STAT 35510
Lecture 6

Spring, 2024

Jingshu Wang

Outline

- RNA velocity
  - Use RNA velocity to improve trajectory inference

RNA velocity

- RNA velocity (La Manno et. al. Nature 2018): the time derivative of the gene expression state

- Most scRNA-seq protocols can capture both spliced and unspliced mRNAs
- Cell observed at time t, abundance of spliced RNA at time $t+1$ can be predicted by the unspliced mRNA at time t
  - For a particular gene, assume

$$\frac{du}{dt} = \alpha(t) - \beta(t)\,u(t)$$

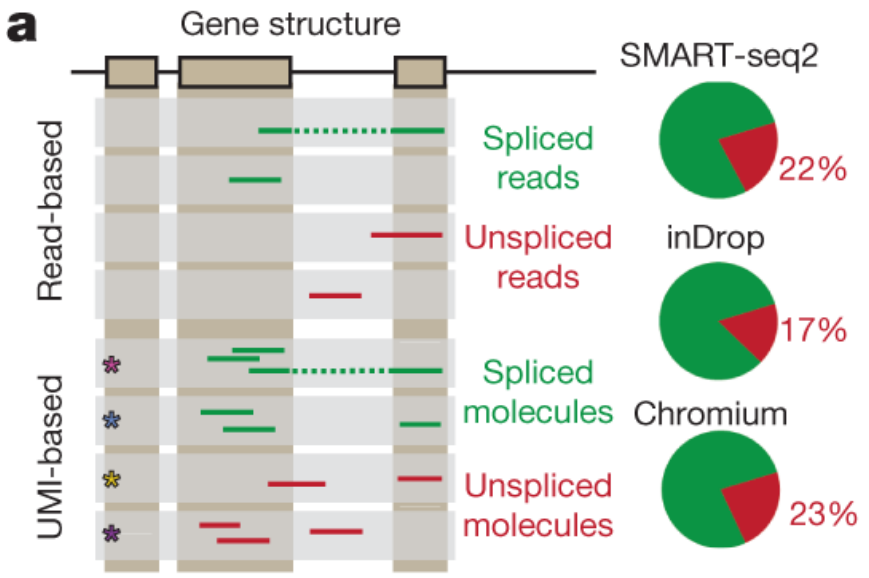
$$\frac{ds}{dt} = \beta(t)\,u(t) - \gamma(t)s(t)$$

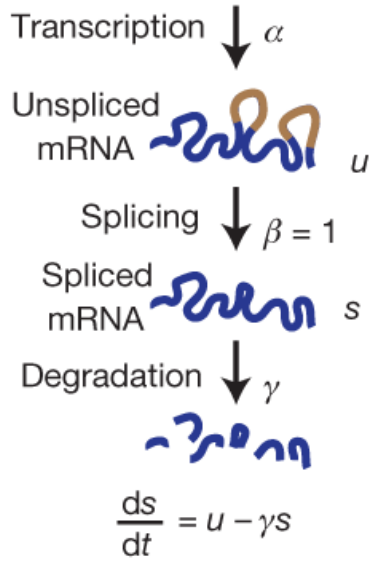
- Assuming $\beta(t) = 1$, $\alpha(t) \equiv \alpha$ and $\gamma(t) \equiv \gamma$, we have

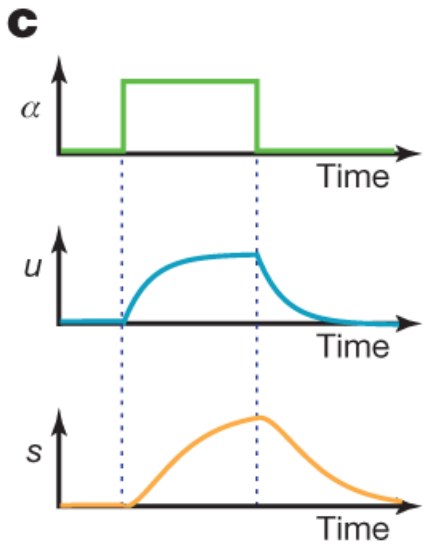
$$u(t) = \alpha(1 - e^{-t}) + u_0 e^{-t}$$

$$s(t) = \frac{e^{-t(1+\gamma)}\left[e^{t(1+\gamma)}\alpha(\gamma-1) + e^{t\gamma}(u_0-\alpha)\gamma + e^{t}(\alpha-\gamma(s_0+u_0+s_0\gamma))\right]}{\gamma(\gamma-1)}$$

- u(t) and s(t) are the expected (non-random) abundance, instead of the actual mRNA copies in the cell

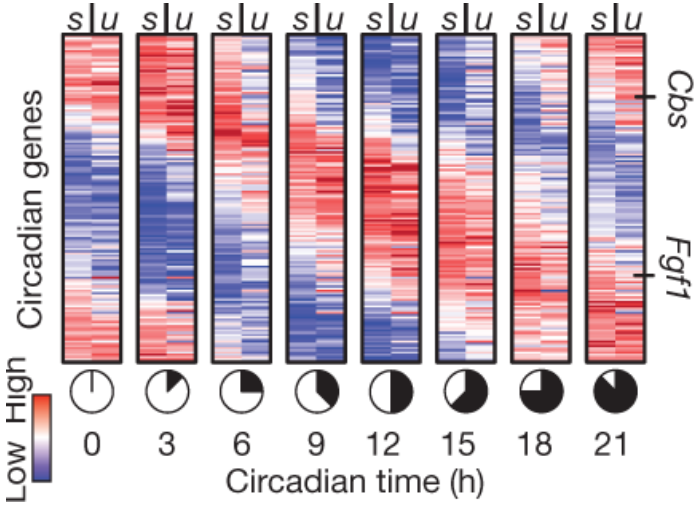




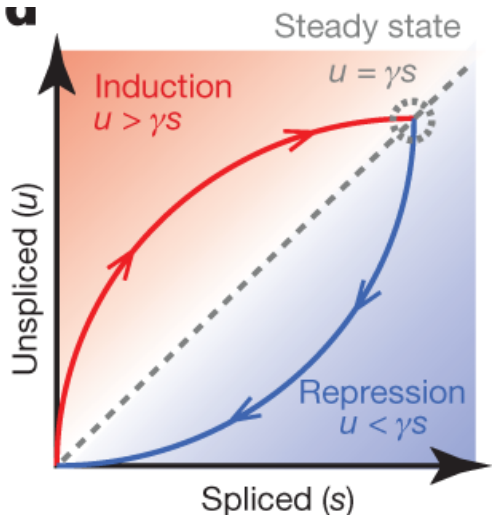


RNA velocity

- Amount of unspliced mRNAs can be predictive for the amount of spliced mRNA at the next time point
  - Intuition: more unspliced mRNA at time t, more spliced mRNA will be generated at time $t + 1$
  - On the other hand, $u(t)$ and $s(t)$ should be highly correlated across time



- Goal: predict the gene expression profile for any cell at the next time point
- Challenge: how to estimate the transcription rate and degradation rate?
  - Gene specific
  - May not be a constant over time



- Core idea: assume constant rates, if a cell is at the steady state (ds/dt = 0, du/dt = 0), then by definition

