Review

RNA-velocity: An Example of Model aided Analysis

Lumi Zhang
12112618

Abstract

RNA splicing dynamics is a powerful indicator of individual cell states. Single-cell RNA sequencing can capture the distinct number of spliced and un-spliced molecules with high throughput, accuracy and sensitivity. But the single-cell snapshot of RNA transcription losses the temporal component, making inference of developmental trajectory challenging. Here, we reviewed a model for RNA splicing dynamics called RNA velocity, which is a set of differential equations used to infer the developmental trajectory from single-cell RNA sequencing data. By distinguishing spliced and un-spliced molecules, the model infers the velocity of the cell, and provide insight on the possible next stages of development. With RNA velocity, single-cell RNA sequencing analysis can provide more valuable information on developmental lineages and cellular dynamics. The application of this simple model in sequencing analysis also showed the great potential of systems biology.

Introduction

Single-cell RNA sequencing (scRNA-seq) techniques have enabled researchers to capture snapshots of cellular transcription. With these snapshots, differences in gene expression level during cell differentiation and development can be examined at an unprecedented level. Sequenced cells can be clustered into different groups representing different cell types after performing dimension-reduction with algorithms such as UMAP (uniform manifold approximation and projection) or principal component analysis (PCA) (Luecken & Theis, 2019). Different cluster of cells can be annotated with cell type gene markers, and the level of gene expression can be studied. The technique of scRNA-seq have provided more insights into the biological processes on the cellular level.

Developmental trajectory of cells is an important aspect of cell development, with the information of developmental trajectory, cell fates could be inferred. Since scRNA-seq only provides a snapshot of the cells, temporal information of the cells are lost. Other methods are needed to infer the developmental trajectory from data obtained with scRNA-seq. RNA velocity is a well-known model for trajectory inference, which have a firm theoretical background in systems biology. Base on the relationship of un-spliced and spliced RNAs, RNA velocity uses a system of differential equations to explain changes in RNA composition. This model can infer a satisfying developmental trajectory of cells from scRNA-seq data, and is widely used over the last years for developmental studies.

Developmental trajectory of cells can be inferred from a simple model from systems biology.

In a system like the cell, the change in the concentration of a molecule is governed by its production and degradation rate, as described by a classic differential equation in systems biology called the dynamic equation (Alon, 2019):

$$\frac{dX}{dt} = \alpha - \beta X$$

Where α, β are the production and degradation rate of the molecule of interest X. This equation assumes a constant production rate of the molecule. The degradation rate is proportional to the concentration of the molecule.

Mature mRNA undergoes a process called splicing in eukaryotic cells where certain parts called 'introns' are cut and removed from the transcribed RNA. The process of RNA splicing can also be fitted into the dynamic model (Fig. 1a). Driven by a time-dependent transcription rate $\alpha(t)$, unspliced mRNAs are produced and spliced at a rate β. Spliced mRNAs later undergo degradation at a rate γ. The expected concentration of un-spliced (μ_u) and spliced (μ_s) reads are thus governed by the following ordinary differential equations:

$$\frac{d\mu_u}{dt} = \alpha(t) - \beta\mu_u$$

$$\frac{d\mu_s}{dt} = \beta\mu_u - \gamma\mu_s = v$$

In this system, mRNAs are assumed not to be degraded, and the splicing and degradation rate are constants. The RNA velocity v is the change in abundance of spliced mRNA. Fig. 1b showed a calculated phase portrait of an imaginary gene with $\alpha = 1, \beta = 1, \gamma = 0.6$. During induction, the phase portrait of the gene lies above the steady state line of $\mathrm{u} = \mathrm{s}\,\gamma/\beta$. When the phase arrives at the stead state, the number of spliced molecules doesn't change. If the gene is repressed, the phase portrait of the gene goes below the steady state.

These equations together described a dynamical model of mRNA splicing. Using this dynamical model, the amount of spliced mRNAs can be inferred for a given time. Thus RNA velocity can be used to infer the next state of the cell. In Fig. 1c-e, we demonstrated the inferred RNA velocity of cells from a chommaffin scRNA-seq data (La Manno et al., 2018). The inferred developmental trajectory fits the

ground truth, which is the development from schwann cell precursors (SCPs) to chromaffin cells. A gene Chga which is know to be highly expressed in chromaffin cells shows phase portrait similar to the induction model. Another gene named Serpine2, which had been proved to be expressed in SCPs but not chromaffin cells also showed matching phase portrait in both induction and repression portraits. These examples showed that the model can be used to determine the developmental trajectory of the cell.

RNA velocity inferred from scRNA-seq data.

