



CellRank 2: unified fate mapping in multiview single-cell data

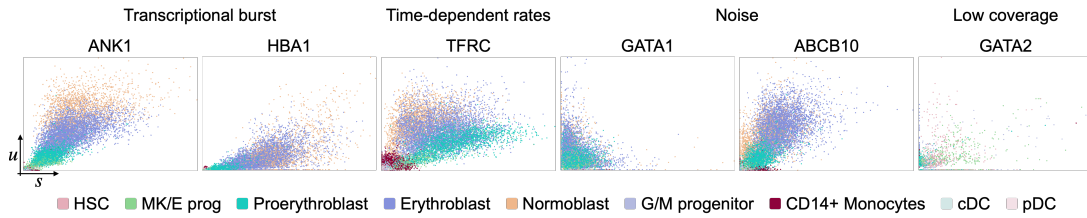
In the format provided by the authors and unedited

Contents

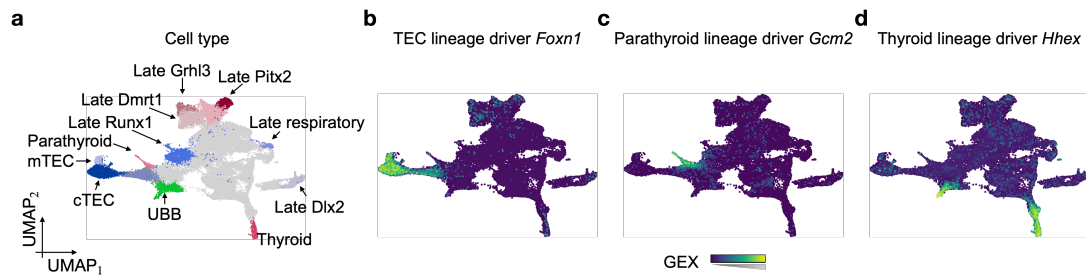
- 1 **Supplementary Figures** **2**

- 2 **Supplementary Note 1** **3**
 - 2.1 Limitations of RNA velocity 3

1 Supplementary Figures



Supplementary Figure 1: RNA velocity model assumptions violate transcriptional kinetics in adult hematopoiesis. Exemplary phase portraits of genes violating assumptions of the inference algorithm in scVelo’s *dynamical model*¹ on the adult hematopoiesis dataset² of Fig. 2. Transcriptional bursts (ANK1, HBA1) and time-dependent kinetic rates (TFRC) violate scVelo’s constant-rate assumption. Very high noise levels (GATA1, ABCB10) and low coverage in either spliced or unspliced transcripts (GATA2) render the inference problem ambiguous (Supplementary Note 1). Each dot is a cell colored by cell type (HSC: hematopoietic stem cell, MK/E prog: megakaryocyte/erythrocyte progenitors, G/M progenitor: Granulocyte/Myeloid progenitor, pDC: plasmacytoid dendritic cell, cDC2: classical dendritic cells) according to original cluster annotations²; the x-axis (y-axis) shows the abundance of spliced s (unspliced u) counts.



Supplementary Figure 2: Driver genes of pharyngeal endoderm development. **a.** UMAP embedding of pharyngeal endoderm development dataset³. Cells are colored according to the original study³. **b.-d.** UMAP embedding colored by CellRank 2-identified lineage drivers Foxn1 (TEC, **b.**), Gcm2 (parathyroid; **c.**), Hhex (thyroid; **d.**).

2 Supplementary Note 1

2.1 Limitations of RNA velocity

Although RNA velocity has been tremendously successful since its introduction⁴, several recent publications^{5–12} highlight the limitations of traditional RNA velocity models (velocyto⁴ and scVelo¹). These shortcomings include

- **Gene structure bias:** RNA velocity is a proxy for the current state of transcriptional regulation fundamentally based on faithfully recovering both spliced and unspliced counts from the same cell. However, recovering sufficient unspliced transcripts across many genes is difficult in practice since polyadenylation and splicing happen mostly simultaneously¹³. In fact, the majority of captured unspliced counts result from internal priming events where the poly(T) primers bind to poly(A) stretches in intronic regions of the unspliced transcript^{4,14,15}. Consequently, unspliced counts of a certain gene can, for example, depend on
 - the length of the gene¹⁴,
 - the number, length, and position of intronic regions within the gene⁵,
 - the amount of poly(A) stretches within introns, or
 - the relative position of poly(A) stretches within an intron.

Each case results in biases that are difficult to identify and impossible to control externally. Specifically, if key driver genes for a biological process lack sufficient unspliced counts for any of these reasons, the corresponding velocity estimates will most likely differ significantly from the ground truth. In this case, genes with large numbers of unspliced reads dominate the overall velocity of cells. Contrastingly, for spliced transcripts, the poly(T) primers bind to the poly(A) tail.