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

$$\alpha = \mathrm{u}$$

Velocyto (La Manno et. al. Nature 2018)

- Goal: For each cell, predict the amount of spliced mRNA for each gene at the next time point

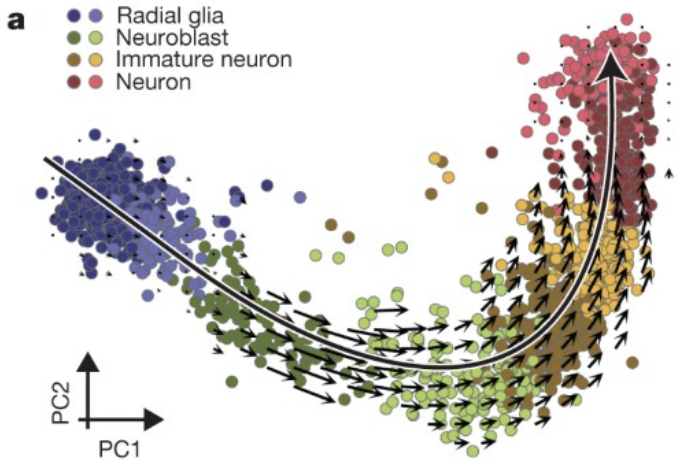
- Core ideas:
  - Assume that all genes have the same constant splicing rate ($\beta_g(t) \equiv 1$)
  - For each gene, estimate the degradation rate γ_g by linear regression regressing observed u on s
    - Only use "steady-state" cells: cells whose s are at the left/right tail of the gene expression
    - An offset for the intercept is introduced

  - Predict future $s_g(t)$ given initial (observed) s_{g0} and u_{g0}
    - Approximation 1: $\frac{ds_g(t)}{dt} \approx v_g \approx u_{g0} - \hat{\gamma}_g s_{g0}$, then $s_g(t) \approx s_{g0} + v_g t$
    - Approximation 2: $\frac{ds_g(t)}{dt} \approx u_{g0} - \hat{\gamma}_g s_g(t)$, then $s_g(t) \approx s_{g0} e^{-\hat{\gamma}_g t} + u_{g0}/\hat{\gamma}_g (1 - e^{-\hat{\gamma}_g t})$
    - Two approximations are similar when t is small, by default only predict $s_g(1)$
    - $v_g(t) = u_g(t)/s_g(t)$ are named as velocities (or $\frac{ds_g(t)}{dt}$)

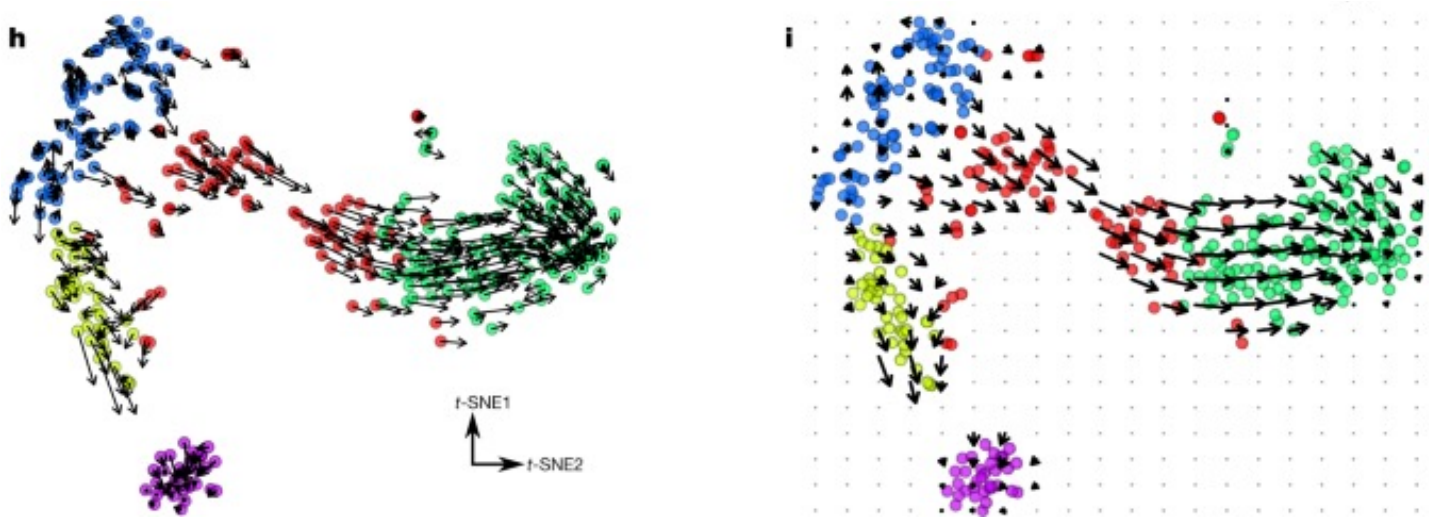
  - Strategies to improve accuracy in the prediction:
    - Pool over similar cells, pool over similar genes
    - Find cell whose spliced mRNA profiles are closest to the predicted profiles and build cell-cell pairs (cell j is the future state of cell i)

Velocyto (La Manno et. al. Nature 2018)

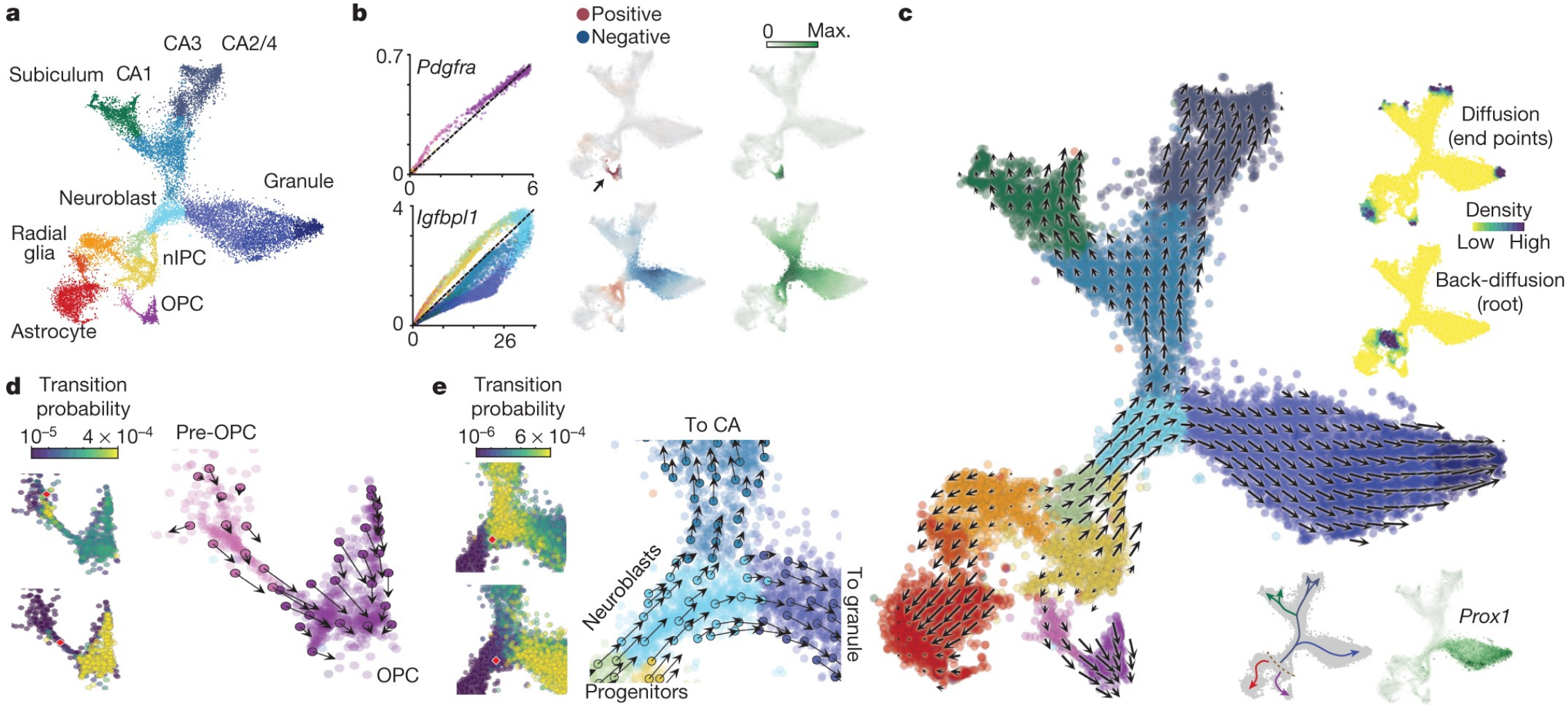
- Automatically identify directed cell lineages