To determine the rate coefficients in the RNA velocity model, we need the sequencing data from scRNA-seq (Fig. 2a). In the first step of the analysis, raw data are processed to distinguish spliced and un-spliced molecules. This step includes aligning the sequenced reads in fastq format to the reference genome to obtain the mapped reads in bam format. This mapping result is fed into the command line tool *velocyto* (La Manno et al., 2018) to distinguish the reads. Different scRNA-seq methods were compared under the same pipeline, and all have 20% of the reads being un-spliced mRNAs. This result indicates that we can successfully identify un-spliced reads (Fig. 2b), and use them for further analysis.

With the count matrix for spliced and un-spliced molecules for each gene, phase portrait of these genes can be inferred. In the 2018 paper by La Manno et al., *velocyto* used a quantile-based method to infer the rate coefficients. By assuming the cells in high and low quantiles as steady state cells for a gene, rate coefficients could be caluculated efficiently (Fig, 2c). Since *velocyto* assumes $\beta = 1$, the remaining two rate constants all have clear numerical solutions as shown below:

$$\gamma = \frac{u}{s}, \alpha = u$$

These steady state assumption is sufficient for velocity inference. However, the assumption of steady state cells caused debates. In 2020, Bergen et al. introduced a new likelihood-based dynamical model. By maximizing the likelihood function, the model iterates through possible values of the rate coefficients and finally obtain a result with satisfying performance (Fig. 2d).

RNA velocity are often calculated with a group of cells clustered with n-nearest neighbour to reduce computational costs. The resulting velocity is a large vector with length equal to the total number of genes. To visualize the velocity, the velocity vector need to be embedded. If we perform dimensionality reduction with linear methods like PCA for the cells, the velocity can be embedded into the lower diension. But PCA often doesn't return good clustering result for the cells. In scRNA-seq analysis, the non-linear embedding method like UMAP is often used. In this case, cells clustered around the end of the velocity vector are used to infer a low dimension direction for the velocity, as described with the following equation:

$$V_i = \sum_{q=1}^{k} (E(s_q) - E(s_i)) w(s_q - s_i, \Delta s_i)$$

For a cell cluster i, all the neighbour q of the inferred state are considered. The difference between the two clusters is compared to the inferred difference and passed into a softmax function to obtain a weight. Weighed cluster location difference after dimensionality reduction are averaged for all the neighbours. The mean is assigned to be the velocity vector in the lower dimension. Thus, the velocity of cells could be plotted on a clustering plot after non-linear dimension reduction for development trajectory inference.

RNA velocity model have limitations.

Despite successful applications in the scRNA-seq field, the RNA velocity model still have some drawbacks due to limits of sequencing technologies and false assumptions. The first problem due to technical limits is the confusion caused by alternative splicing of genomic regions. scRNA-seq data cannot be accurately assigned to a gene after alternative splicing, since the un-spliced molecule is the same for the resulting genes.

A modelling problem worth mentioning is the existence of genes with abnormal phase portraits. In all the publications, phase portraits were shown for a limited amount of genes. These chosen phase portraits fit the calculated phase portrait well. However, using the chromaffin data again, phase portrait for the top five marker genes can be plotted. These marker genes are highly expressed in the corresponding unique type of cell (Fig. 3a), and should have a phase portrait similar to the induction line in Fig. 1b. By observing the phase portrait of these genes (Fig. 3b), some of them seem to follow the trend of the phase portrait, while most genes behave unexpectedly. For most genes, the number of spliced and un-spliced molecules vary significantly. This indicates a very small γ/β value, indicating the degradation rate is much slower than splicing. However, the RNA velocity phase portrait with small γ/β value ($\alpha = \beta = 1, \gamma = 0.01$) in Fig. 3c is still different from the actual phase portraits of these genes. The fitting quality of these type of genes are worrying, and their contribution to the calculation of velocity is unknown. These genes might be pure noise and are not important for velocity inference, or there are possible improvements for the RNA velocity model. Since there is currently no accurate model for fitting assessment, a more general criteria is required to assess the fitting of sequenced data.

The RNA velocity model assumed constant and sudden transcription initiation rate change across all genes. But gene activation is not a sudden process, but a continuous process. Logistic functions such as the Hill function can be used to approximate the continuous activation procedure, as demonstrated with the following equation:

$$\alpha(t) = \alpha_0 \frac{t^n}{K^n + t^n}$$

Where α_0 is the final trascription rate, K is the activation constant which influence the activation time, and n is the rate constant for activation which influence the rate of activation. Although this approximation might not be accurate, but the resulting plot can already show the power of assuming a

logistic activation function. As shown by an example set of parameters of $K = 1, n = 2, \alpha_0 = 1, \beta = 1, \gamma = 0.6$ in Fig. 4a, the phase portrait becomes more curly and is different from the phase portrait obtained from the simple RNA velocity model.

