s t_k})}{\gamma_s(1-e^{-\beta_j(t_k)t_k})} + \frac{\beta_j(t_k)(e^{-\gamma_s t_k} - e^{-\beta_j(t_k)t_k})}{(\gamma_s-\beta_j(t_k))(1-e^{-\beta_j(t_k)t_k})}. \tag{61}$$

To solve $\beta_j(t_k)$, we set its initial value as previously inferred β by global constant assumption. We then call the *foot* function in SciPy to solve the nonlinear equation (61) to get $\beta_j(t_k)$. The $\alpha_j(t_k)$ has a closed form solution

$$\alpha_j(t_k) = \hat{u}_{l,j}(t_k)\beta_j(t_k)/(1-e^{-\beta_j(t_k)t_k}). \tag{62}$$

In summary, in Model 2, we can infer the cell-specific transcription rate $\alpha_j(t_k)$ and splicing rate $\beta_j(t_k)$.

In Model 3, we have

$$l_i(t_k) \sim \text{ZIP}(p_i a_j(t_k), p_{\text{off},j}(t_k)), \quad \forall i \in \mathcal{N}_{j,t_k}. \tag{63}$$

When computing RNA velocity, we only need to know $\alpha_j(t_k)(1-p_{\text{off},j}(t_k))$ as a whole, and not their respective values (see next subsection). To simplify the computation, we used the moment estimation instead of MLE, and got

$$\hat{l}_j(t_k) = (1-p_{\text{off},j})a_j(t_k) = \frac{(1-p_{\text{off},j}(t_k))\alpha_j(t_k)}{\gamma_t}(1-e^{-\gamma_t t_k}). \tag{64}$$

Similarly, assuming γ_t has been inferred, $\alpha_j(t_k)(1-p_{\text{off},j}(t_k))$ has a closed form solution

$$\alpha_j(t_k)(1-p_{\text{off},j}(t_k)) = \hat{l}_j(t_k)\gamma_t/(1-e^{-\gamma_t t_k}). \tag{65}$$

Reduction from stochastic to deterministic models for RNA velocity

We used discrete counts data in the proposed parameter inference and goodness-of-fit calculation via stochastic models. However, when we need to compute and visualize the RNA velocity, we should take the reduction from stochastic to deterministic models to get the mean velocity. Below we would show the reduction process and reveal the connection between the stochastic and their corresponding deterministic models.

In Model 1, let us denote the mean value of $\tilde{l}(t)$ by $\langle\tilde{l}(t)\rangle$, which is defined as $\langle\tilde{l}(t)\rangle = \sum_{n=1}^{\infty} n\tilde{P}_n(t)$. From Eq. (12) we can obtain the deterministic equation after suitable algebraic manipulations

$$\begin{aligned}\frac{d\langle\tilde{l}(t)\rangle}{dt} =& \sum_{n=1}^{\infty} n\frac{d\tilde{P}_n(t)}{dt}\\ =& \sum_{n=1}^{\infty} n(-(\alpha+n\gamma_t)\tilde{P}_n + \alpha\tilde{P}_{n-1} + \gamma_t(n+1)\tilde{P}_{n+1})\\ =& \alpha - \gamma_t\langle\tilde{l}(t)\rangle.\end{aligned} \tag{66}$$

Similarly, the mean value of total RNA $\tilde{r}(t)$ satisfies the equation

$$\frac{d\langle\tilde{r}(t)\rangle}{dt} = \alpha - \gamma_t\langle\tilde{r}(t)\rangle. \tag{67}$$

Since the initial value of $\tilde{l}(t)$ is zero, we got

$$\langle\tilde{l}(t)\rangle = a(t) = \frac{\alpha}{\gamma_t}(1-e^{-\gamma_t t}). \tag{68}$$

In Model 2, the marginal PMFs of $\tilde{u}_l(t)$ and $\tilde{s}_l(t)$ are

$$\begin{aligned}\tilde{P}_{m,\cdot}(t) :=& \text{Prob}\left(\tilde{u}_l(t)=m\right) = \sum_{n=0}^{\infty}\tilde{P}_{m,n}(t),\\ \tilde{P}_{\cdot,n}(t) :=& \text{Prob}\left(\tilde{s}_l(t)=n\right) = \sum_{m=0}^{\infty}\tilde{P}_{m,n}(t),\end{aligned} \tag{69}$$

respectively. The mean values of $\tilde{u}_l(t)$ and $\tilde{s}_l(t)$ have the form $\langle\tilde{u}_l(t)\rangle = \sum_{m=1}^{\infty} m\tilde{P}_{m,\cdot}(t)$ and $\langle\tilde{s}_l(t)\rangle = \sum_{n=1}^{\infty} n\tilde{P}_{\cdot,n}(t)$. From the CME (17), we can obtain

