

Methods for dynamical inference from high-throughput single-cell omics data

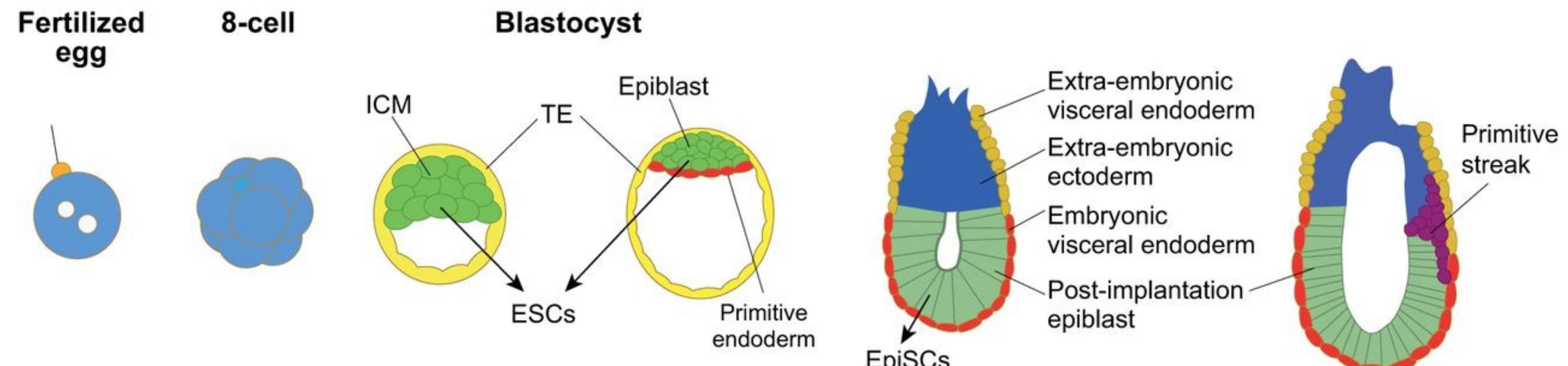
LALEH HAGHVERDI

SHARIF UNIVERSITY OF TECHNOLOGY
STATISTICAL PHYSICS SEMINAR

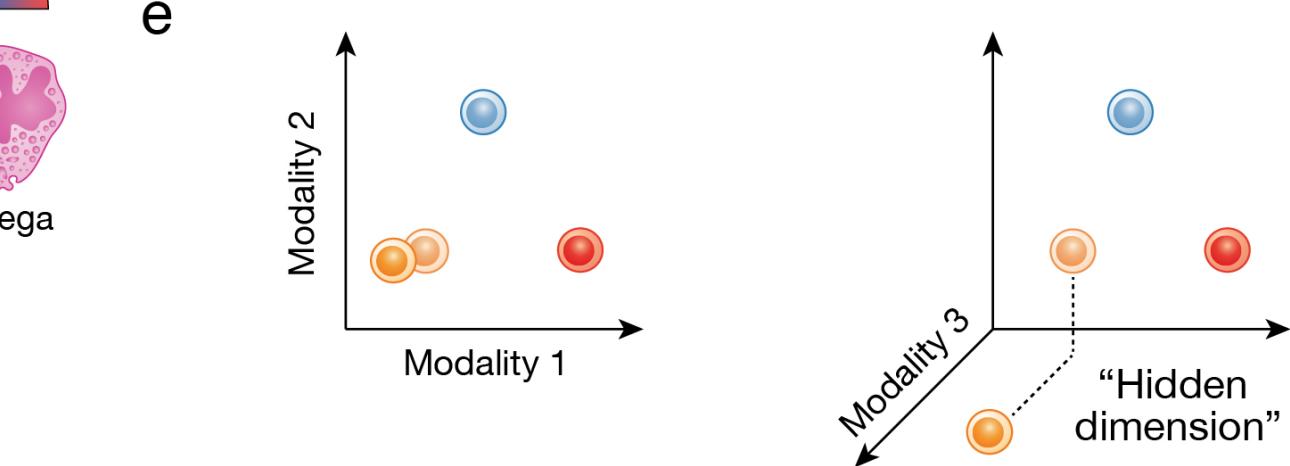
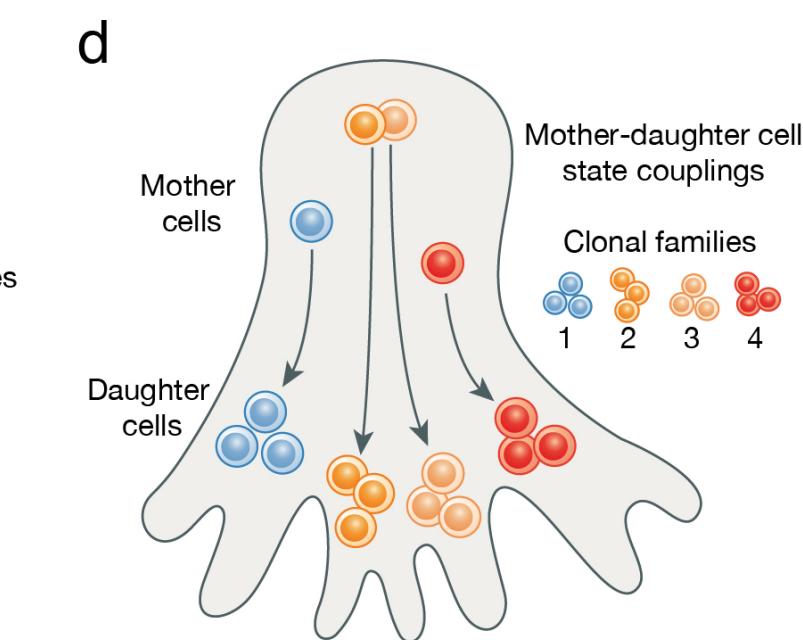
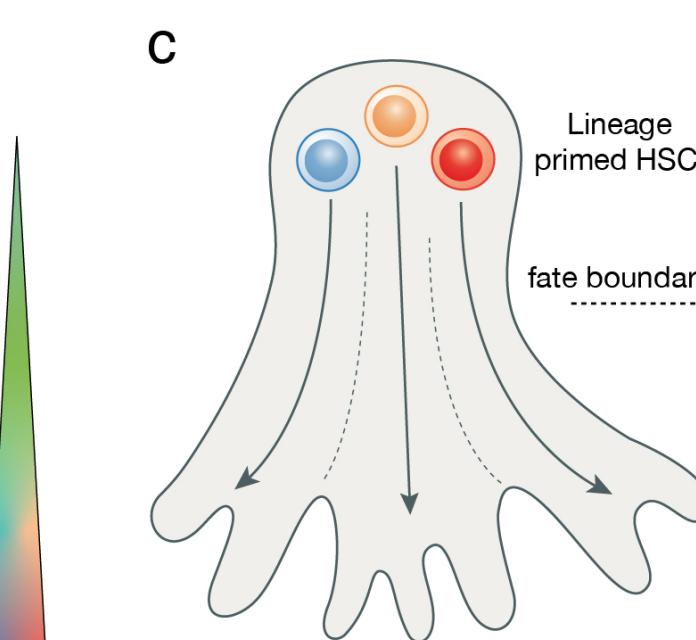
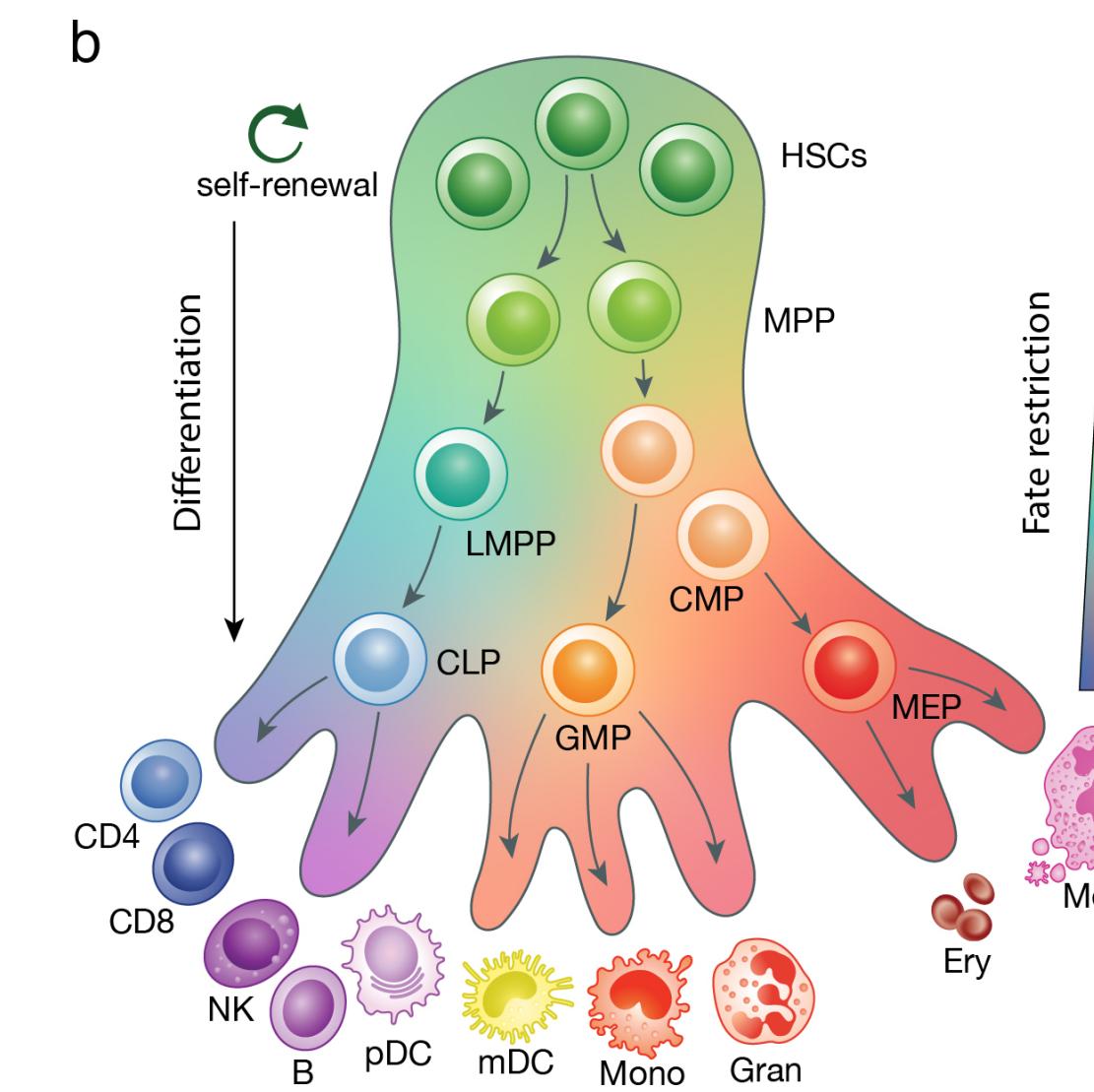
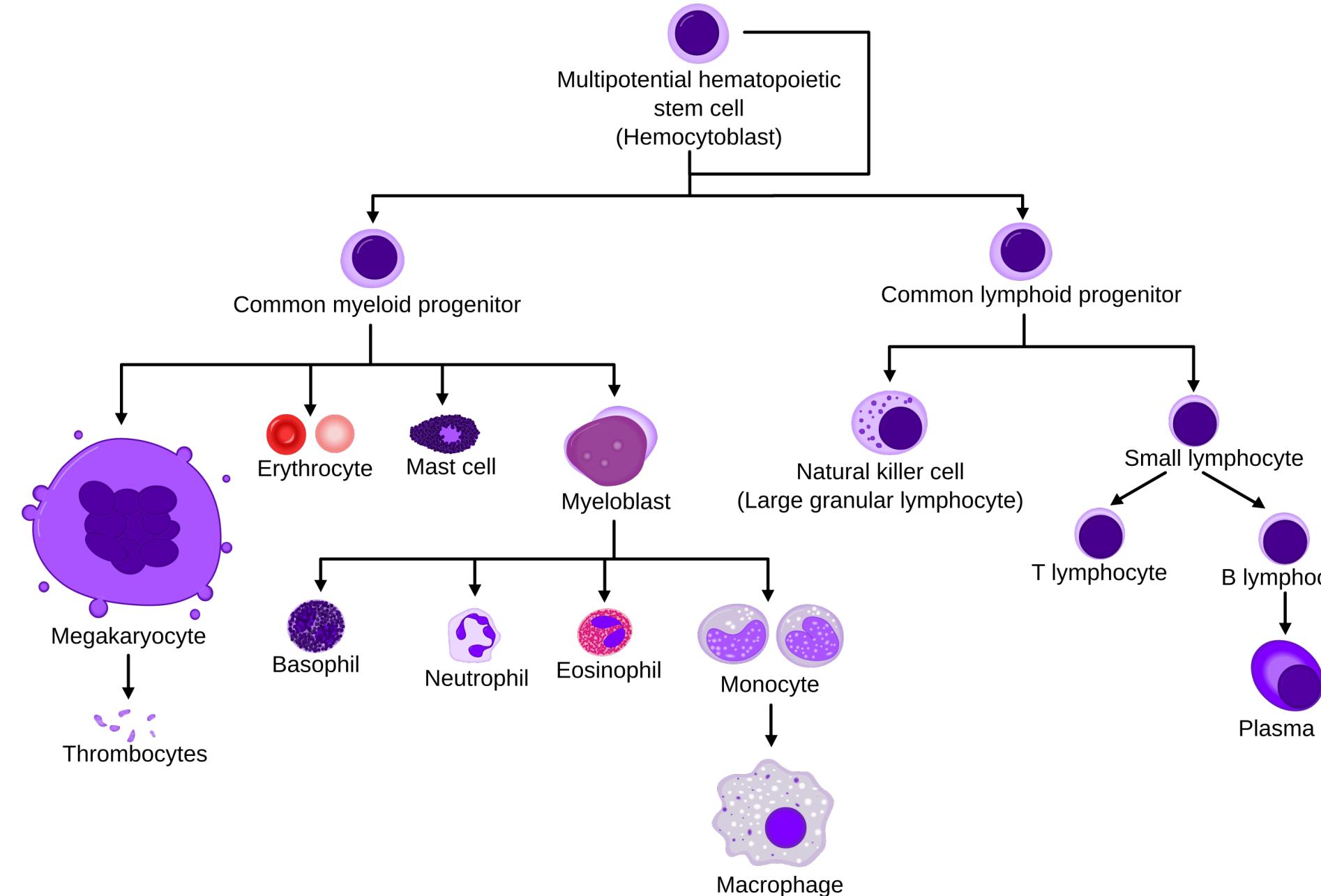
27.04.25

Same DNA in all cells but different gene activities and cell types

- Embryogenesis

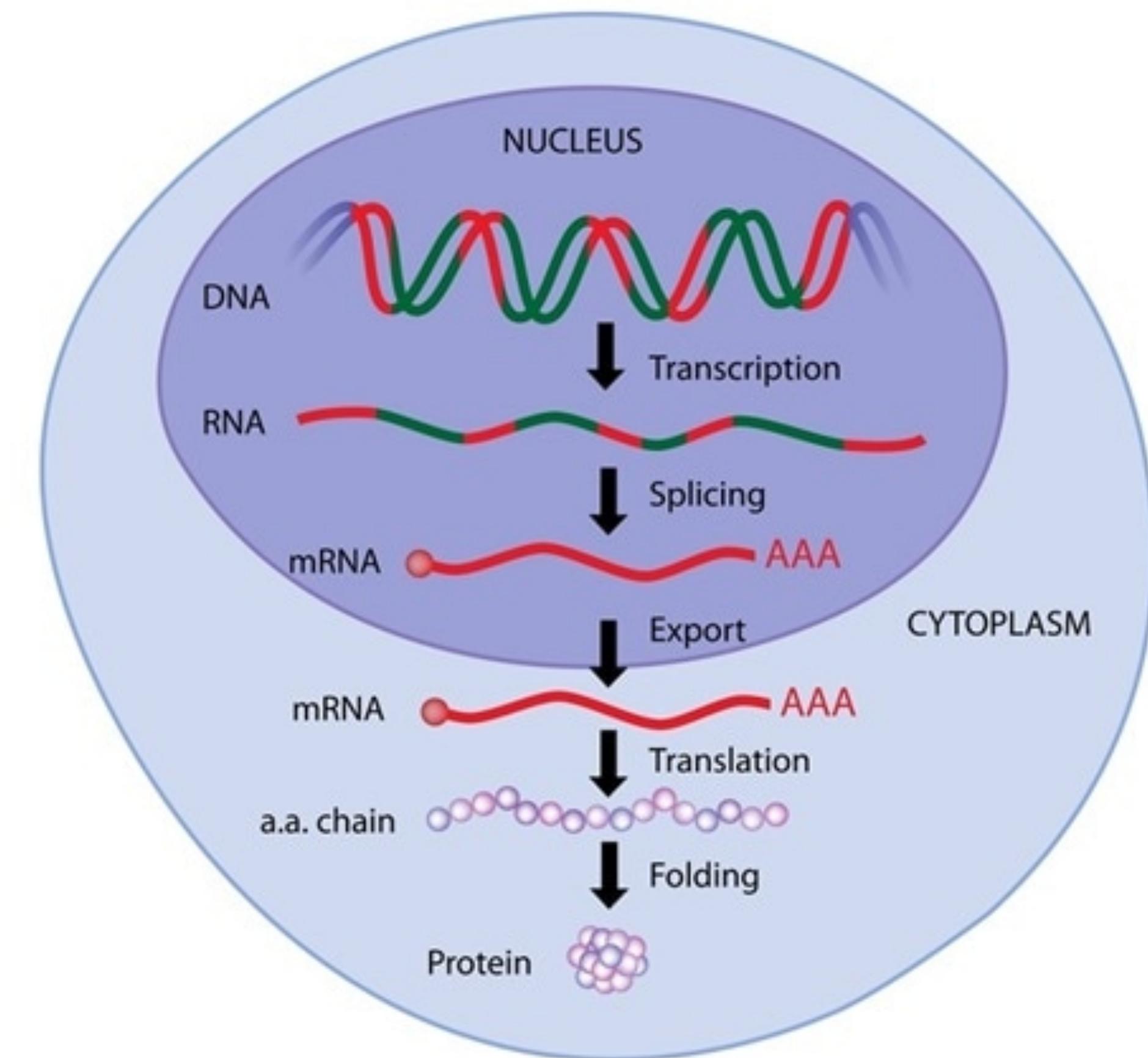


- Haematopoiesis



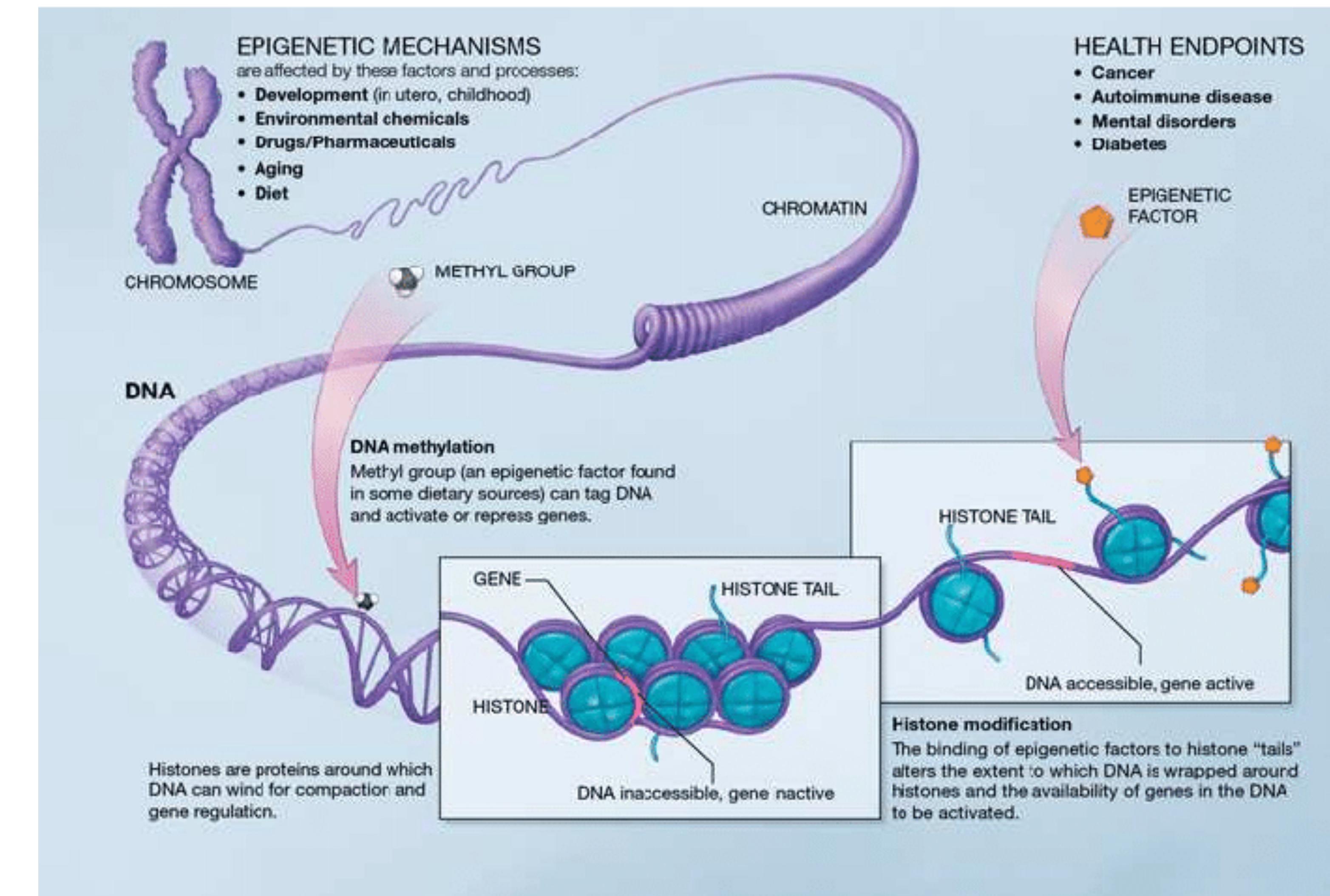
DNA → nascent mRNA → spliced mRNA → protein