- Improved visualization of the directed trajectory structure



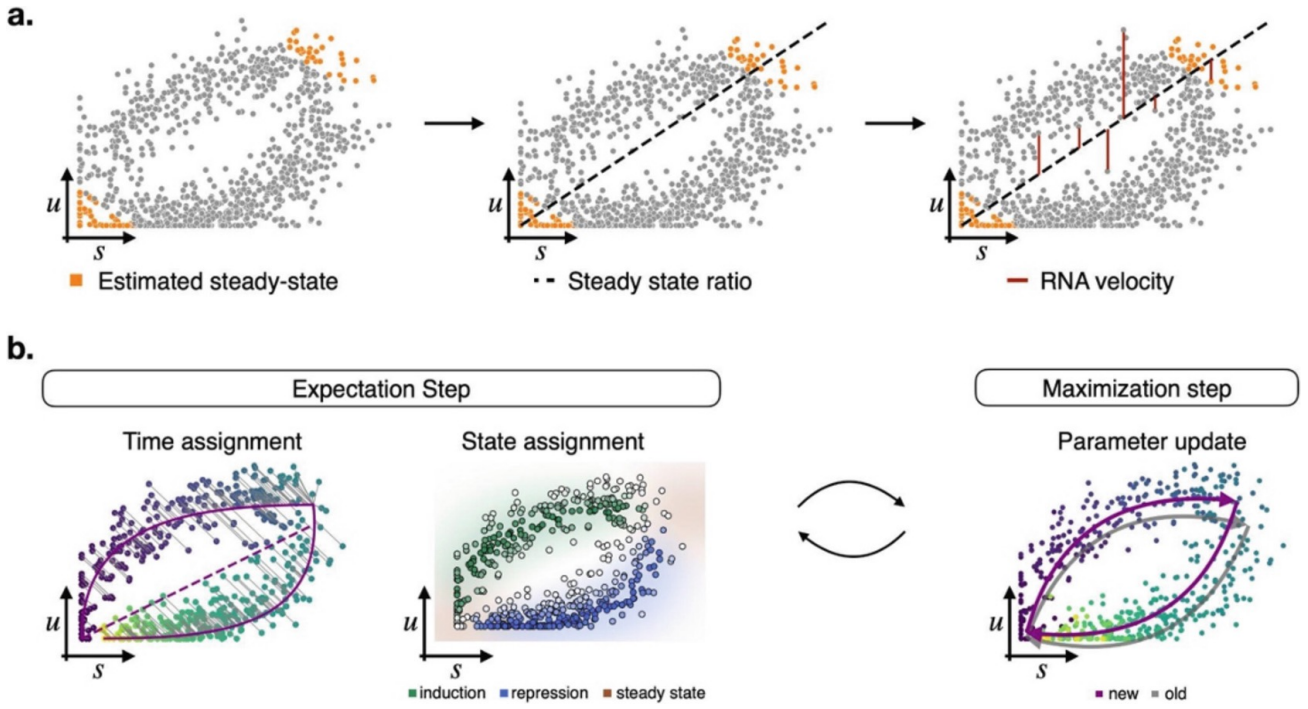
Velocyto (La Manno et. al. Nature 2018)



scVelo (Bergen et. al. Nature Biotech 2020)

- Velocyto heavily depend on steady-state modeling of the cells to simplify estimation
  - Select steady-state cells for estimating the gene-specific degradation rate
  - Assume that all cells have steady-state unspliced mRNAs ($u(t) \equiv u_{g0}$) in the prediction step

- scVelo assumes that the transcription rate $\alpha(t)$ is not constant, but can switch between k unknown values (latent states)
  - Incorporate the transcriptional bursting model (lecture 2) into the dynamics



scVelo (Bergen et. al. Nature Biotech 2020)

- Core ideas
  - Assume the following dynamic model

$$\frac{du(t)}{dt} = \alpha^{(k)}(t) - \beta u(t)$$
$$\frac{ds(t)}{dt} = \beta u(t) - \gamma s(t)$$

  - Four latent states: on, off and two steady states
  - Assume that each cell has a gene-specific latent time t_{ig}, and the observed (u_g, s_g) should be close to the model predicted $(u_g(t_{ig}), s_g(t_{ig}))$ for each cell and gene
    - Construct a likelihood of observed data given t_{ig}

  - Joint estimate $(t_{1g}, \cdots, t_{ng})$ and model parameters $(\alpha_g, \beta_g, \gamma_g, t_g^s)$ for all genes
    - Claimed using an EM algorithm assuming Gaussian data for each gene separately

  - Estimate the velocities of each gene and cell as $\hat{\beta}_g \hat{u}_g(\hat{t}_{ig}) - \hat{\gamma}_g \hat{s}_g(\hat{t}_{ig})$
  - Predicted spliced mRNA level at next time point: $s_g + \hat{\beta}_g \hat{u}_g(\hat{t}_{ig}) - \hat{\gamma}_g \hat{s}_g(\hat{t}_{ig})$

scVelo (Bergen et. al. Nature Biotech 2020)

- Some details of the four states model
  - Assume four transcriptional states changing sequentially:
    induction state, induction steady state, repression state and repression steady state
  - Denote the change point of each state as $t_{g0}^{(k)}$. $t_{g0}^{(1)} = 0$
  - Induction state:
    - Initialization: $u_{g1}^0 = 0, s_{g1}^0 = 0, \alpha_{g1} > 0$ and $t_{g1}^0 = 0$

$$\bar{u}^{(g)}(t_{ng}, k=1) := \frac{\alpha_{g1}}{\beta_g}\left(1 - e^{-\beta_g t_{ng}}\right)$$

$$\bar{s}^{(g)}(t_{ng}, k=1) := \frac{\alpha_{g1}}{\gamma_g}\left(1 - e^{-\gamma_g t_{ng}}\right) + \frac{\alpha_{g1}}{\gamma_g - \beta_g}\left(e^{-\gamma_g t_{ng}} - e^{-\beta_g t_{ng}}\right)$$

