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 inferrence of developmental trajectory challenging. Here, we reviewed a model of RNA splicing dynamics called RNA velocity, which is 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 aided studies.

# Introduction

Single-cell RNA sequencing (scRNA-seq) techniques have enabled researchers to capture cellular transcription snapshots. 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 of specific cell types can be studied. The technique of scRNA-seq have provided more insights into the biological process of cell development.

Developmental trajectory of cells is an important aspect of cell development, with the information of developmental trajectory, we can infer the fates of cells. Since scRNA-seq only provides a snapshot of the cell, temporal information of the cell is lost. We need other methods to infer the developmental trajectory for data obtained from scRNA-seq. RNA velocity is a well-known model for trajectory inference, which have a firm theoretical background of systems biology. Base on the relationship of pre-mRNAs and mature mRNAs, RNA velocity can give a satisfying inference of the developmental trajectory of cells from scRNA-seq data.

# 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):

descript

Where descript are the production and degradation rate of our molecule of interest descript.

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 this model. Driven by a time-dependent transcription rate descript, unspliced mRNAs are produced and spliced at a rate descript. Spliced mRNAs later undergo degradation at a rate descript (Fig. 1a). The expected concentration of un-spliced (descript) and spliced (descript) reads are thus governed by the following ordinary differential equations:

descript

descript

In this system, mRNAs are assumed not to be degraded, and the splicing and degradation rate are constants. The RNA velocity descript is the change in abundance of spliced mRNA.

These equations together described a dynamical model of mRNA splicing. Using these equations, the amount of spliced mRNA can be inferred at a given time. The RNA velocity can also 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. 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. Phase portrait in SCPs corresponds to induction phase, and in chromaffin cells it corresponds to repression phase. These examples showed that the model can be used to determine the developmental trajectory of the cell.

# RNA velocity inferred from scRNA-seq data.

In the first step of RNA velocity 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 can be used to distinguish spliced and un-spliced molecules using the command line tool *velocyto* (La Manno et al., 2018). Different sequencing methods were analyzed with the same pipeline, and all returned a similar result of 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). With steady-state population of cells, descript. Since *velocyto* assumes descript, the remaining two constants all have clear numerical solutions as shown below:

descript

These steady state assumption is sufficient for velocity inference. However, the assumption of steady state cells are not always accurate. 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 best performance. Both models can return a rather accurate approximation for the equations, and thus can calculate the RNA velocity for trajectory inference.

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 the dimension 1 times total number of genes. To visualize the velocity, we need to reduce the dimension of the velocity vector. If we perform dimensionality reduction with linear methods like PCA, the velocity can be linearly transformed into the lower diension. But PCA often doesn't return good clustering result for the cells. In scRNA-seq, the most commonly used dimensionality reduction method is UMAP, which is non-linear. Here we used the cells clustered around the end of the velocity vector to infer a low dimension direction for the velocity, as described with the following equation:

descript

For a cell cluster descript, all the neighbours descript of the inferred state is considered. The difference between the two clusters was compared to the inferred difference and passed into a softmax function to obtain a weight. Cell cluster location difference after dimensionality reduction times weight are averaged for all the neighbours and outputs the proposed velocity vector in the lower dimension. Thus, the velocity of cells could be plotted on a clustering plot after non-linear dimension reduction.

# 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.

Another problem worth mentioning is the existance of genes with abnormal phase portraits. In all the publications, phase portraits were shown for a limited amount of genes. These chose phase portraits seem to fit the calculated phase portrait well. However, using the chromaffin data again, we can plot the phase portrait for the top five marker genes which are highly expressed in each cell type. These marker genes are highly expressed in a 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 descript value, indicating the degradation rate is much slower than splicing. However, the RNA velocity phase portrait with small descript value (descript) 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 imporvements for the RNA velocity model. Since there is currently no accurate model for fitting acessment, 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:

descript

Where descript is the final trascription rate, descript is the activation constant which influence the activation time, and descript is the rate constant for activation which influence the rate of activation. As shown by an example set of parameters of descript in Fig. 4a, the phase potrait 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 terminated. In Fig. 4b, we set the descript value for inhibition to 2, and resulted in an interesting phase portrait. 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 descript and descript need to be inferred from the data.