- A defect in any step may lead to a disease e.g.
 - Mutations in the gene body → broken protein
 - Mutations in other regulatory regions → too little or too much production of an mRNA or protein
- Human and mouse ~20k genes
- Only <10% of the DNA is genes



Genome organisation: epigenetic

- Some diseases arise because of incorrect chromatin unfolding and folding



Single-cell measurements

- Heterogenous cell populations:

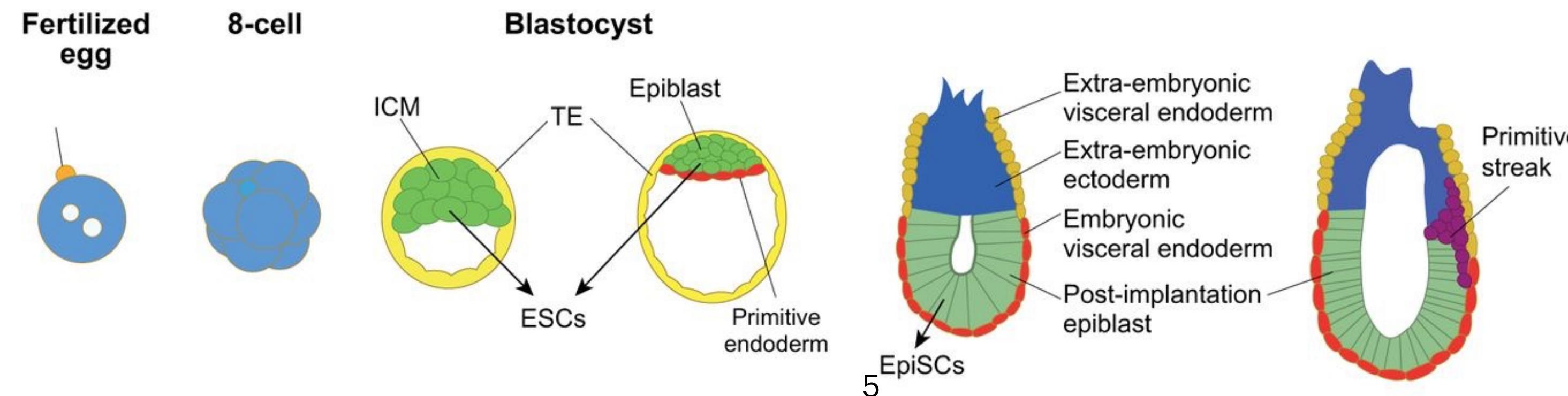
- Cancer and tumour cell populations
- Rare cell types
- Several seemingly homogenous cell populations
- Transitory and asynchronous states e.g. in development

Tissue (e.g. tumor)



- High-throughput single-cell omics

- Destroys the cells to extract the mRNA, DNA, etc molecules → no time-series
- Dynamical Inference methods such as pseudotime inference, optimal transport , etc.



Some of the questions we try to address

- Differentiation maps and manifolds
- Data integration and comparisons
- Differentiation Dynamics on the inferred the manifold
- The regulatory circuits that generate that dynamics and Gene Regulatory Networks (GRN)
- Network perturbation and response prediction
- How do diseased cells go out of natural trajectories? Can we return them back?
- How can we efficiently turn one cell type into another?
- Cell-cell communication and signalling
- Interaction with the environment

Overview

- Our methods for:
 - Pseudotime inference
 - Velocity estimation
 - Data integration and comparison
 - Clonal tracing from scRNA-seq

- Diffusion maps for high-dimensional single-cell analysis of differentiation data, L. Haghverdi et al. Bioinformatics (2015)
- Diffusion pseudotime robustly reconstructs lineage branching, L. Haghverdi et al. Nature methods (2016)
- *destiny*: diffusion maps for large-scale single-cell data in R, P. Angerer et al. (2016)
- Towards reliable quantification of cell state velocities, V. Marot-Lassauzaie et al. PLoS Computational Biology (2022)
- Single-cell time series analysis reveals the dynamics of in vivo HSPC responses to inflammation, B. J. Bouman et al. Life Science Alliance (2023)
- Single-cell multi-omics and lineage tracing to dissect cell fate decision-making, L. Haghverdi & L. S. Ludwig, Stem Cell Reports (2023)



Dynamical Inference



Regulatory Networks

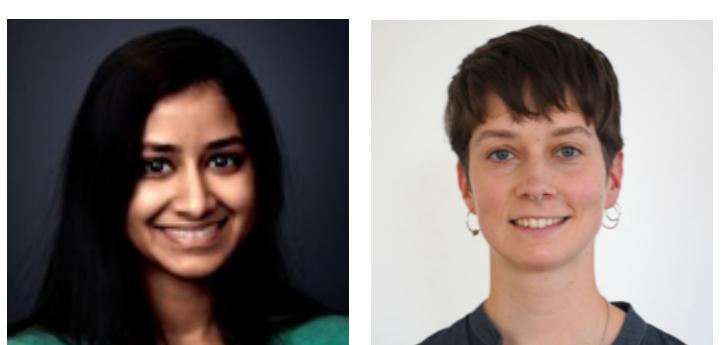
- Decoding the regulatory network of early blood development from single-cell gene expression measurements, V. Moignard et al. Nature biotechnology (2015)
- Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data, A. Ocne et al. Bioinformatics (2015)
- Seyed Amir Malekpour, et al, Single-cell multi-omics analysis identifies context-specific gene regulatory gates and mechanisms, *Briefings in Bioinformatics* (2024)



- Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, L. Haghverdi et al. Nature biotechnology (2018)
- Adjustments to the reference dataset design improve cell type label transfer, C. Moelbert & L. Haghverdi *Frontiers in Bioinformatics* (2023)
- Colin G Cess, Laleh Haghverdi, Compound-SNE: comparative alignment of t-SNEs for multiple single-cell omics data visualization, *Bioinformatics* (2024)
- Valérie Marot-Lassauzaie, Sergi Beneyto-Calabuig, Benedikt Obermayer, Lars Velten, Dieter Beule, Laleh Haghverdi, Identifying cancer cells from calling single-nucleotide variants in scRNA-seq data, *Bioinformatics* (2024)

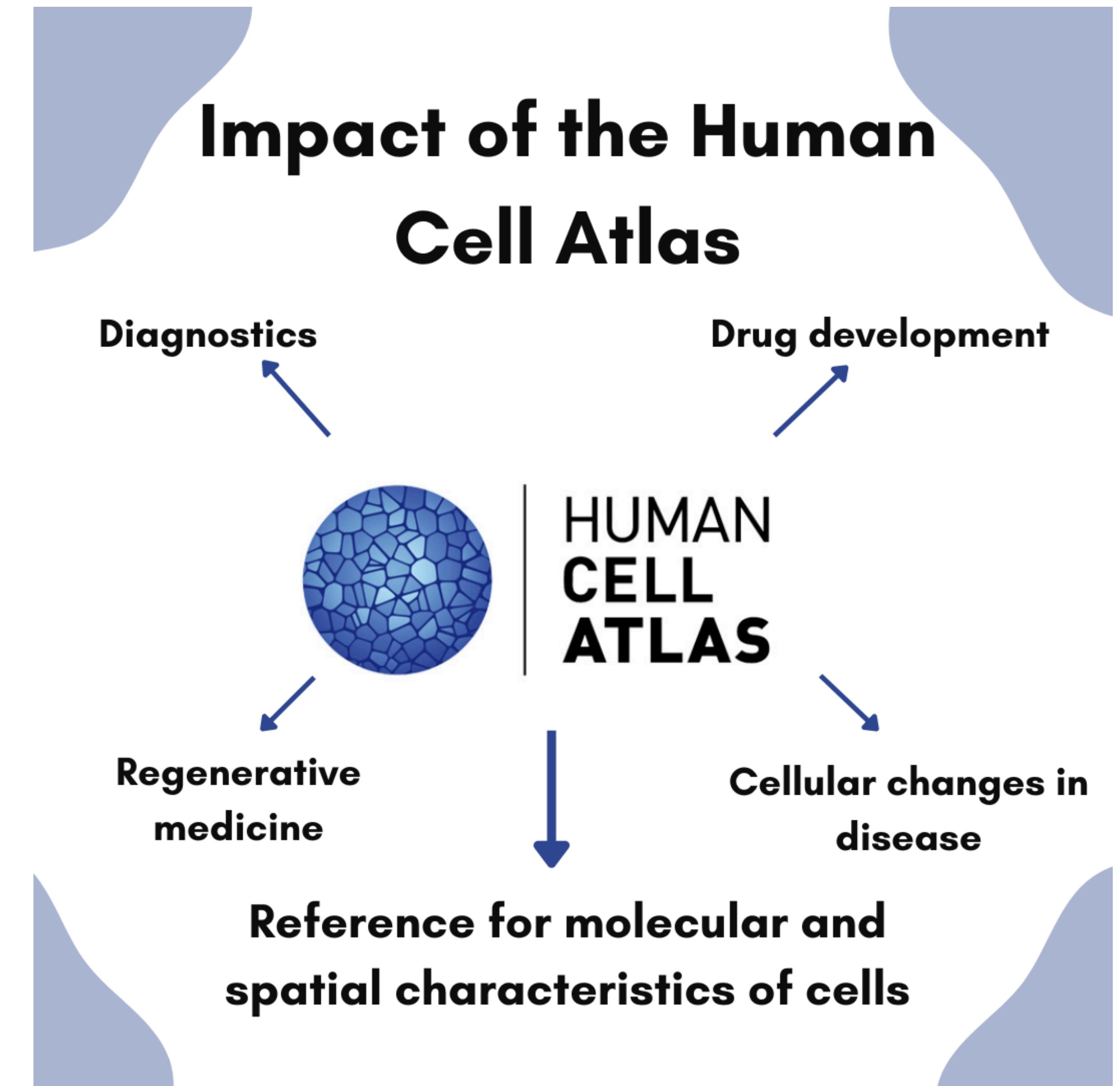
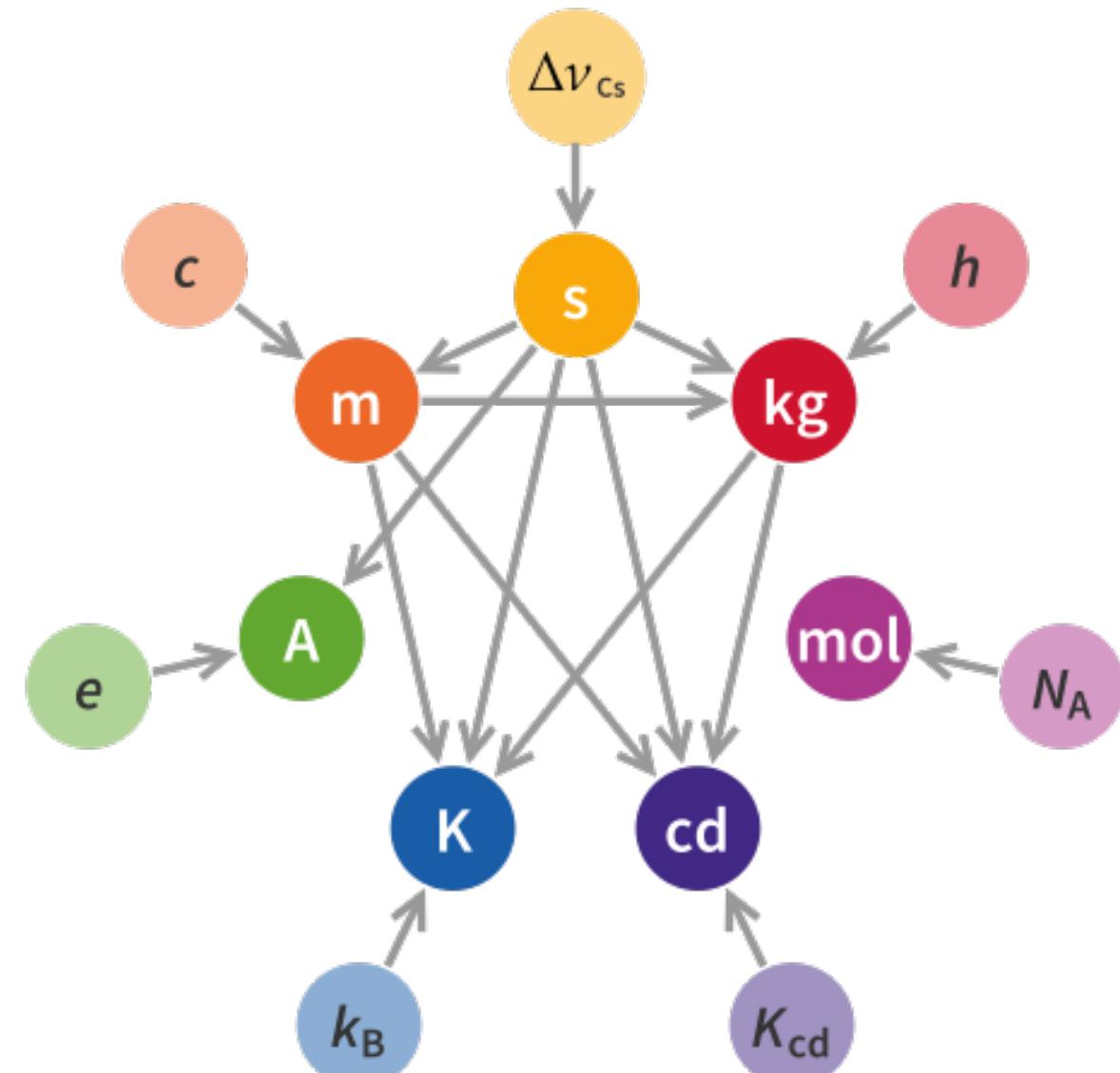


Data integration & latent spaces

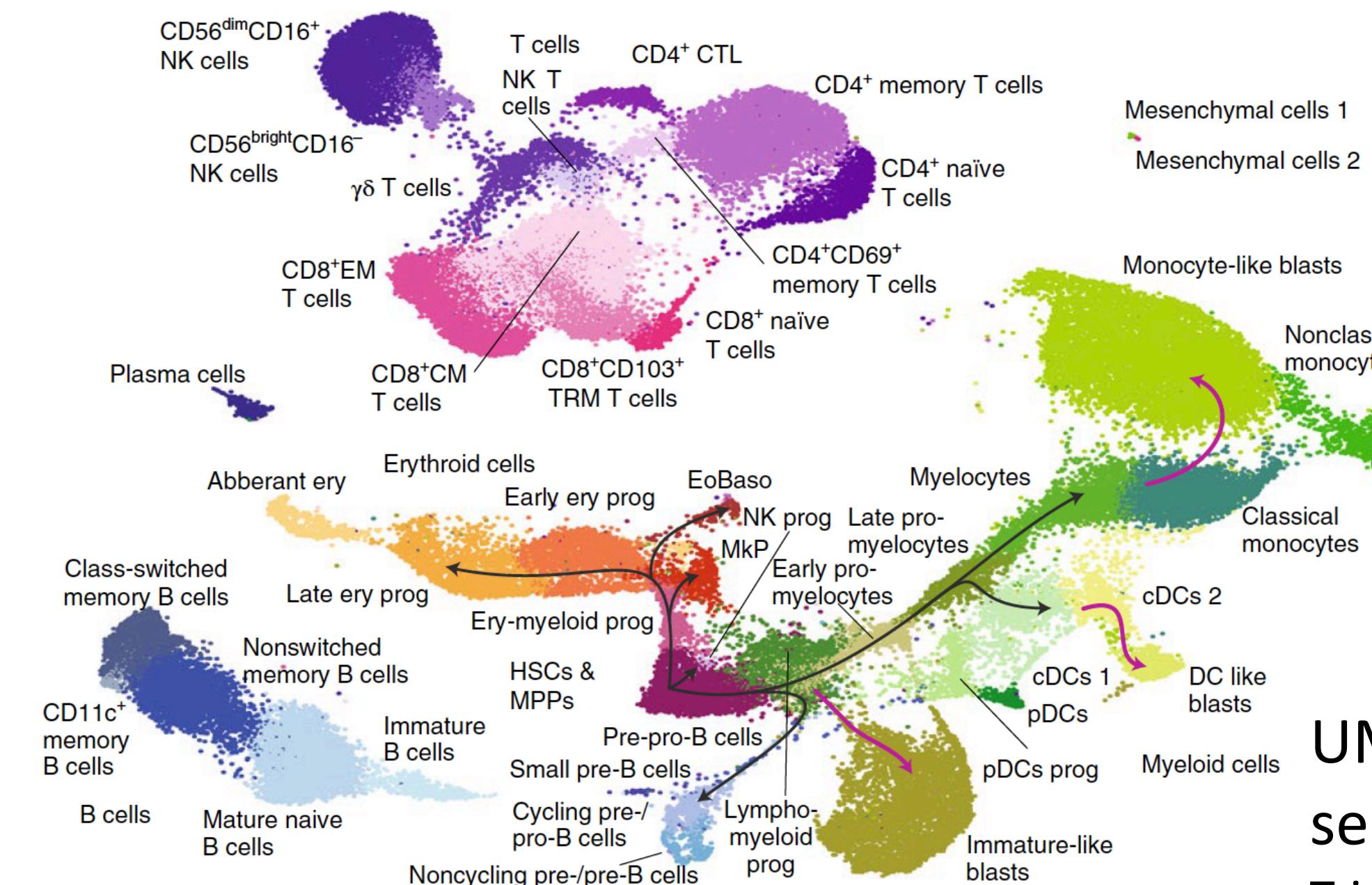
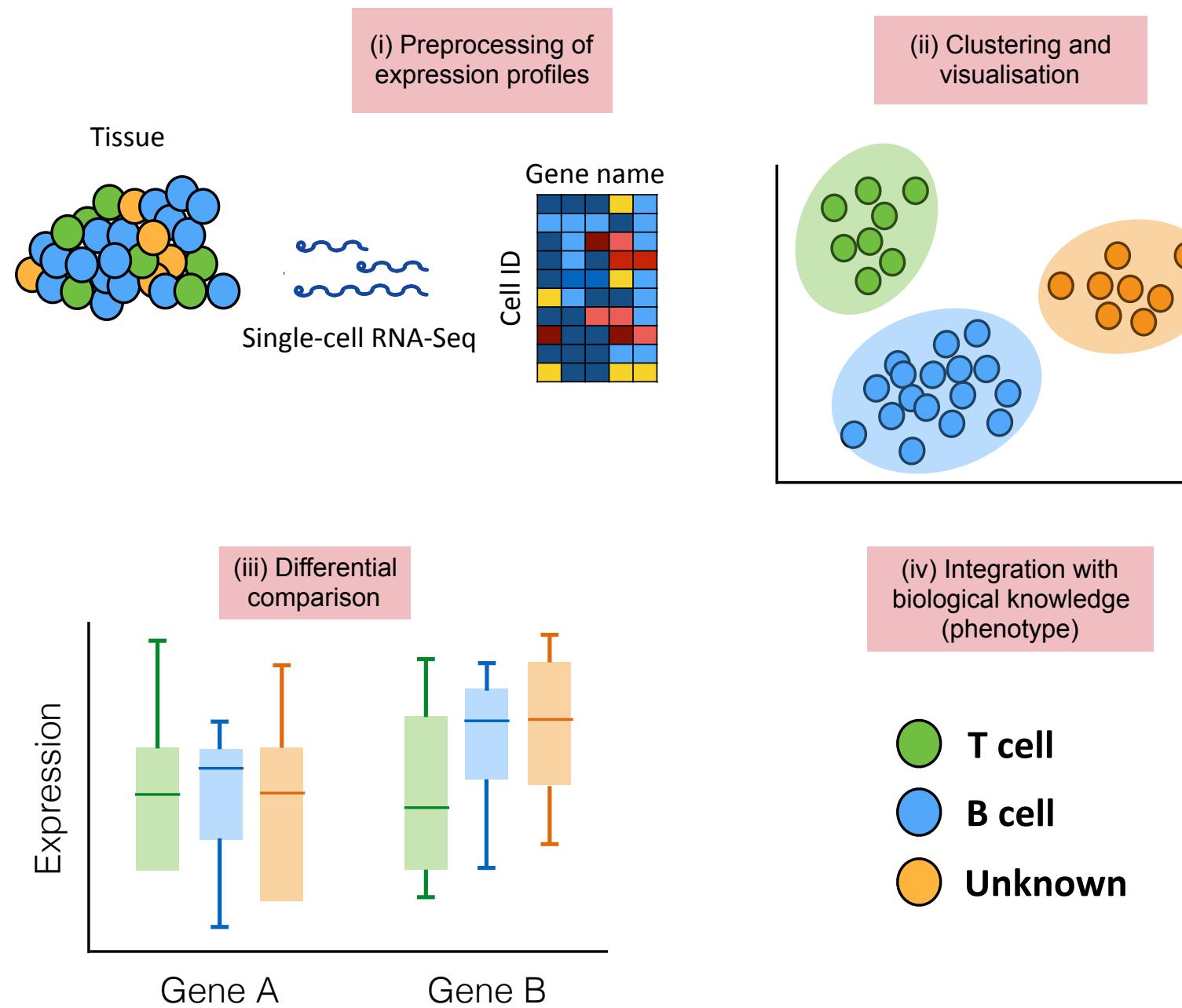


HCA consortium for standardising data collection and processing

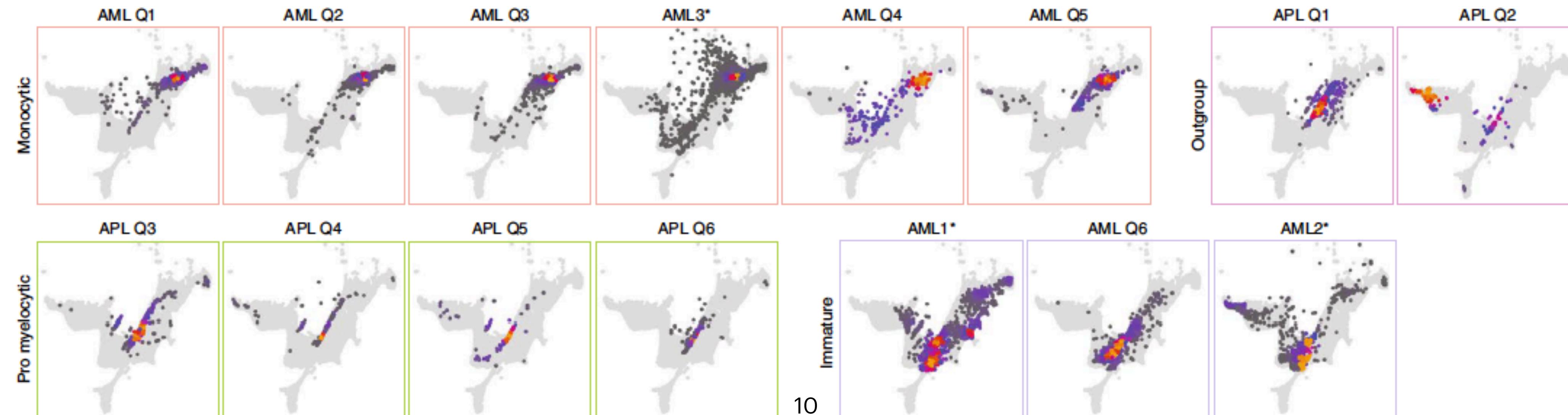
- An ~SI system for omics data?
- International System of Units, SI (from French Système international d'unités) ~1875-1975
- Reverse dependencies of the SI base units on seven physical constants, which are assigned exact numerical values in the 2019 redefinition.