$$\begin{aligned}\frac{d\langle\tilde{u}_l(t)\rangle}{dt} =& \sum_{m=1}^{\infty} m\partial_t\tilde{P}_{m,\cdot}(t) = \sum_{m=1}^{\infty} m\sum_{n=0}^{\infty}\partial_t\tilde{P}_{m,n}(t)\\ =& \sum_{m=1}^{\infty} m\sum_{n=0}^{\infty}\alpha(\tilde{P}_{m-1,n}-\tilde{P}_{mn}) + \beta((m+1)\tilde{P}_{m+1,n-1} - m\tilde{P}_{mn})\\ &+\gamma_s((n+1)\tilde{P}_{m,n+1} - n\tilde{P}_{mn})\\ =& \alpha - \beta\langle\tilde{u}_l(t)\rangle,\end{aligned} \tag{70}$$

and

$$\begin{aligned}\frac{d\langle\tilde{s}_l(t)\rangle}{dt} =& \sum_{n=1}^{\infty} n\partial_t\tilde{P}_{\cdot,n}(t) = \sum_{n=1}^{\infty} n\sum_{m=0}^{\infty}\partial_t\tilde{P}_{m,n}(t)\\ =& \sum_{n=1}^{\infty} n\sum_{m=0}^{\infty}\alpha(\tilde{P}_{m-1,n}-\tilde{P}_{mn}) + \beta((m+1)\tilde{P}_{m+1,n-1} - m\tilde{P}_{mn})\\ &+\gamma_s((n+1)\tilde{P}_{m,n+1} - n\tilde{P}_{mn})\\ =& \beta\langle\tilde{u}_l(t)\rangle - \gamma_s\langle\tilde{s}_l(t)\rangle.\end{aligned} \tag{71}$$

Similarly, we can derive the equations for the mean values of total unspliced and spliced mRNA $(\tilde{u}(t), \tilde{s}(t))$:

$$\begin{aligned}\frac{d\langle\tilde{u}(t)\rangle}{dt} &= \alpha - \beta\langle\tilde{u}(t)\rangle,\\ \frac{d\langle\tilde{s}(t)\rangle}{dt} &= \beta\langle\tilde{u}(t)\rangle - \gamma_s\langle\tilde{s}(t)\rangle.\end{aligned} \tag{72}$$

Since the initial value of $(\tilde{u}_l(t), \tilde{s}_l(t))$ is $(0,0)$, we got

$$\langle\tilde{u}_l(t)\rangle = b(t) = \frac{\alpha}{\beta}(1-e^{-\beta t}) \tag{73}$$

and

$$\langle \tilde{s}_l(t)\rangle = c(t) = \begin{cases} \frac{\alpha}{\gamma_s}(1-e^{-\gamma_s t}) + \frac{\alpha}{\gamma_s-\beta}(e^{-\gamma_s t}-e^{-\beta t}), & \beta \neq \gamma_s, \\ \frac{\alpha}{\beta}(1-e^{-\beta t}) - \alpha t e^{-\beta t}, & \beta = \gamma_s. \end{cases} \tag{74}$$

Similar to Model 1, in Model 3, $d\langle\tilde{l}(t)\rangle/dt$ and $d\langle\tilde{r}(t)\rangle/dt$ satisfy the equations

$$\begin{aligned} \frac{d\langle\tilde{l}(t)\rangle}{dt} &= (1-p_{\text{off}})\alpha - \gamma_t\langle\tilde{l}(t)\rangle, \\ \frac{d\langle\tilde{r}(t)\rangle}{dt} &= (1-p_{\text{off}})\alpha - \gamma_t\langle\tilde{r}(t)\rangle. \end{aligned} \tag{75}$$

Since the initial value of $\tilde{l}(t)$ is zero, we got

$$\langle\tilde{l}(t)\rangle = \frac{(1-p_{\text{off}})\alpha}{\gamma_t}(1-e^{-\gamma_t t}). \tag{76}$$

Computation of RNA velocity

To ease the notation, we denoted the new mRNA after data preprocessing by $\bar{l}(t)$, defined as

$$\bar{l}_j(t_k) = \frac{1}{|\mathcal{N}_{j,t_k}|}\sum_{i\in\mathcal{N}_{j,t_k}}\frac{l_i(t_k)}{p_i(t_k)},$$

which is different from the true expression $\tilde{l}(t)$, the discrete counts data $l(t)$, and the notation $\hat{l}(t)$ in the post-processing subsection. We would also use the notation $\bar{u}(t)$, $\bar{s}(t)$ and $\bar{r}(t)$ with similar definition.