  - Induction steady state

$$\bar{u}^{(g)}(t_{ng}, k=1) := \lim_{t_{ng}\to\infty} \bar{u}^{(g)}(t_{ng}, k=1) = \frac{\alpha_{g1}}{\beta_g}$$

$$\bar{s}^{(g)}(t_{ng}, k=2) := \lim_{t_{ng}\to\infty} \bar{s}^{(g)}(t_{sg}, k=1) = \frac{\alpha_{g1}}{\gamma_g}.$$

scVelo (Bergen et. al. Nature Biotech 2020)

- Some details of the four states model
  - Induction state
  - Induction steady state
  - Repression state:

$$\alpha_{g3} = 0 \text{ and } t^0_{g3} = t^s_g$$

$$u^0_{g3} = \bar{u}^{(g)}(t_{sg}, k=2)$$

$$s^0_{g3} = \bar{s}^{(g)}(t_{sg}, k=2)$$

    - Change of $u_g(t)$ and $s_g(t)$

$$\bar{u}^{(g)}(t_{ng}, k=3) := u^0_{g3} e^{-\beta_g\left(t_{ng} - t^0_{g3}\right)}$$

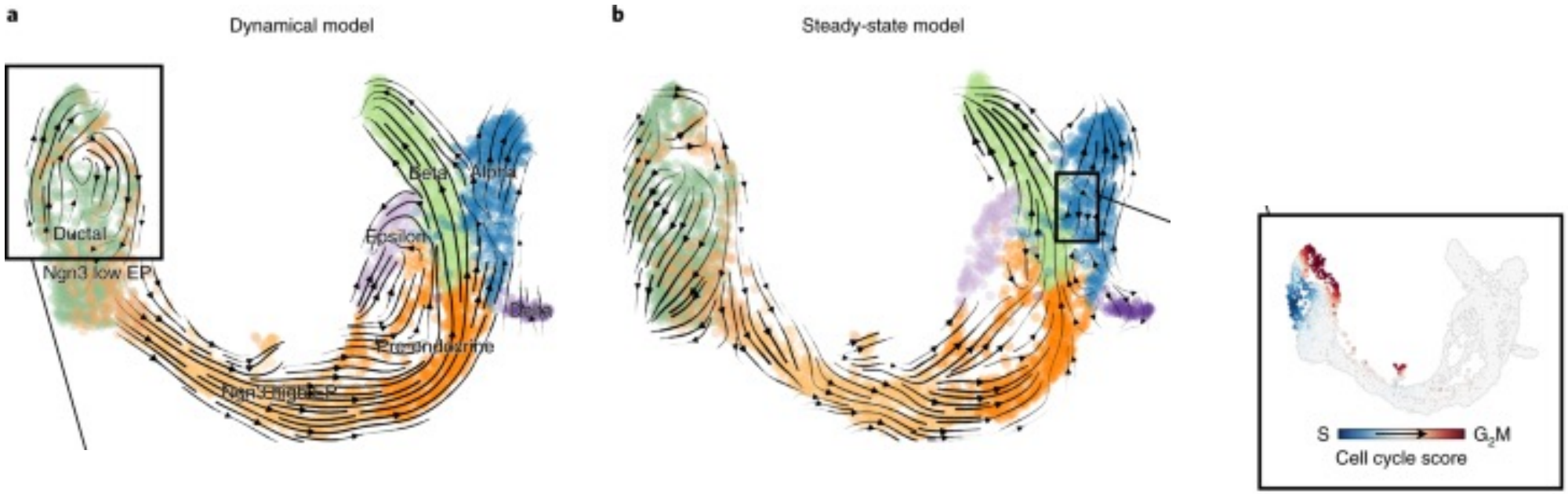
$$\bar{s}^{(g)}(t_{ng}, k=3) := s^0_{g3} e^{-\gamma_g\left(t_{ng} - t^0_{g3}\right)} - \frac{\beta_g u^0_{g3}}{\gamma_g - \beta_g}\left(e^{-\gamma_g \tau} - e^{-\beta_g (t_{ng} - t^0_{g3})}\right)$$

  - Repression steady state $\bar{u}^{(g)}(t_{ng}, k=4) := 0$

$$\bar{s}^{(g)}(t_{ng}, k=4) := 0$$

  - Only parameter for the change points is t^s_g

scVelo (Bergen et. al. Nature Biotech 2020)

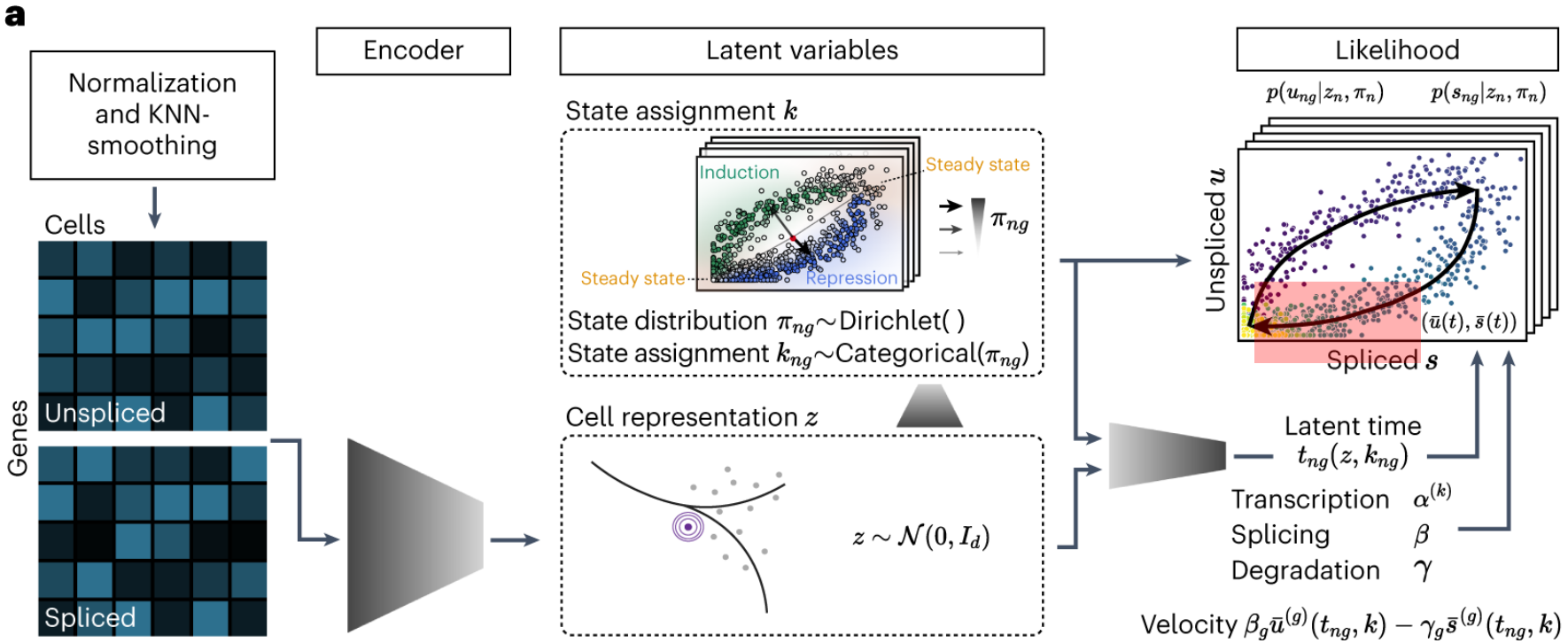


- (Gorin et. al. PLOS Computational Biology 2022) pointed out that the arrows generated by Velocyto or scVelo may reverse the true direction in worst case scenarios

veloVI (Li et. al. Nature Methods 2024)