Dimension Reduction on single-cell omics data

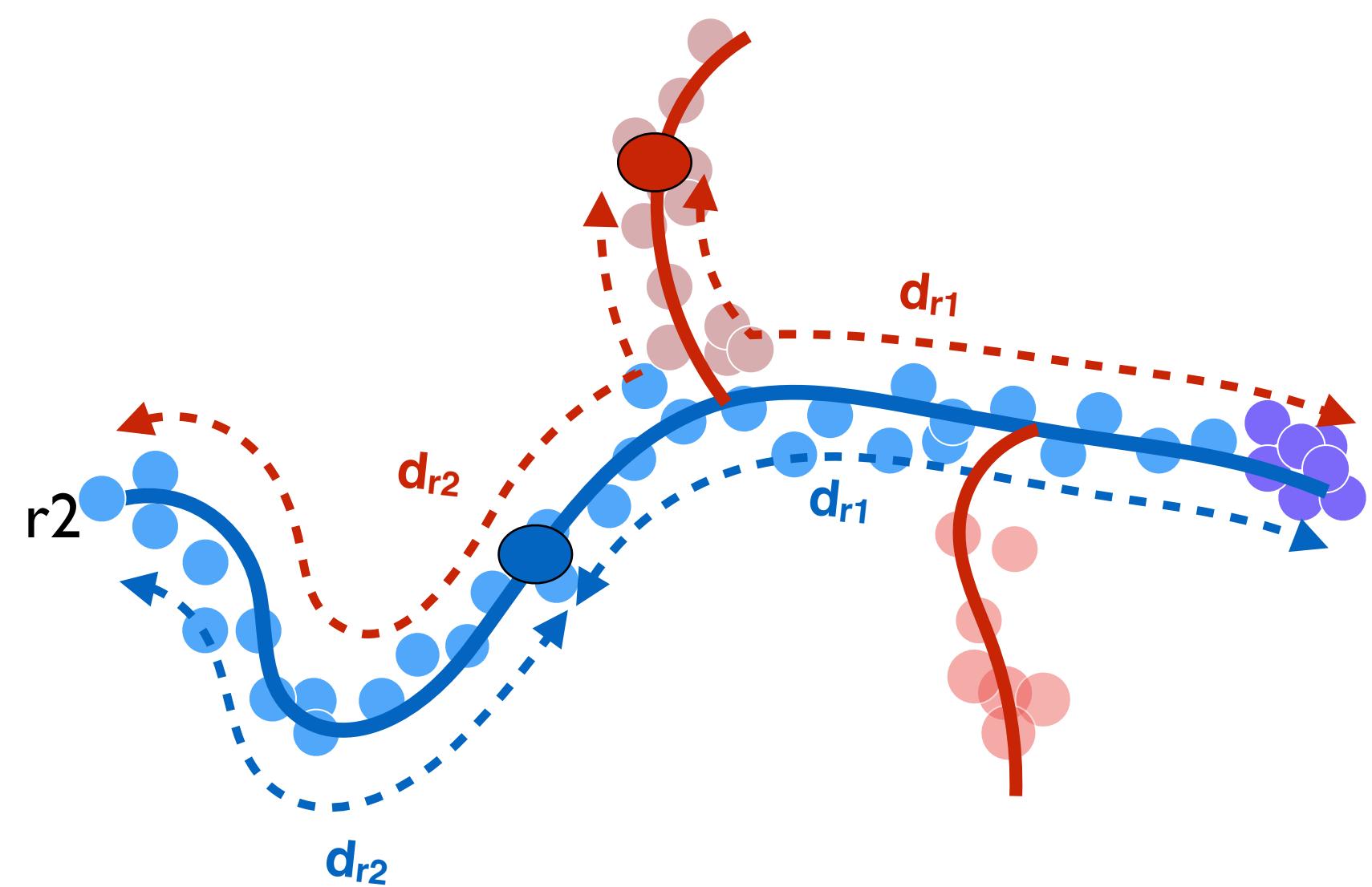


UMAP on bone marrow scRNA-seq & surface markers data
Triana et al. *Nature Immunology* 2021

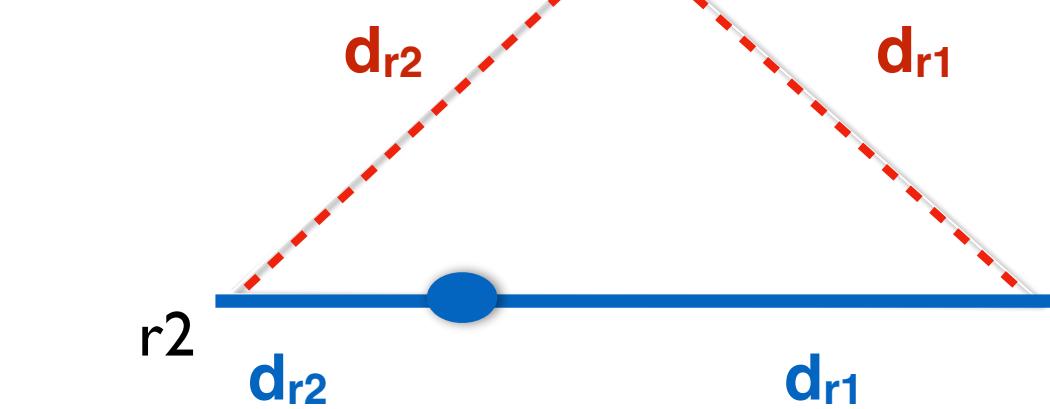


Trajectory analysis: pseudotime ordering and branch identification

- Pseudotime: on-manifold distance from the root cell
- Branch identification: generalised triangle inequality



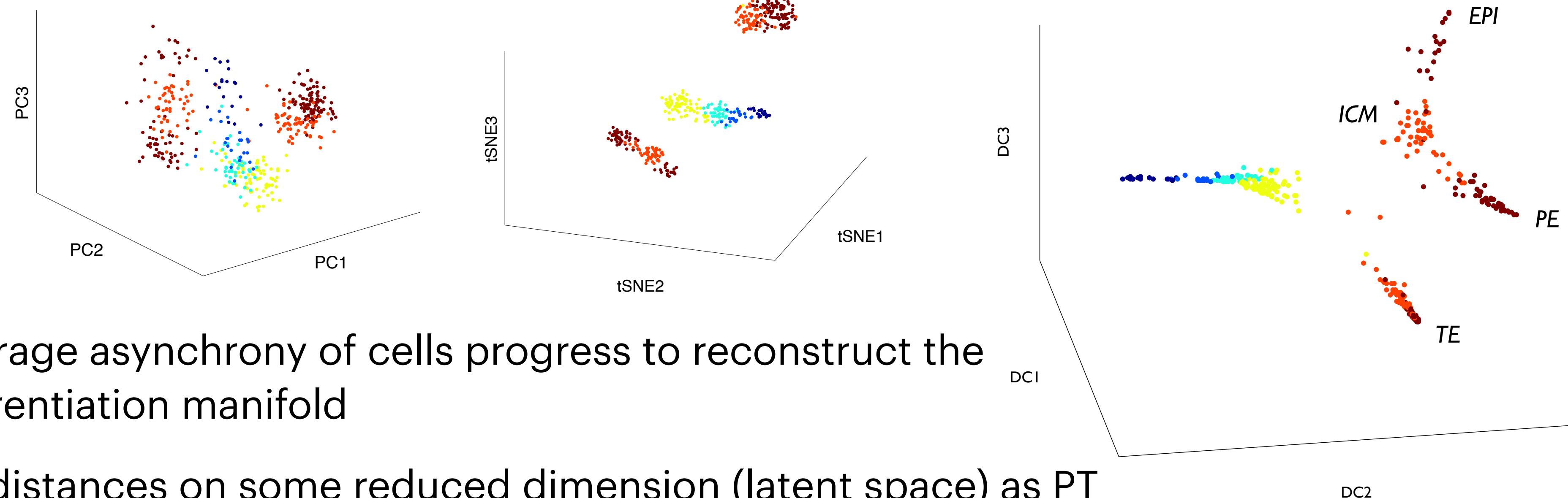
We need a continuity keeping DR method



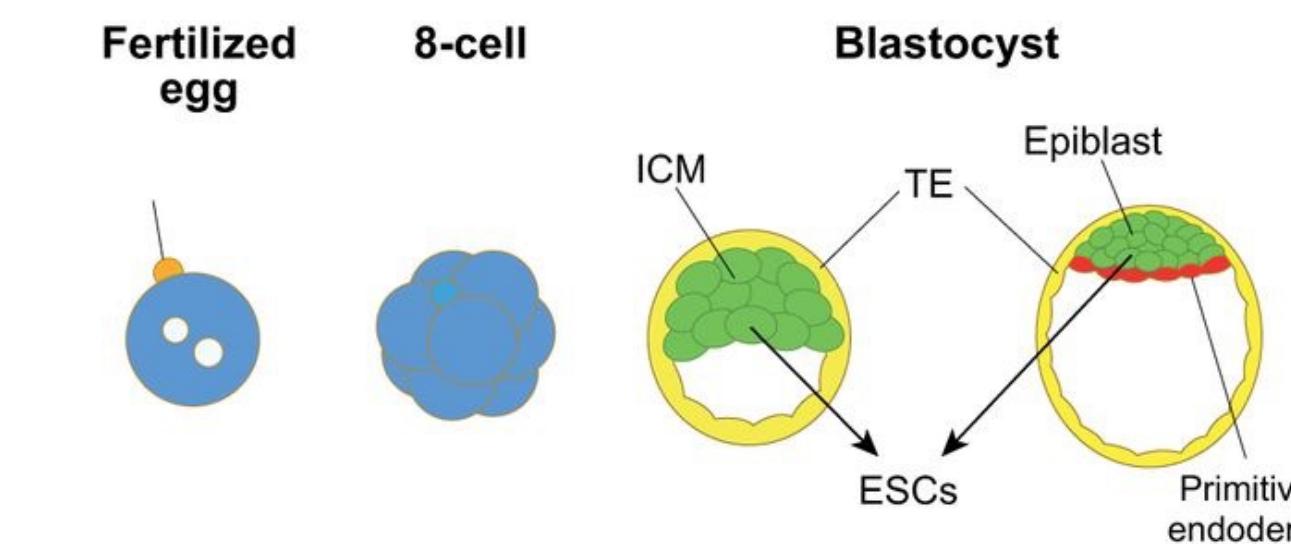
Haghverdi et al. *Nature Methods* 2016
(Supplement)

Diffusion Maps for single-cell omics data

- sc-qPCR ~500*48 mouse embryonic stem cells data set



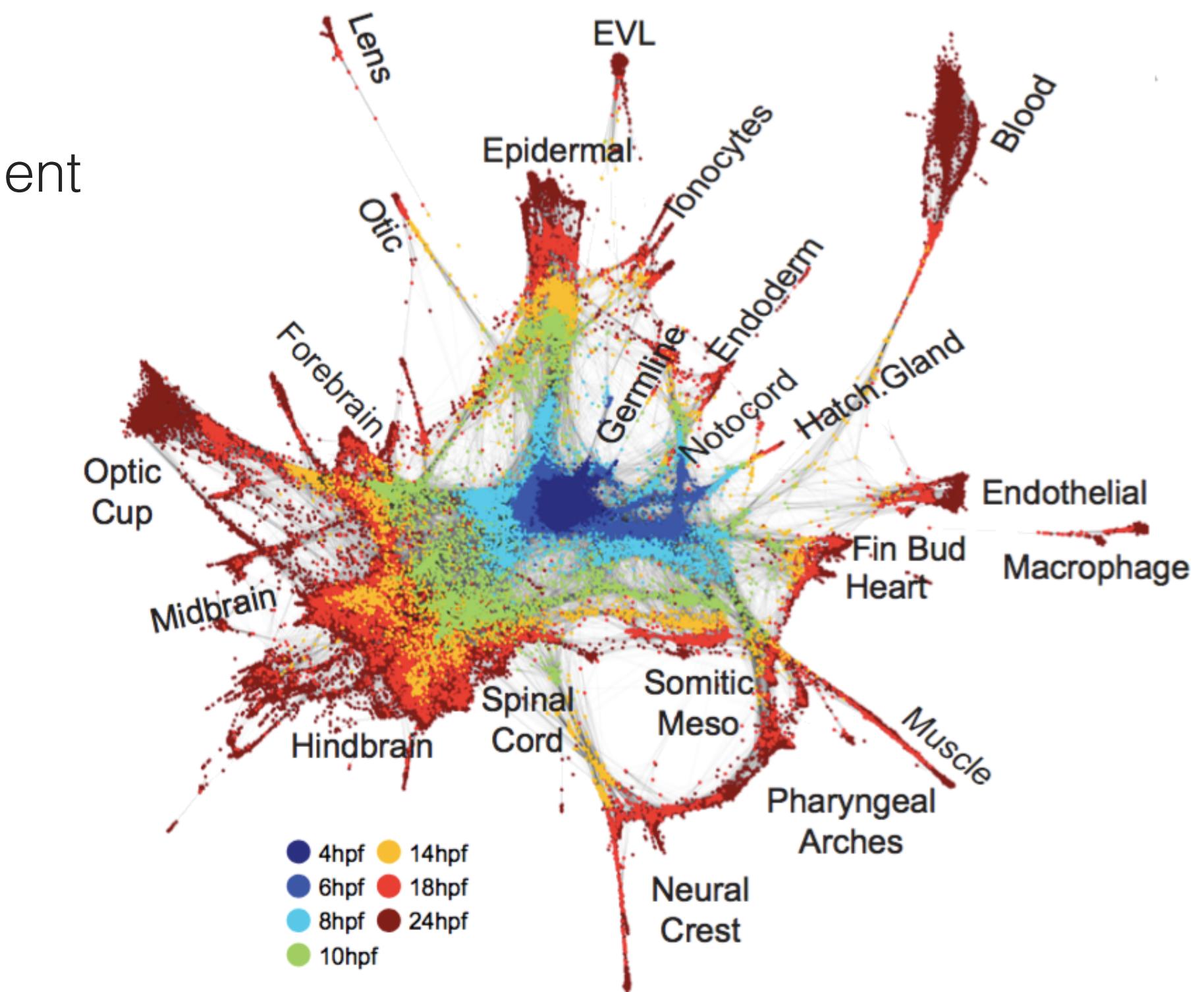
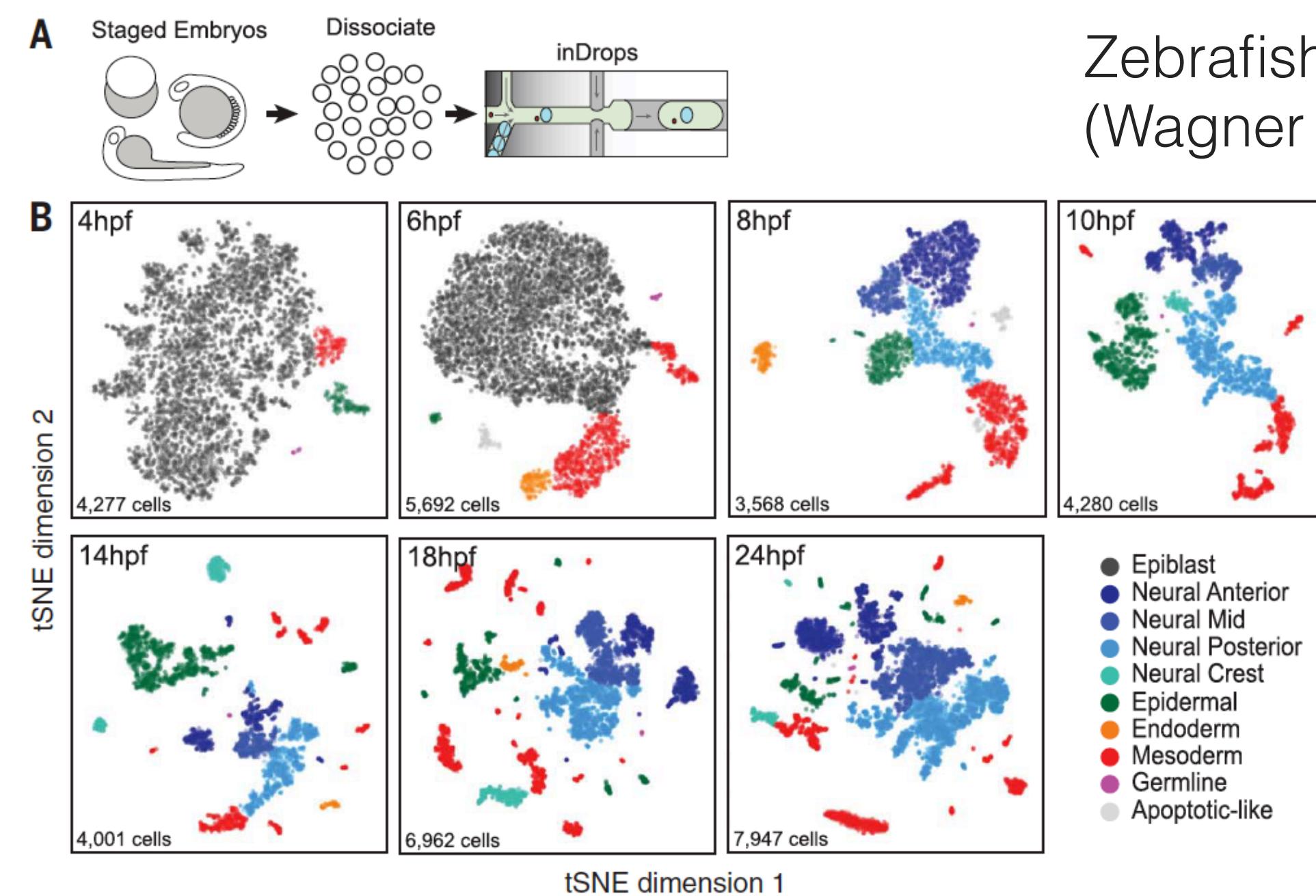
- Leverage asynchrony of cells progress to reconstruct the differentiation manifold
- Use distances on some reduced dimension (latent space) as PT
- Among DR methods, diffusion maps/pseudotime are especially suitable for differentiation data analysis
 - Guaranteed to keep the manifold continuity
 - Robust to noise
 - Biological interpretation of diffusion



- Haghverdi, et al.
Bioinformatics (2015)
- Guo et al.
Developmental cell 2010

Typical scRNA-seq data sizes and properties

- $\sim 10^6$ cells * $\sim 20k$ genes
- Sparse and noisy data
- The volume of 2-state genes space $2^{20k} \gg$ number of sampled cells
- The higher the number of dimensions, the more severe artefacts of curse of dimensionality → feature selection
- Several branches and cell types
- No ideal DR method yet for highly complex manifolds (2D visualisation, automatic and reproducible, preserve continuity, etc.)