- **State and time-dependent reaction parameters:** RNA velocity is based on a model of the mRNA lifecycle in which transcripts are created at rate α , spliced at rate β , and degraded at rate γ . Traditional velocity models assume gene-specific constant rates during induction and repression, respectively. However, practical examples, including some applications from the original publications^{1,4}, have shown that both splicing and degradation rates vary across cell types and states^{5,16}. Further, transcription rates vary over time, for example, in the case of transcriptional bursting events during erythroid maturation⁸. Such time and state-dependent kinetic parameters violate assumptions of traditional RNA velocity models leading to velocity trends inconsistent with prior biological knowledge, for example, in hematopoiesis^{5,8}.
- **Gene-gene interactions:** Current velocity methods^{4,5} fit rate parameters in a gene-specific fashion. This assumption ignores the fundamentals of gene-gene regulation.
- **Batch effects:** RNA velocity is based on the ratio of spliced to unspliced counts. This ratio is not preserved across batches or data integration approaches. Thus, spliced and unspliced layers cannot be processed independently, making it difficult to infer RNA velocity for frequently generated atlas-scale datasets^{17,18}.
- **Noisy, stochastic expression kinetics:** State-of-the-art single-cell sequencing protocols yield noisy and sparse mRNA transcript counts^{19–21}. This aspect holds in particular for unspliced transcripts, which are harder to detect than their spliced counterparts. Currently, most RNA velocity approaches handle the high noise level by smoothing gene expression before inference^{1,4}, thereby ignoring the stochastic nature of gene expression.

- **Time scales of splicing:** Splicing kinetics determine the time delay between unspliced and spliced transcripts. However, this temporal delay cannot be controlled externally. Therefore, the time horizon over which RNA velocity extrapolates cellular states is fixed and may be incompatible with typical time scales in a biological system of interest.
- **Parameter non-identifiability:** RNA velocity is traditionally inferred based on a two-dimensional system of ordinary differential equations (ODEs). In the steady-state setting proposed by velocity⁴, the first method estimating RNA velocity, gene-specific parameters could only be estimated up to a multiplicative constant and are thus not comparable across genes. The scVelo¹ approach partially addressed this issue through dynamical modeling and post-hoc estimation of a gene-shared latent time. While scVelo’s inference scheme rendered kinetic parameters comparable across genes, it left a global multiplicative constant that precludes estimation of kinetic rates in their native units.

Several recently proposed models attempt to overcome some of these limitations and can be grouped as follows:

- **Model extension towards other modalities:** While the original RNA velocity model is based on the time delay between spliced and unspliced mRNA processing stages, emerging models include other molecular layers, including
 - mRNA and surface proteins²²,
 - mRNA and chromatin accessibility^{23–25},
 - open and closed chromatin²⁶,
 - new and total transcripts in single-cell metabolic labeling assays^{6,27–29}.

By including other modalities, these approaches relax the dependency on gene structure and try to estimate kinetic rate parameters more accurately through more inclusive models.

- **Stochastic models of splicing kinetics:** While both scVelo¹ and velocity⁴ model average smoothed expression kinetics, recent approaches^{14,30} formulate stochastic extensions to account for high noise levels in unspliced transcript abundance. However, these approaches are theoretical frameworks with no practical implementation.
- **Adapted velocity models:** scVelo and velocity solve a system of two coupled ODEs per gene. More recent approaches formulate different models for extracting directional information from the spliced/unspliced-count ratio. Approaches include radial basis function fitting³¹, a density-adaptive model to overcome parameter non-identifiability¹¹, and graph-convolutional neural networks to estimate state and time-dependent reaction parameters³².

Although these approaches address some of the velocity issues we outlined above and may be used directly through CellRank 2’s VelocityKernel, they often do so at the cost of introducing new heuristics unlikely to generalize across datasets. These heuristics include estimation based on scVelo’s steady-state model³² or assuming that time differences between two cell states are inversely proportional to the density of sampled cells that lie between them¹¹.