As a mathematical model, RNA velocity will never 100% explain all the data. When using RNA velocity to perform trajectory inference tasks, 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. Indicating that 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 have aimed at understanding complex biological systems with elegant explainations. This field of 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 will thrive in the future.

# **References**

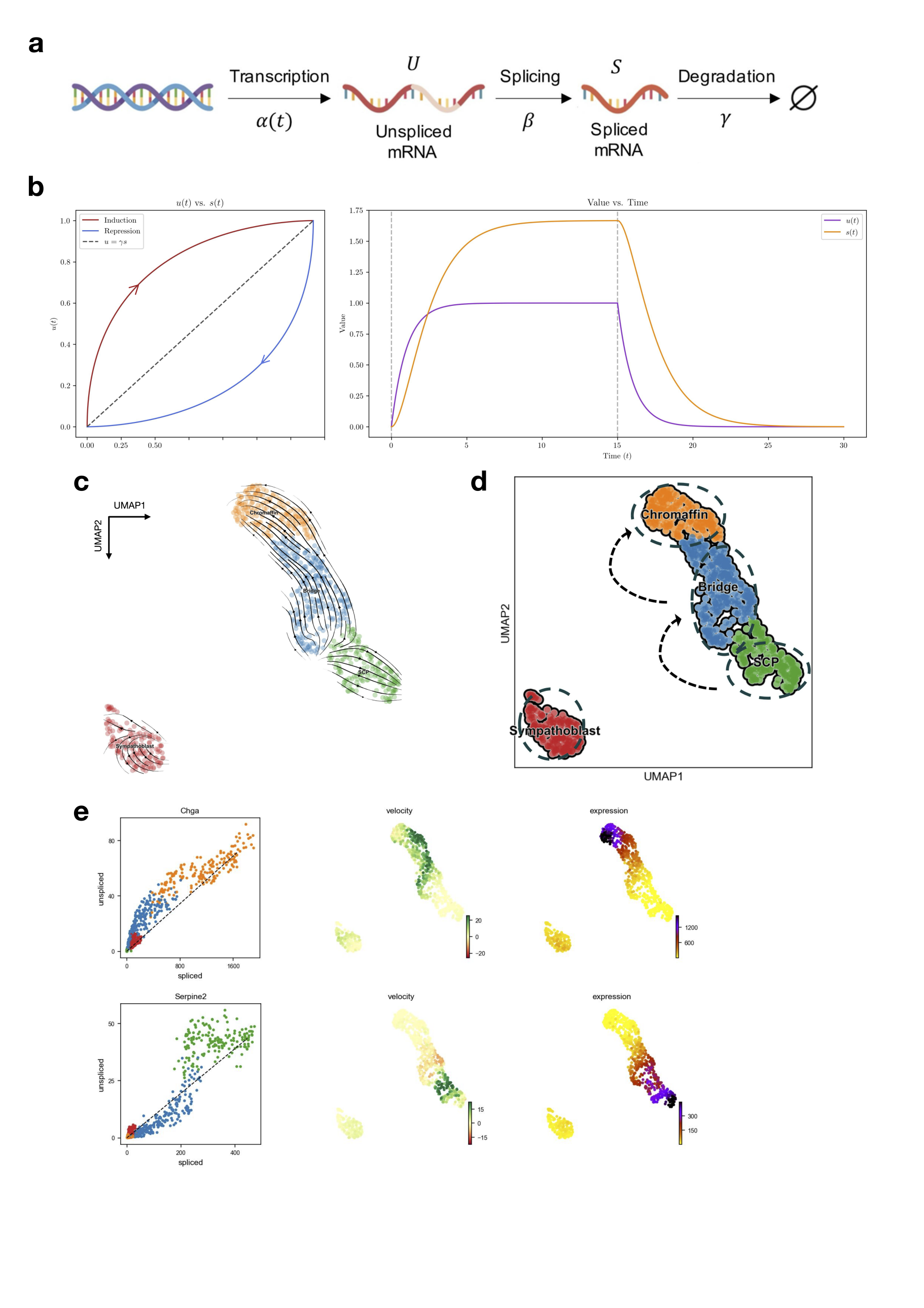
Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*, 1408–1414. doi:10.1038/s41587-020-0591-3