- Solve the scVelo model four-state model with variational autoencoder

- Assume a shared latent space for all genes describing changes between four states
  - Use the latent variable to approximate posterior distribution of the gene-specific latent time t_{ig} for each cell i
  - Sample z to approximate the posterior distribution of the velocities
  - Has the flexibility to allow non-constant transcription rate in the induction state



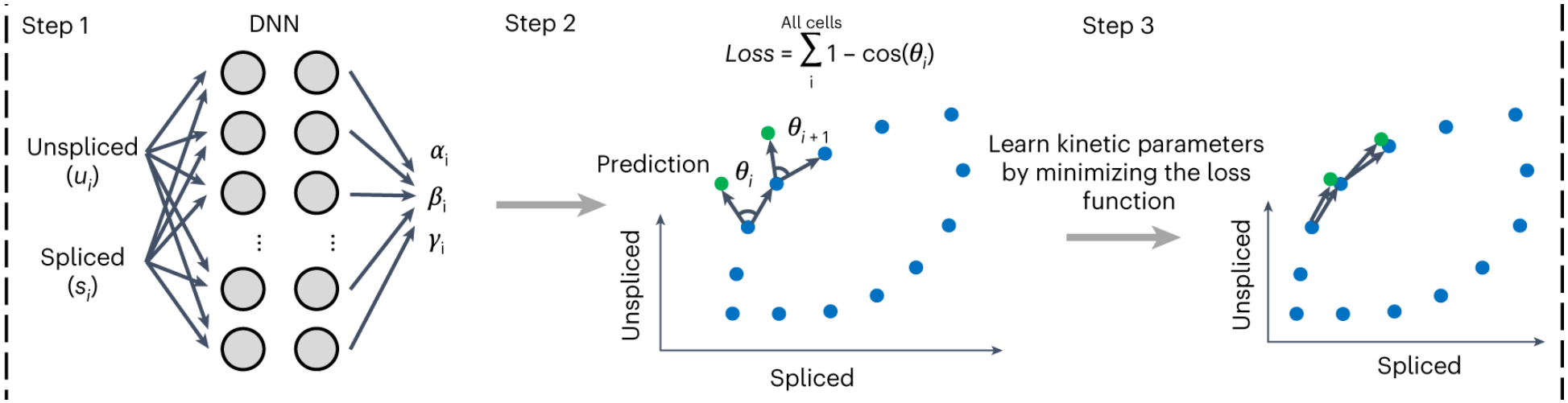
cellDancer (Li et. al. Nature Biotech 2024)

- Allow time-varying and gene-specific transcription rate, splicing and degradation rates

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = \alpha(t) - \beta(t)\,u(t)$$

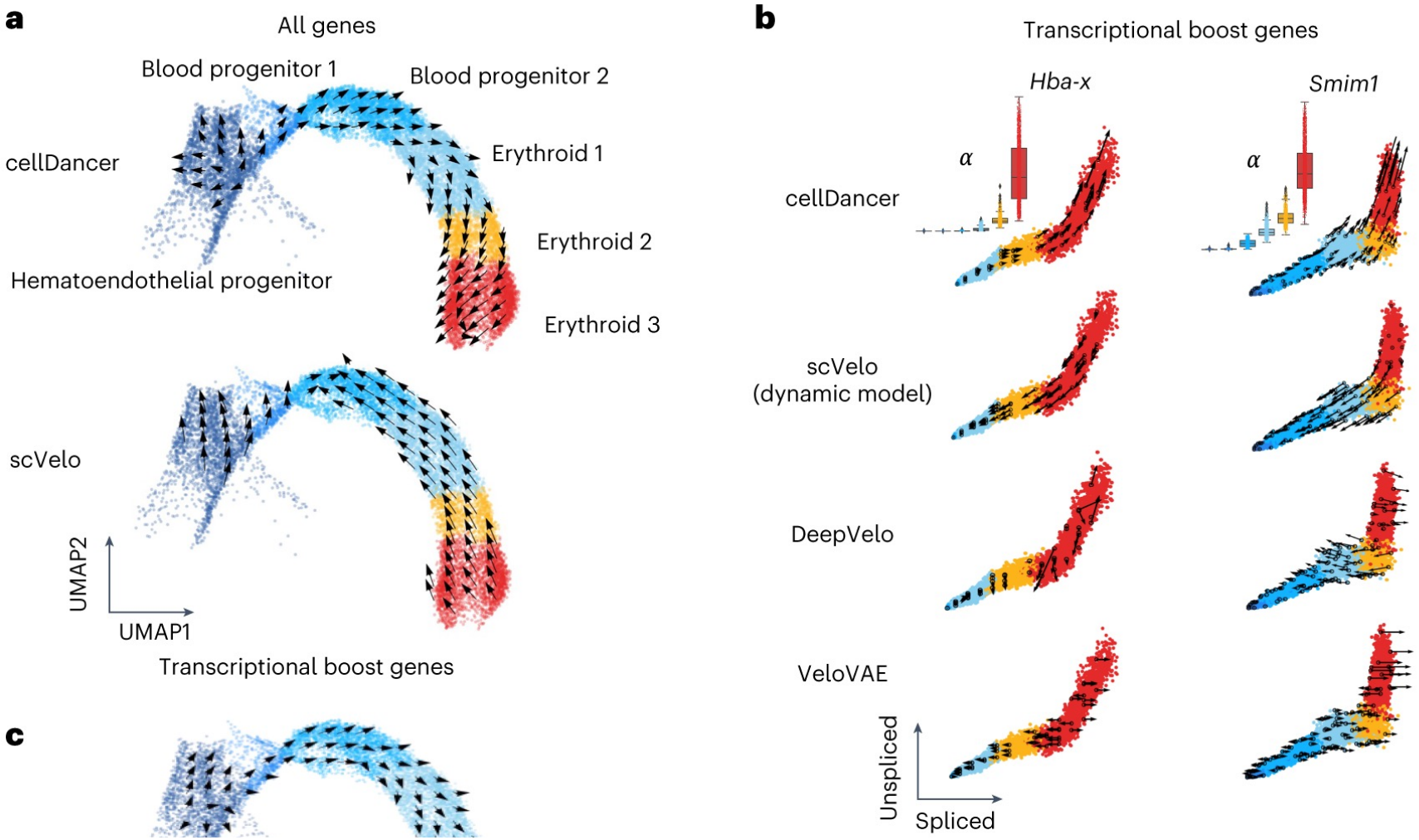
$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = \beta(t)\,u(t) - \gamma(t)\,s(t)$$