Other adjustments could be made to improve the model, but in the cost of extremely large calculation costs. In all the previous models, we assumed the genes initiate and terminate transcription at the same speed. But what if there is a difference between these two rates? For example, the gene transcription might be easily initiated, but takes a longer time to be repressed. In Fig. 4b, a larger K value of 2 is chosen for inhibition. This phase portrait is narrower, due to the slower rate of unspliced molecule production. The new phase portrait fits the gene expression phase portrait in scRNA-seq data better than the old RNA velocity model. But is computationally more demanding since K and α_0 need to be inferred from the data.

As a mathematical model, RNA velocity will never 100% explain all the data. When development trajectories are obtained from any method, the results need to be handled with care. In Fig. 4c, we see the different results of *scVelo* and *velocyto* on the same chromaffin dataset. In the *scVelo* output, cells in SCP have a general velocity flow towards the top left corner of the graph. This phenomenon is contradicting with the ground truth that SCPs are suppose to develope to the bridge cells and later the chromaffin cells. The reason for this type of error is unknown. The trajectory inference tasks with RNA velocity should be anlayzed and treated with care.

The bright future of systems biology aided biological studies.

RNA velocity is a great example of applying systems biology studies in biological studies. System biology can explain complex biological systems with elegant equations. RNA velocity showed us that these systems can be utilized to guide biological studies as well. Other systems biology fields such as promoter-transcription relations, protein-protein interactions can be used to design new and controllable transcription systems in cells, or new molecule sensing report proteins. With the development of artificial intelligence and biology, systems biology aided biological studies have a promising future.

References

Alon, U. (2019). An Introduction to Systems Biology: Design Principles of Biological Circuits (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429283321

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*, 1408–1414. doi:10.1038/s41587-020-0591-3

Bergen, V., Soldatov, R. A., Kharchenko, P. V., & Theis, F. J. (2021). RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, *17*. doi:10.15252/msb.202110282

Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled. *PLoS Computational Biology*, *18*. doi:10.1371/journal.pcbi.1010492

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*. doi:10.15252/msb.20188746

Manno, G. L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, Vol. 560, pp. 494–498. doi:10.1038/s41586-018-0414-6