Diffusion maps

Coifman et al. PNAS 2005

row normalised transition matrix P (from the cells' pairwise Gaussian kernel matrix W)

$$P_{n \times n} = D^{-1}W = D^{-1/2}(D^{-1/2}WD^{-1/2})D^{1/2}$$

$$D_{ii} = \sum_j W_{ij}$$

P^t transition probabilities in t steps

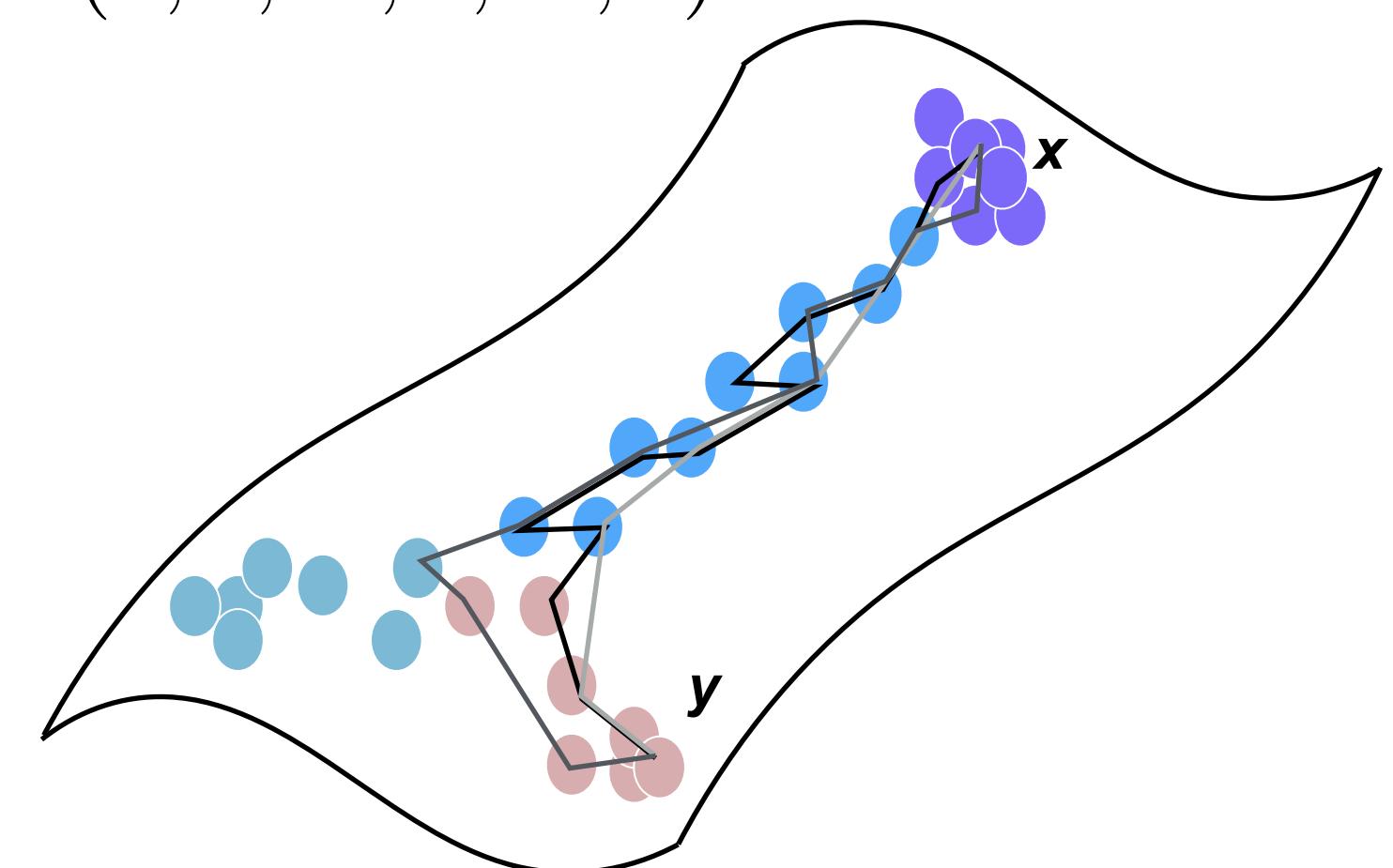
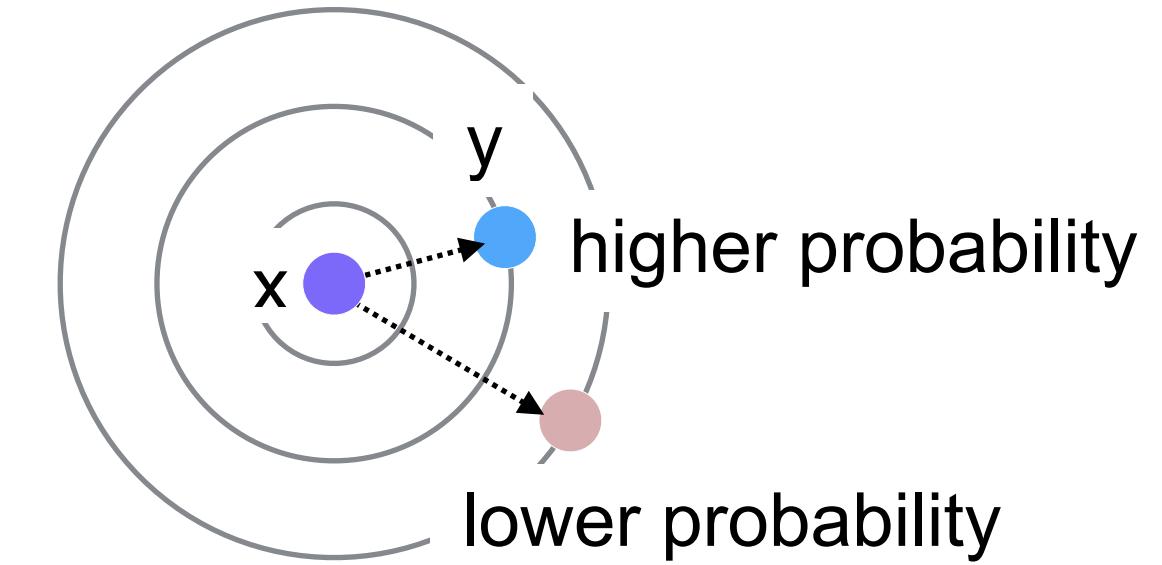
$$\phi_i P = \lambda_i \phi_i, \quad P \psi_i = \lambda_i \psi_i, \quad \lambda_i \leq 1$$

The state vector X can represent one cell (with only one nonzero value at position x) but can a distribution of cells (with more 1 enters)

$$\begin{aligned} D_t^2(x, y) &= \|P^t(x, \cdot) - P^t(y, \cdot)\|_{1/\phi_0}^2 \\ &= \sum_{i=1}^{n-1} \lambda_i^{2t} (\psi_i(x) - \psi_i(y))^2 \end{aligned}$$

$$P^t(x, \cdot) = X \cdot P^t \quad X = (0, 0, \dots, 1, \dots, 0)$$

For large t we can cut the sum at $i=K$ for low-dimensional approximation



Diffusion pseudotime

$$T_{n \times n} = D^{-1/2} W D^{-1/2}$$

Symmetric transition matrix

T^t Transition probabilities in t steps

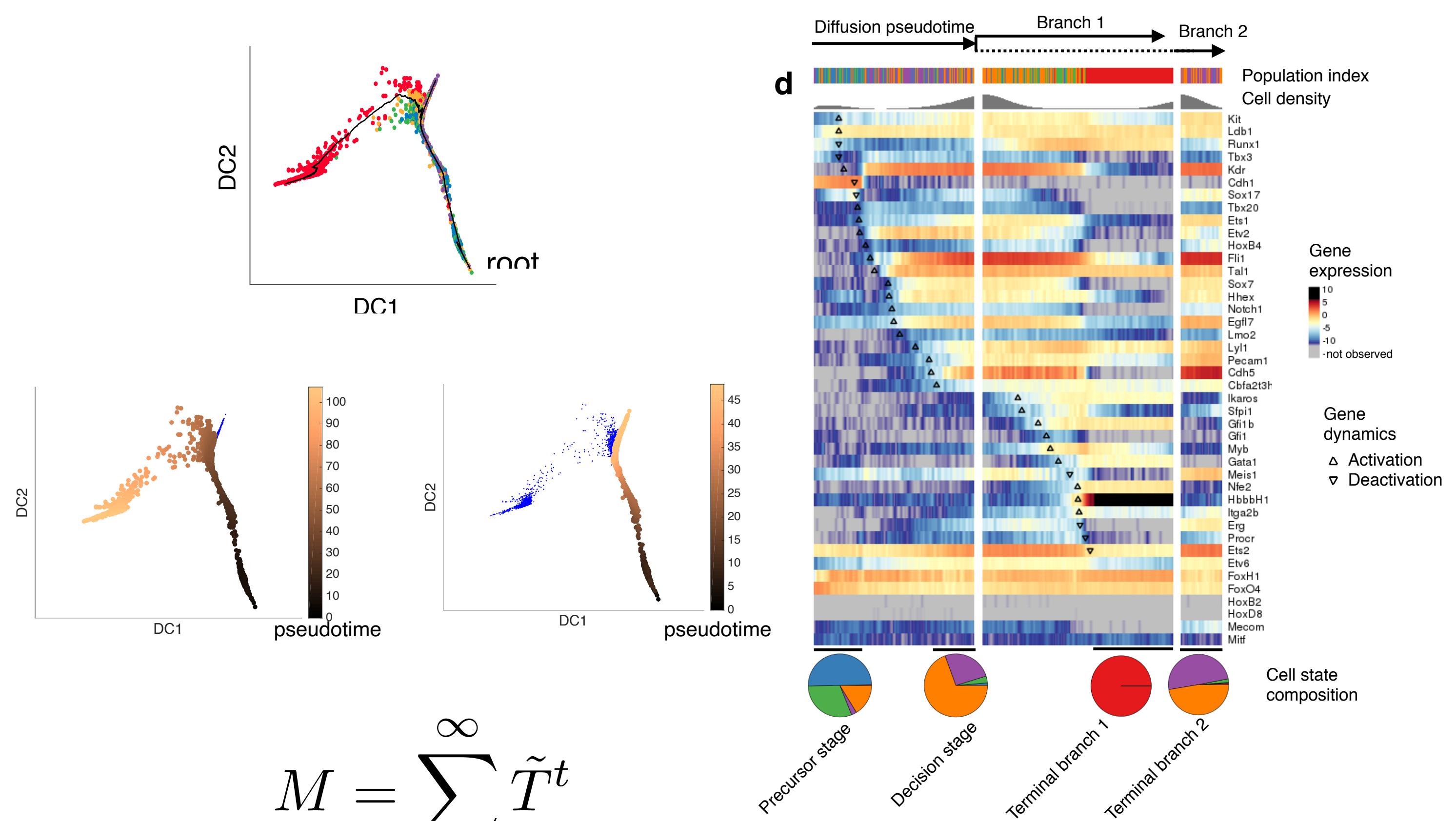
$$\phi_i T = \lambda_i \phi_i, \quad T \phi_i = \lambda_i \phi_i, \quad \lambda_i \leq 1$$

$$\tilde{T} = T - \phi_0 \phi_0^T \quad \text{Remove the stationary state}$$

$$dpt_t^2(x, y) = ||M(x, .) - M(y, .)||^2 = XMX^T + YMY^T - XMY^T - YM X^T$$

$$= \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_i} \right)^2 (\phi_i(x) - \phi_i(y))^2$$

- "Diffusion pseudotime robustly reconstructs lineage branching" Haghverdi et al. Nature methods 2016
 - "Geometric diffusions for reconstruction of cell differentiation dynamics".
- 15 Haghverdi, L., Diss. Technische Universität München, 2016.



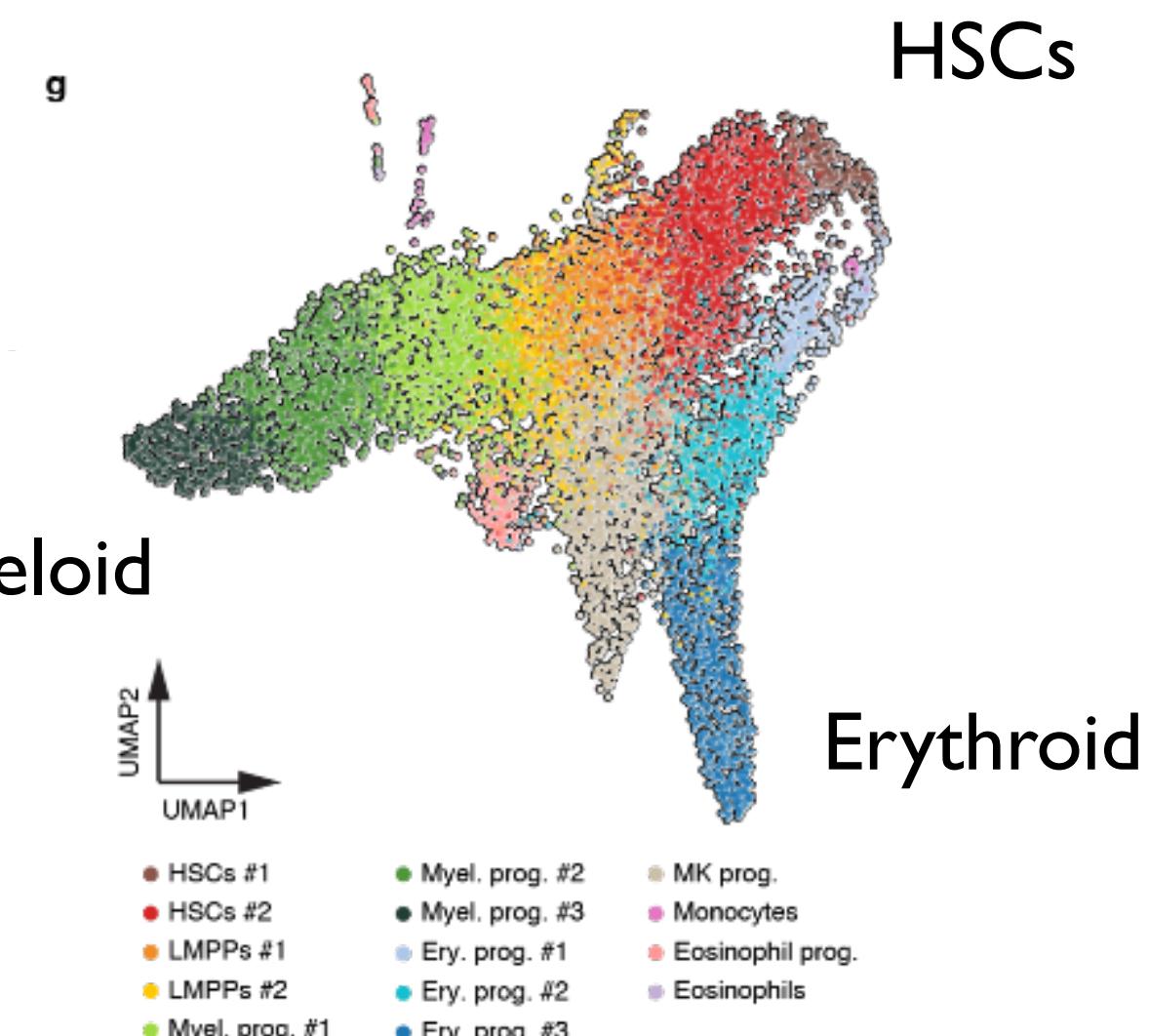
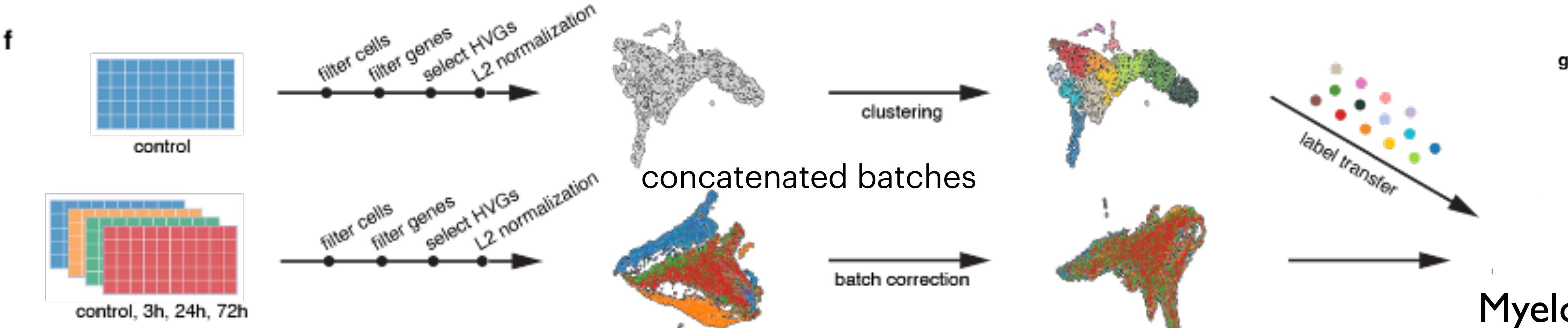
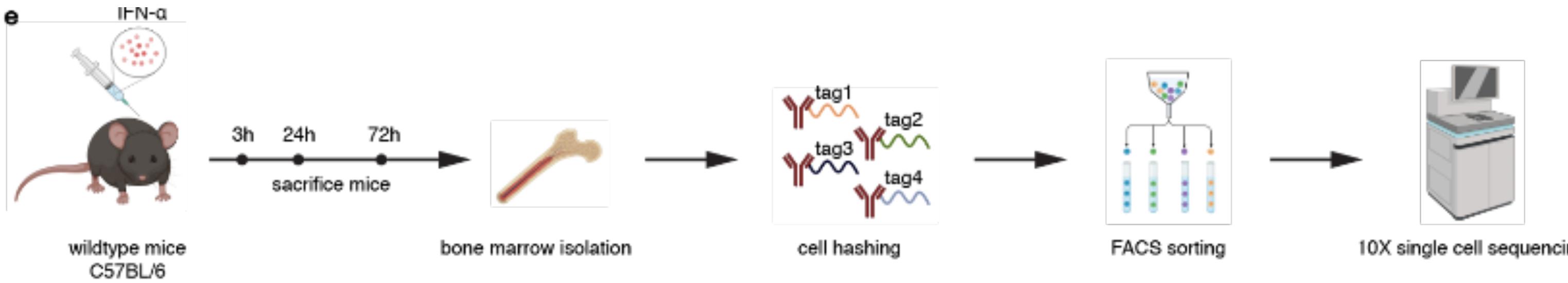
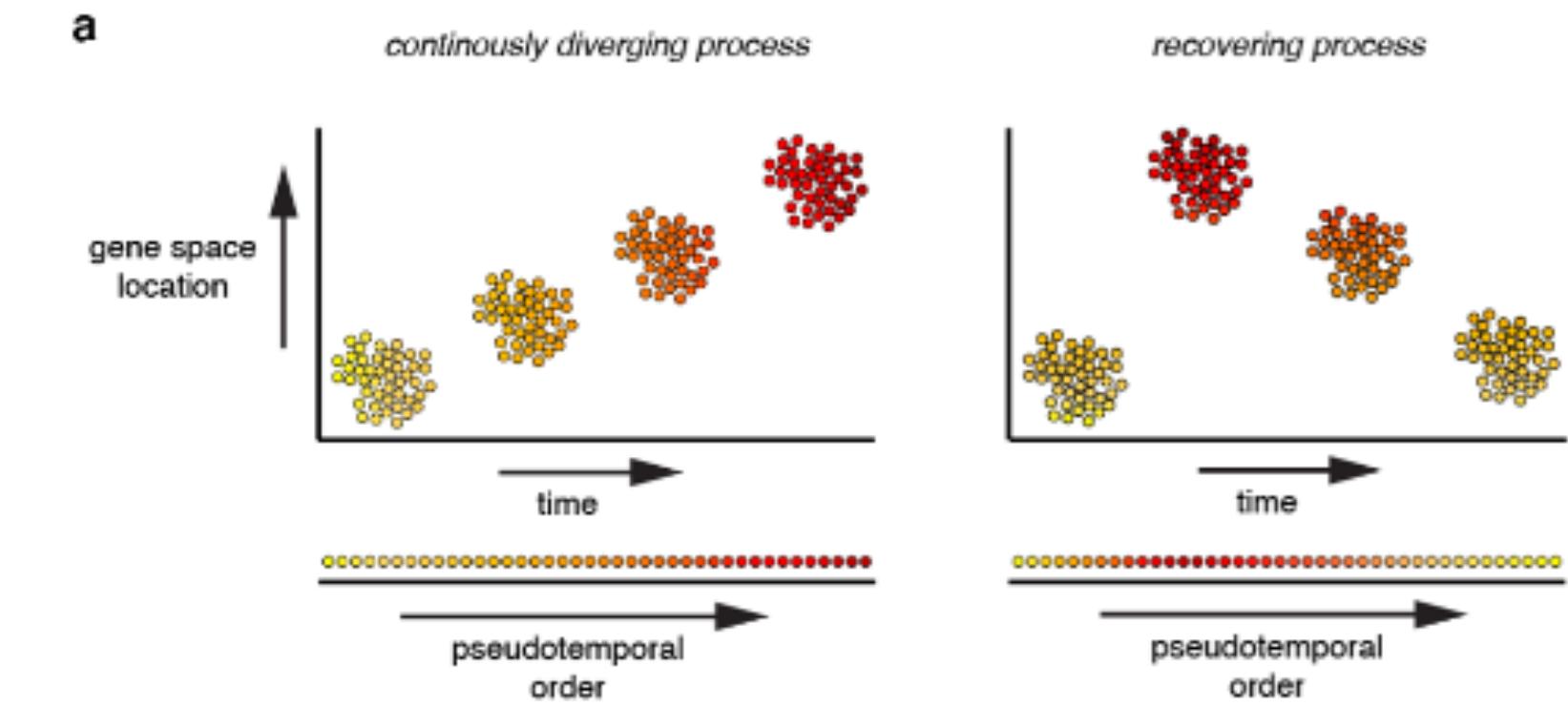
Single-cell time series analysis and response pseudotime