- Train a neural network for each gene using the cells as samples to estimate



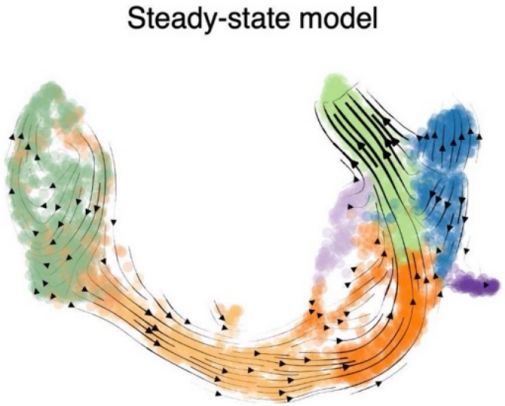
- Each cell i can have different rates as the underlying time t_i is different (like assumed in scVelo)
- Predict $(u(t_i + \Delta t), v(t_i + \Delta t))$, minimizing the different between predicted values and best observed nearest neighbor

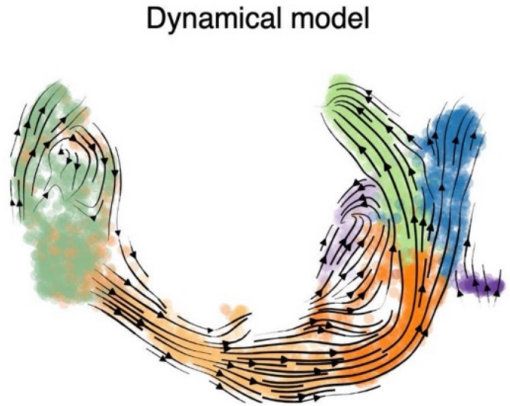
cellDancer (Li et. al. Nature Biotech 2024)

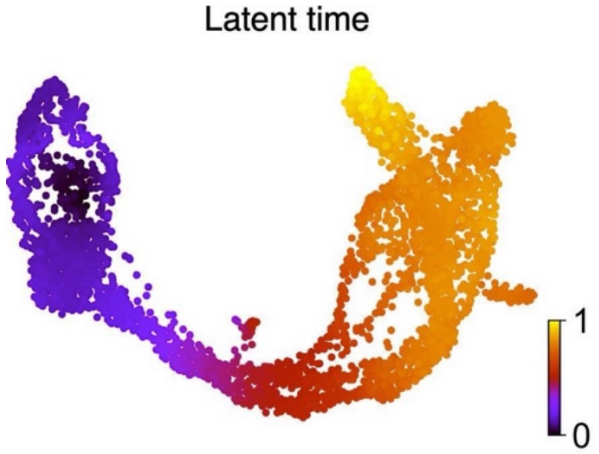


RNA velocity and trajectory inference

- An example from Weiler et. al. , methods Mol Biol. 2023

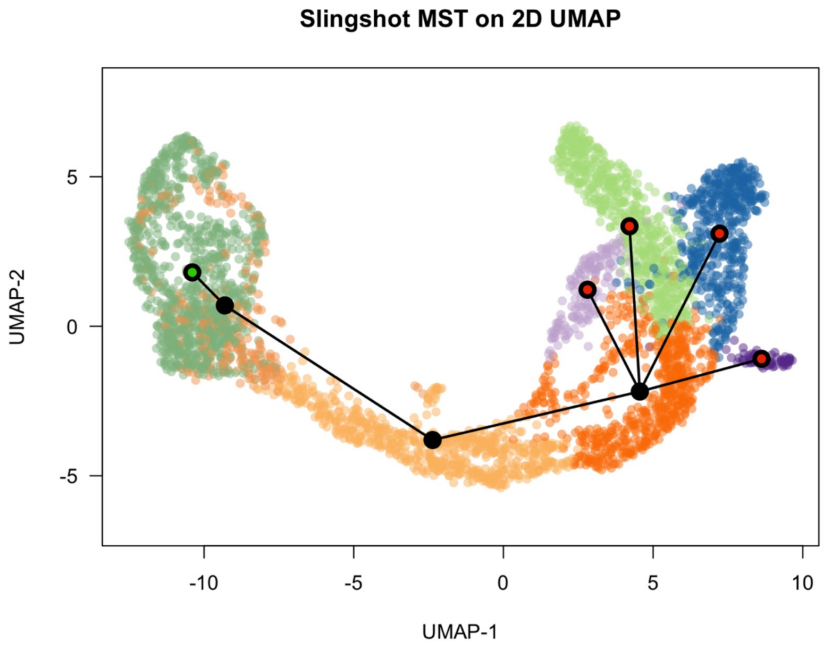


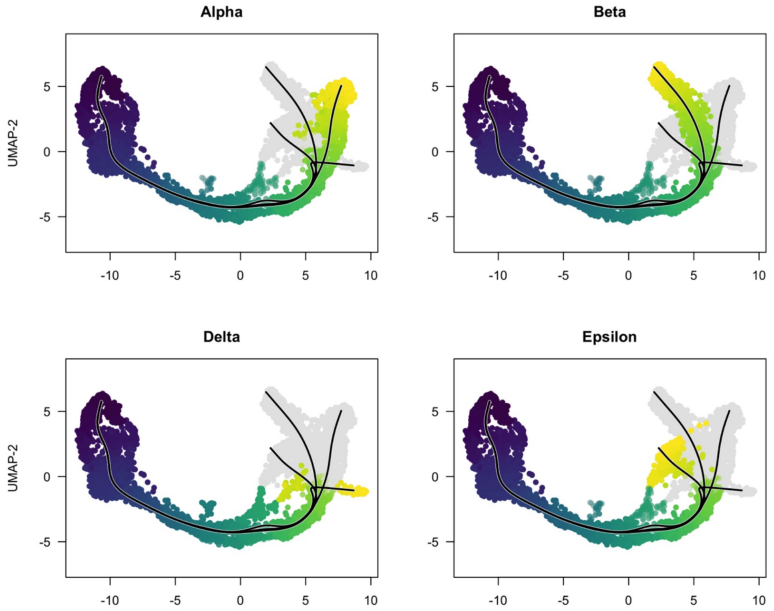




Estimated gene-shared latent time

Ductal Ngn3 low EP Ngn3 high EP Pre-endocrine Alpha Beta Delta Epsilon





Estimated pseudotime for each lineage

- RNA velocity estimates are directional, cell specific but noisier
- Trajectory inference may capture the global trend better

CellRank (Lange et. al., Nature Methods, 2022)

- Use RNA velocity to improve trajectory inference and estimate the cell pseudotime

- Core steps (mathematical details not provided in the paper):
  - Data input: Gene expression matrix X and velocity matrix V
  - Compute a directed KNN graph incorporating both RNA velocity and cell-cell similarity
    - Undirected KNN graph using matrix X
    - Make the KNN graph directed and weighted by computing similarity (pearson correlation $\boldsymbol{c}_i$) of neighboring cell difference with estimated velocity vectors for each cell i