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



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

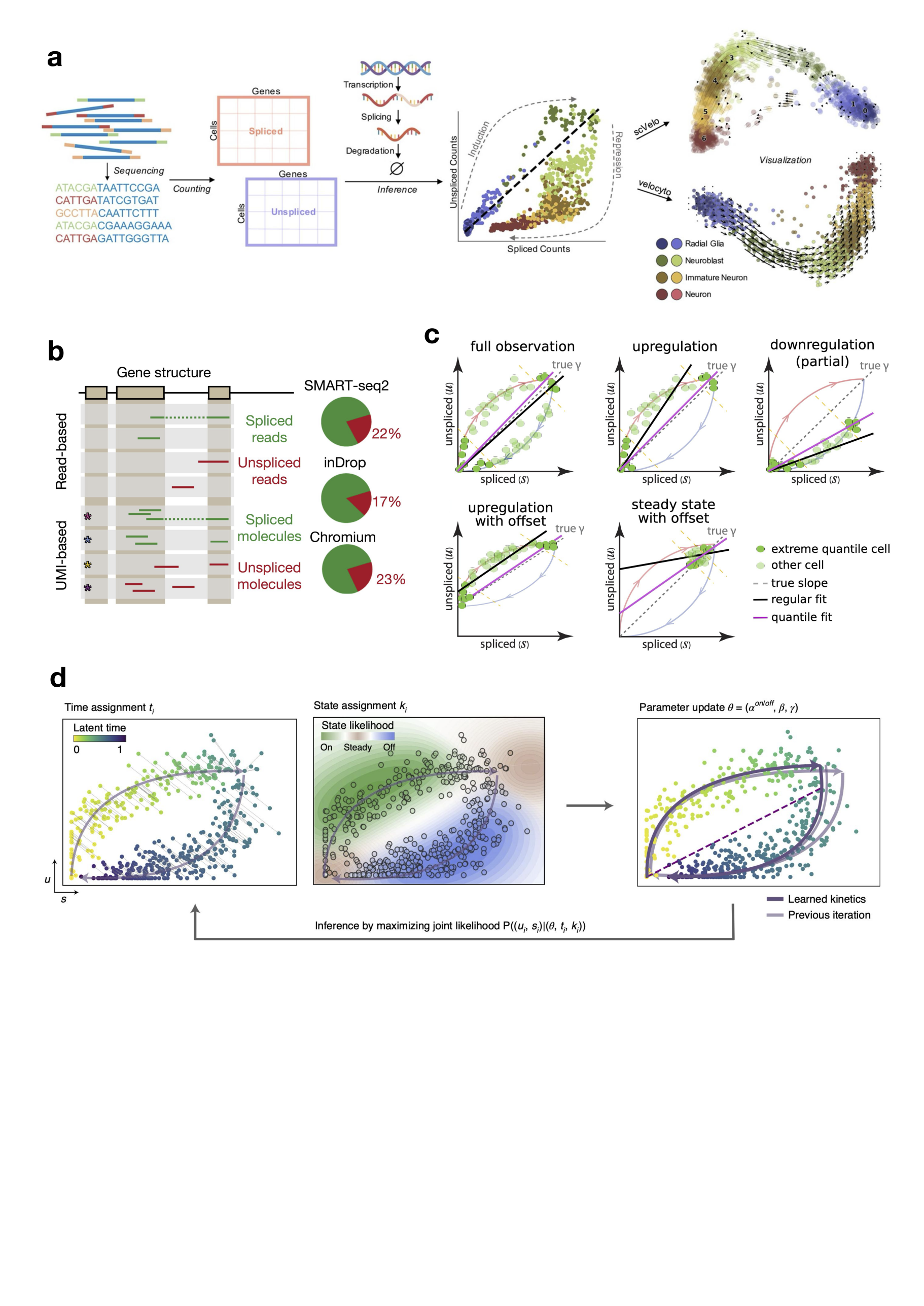
**a,** Transcription model of mRNA processing in cell involving transcription, splicing and degradation.