Resource



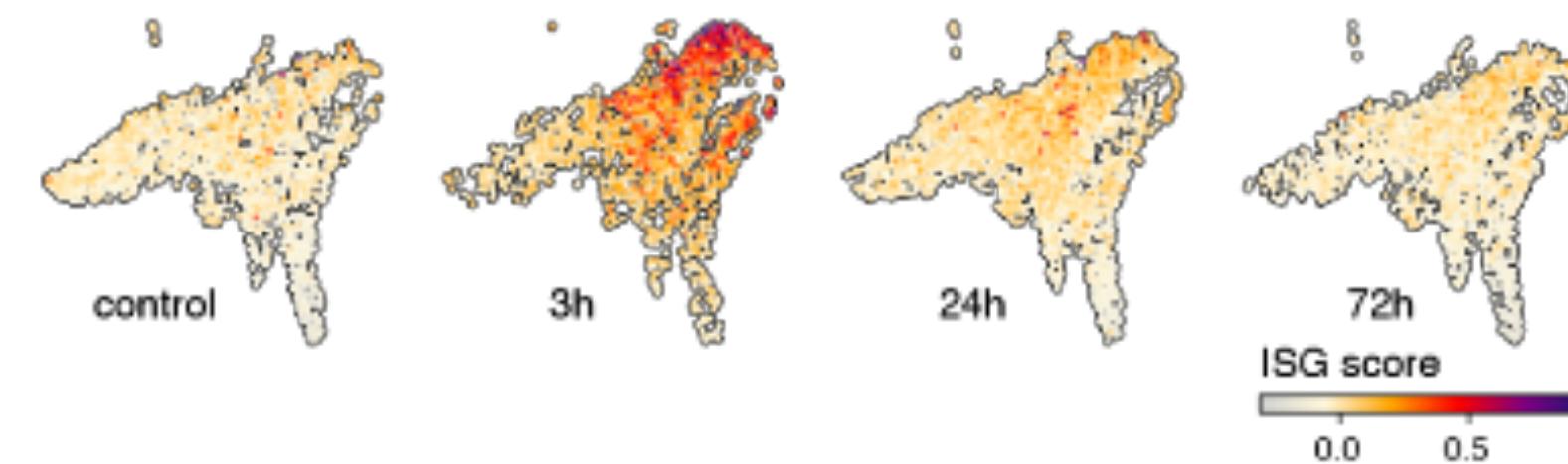
Single-cell time series analysis reveals the dynamics of HSPC response to inflammation

Brigitte J Bouman^{1,2,*}, Yasmin Demerdash^{3,4,5,*}, Shubhankar Sood^{3,4,5}, Florian Grünschläger^{4,5,6}, Franziska Pilz^{3,4}, Abdul R Itani^{3,4,5} , Andrea Kuck^{3,4}, Valérie Marot-Lassauzaie^{1,11} , Simon Haas^{1,4,7,8,9,10,11}, Laleh Haghverdi¹ , Marieke AG Essers^{3,4,12}

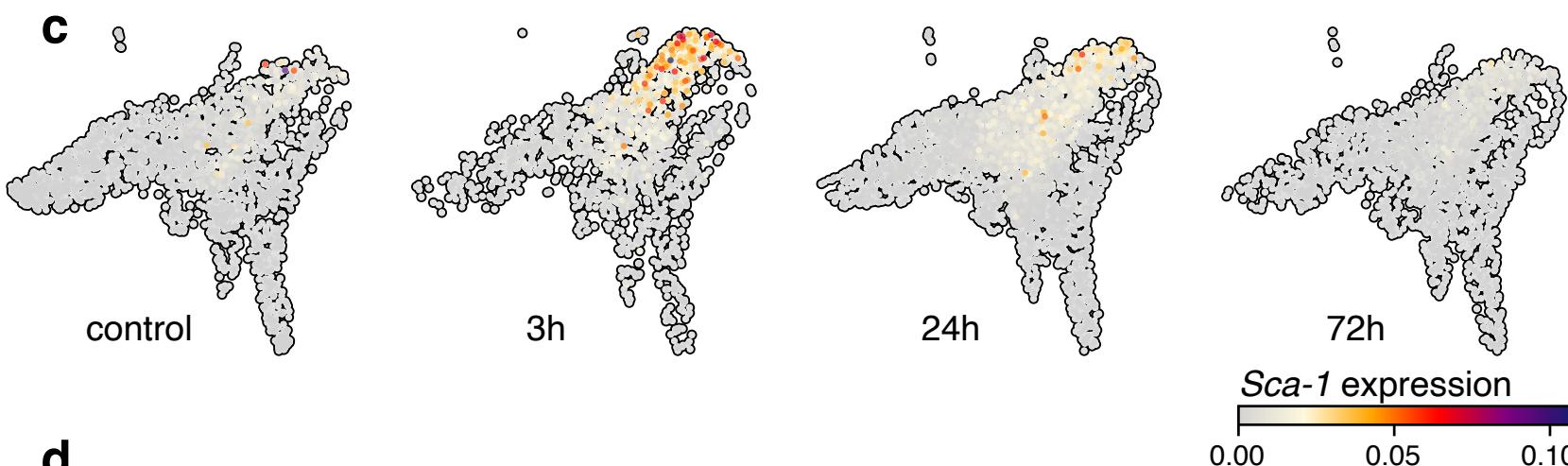


Inflammation response of gene expression

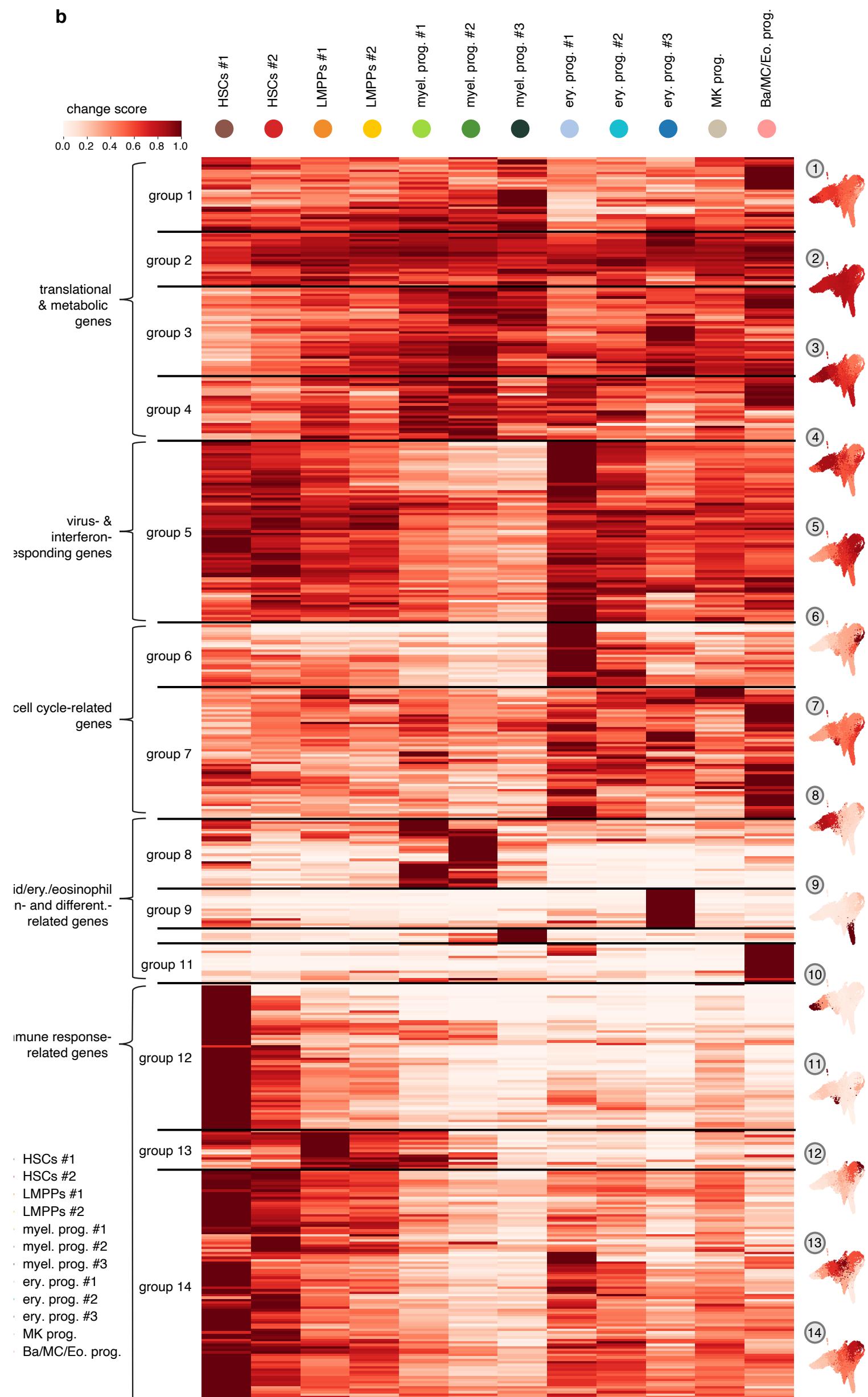
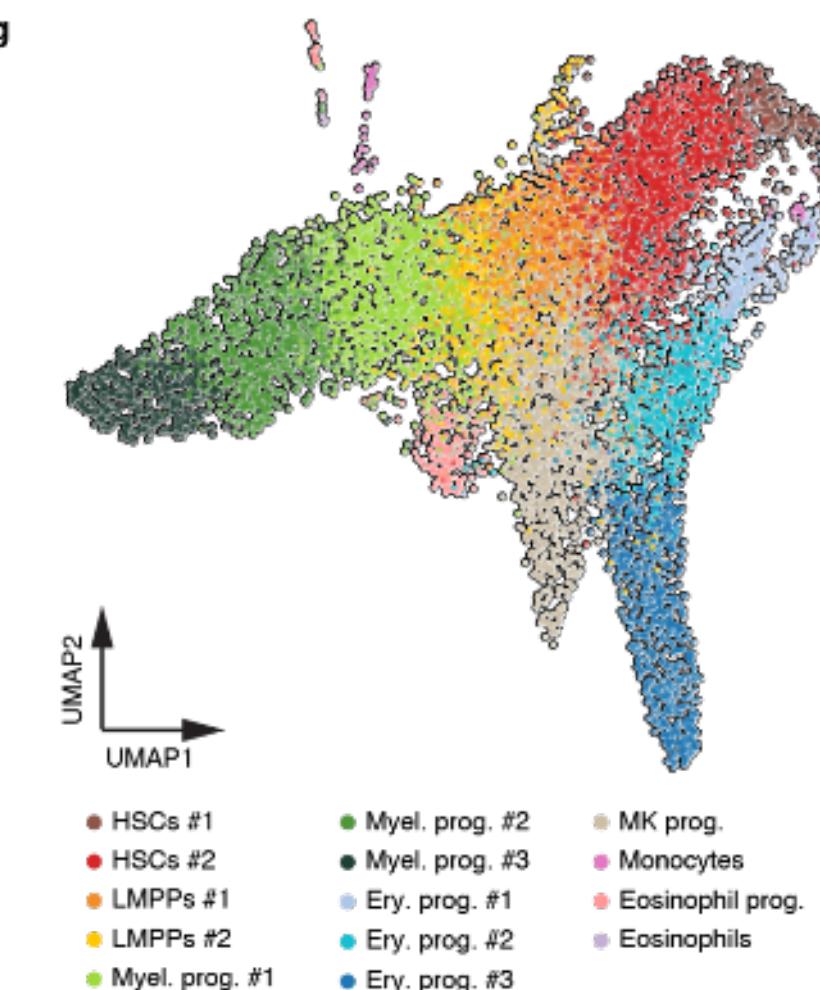
Interferon stimulated genes



(Sca-1) Stem Cell marker

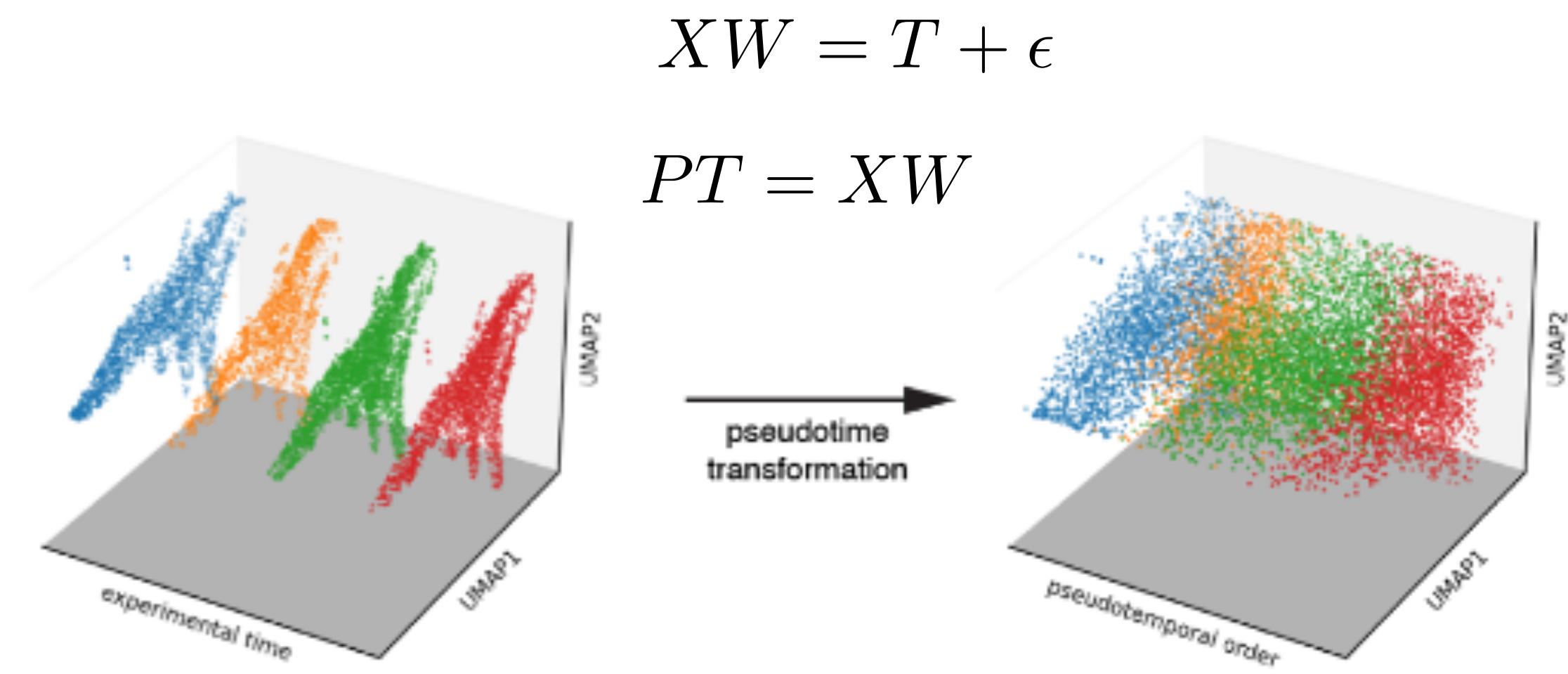


- Locally changing → cell type specific responses
- Globally changing → allows a pseudo time ordering of cells irrespective of their cluster

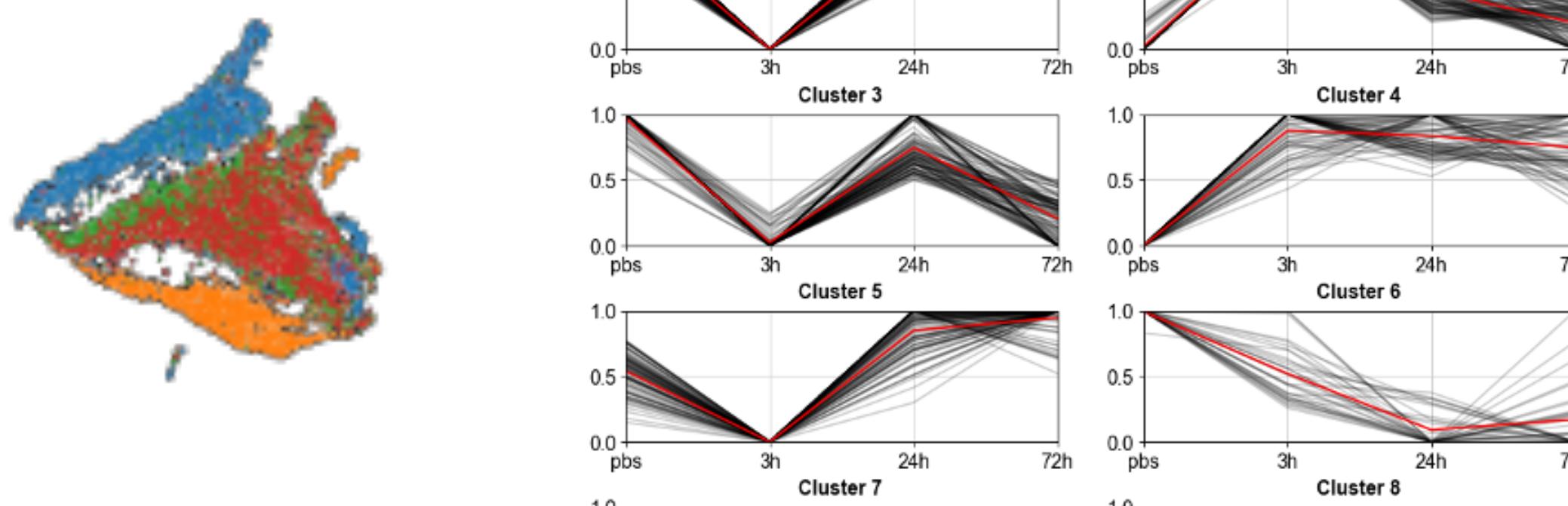


Response pseudotime (semi-supervised)

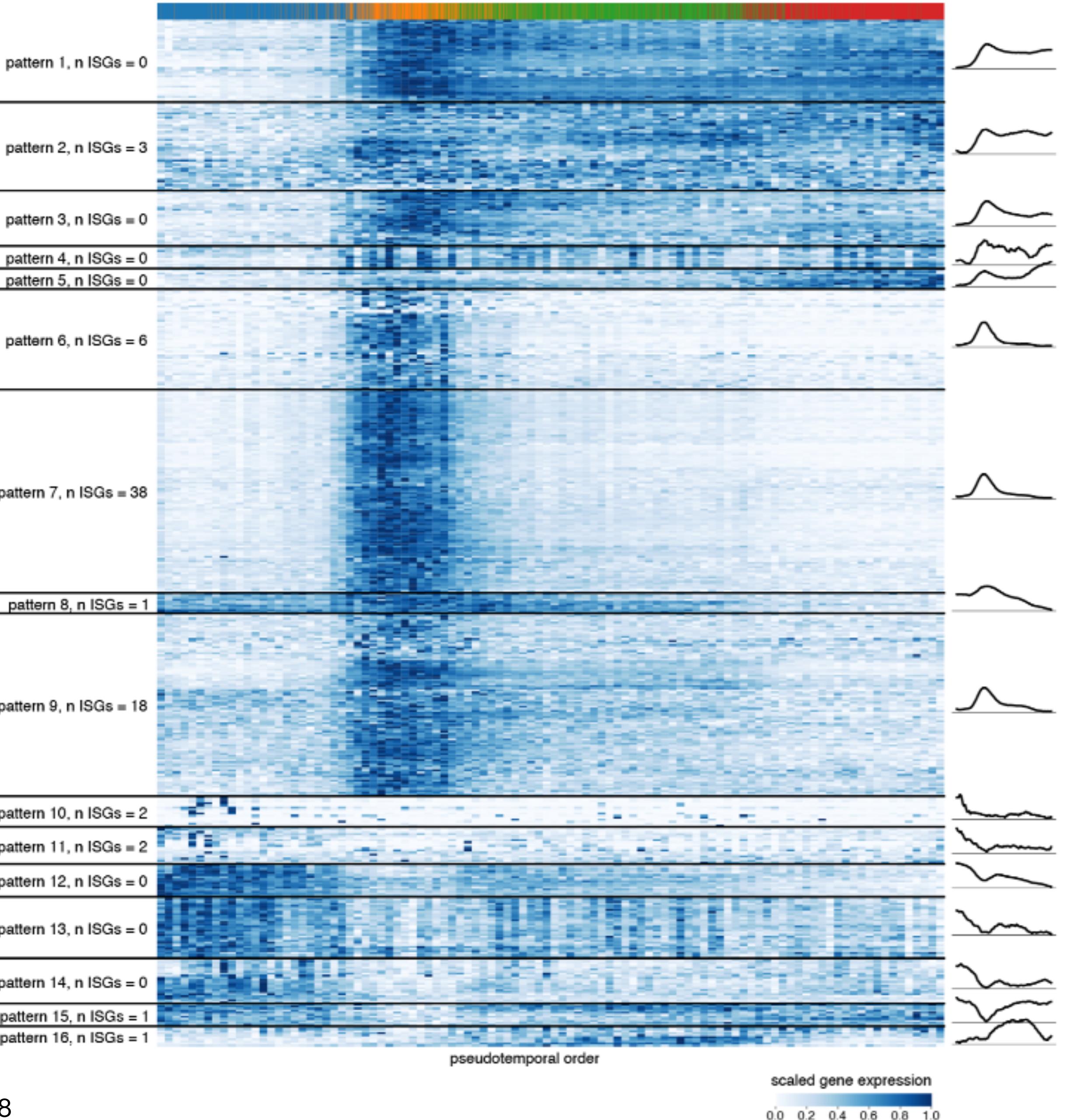
- Use the globally (across all cell types) changing genes for response pseudotime estimation