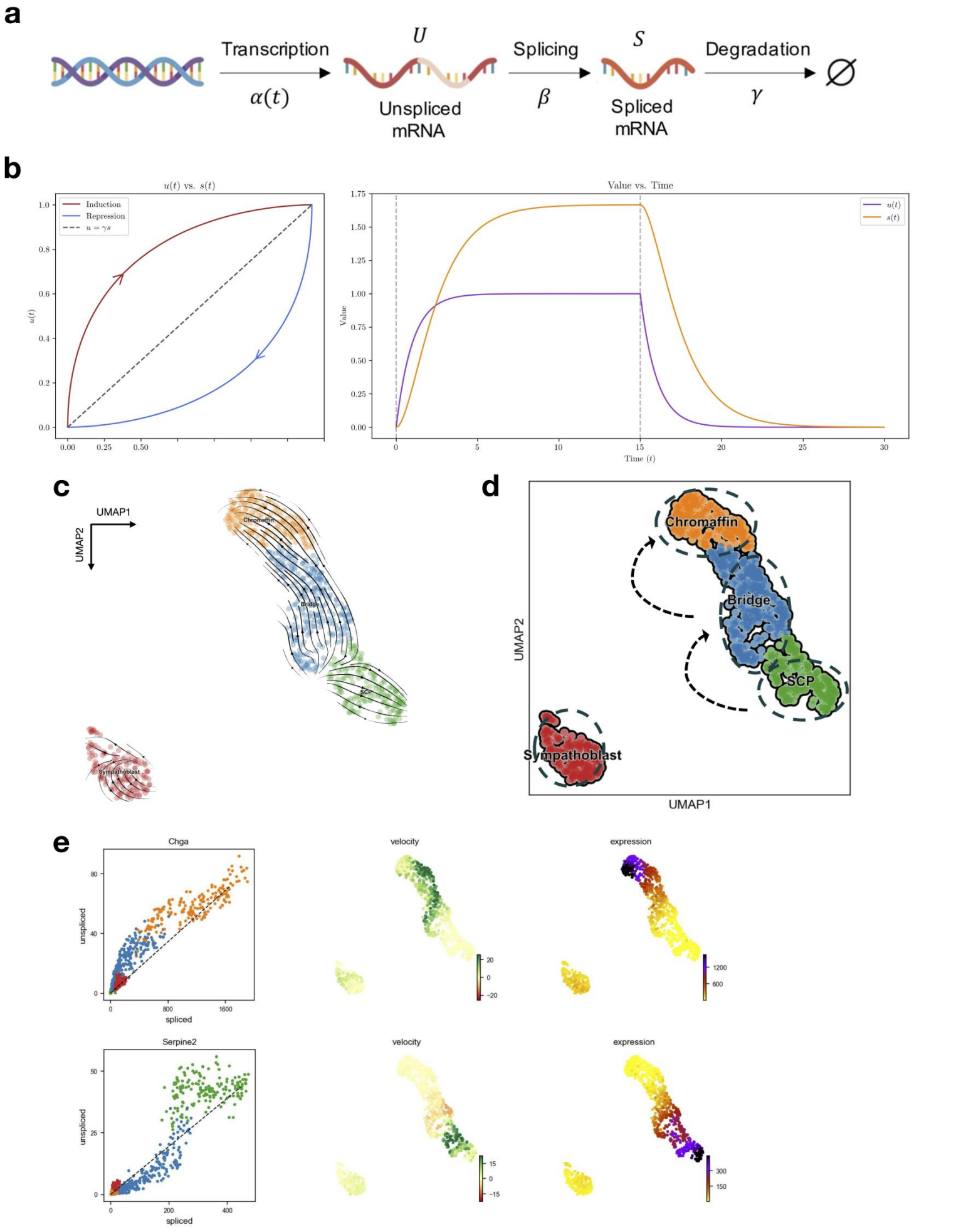

a
Transcription
$\alpha(t)$
U
Unspliced mRNA
Splicing
β
S
Spliced mRNA
Degradation
γ
b
$u(t)$ vs. $s(t)$
Induction
Repression
$u = \gamma s$
$u(t)$
Value vs. Time
$u(t)$
$s(t)$
Value
Time (t)
c
UMAP1
UMAP2
Chromaffin
Bridge
SCP
Sympathoblast
d
Chromaffin
Bridge
SCP
Sympathoblast
UMAP2
UMAP1
e
Chga
unspliced
spliced
velocity
expression
Serpine2
unspliced
spliced
velocity
expression

Fig.1 | RNA velocity model applied on chromaffin scRNA-seq data.

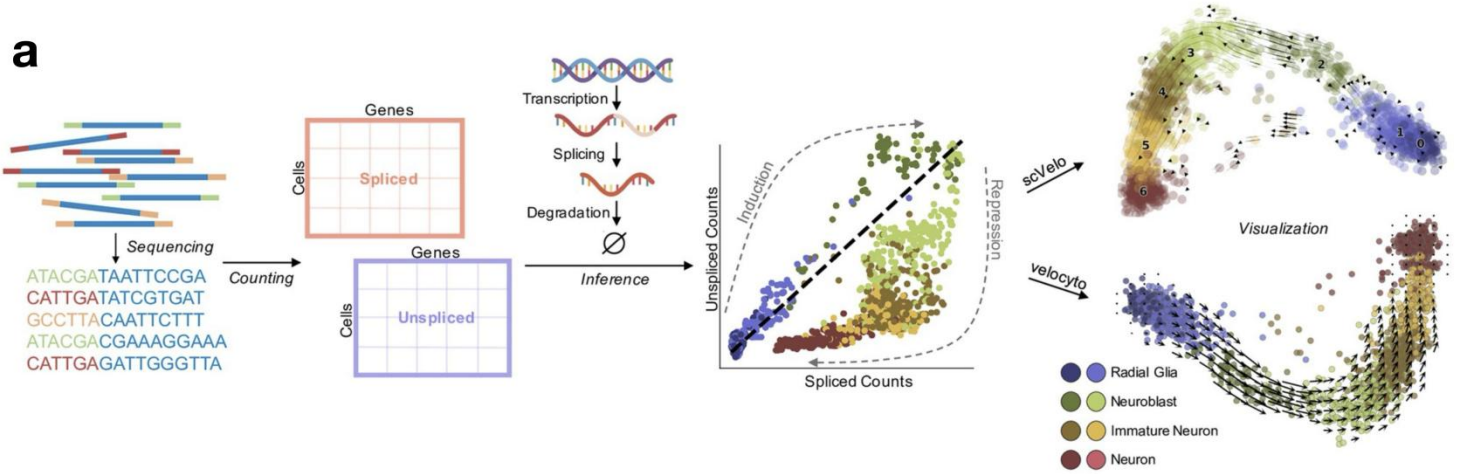
a, Transcription model of mRNA processing in cell involving transcription, splicing and degradation.