In Model 1, only the total RNA velocity can be obtained due to the lack of the splicing stage. From Eq. (67), we have

$$v_{\text{total},r_j(t_k)} = \alpha_j(t_k) - \gamma_t\bar{r}_j(t_k), \tag{77}$$

where $\bar{r}_j(t_k)$ is the number of total mRNA molecules of the jth cell labeled with length t_k after data preprocessing.

In Model 2, we add the two equations in Eq. (72) to obtain

$$\frac{d\langle\tilde{r}(t)\rangle}{dt} = \frac{d\langle\tilde{u}(t)\rangle}{dt} + \frac{d\langle\tilde{s}(t)\rangle}{dt} = \alpha - \gamma_s\langle\tilde{s}(t)\rangle, \tag{78}$$

and thus get the equation for total RNA velocity

$$v_{\text{total},r_j(t_k)} = \alpha_j(t_k) - \gamma_s\bar{s}_j(t_k). \tag{79}$$

In addition, in Model 2, we can also calculate the spliced RNA velocity by the following equation

$$v_{\text{spliced},s_j(t_k)} = \beta_j(t_k)\bar{u}_j(t_k) - \gamma_s\bar{s}_j(t_k). \tag{80}$$

Similar to Model 1, the total RNA velocity in Model 3 can be obtained by the equation

$$v_{\text{total},r_j(t_k)} = (1-p_{\text{off},j}(t_k))\alpha_j(t_k) - \gamma_t\bar{r}_j(t_k). \tag{81}$$

Relationship between γ_t and γ_s and its implications

The difference between Eqs. (67) and (78) implies the difference between the total mRNA degradation rate γ_t and spliced mRNA degradation rate γ_s. After suitable manipulations, we had the relation between γ_t and γ_s as below

$$\frac{\gamma_s}{\gamma_t} = \frac{\langle\tilde{r}(t)\rangle}{\langle\tilde{s}(t)\rangle}. \tag{82}$$

Therefore, we naturally got a method to infer γ_t when γ_s is known. Specifically, we first performed a zero-intercept linear regression

$$\bar{r}_j(t_k) = k\bar{s}_j(t_k) \tag{83}$$

to get the slope k. Then we computed γ_t by $\gamma_t = \gamma_s/k$. Therefore, we can also infer γ_t and compute the total RNA velocity by Eq. (77) in Model 2.

We would also like to point out that Model 1 and 3 are incompatible upon assuming that γ_t and γ_s are both constants. These two assumptions usually do not hold simultaneously. Otherwise, from Eq. (82) we knew that $\langle\tilde{s}(t)\rangle/\langle\tilde{r}(t)\rangle$ is a constant, which is equivalent to that $\langle\tilde{u}(t)\rangle/\langle\tilde{r}(t)\rangle$ is a constant, i.e., $\gamma_t(1-e^{-\beta t})/(\beta(1-e^{-\gamma_t t}))$ is a constant. But this is only true when β and γ_t are equal.

Definition of correctness and consistency of velocity

The correctness of cell velocities is defined as follows: Consider the cell i with position x_i and velocity v_i. Define its one-step extrapolated position as $x_i + v_i$. We say that v_i is correct (correctness index $= 1$) if the cell j closest to the extrapolated position $x_i + v_i$ ranks after i in the temporal ordering. Otherwise the correctness does not hold and we set the correctness index to be 0. Thus the average correctness refers to the percentage of correct velocities.

The consistency means the extent to which the velocity of one cell is consistent with the velocities of its neighboring cells, and we use the average cosine similarity proposed in scVelo Bergen et al. (2020) to measure this consistency.

Calculation of cell cycle time

After the total RNA velocities are obtained, we can evaluate the time of each phase of a cell cycle based on them. Specifically, we first pick k cells x_i^0 $(i = 1, 2, \ldots, k)$ whose relative positions are closest to 0 as a cell group, calculate their average expression $\bar{x}^0$ and velocity $\bar{v}^0$ as the initial expression x^0 and velocity v^0, and extrapolate the state of the cell group with a short time step dt, that is, $x^1 = x^0 + v^0 dt$. We then search for another k cells x_i^1 $(i = 1, 2, \ldots, k)$ which are closest to the extrapolated state x^1, set the majority of the phase of these k cells to the phase of x^1, and set their average velocity $\bar{v}^1$ as v^1 for the second cell group. Next, the extrapolation and local k-cells group identification step can be repeated until a given threshold of the relative position is exceeded. In the actual calculation, we set $k = 300$, $dt = 0.01$, and the threshold of the relative position to be 88% quantile of all relative positions. The above approach for processing the cell groups instead of cells themselves is to reduce the data noise by local averaging.