Discrete Dynamics

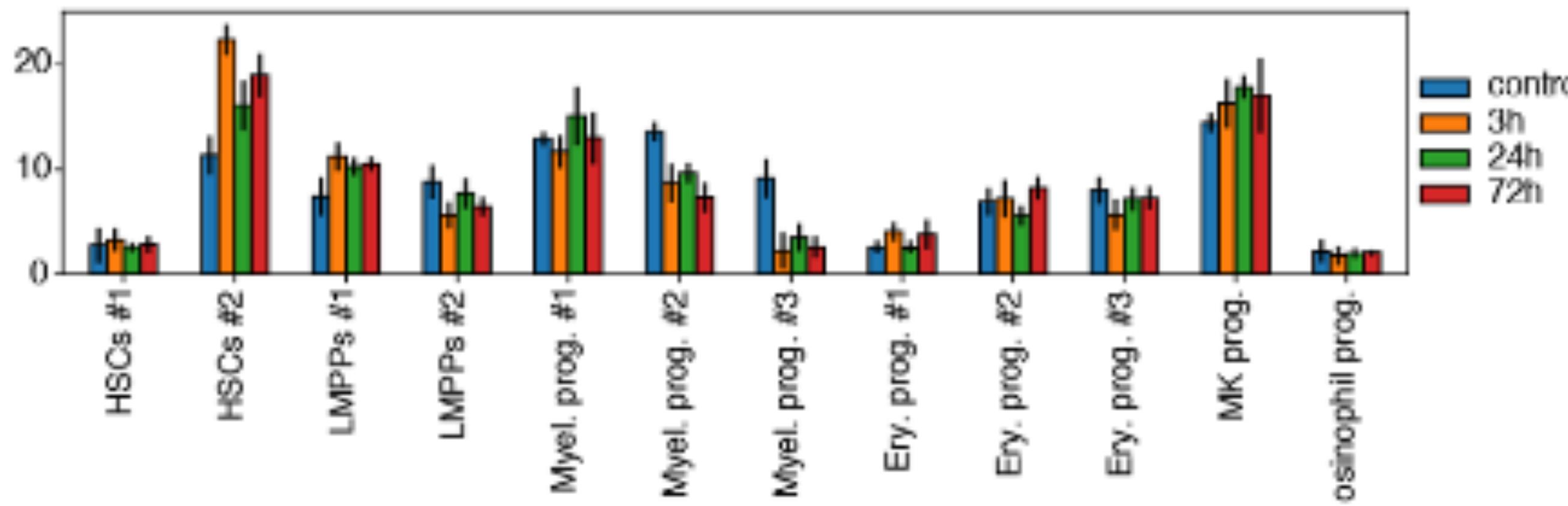
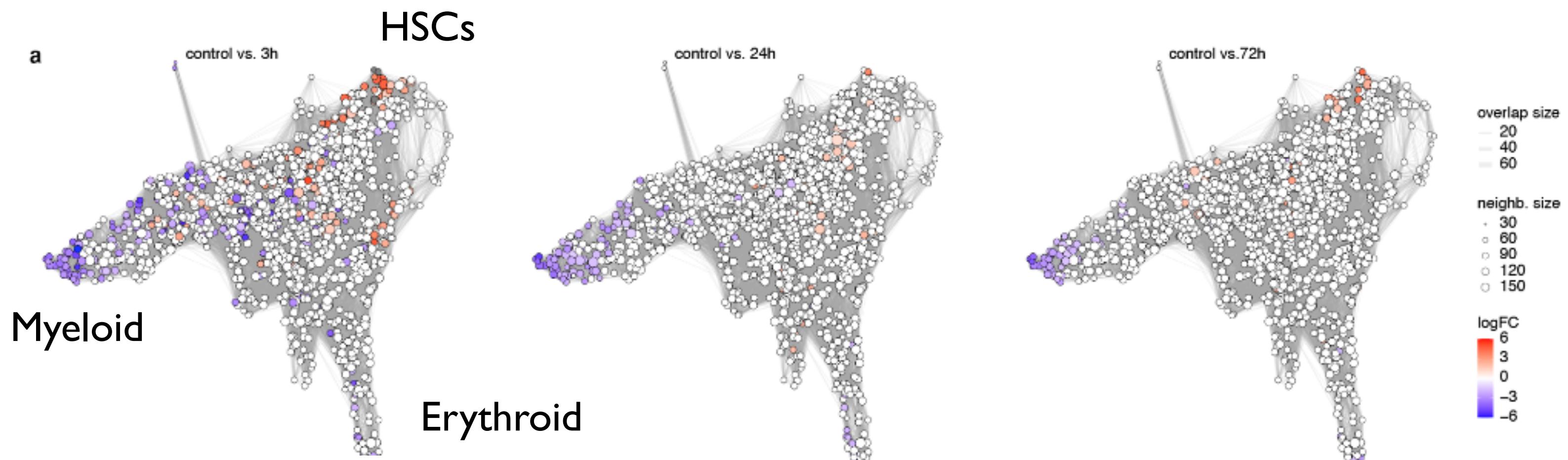


Continuous response pseudotime

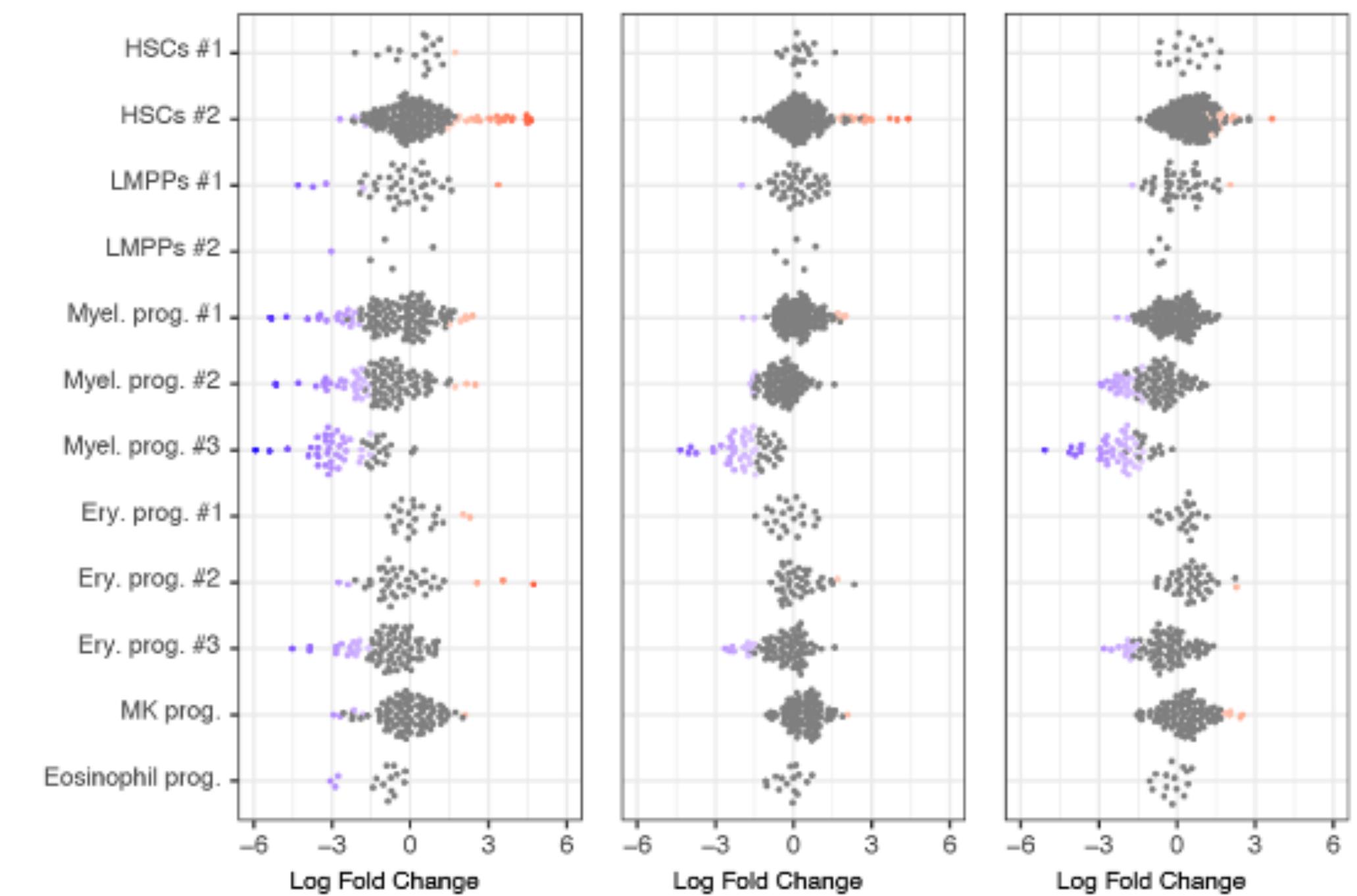


Abundance analysis

- HSCs increase
- Myeloids decrease

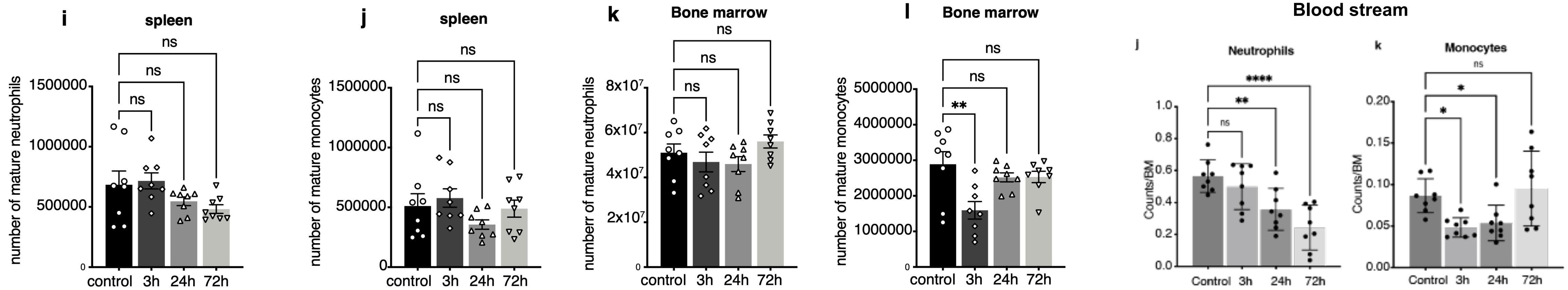
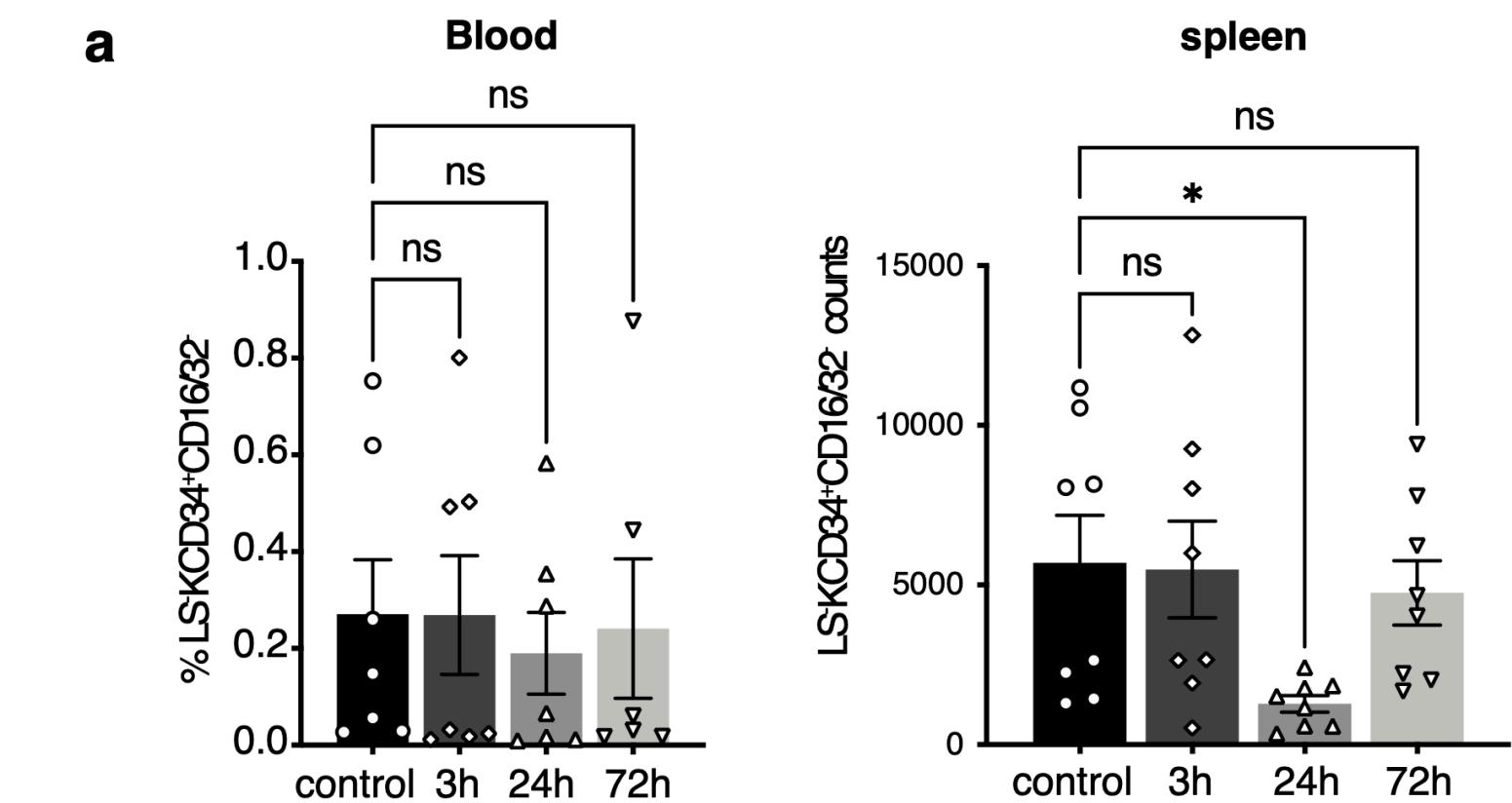


- Change in drift (i.e., differentiation bias)?
- Change in proliferation/death rates? — cells migration?

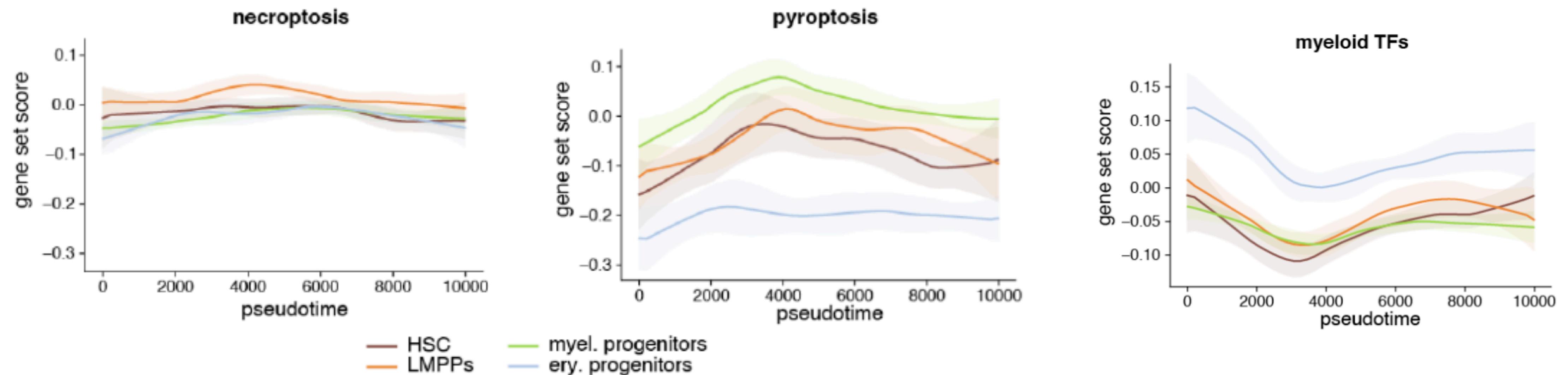


Myeloids and progenitors migrate? FACs checks

- Myeloids decrease in blood and spleen (in addition to BM)
- Monocytes and Neutrophils as well
- No significant change in migration pattern
- Overall depletion of myeloid population

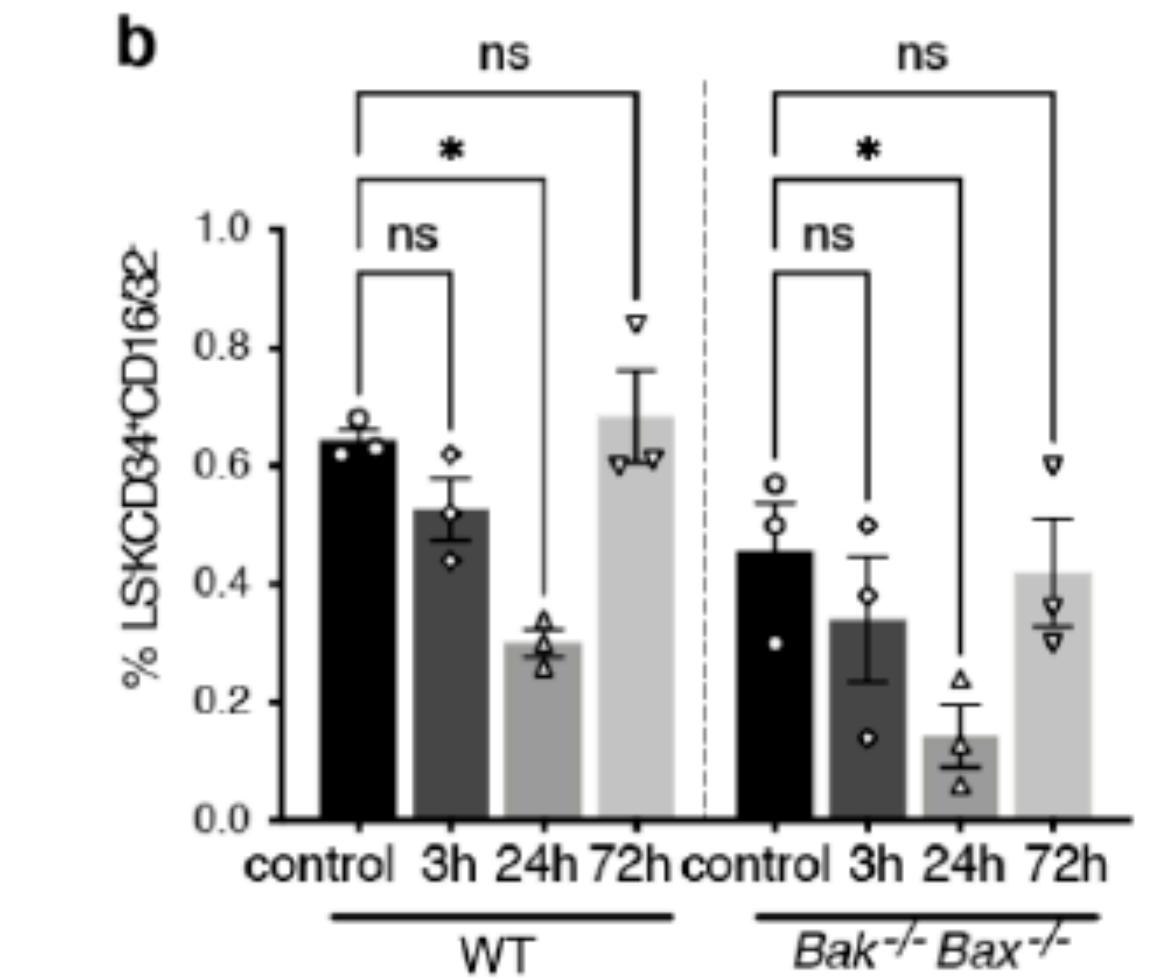


Gene expressions over pseudotime



- Ruled out apoptosis
- Change in proliferation/death rates? — Probably yes
- Cells migration? — Probably not
- Change in drift? — Probably yes, a global change post treatment
- Computational validation by parallel methods?

Same density change pattern in apoptosis knock-out mice



Understanding cell fate decision-making

Stem Cell Reports

ISSCR

Perspective

Single-cell multi-omics and lineage tracing to dissect cell fate decision-making

Laleh Haghverdi^{1,3,*} and Leif S. Ludwig^{1,2,3,*}

¹Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany

²Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

³These authors contributed equally

*Correspondence: laleh.haghverdi@mdc-berlin.de (L.H.), leif.ludwig@bih-charite.de (L.S.L.)