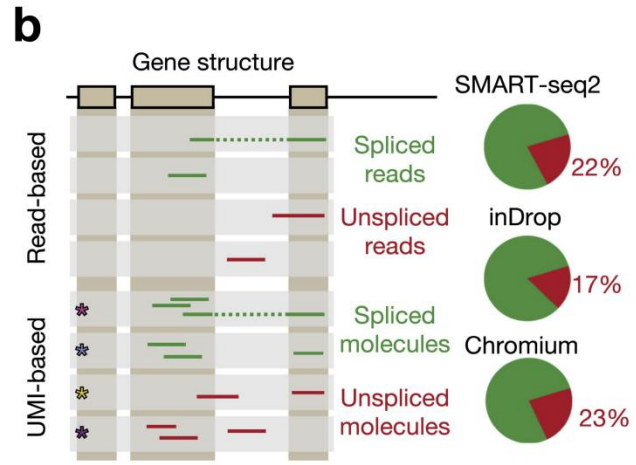
b, Plots of un-spliced and spliced RNA phase portrait with coefficients $\alpha = 1, \beta = 1, \gamma = 0.6$. Left, The phase portrait plot for the model under the assumption of $\beta = 1$ in all cells. Dashed lines indicates the steady state, arrows indicates the development of time. Right, Dashed line indicates the time of sudden transcription initiation and termination.

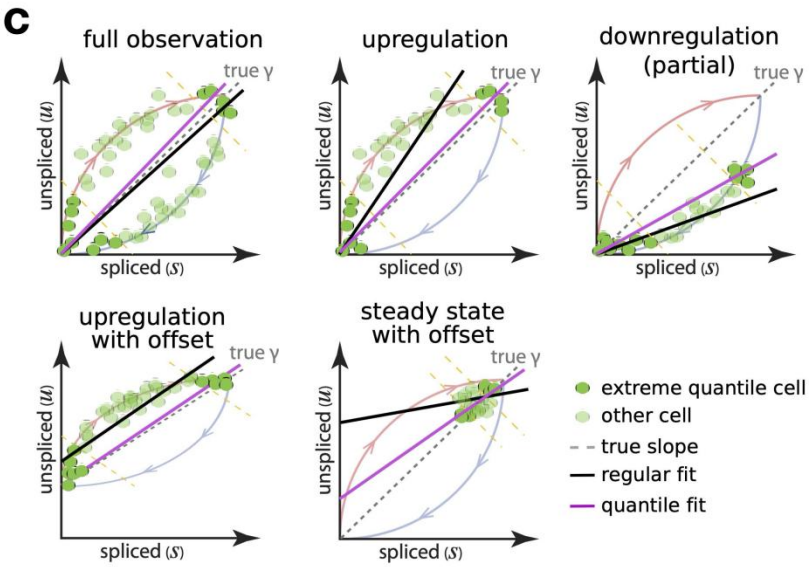
c, RNA velocity graph with *scVelo* of mouse chromaffin dataset (La Manno et al., 2018).

d, Manual cell-type annotation results for the clusters and ground-truth devlopmental knowledge indicated with dashed arrows. SCP: schwann cell precursors.

e, Gene expression portrait of Chga (up) and Serpine2 (down). Chga is expressed in chromaffin cells, and Serpine2 in schwann cell precursors.


a
Sequencing
ATACGATAATTCCGA
CATTGATATCGTGAT
GCCTTACAATTCTTT
ATACGACGAAAGGAAA
CATTGAGATTGGGTTA
Counting
Genes
Cells
Spliced
Genes
Cells
Unspliced
Transcription
Splicing
Degradation
Inference
Unspliced Counts
Induction
Repression
Spliced Counts
scVelo
velocyto
Visualization
Radial Glia
Neuroblast
Immature Neuron
Neuron


b
Gene structure
Read-based
UMI-based
Spliced reads
Unspliced reads
Spliced molecules
Unspliced molecules
SMART-seq2
22%
inDrop
17%
Chromium
23%


c
full observation
upregulation
downregulation (partial)
upregulation with offset
steady state with offset
true γ
unspliced (u)
spliced (s)
extreme quantile cell
other cell
true slope
regular fit
quantile fit