Data availability

In this study, we used the following public tscRNA-seq datasets from scSLAM-seq Erhard et al. (2019), scNT-seq Qiu et al. (2020), sci-fate Cao et al. (2020) and scEU-seq Battich et al. (2020). These datasets can be downloaded directly through the Python package Dynamo.

Code availability

Storm is implemented in Python and is available at `https://github.com/aristoteleo/CSP4ML`.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

TL and XQ designed the research. QP performed the research. All of the authors analyzed the data and wrote the paper.

Acknowledgements

We thank Prof. Fang Yao for helpful discussions. TL and QP acknowledge the support from NSFC and MSTC under Grant No.s 11825102, 12288101 and 2021YFA1003300.

References

Battich, N., Beumer, J., de Barbanson, B., Krenning, L., Baron, C. S., Tanenbaum, M. E., Clevers, H., and van Oudenaarden, A. (2020). Sequencing metabolically labeled transcripts in single cells reveals mrna turnover strategies. *Science*, 367(6482):1151–1156.

Benhamou, E. and Melot, V. (2018). Seven proofs of the pearson chi-squared independence test and its graphical interpretation. *arXiv preprint arXiv:1808.09171*.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414.

Bergen, V., Soldatov, R. A., Kharchenko, P. V., and Theis, F. J. (2021). Rna velocitycurrent challenges and future perspectives. *Molecular systems biology*, 17(8):e10282.

Cao, J., Zhou, W., Steemers, F., Trapnell, C., and Shendure, J. (2020). Sci-fate characterizes the dynamics of gene expression in single cells. *Nature biotechnology*, 38(8):980–988.

Chao, H. X., Fakhreddin, R. I., Shimerov, H. K., Kedziora, K. M., Kumar, R. J., Perez, J., Limas, J. C., Grant, G. D., Cook, J. G., Gupta, G. P., et al. (2019). Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Molecular systems biology*, 15(3):e8604.

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326.

Cui, H., Maan, H., and Wang, B. (2022). Deepvelo: Deep learning extends rna velocity to multi-lineage systems with cell-specific kinetics. *bioRxiv*.

Erhard, F., Baptista, M. A., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., and Dölken, L. (2019). scslam-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423.

Farrell, S., Mani, M., and Goyal, S. (2022). Inferring single-cell transcriptomic dynamics with structured latent gene expression dynamics. *bioRxiv*.

Gao, M., Qiao, C., and Huang, Y. (2022). Unitvelo: temporally unified rna velocity reinforces single-cell trajectory inference. *Nature Communications*, 13(1):6586.

Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A. M., Theis, F. J., and Yosef, N. (2022). Deep generative modeling of transcriptional dynamics for rna velocity analysis in single cells. *bioRxiv*.

Gorin, G., Svensson, V., and Pachter, L. (2020). Protein velocity and acceleration from single-cell multiomics experiments. *Genome biology*, 21(1):1–6.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640.

Gu, Y., Blaauw, D., and Welch, J. D. (2022). Bayesian inference of rna velocity from multi-lineage single-cell data. *bioRxiv*, pages 2022–07.

Hendriks, G.-J., Jung, L. A., Larsson, A. J., Lidschreiber, M., Andersson Forsman, O., Lidschreiber, K., Cramer, P., and Sandberg, R. (2019). Nasc-seq monitors rna synthesis in single cells. *Nature communications*, 10(1):1–9.

Koehler, K. J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75(370):336–344.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). Rna velocity of single cells. *Nature*, 560(7719):494–498.

Li, C., Virgilio, M. C., Collins, K. L., and Welch, J. D. (2022). Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature Biotechnology*, pages 1–12.

Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K. W., Le, N.-T., Abe, J.-i., Cooke, J. P., and Wang, G. (2023). A relay velocity model infers cell-dependent rna velocity. *Nature Biotechnology*, pages 1–10.

Li, T., Shi, J., Wu, Y., and Zhou, P. (2021). On the mathematics of rna velocity i: Theoretical analysis. *CSIAM Transactions on Applied Mathematics*, 2(1):1–55.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G., and Wu, H. (2020). Massively parallel and time-resolved rna sequencing in single cells with scnt-seq. *Nature methods*, 17(10):991–1001.

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Min, K. H. J., Wang, L., et al. (2022). Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711.

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E. R., Botrugno, O. A., et al. (2022). Chromatin velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*, 40(2):235–244.

Vallender, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Zhang, J., Zhong, W., and Ma, P. (2021). A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, pages 279–300.

Additional Files

Additional file 1 — Supplementary information with supplemental Figures S1-S3.