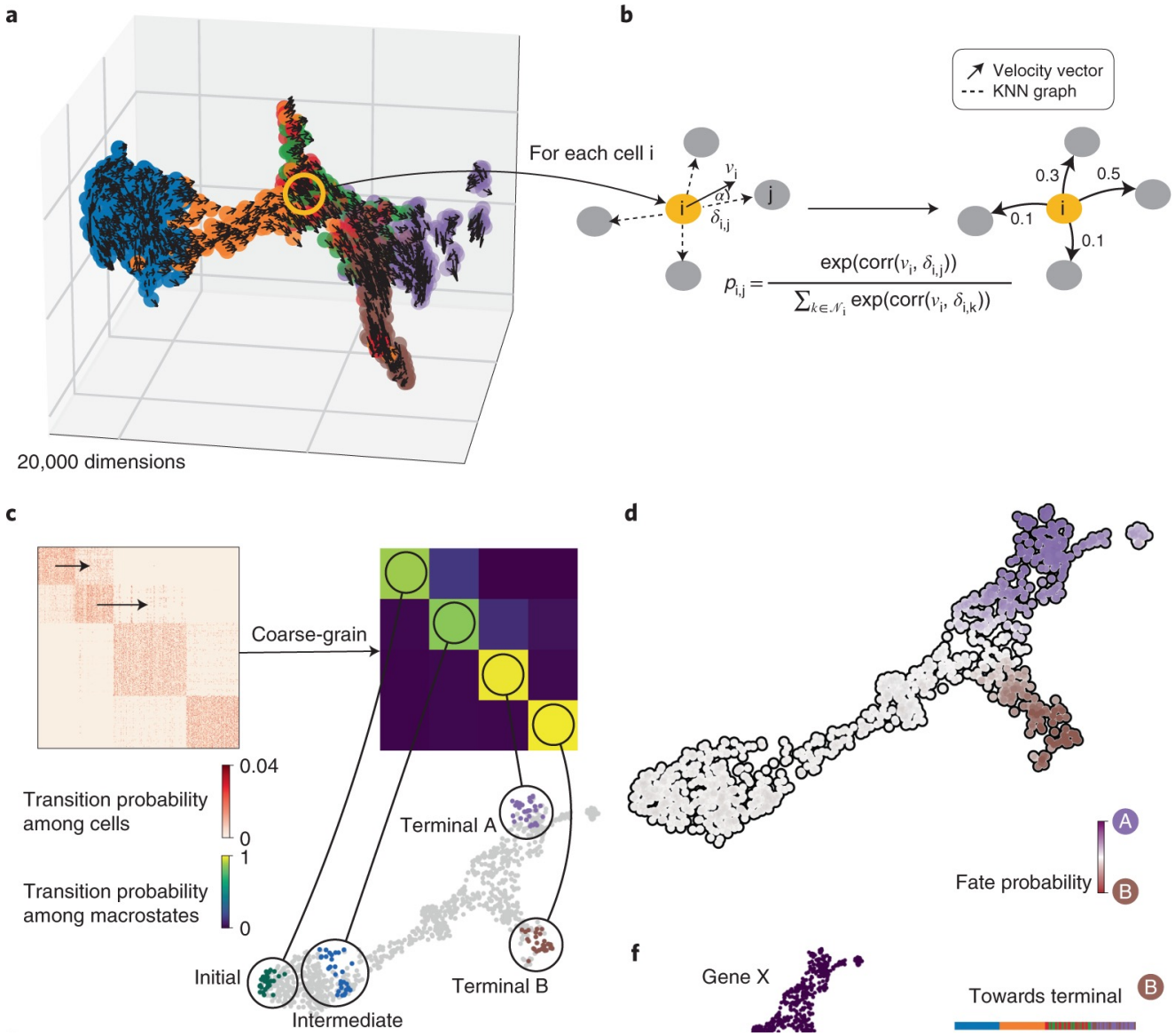
  - Compute cell-cell transition probabilities

$$P = (1-\lambda)P_{\mathrm{v}} + \lambda P_{\mathrm{s}} \text{ for } \lambda \in [0,1].$$

    - For $P_{\boldsymbol{v}}$: $p_{\mathrm{ik}} = \frac{e^{\sigma c_{\mathrm{ik}}}}{\sum_{\mathrm{l}} e^{\sigma c_{\mathrm{il}}}}$

  - In principle, we can feed the new KNN graph and cell-cell transition probabilities into the original PAGA algorithm
  - In CellRank, instead of using clustering and calculate connectivity score, they reduce cell-cell transition matrix to coarse-grain transition matrix of macrostates (something like soft-clustering) using a method called GPCCA
  - Automatically identify initial state by finding the stationary distribution of a Markov chain with transition probability P → initial state has the smallest probability in the stationary distribution

CellRank (Lange et. al., Nature Methods, 2022)



CellPath (Zhang and Zhang, Cell Reports Methods, 2021)

Core ideas:

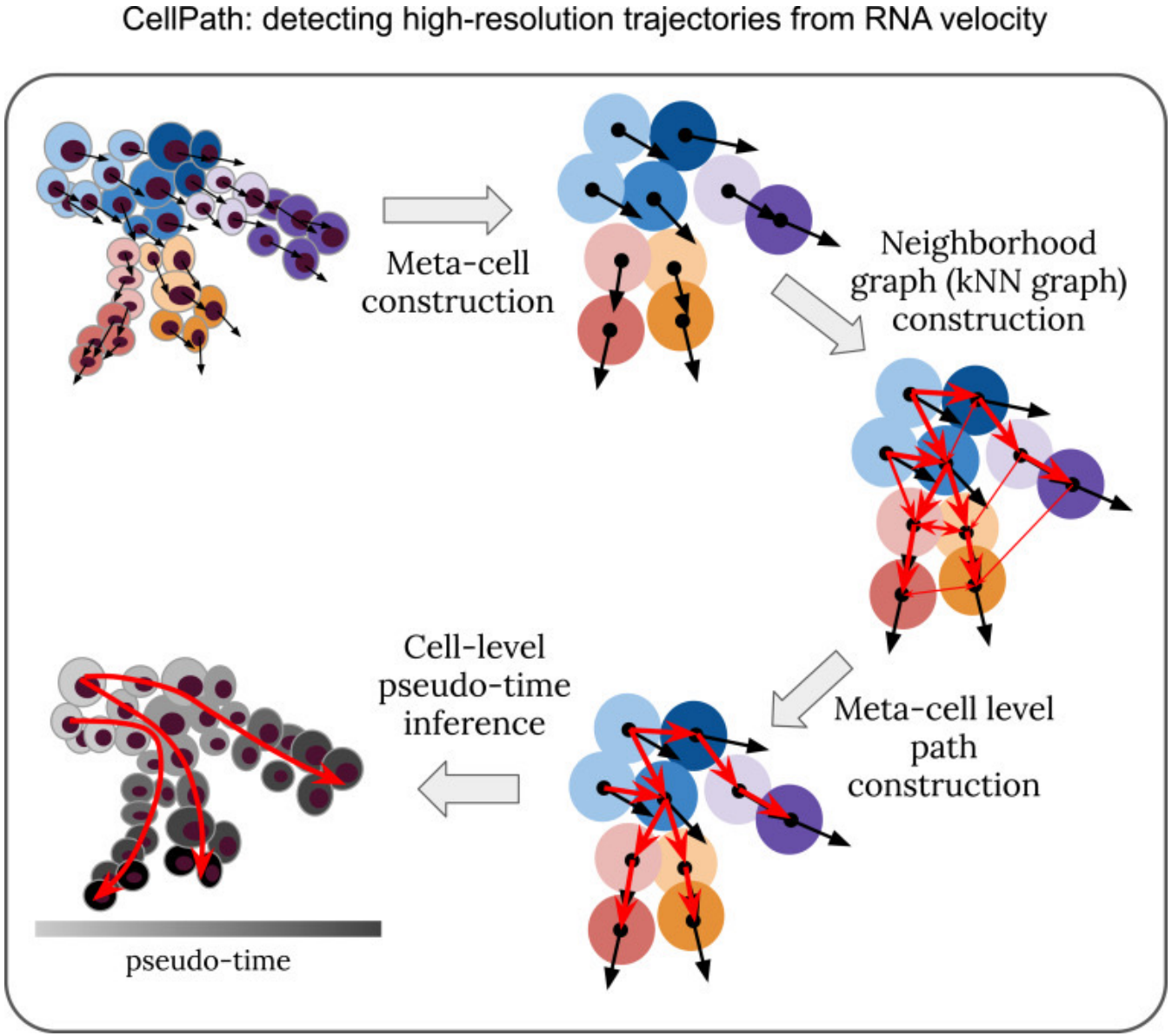
- Data input: Gene expression matrix X and velocity matrix V
- Meta-cell construction: clustering of the cells and treat each cluster as a meta-cell
  - For each meta-cell, get an average gene expression vector and a smoothed RNA velocity vector
    - Instead of simply averaging the velocity vectors use kernel regression $v_i = f(x_i)$
  - Construct a cell-cell directed KNN graph on the meta-cells using a similar idea as cellRank
    - Weight between two cells calculated as (details omitted)