<https://doi.org/10.1016/j.stemcr.2022.12.003>

- Supplemental Note → Mathematical relation between Dynamical Inference
 - Pseudotime
 - Optimal Transport
 - Cell state velocities

$$\frac{\partial}{\partial t} p(s, t) = \nabla \cdot \left(\underline{\underline{\nabla D(s)p(s, t)}} + \underline{p(s, t)\nabla U(s)} + \underline{\nabla S(s, t)p(s, t)} \right)$$

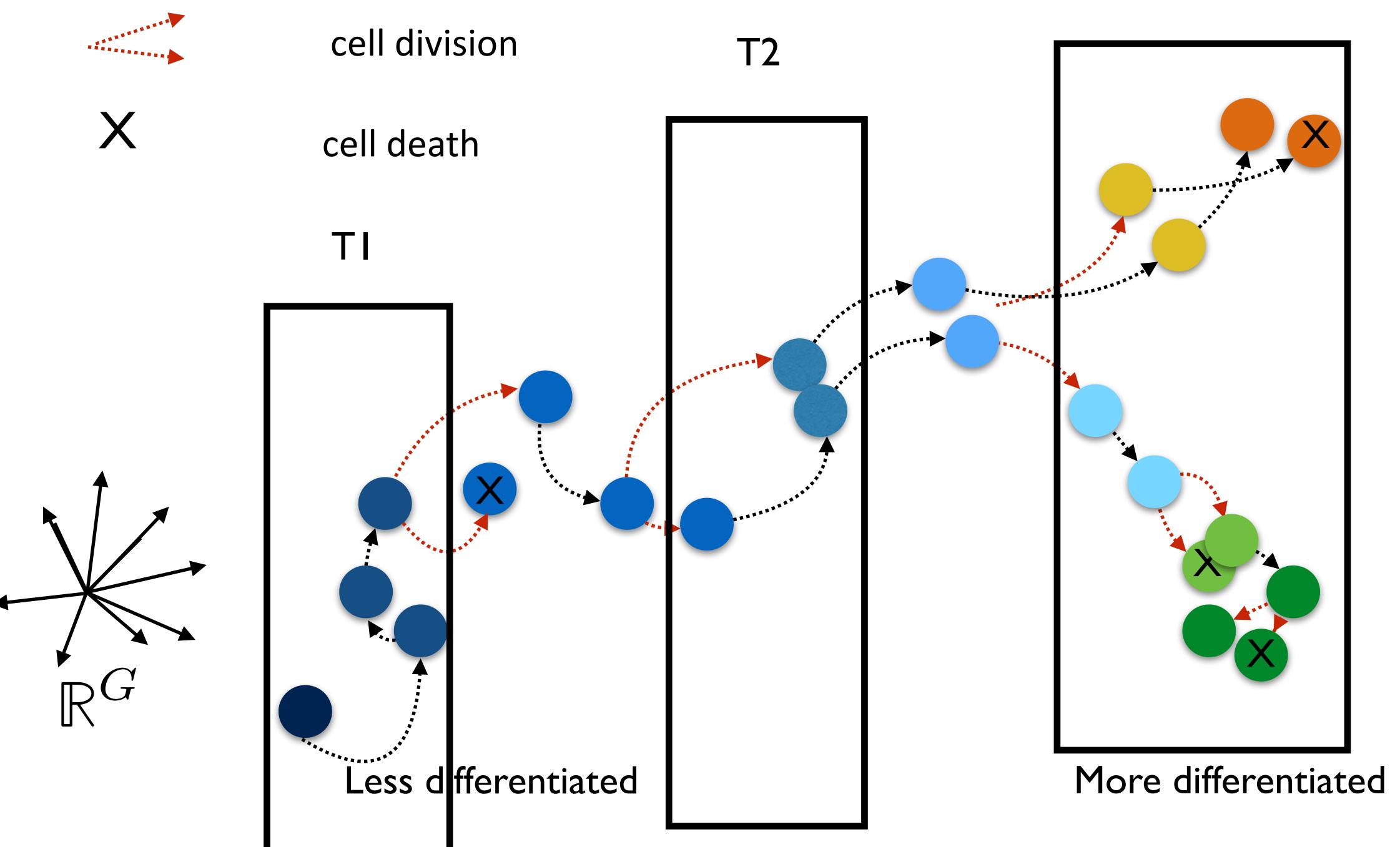
Probability density at position s, time t

Diffusion coefficient (Position dependent)

Drift force (Position dependent)

Birth/death (Position dependent)

OPEN ACCESS



Diffusion-drift relation with cell state velocities

$$\frac{\partial}{\partial t} p(s, t) = \nabla \cdot \left(\nabla D(s)p(s, t) + p(s, t)\nabla U(s) + \nabla S(s, t)p(s, t) \right)$$

We can rewrite equation 1 as:

- J: probability flux

$$\frac{\partial}{\partial t} p(s, t) = \nabla \cdot \vec{J}(s, t)$$

$$\begin{aligned} J(s, t) &= \nabla D(s)p(s, t) + p(s, t)\nabla U(s) + \nabla S(s, t)p(s, t) \\ &= \vec{V}(s)p(s, t) + \nabla S(s, t)p(s, t) \end{aligned}$$

- V: Cell state velocity

- Langevin equation

$$\vec{V}(s)p(s, t) = \nabla D(s)p(s, t) + p(s, t)\nabla U(s)$$

Optimal Transport relation with the diffusion-drift drift model

Π : transition matrix

C : distance matrix

\mathbb{P} : vector estimate of birth-death rate for the cells in T1

\mathbb{Q} : vector of 1 entries for all cells in T2

$$\pi_{ij} = \operatorname{argmin}_{\pi} \left(\sum_{i \in 1:N1, j \in 1:N2} c(s_i, s_j) \pi_{ij} - \epsilon \sum_{i \in 1:N1, j \in 1:N2} \pi_{ij} \log \pi_{ij} \right)$$
$$\beta_1 \tilde{KL}\left(\sum_{i \in 1:N1} \pi_{ij} || \mathbb{Q}\right) + \beta_2 \tilde{KL}\left(\sum_{j \in 1:N2} \pi_{ij} || \mathbb{P}\right)$$

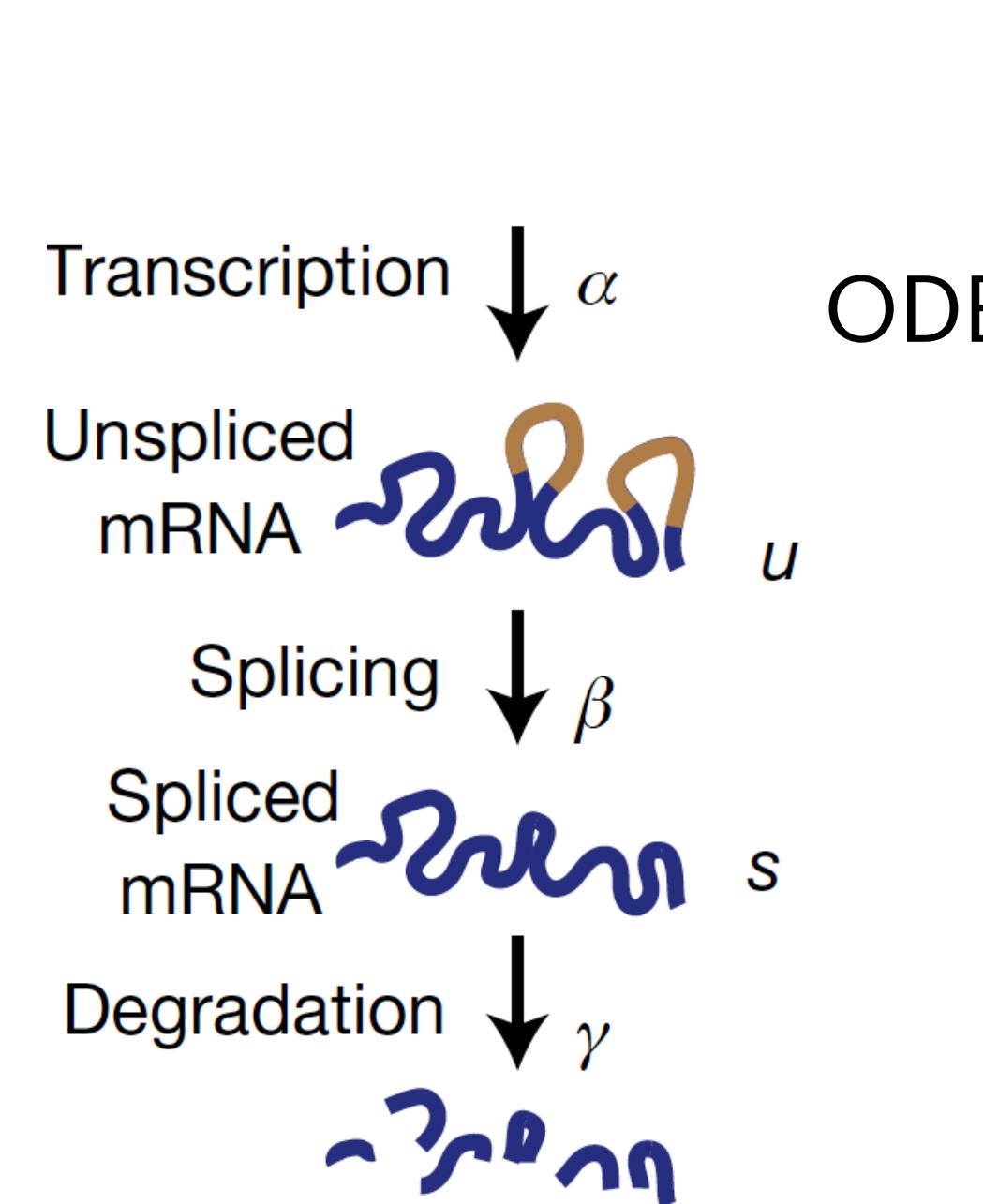
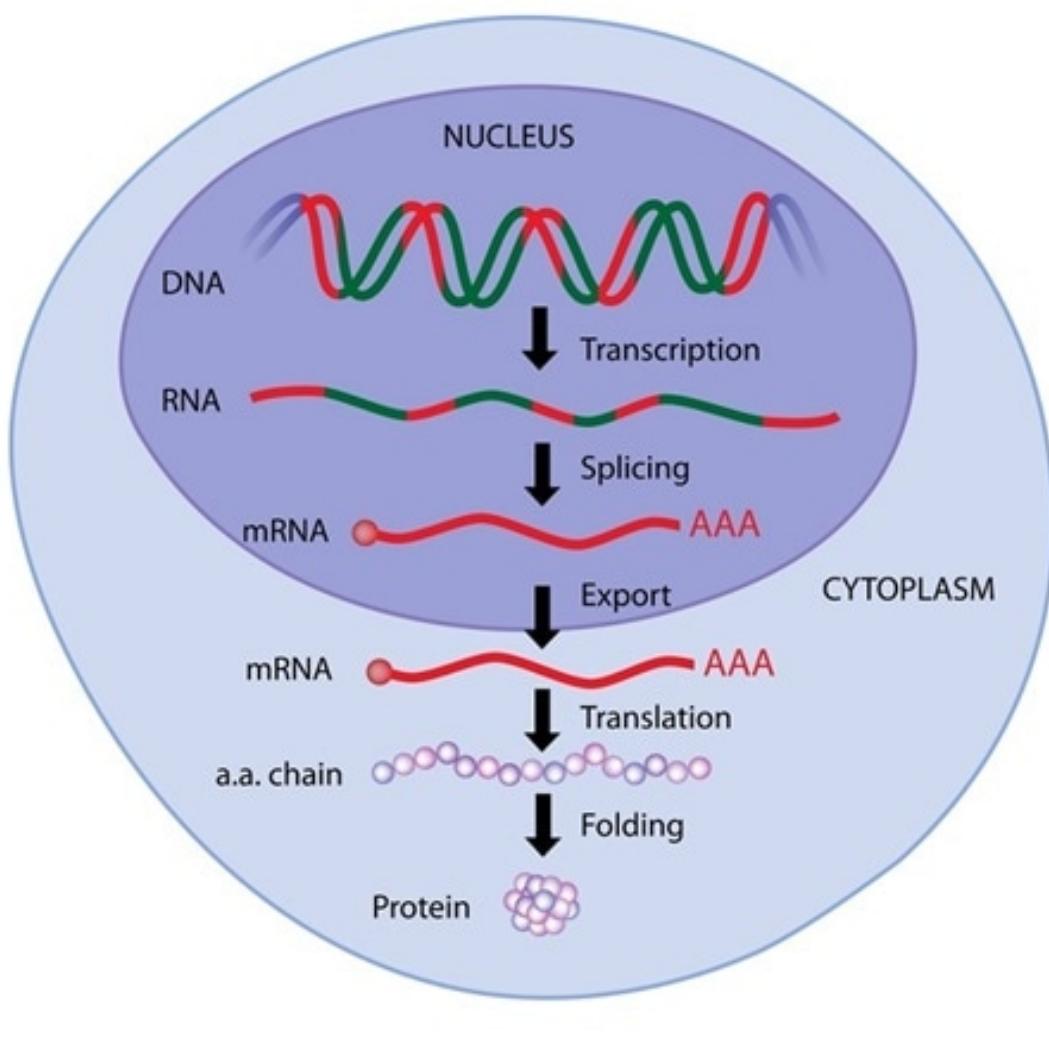
Birth/date

Log-likelihood optimisation of discrete diffusion-drift \rightarrow Optimal Transport Under certain assumptions:

- Change in energy \sim Euclidean distance between cell states
- Same diffusion parameter for all cell states (i.e, positions)

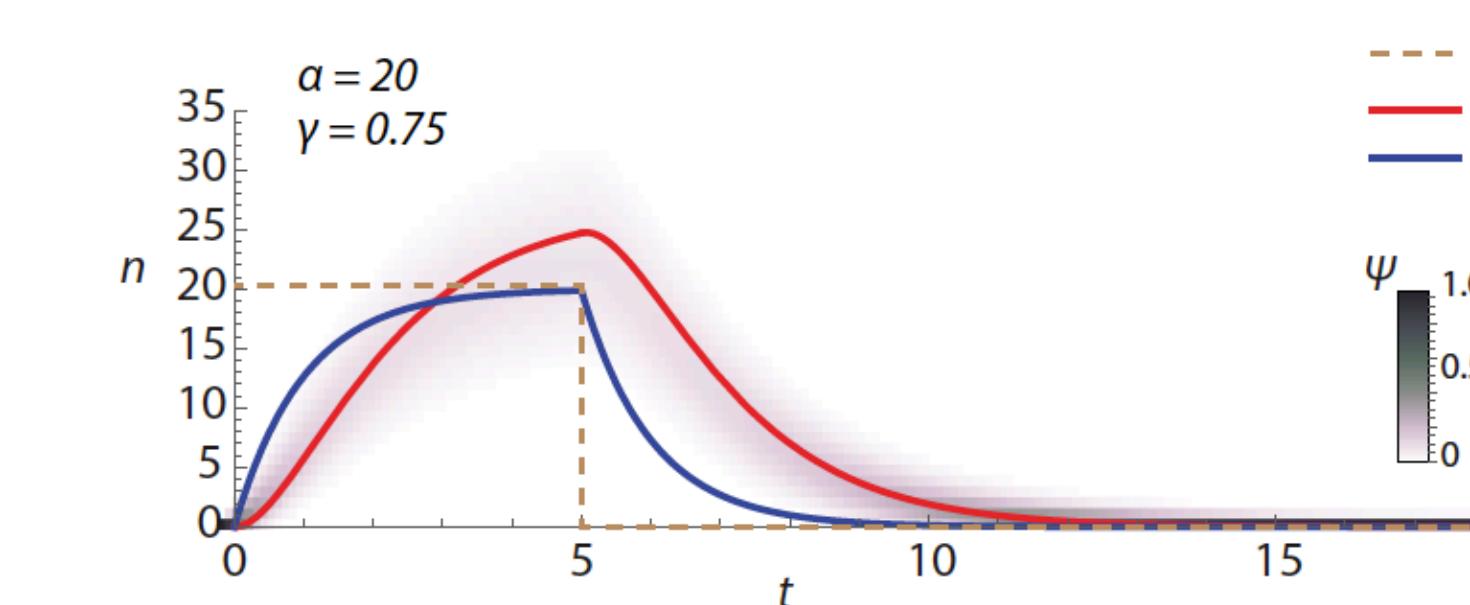
Cell state velocities

La Manno et al,
Nature 2018

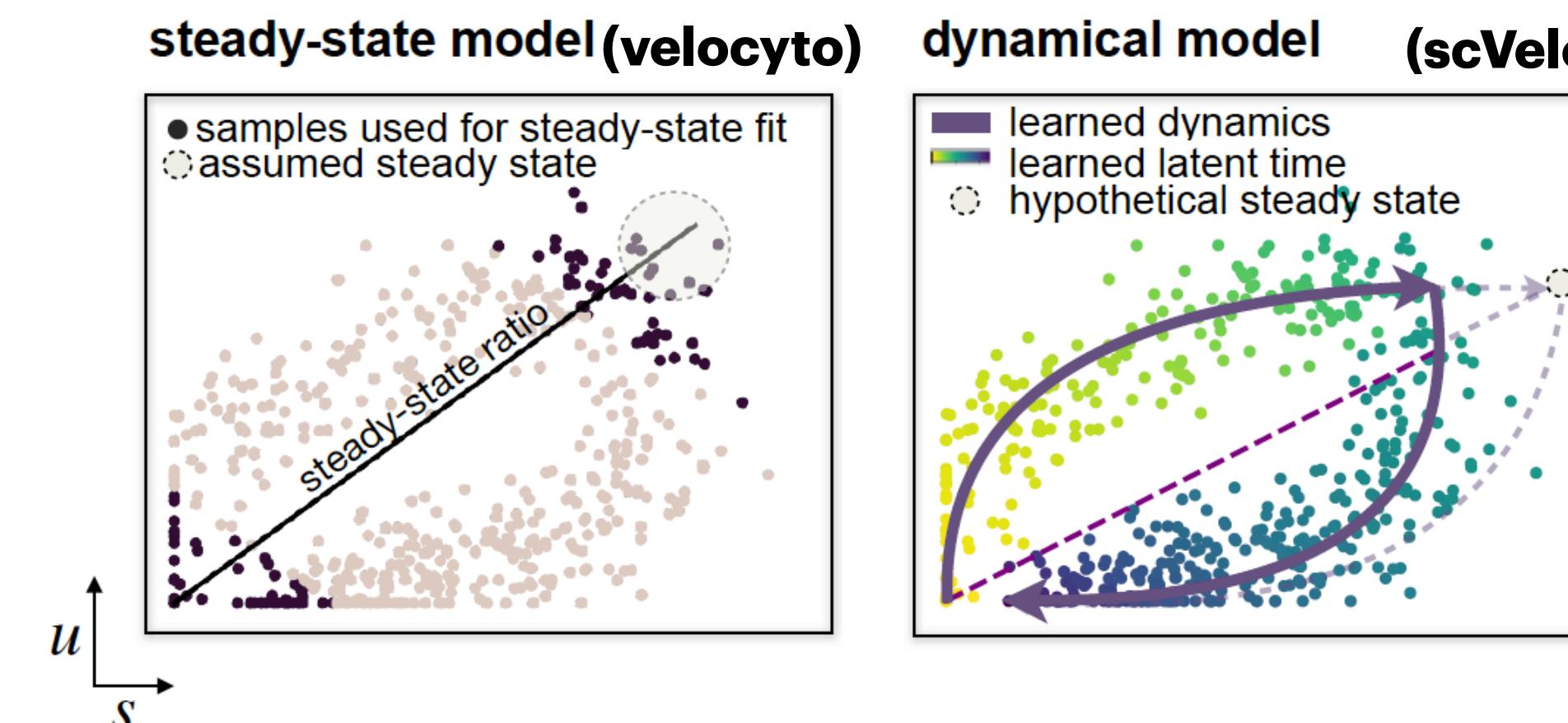
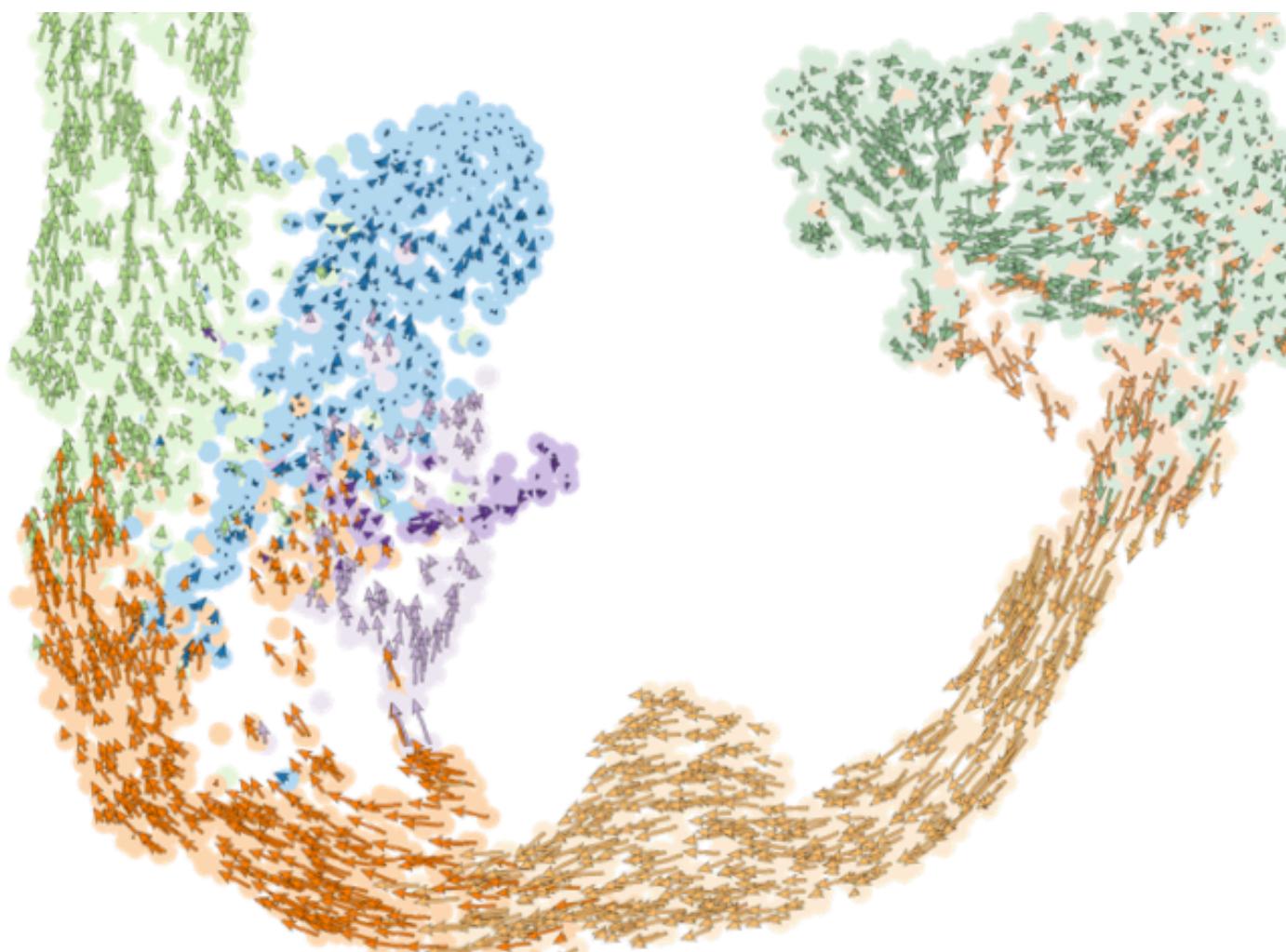
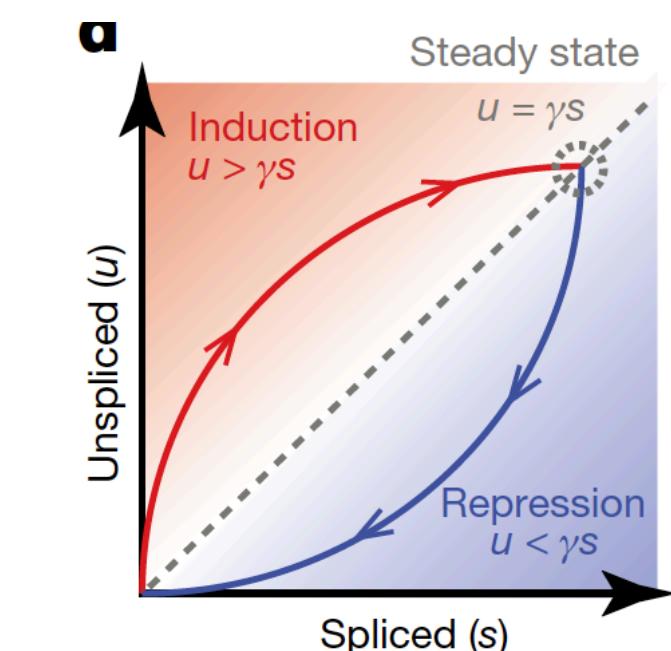