**b,** Plots of un-spliced and spliced RNA phase portrait with coefficients descript. Left, The phase portrait plot for the model under the assumption of *descript* 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.



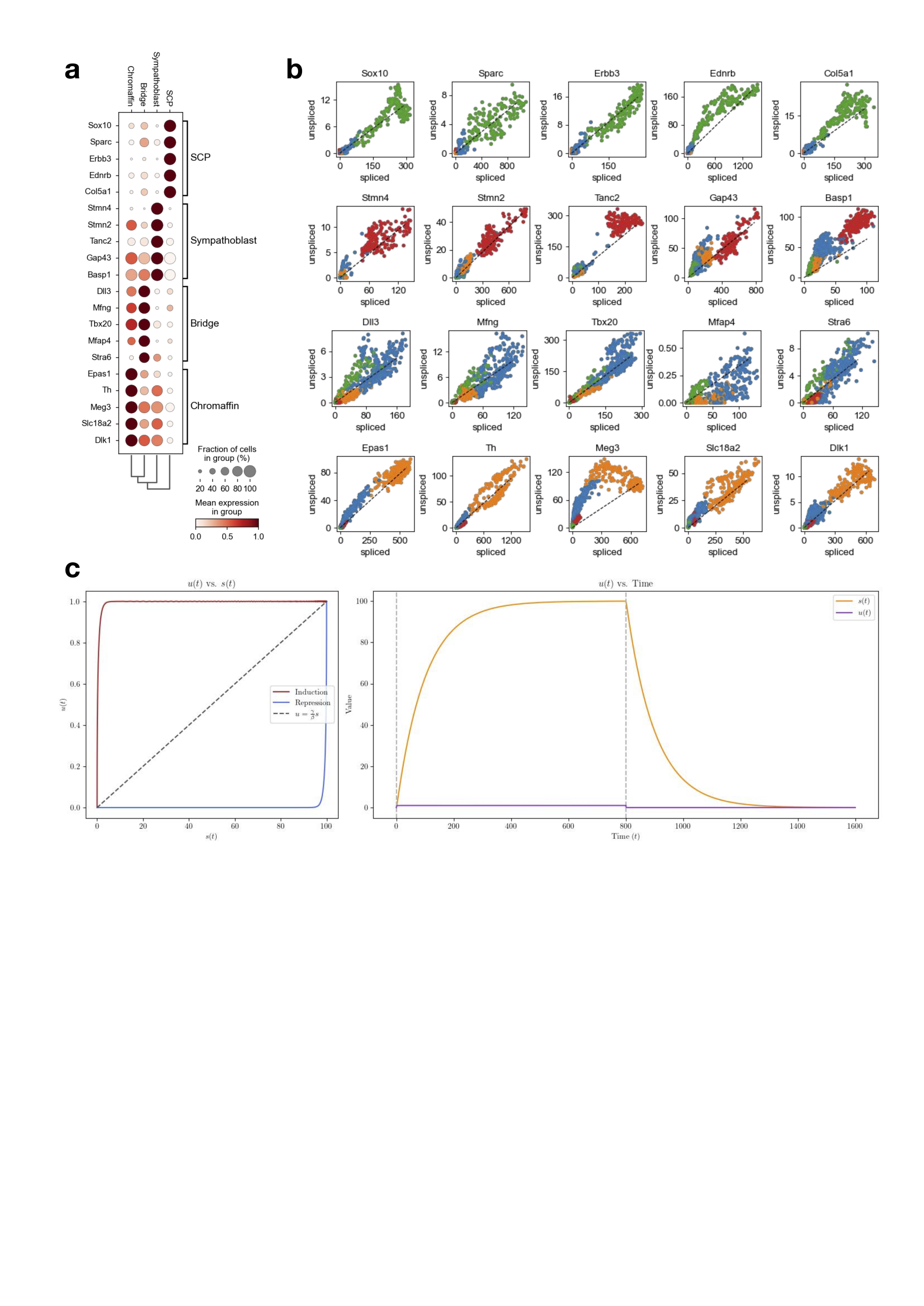
**Fig.2 | RNA velocity inference from scRNA-seq data.**