$$e(i,j) = [\lambda(\beta \ell_{dist}(i,j) + \ell_{\theta}(i,j))]^{\lambda}$$

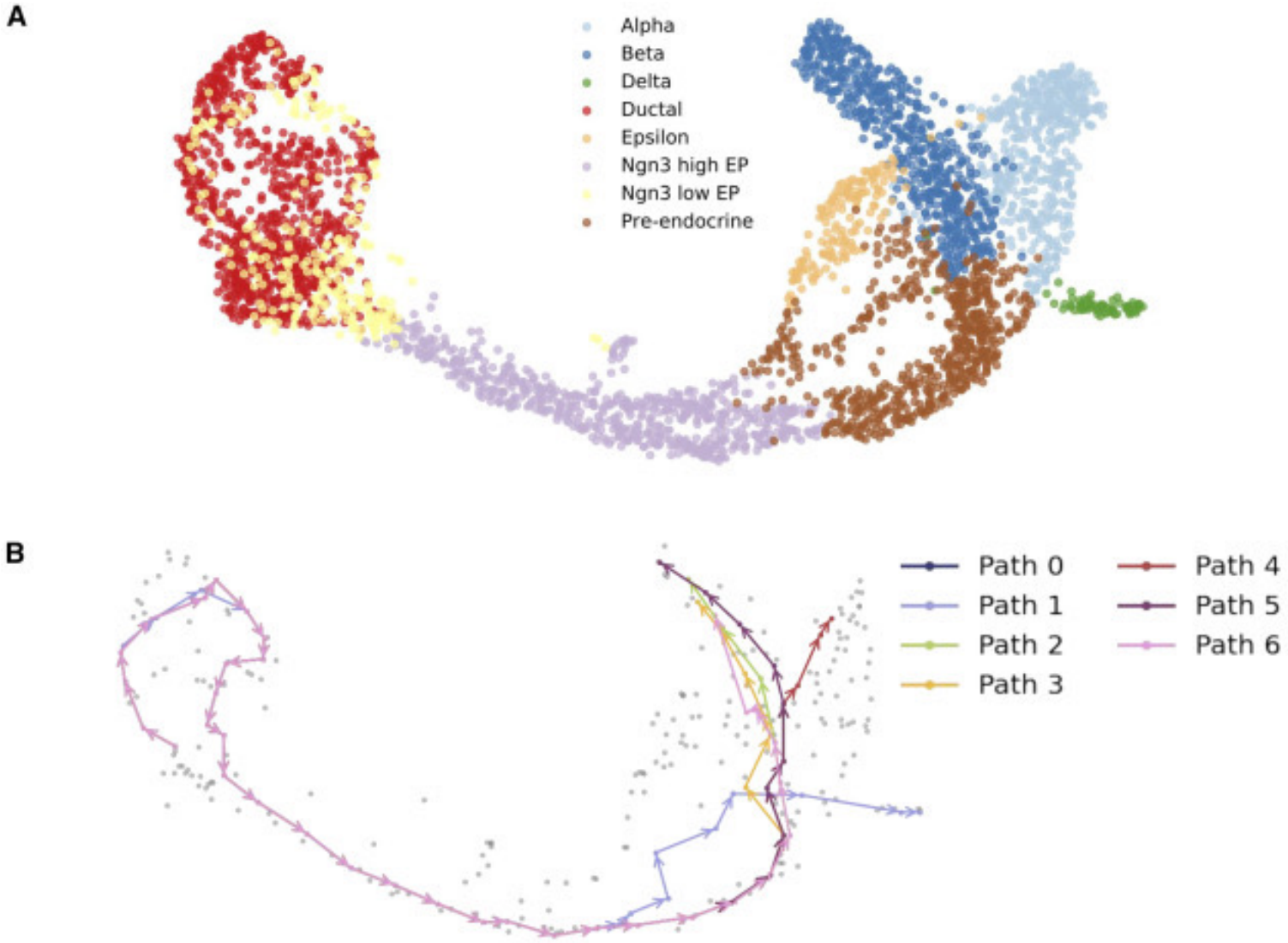
  - Find shortest directed path between any two meta-cells that are within 3degree in the KNN graph
  - Calculate pseudotime for each cell j within meta-cell i
    - Order cells based on the projection

$$\frac{(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{v}_i}{\|\mathbf{v}_i\|_2}$$

CellPath (Zhang and Zhang, Cell Reports Methods, 2021)



CellPath (Zhang and Zhang, Cell Reports Methods, 2021)



Related papers

- La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., ... & Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494-498.
- 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*(12), 1408-1414.
- Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled. *PLOS Computational Biology*, 18(9), e1010492.
- Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., ... & Yosef, N. (2024). Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature methods*, *21*(1), 50-59.
- Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K. W., Le, N. T., ... & Wang, G. (2024). A relay velocity model infers cell-dependent RNA velocity. *Nature biotechnology*, *42*(1), 99-108.

- Weiler, P., Van den Berge, K., Street, K., & Tiberi, S. (2022). A guide to trajectory inference and RNA velocity. In *Single Cell Transcriptomics: Methods and Protocols* (pp. 269-292). New York, NY: Springer US.
- Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., ... & Theis, F. J. (2022). CellRank for directed single-cell fate mapping. *Nature methods*, *19*(2), 159-170.
- Zhang, Z., & Zhang, X. (2021). Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity. *Cell Reports Methods*, *1*(6).