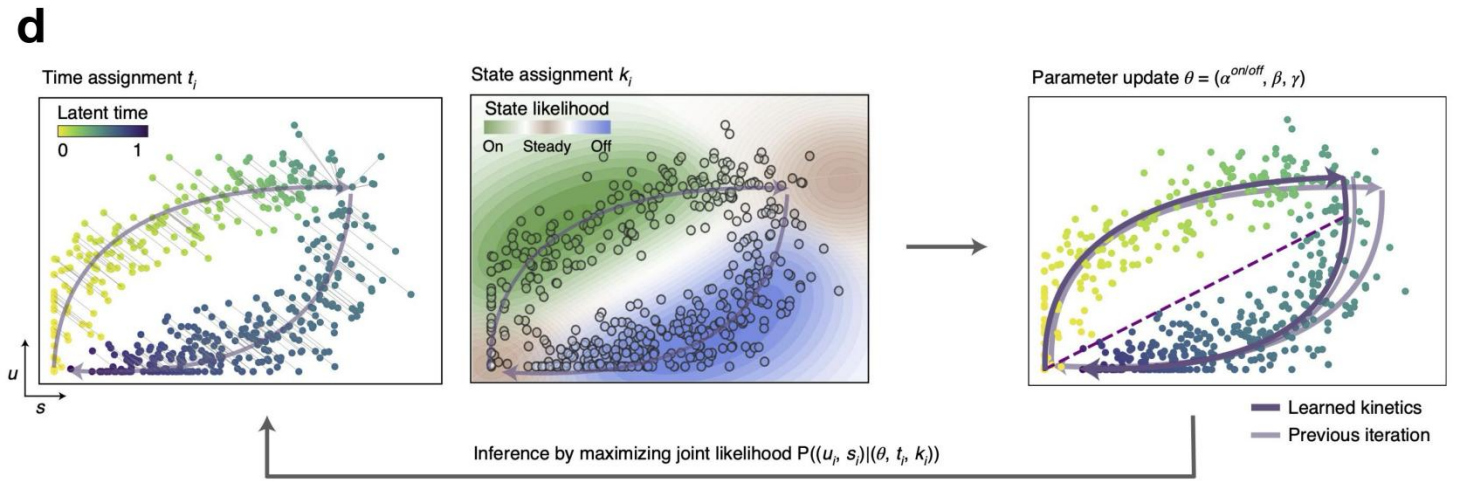

d
Time assignment t_i
Latent time
0
1
State assignment k_i
State likelihood
On
Steady
Off
Parameter update $\theta = (\alpha^{on/off}, \beta, \gamma)$
u
s
Learned kinetics
Previous iteration
Inference by maximizing joint likelihood $P((u_i, s_i)|(\theta, t_i, k_i))$

Fig.2 | RNA velocity inference from scRNA-seq data.

a, Summary graph of a typical RNA velocity workflow (Gorin et al., 2022).

b, Spliced and un-spliced reads are mapped to different part of the gene. Pie charts indicates the percentage of un-spliced genes in different sequencing technologies (LaManno et al., 2018).

c, Percentile fit used by LaManno et al. (2018). The pink line (quantile fit) works better than regular fit under all conditions. By assuming a steady state at the extreme quantiles, rate coefficients can be found mathematically.

d, Maximum likelihood method for rate coefficient determination. Parameters of rate coefficients are updated iteratively with inferred time and state based on the spliced and unspliced gene counts from sequencing data (Bergen et al., 2020).