**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).

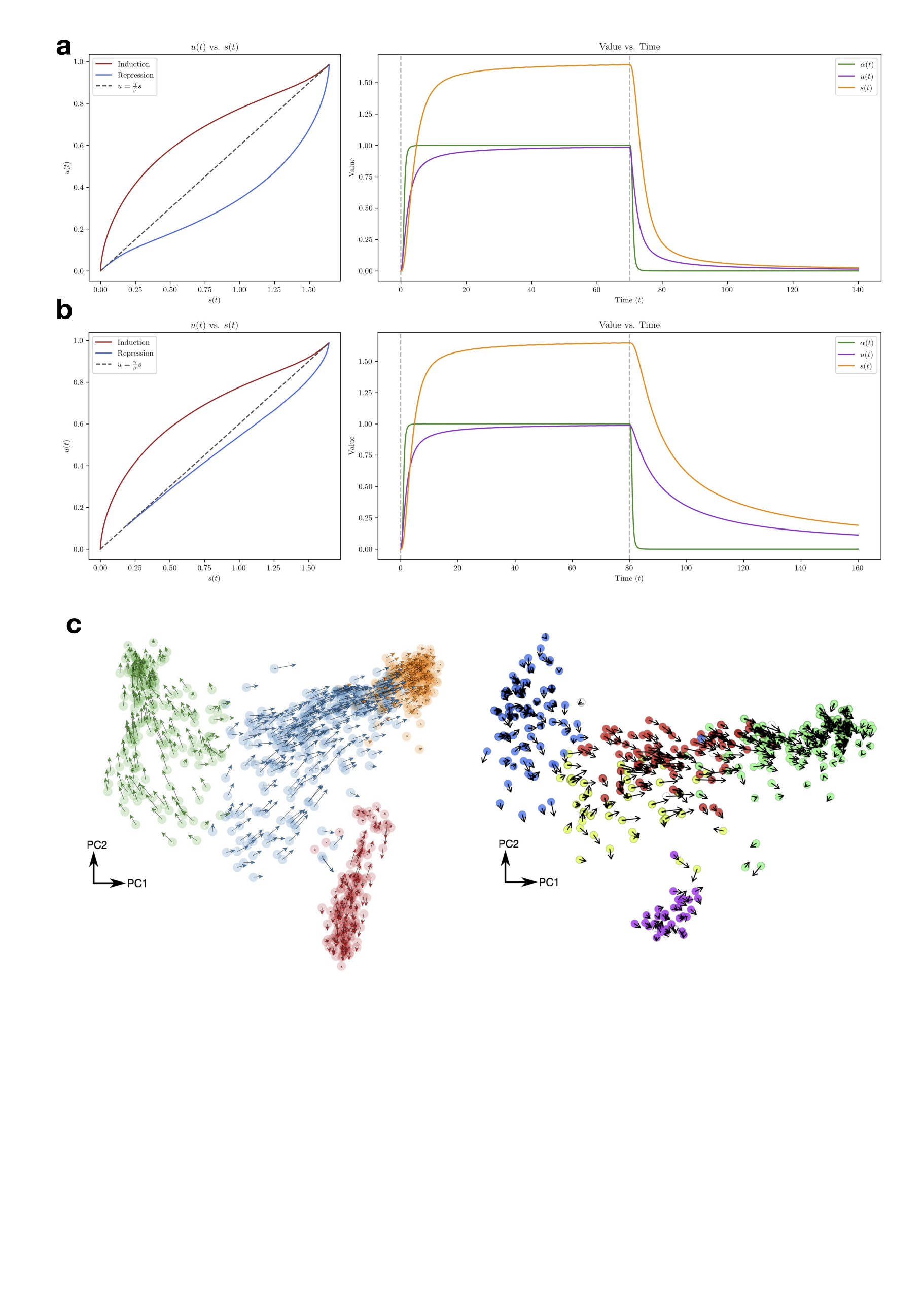


**Fig. 3 | Phase portraits of marker genes in different cells.**

**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 descript. The phase portrait is nearly a square, which still cannot explain the phase portrait of marker genes.



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

**a,** Phase portrait of Hill function transcription rate with descript. Rate constants are the same as in Fig. 1b.

**b,** Phase portrait of Hill function transcription rate with different activation and inhibition rate (descript).

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