References

1. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*. <http://dx.doi.org/10.1038/s41587-020-0591-3> (2020).
2. Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L. & Pe'er, D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*. <http://dx.doi.org/10.1038/s41587-019-0068-4> (2019).
3. Magaletta, M. E., Lobo, M., Kernfeld, E. M., Aliee, H., Huey, J. D., Parsons, T. J., Theis, F. J. & Maehr, R. Integration of single-cell transcriptomes and chromatin landscapes reveals regulatory programs driving pharyngeal organ development. *Nature Communications*. <http://dx.doi.org/10.1038/s41467-022-28067-4> (2022).
4. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M. E., Lönnerberg, P., *et al.* RNA velocity of single cells. *Nature*. <http://dx.doi.org/10.1038/s41586-018-0414-6> (2018).
5. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*. <http://dx.doi.org/10.15252/msb.202110282> (2021).
6. Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi (Joseph) Min, K., *et al.* Mapping transcriptomic vector fields of single cells. *Cell*. <http://dx.doi.org/10.1016/j.cell.2021.12.045> (2022).
7. Sonesson, C., Srivastava, A., Patro, R. & Stadler, M. B. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology* (ed Li, M.) <http://dx.doi.org/10.1371/journal.pcbi.1008585> (2021).
8. Barile, M., Imaz-Rosshandler, I., Inzani, I., Ghazanfar, S., Nichols, J., Marioni, J. C., Guibentif, C. & Göttgens, B. Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology*. <http://dx.doi.org/10.1186/s13059-021-02414-y> (2021).
9. Weiler, P., Van den Berge, K., Street, K. & Tiberi, S. in *Single Cell Transcriptomics* (Springer US, 2022). http://dx.doi.org/10.1007/978-1-0716-2756-3_14.
10. Gorin, G. & Pachter, L. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophysical Reports*. <http://dx.doi.org/10.1016/j.bpr.2022.100097> (2023).
11. Marot-Lassauzaie, V., Bouman, B. J., Donaghy, F. D., Demerdash, Y., Essers, M. A. G. & Haghverdi, L. Towards reliable quantification of cell state velocities. *PLOS Computational Biology* (ed Li, W.) <http://dx.doi.org/10.1371/journal.pcbi.1010031> (2022).
12. Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A. & Hansen, K. D. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biology*. <http://dx.doi.org/10.1186/s13059-023-03065-x> (2023).
13. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA Processing with Transcription. *Cell*. [http://dx.doi.org/10.1016/S0092-8674\(02\)00617-7](http://dx.doi.org/10.1016/S0092-8674(02)00617-7) (2002).
14. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLOS Computational Biology* (ed Nie, Q.) <http://dx.doi.org/10.1371/journal.pcbi.1010492> (2022).
15. Selewa, A., Dohn, R., Eckart, H., Lozano, S., Xie, B., Gauchat, E., Elorbany, R., Rhodes, K., Burnett, J., *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Scientific Reports*. <http://dx.doi.org/10.1038/s41598-020-58327-6> (2020).
16. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell*. <http://dx.doi.org/10.1016/j.cell.2015.11.018> (2015).
17. Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madisson, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., *et al.* An integrated cell atlas of the lung in health and disease. *Nature Medicine*. <http://dx.doi.org/10.1038/s41591-023-02327-2> (2023).
18. Qiu, C., Cao, J., Martin, B. K., Li, T., Welsh, I. C., Srivatsan, S., Huang, X., Calderon, D., Noble, W. S., *et al.* Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nature Genetics*. <http://dx.doi.org/10.1038/s41588-022-01018-x> (2022).

19. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*. <http://dx.doi.org/10.1016/j.molcel.2017.01.023> (2017).
20. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*. <http://dx.doi.org/10.15252/msb.20188746> (2019).
21. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*. <http://dx.doi.org/10.1038/s41576-023-00586-w> (2023).
22. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*. <http://dx.doi.org/10.1186/s13059-020-1945-3> (2020).
23. Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*. <http://dx.doi.org/10.1016/j.cell.2020.09.056> (2020).
24. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature Biotechnology*. <http://dx.doi.org/10.1038/s41587-022-01476-y> (2022).
25. Burdziak, C., Zhao, C. J., Haviv, D., Alonso-Curbelo, D., Lowe, S. W. & Pe’er, D. scKINETICS: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics*. <http://dx.doi.org/10.1093/bioinformatics/btad267> (2023).
26. Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E. R., *et al.* Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*. <http://dx.doi.org/10.1038/s41587-021-01031-1> (2021).
27. Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., *et al.* scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*. <http://dx.doi.org/10.1038/s41586-019-1369-y> (2019).
28. Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G. & Wu, H. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*. <http://dx.doi.org/10.1038/s41592-020-0935-4> (2020).
29. Battich, N., Beumer, J., de Barbanson, B., Krenning, L., Baron, C. S., Tanenbaum, M. E., Clevers, H. & van Oudenaarden, A. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science*. <http://dx.doi.org/10.1126/science.aax3072> (2020).
30. Li, T. On the Mathematics of RNA Velocity I: Theoretical Analysis. *CSIAM Transactions on Applied Mathematics*. <http://dx.doi.org/10.4208/csiam-am.so-2020-0001> (2021).
31. Gao, M., Qiao, C. & Huang, Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nature Communications*. <http://dx.doi.org/10.1038/s41467-022-34188-7> (2022).
32. Cui, H., Maan, H., Vladioiu, M. C., Zhang, J., Taylor, M. D. & Wang, B. DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. *Genome Biology*. <http://dx.doi.org/10.1186/s13059-023-03148-9> (2024).