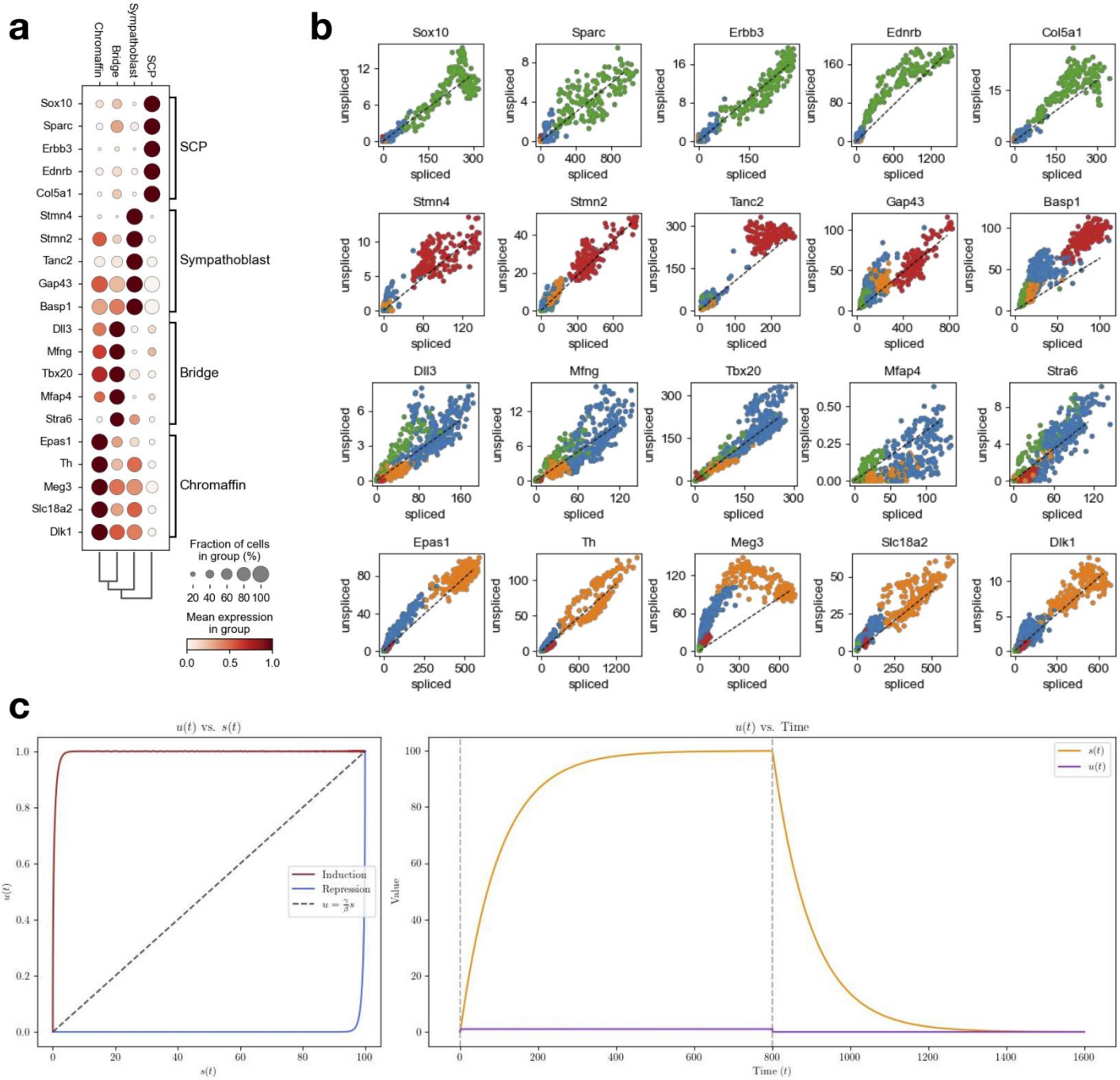

a
Chromaffin
Bridge
Sympathoblast
SCP
Sox10
Sparc
Erbb3
Ednrb
Col5a1
Stmn4
Stmn2
Tanc2
Gap43
Basp1
Dll3
Mfng
Tbx20
Mfap4
Stra6
Epas1
Th
Meg3
Slc18a2
Dlk1
SCP
Sympathoblast
Bridge
Chromaffin
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0.0 0.5 1.0
b
Sox10
Sparc
Erbb3
Ednrb
Col5a1
Stmn4
Stmn2
Tanc2
Gap43
Basp1
Dll3
Mfng
Tbx20
Mfap4
Stra6
Epas1
Th
Meg3
Slc18a2
Dlk1
unspliced
spliced
c
u(t) vs. s(t)
Induction
Repression
u = (β/γ)s
u(t)
s(t)
u(t) vs. Time
s(t)
u(t)
Value
Time (t)

Fig. 3 | Phase portraits of marker genes in different cells.

a, Marker genes for differen cell type clusters. These genes are the top expressed genes in each cluster, as the heatmap shows.

b, The phase portrait of all the marker genes. Most of them do not seem to follow the expected phase portrait graph.

c, Phase portrait with rate coefficients $\alpha = 1, \beta = 1, \gamma = 0.01$. The phase portrait is nearly a square, which still cannot explain the phase portrait of marker genes.

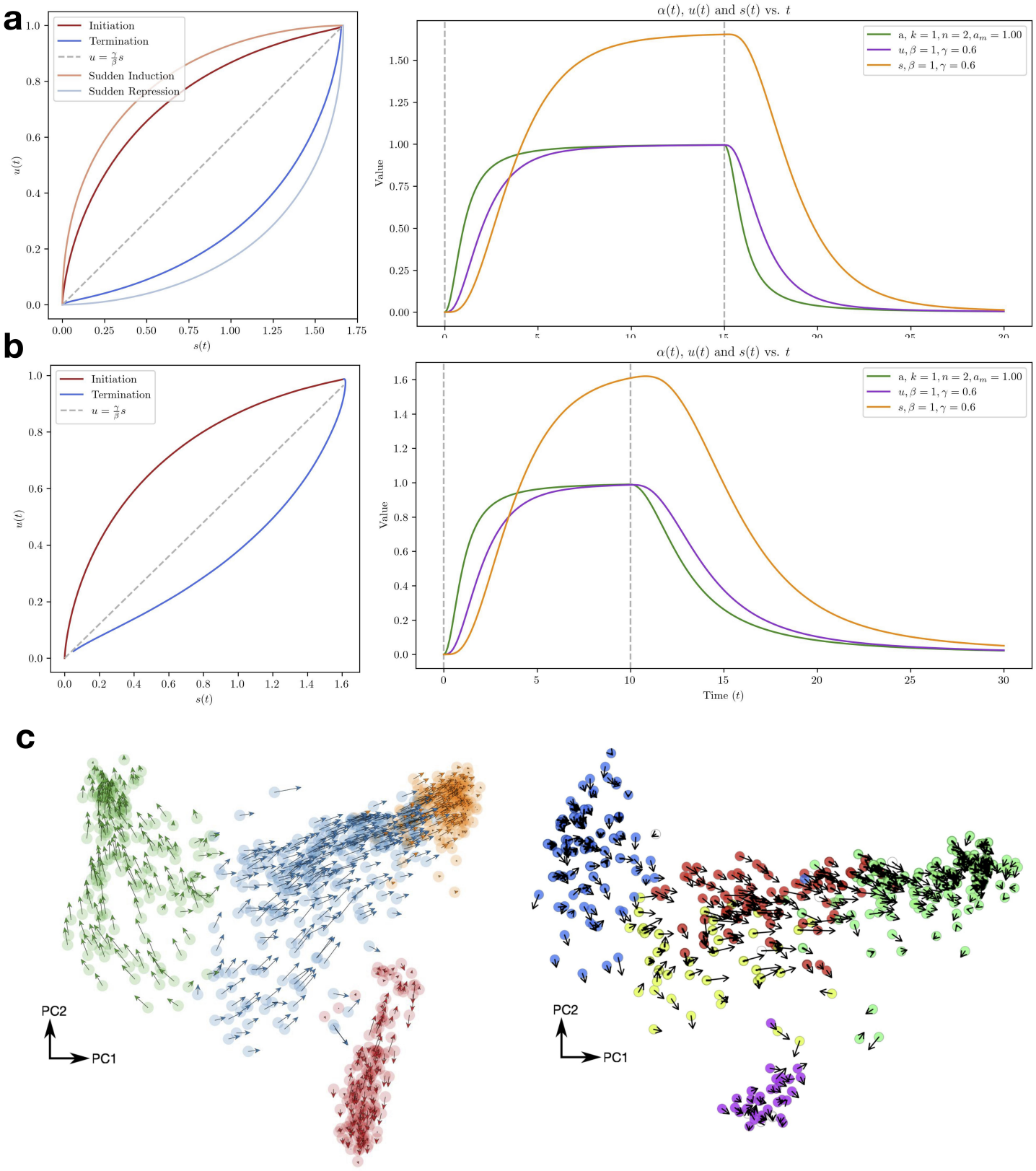

a
Initiation
Termination
$u = \frac{\gamma}{\beta}s$
Sudden Induction
Sudden Repression
$u(t)$
$s(t)$
$\alpha(t)$, $u(t)$ and $s(t)$ vs. t
a, $k = 1, n = 2, a_m = 1.00$
$u, \beta = 1, \gamma = 0.6$
$s, \beta = 1, \gamma = 0.6$
Value
b
Initiation
Termination
$u = \frac{\gamma}{\beta}s$
$u(t)$
$s(t)$
$\alpha(t)$, $u(t)$ and $s(t)$ vs. t
a, $k = 1, n = 2, a_m = 1.00$
$u, \beta = 1, \gamma = 0.6$
$s, \beta = 1, \gamma = 0.6$
Value
Time (t)
c
PC2
PC1
PC2
PC1

Fig. 4 | Improvements and other problems regarding RNA velocity.

a, Phase portrait of Hill function transcription rate with $K = 1, n = 2$. Rate constants are the same as in Fig. 1b. Fainter line in left figure shows phase portrait with constant transcription rate as in Fig. 1b.

b, Phase portrait of Hill function transcription rate with different activation and inhibition rate ($K_1 = 1, K_2 = 10, n = 2$).

c, Chromaffin data analyzed with *scVelo* and *velocyto.* SCP analyzed with *scVelo* (green cluster) showed abnormal velocity.

Code Availability:

Codes for plotting the phase portrait graphs and for *scVelo* analysis can be found in the GitHub repository RNA-velocity_Review (github.com/BHAAA-ZLM/RNA-velocity_Review), which is still updating.

Mouse chromaffin data was obtained from NCBI under the session series GSE99933. Methods for *scVelo* analysis can be found in the GitHub repository.