$$\frac{du_g}{dt} = \alpha_g - \beta_g u_g$$

$$\frac{ds_g}{dt} = \beta_g u_g - \gamma_g s_g = v_g$$

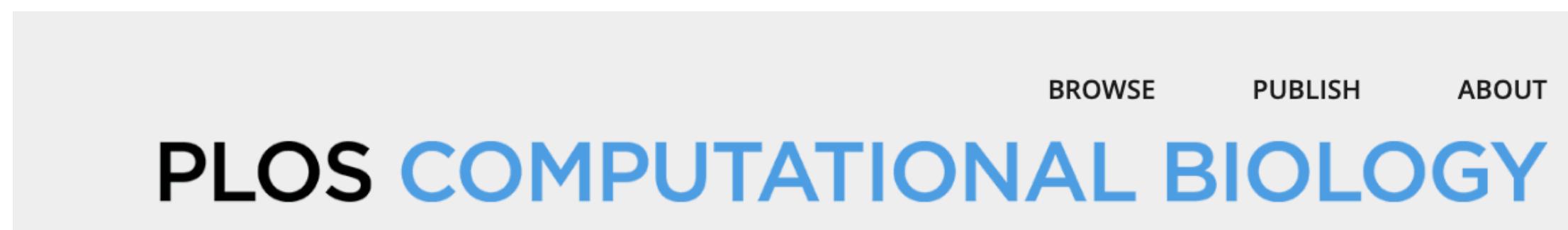


$$\vec{V} = \sum_{g=1}^G \vec{v}_g$$



Bergen et al, Nature Biotechnology 2020

Challenges of cell state velocities estimation



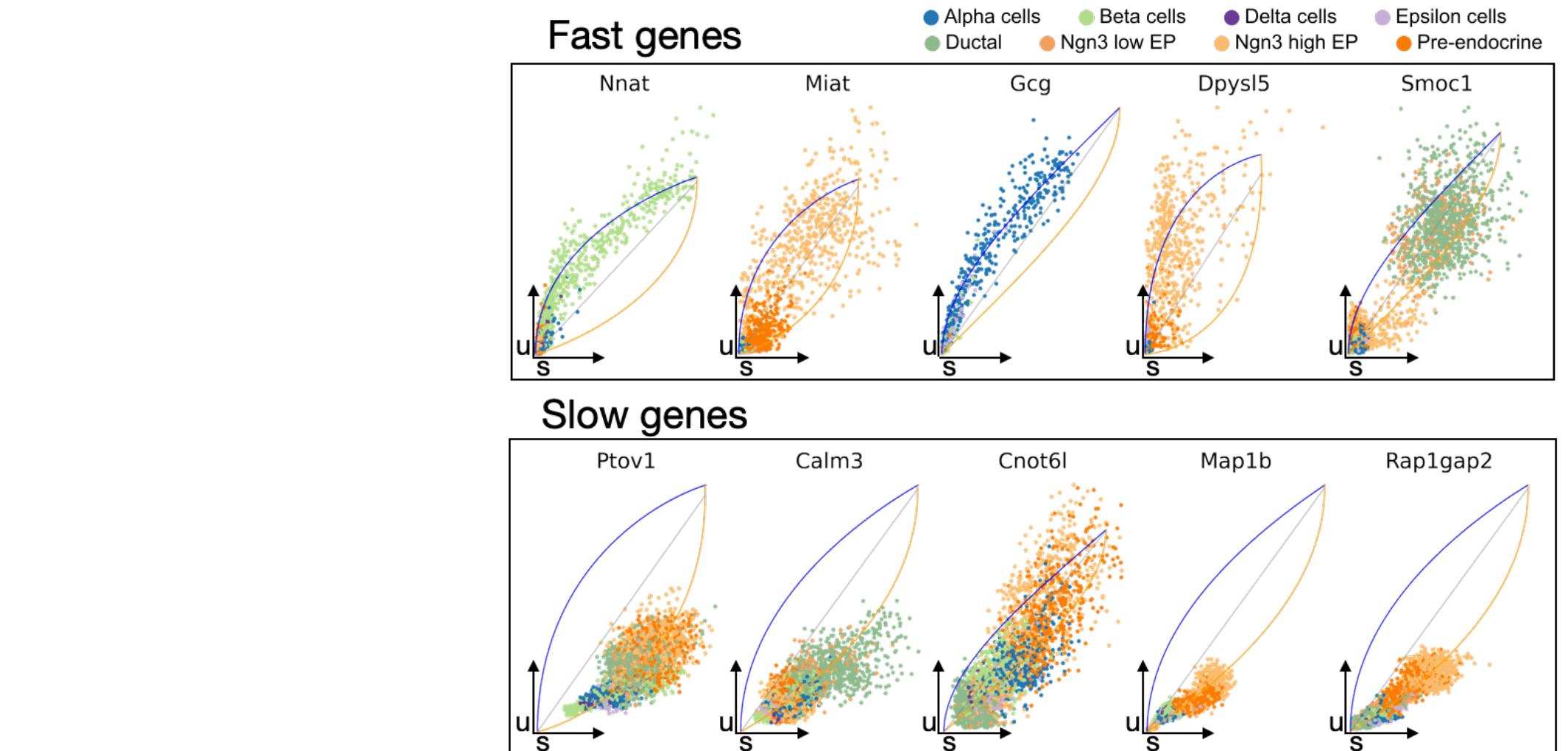
OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Towards reliable quantification of cell state velocities

Valérie Marot-Lassauzaie , Brigitte Joanne Bouman , Fearghal Declan Donaghy, Yasmin Demerdash, Marieke Alida Gertruda Essers, Laleh Haghverdi

Version 2 Published: September 28, 2022 • <https://doi.org/10.1371/journal.pcbi.1010031>

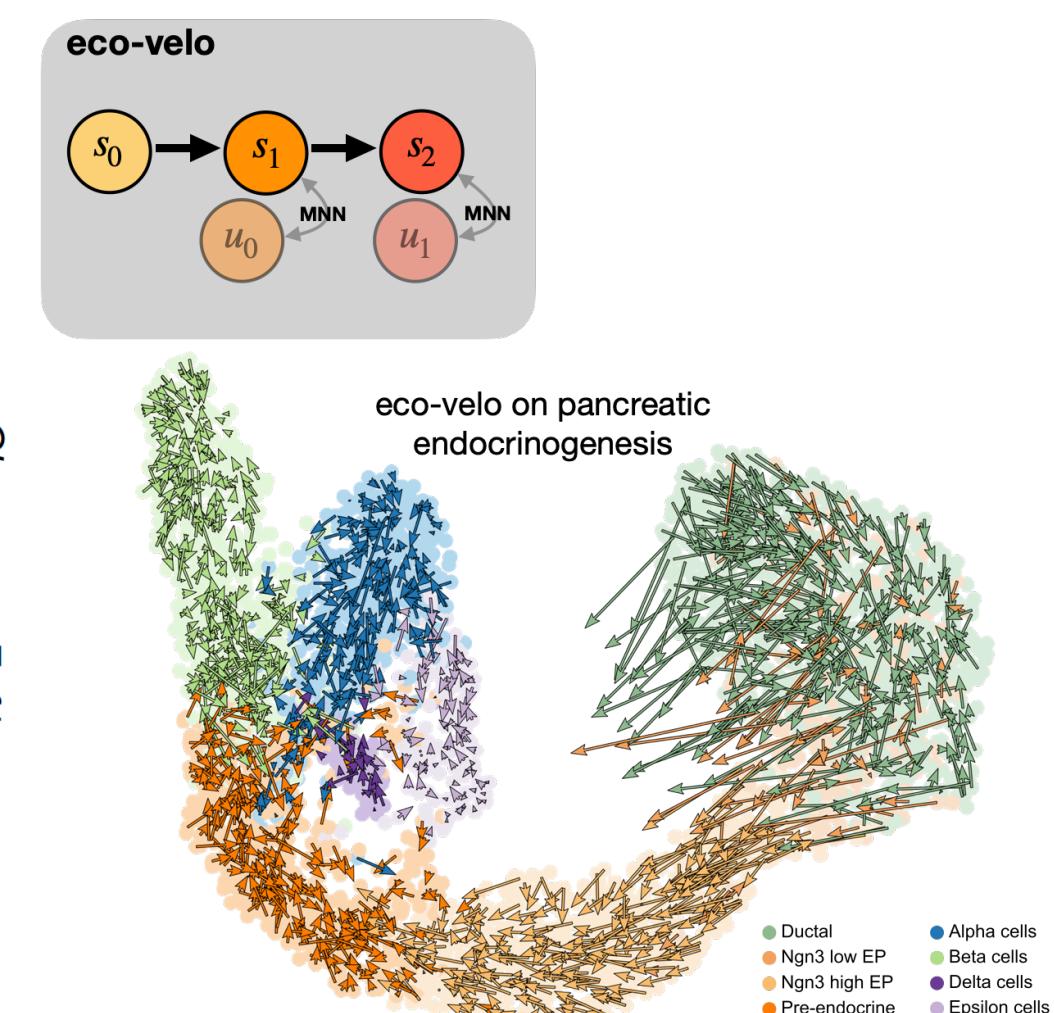
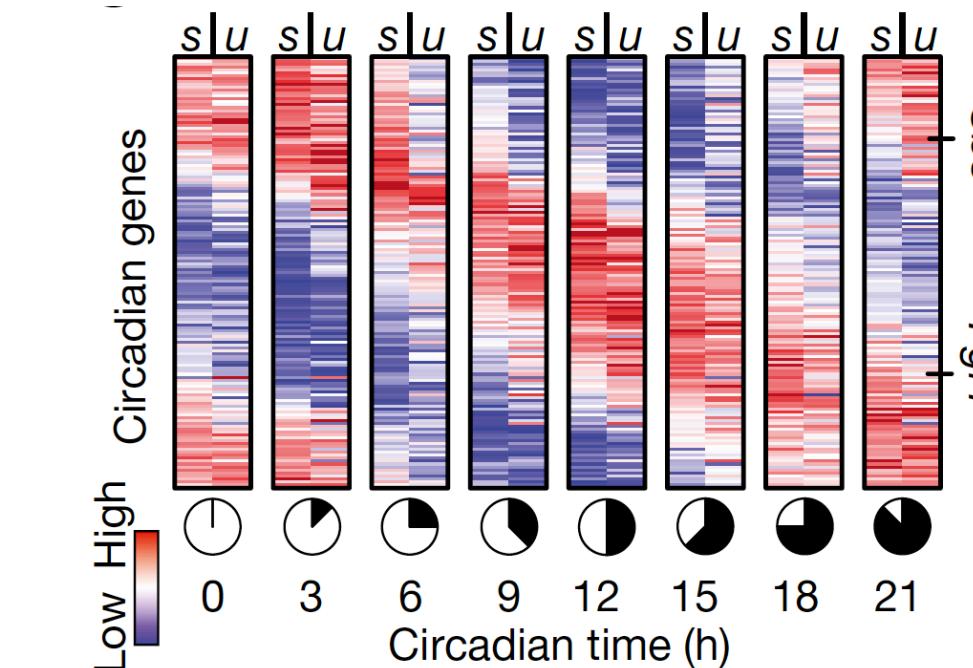
- Kappa-velo: address scale invariance by using cell densities



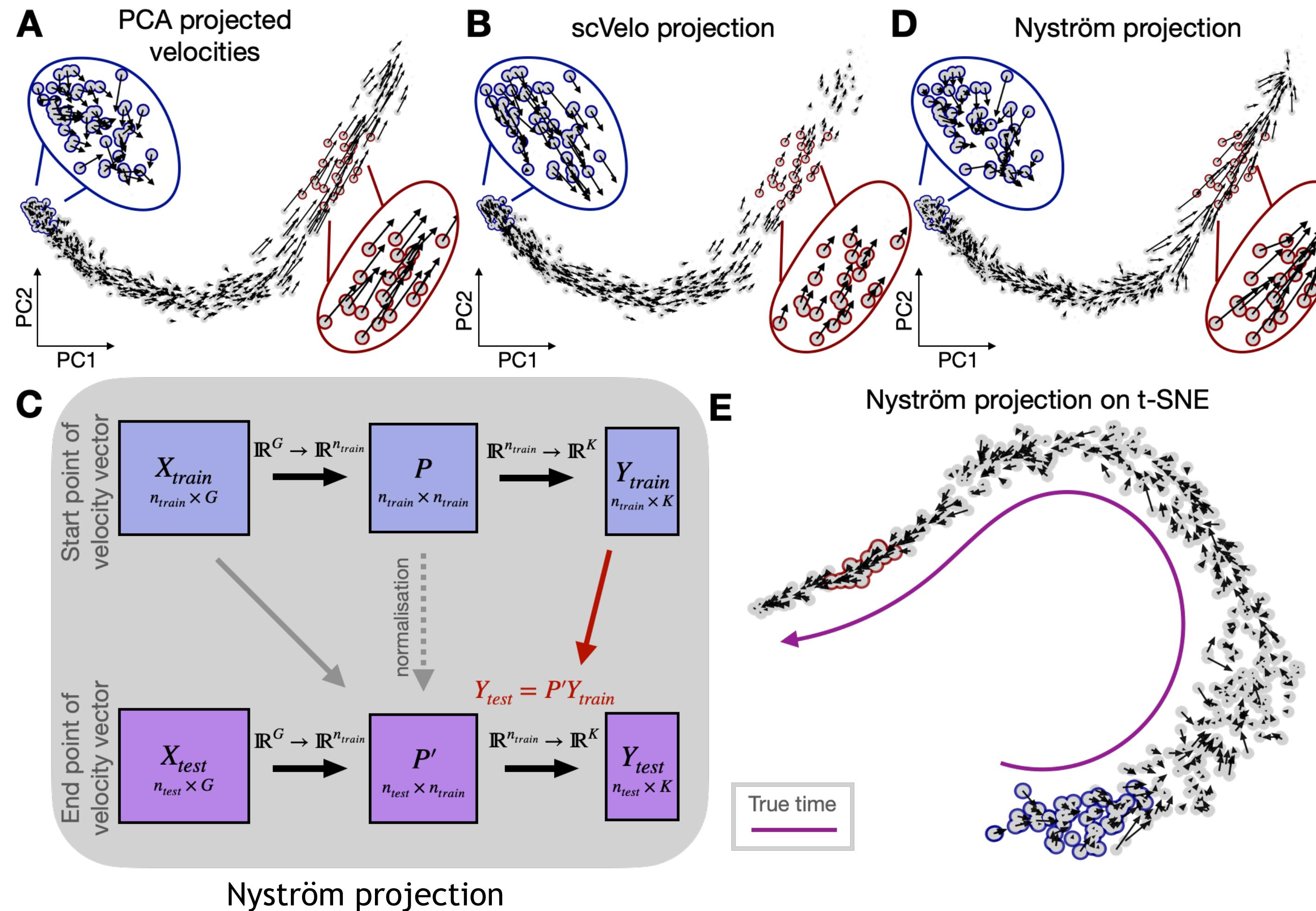
- Eco-velo:

- Skip error-prune expensive gene-wise curve fitting
- u/s MNN matching in genes space
- Select gene set with similar splicing and degradation time scales

Circadian-associated genes in the mouse liver over a 24-h time course [La Manno et al, Nature 2018]



Visualisation of cell state velocities



- Previous (velocyto, scVelo) methods:

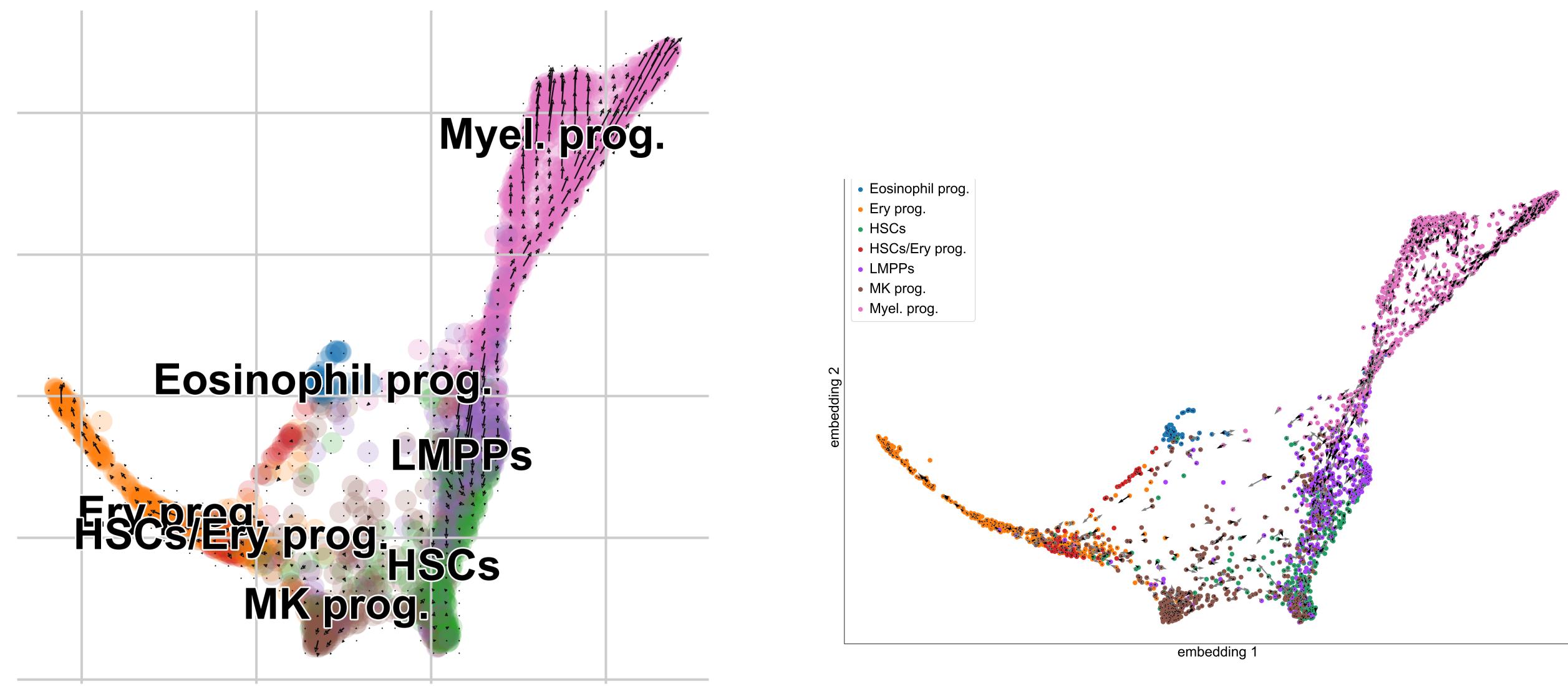
$$\Delta \vec{Y}_i = \sum_j (P_{ij} - \frac{1}{n}) \frac{\vec{Y}_j - \vec{Y}_i}{\|\vec{Y}_j - \vec{Y}_i\|}$$

$$P_{ij} = \exp\left(\frac{\cos \angle(\vec{s}_j - \vec{s}_i, \vec{v}_i)}{\sigma^2}\right)$$

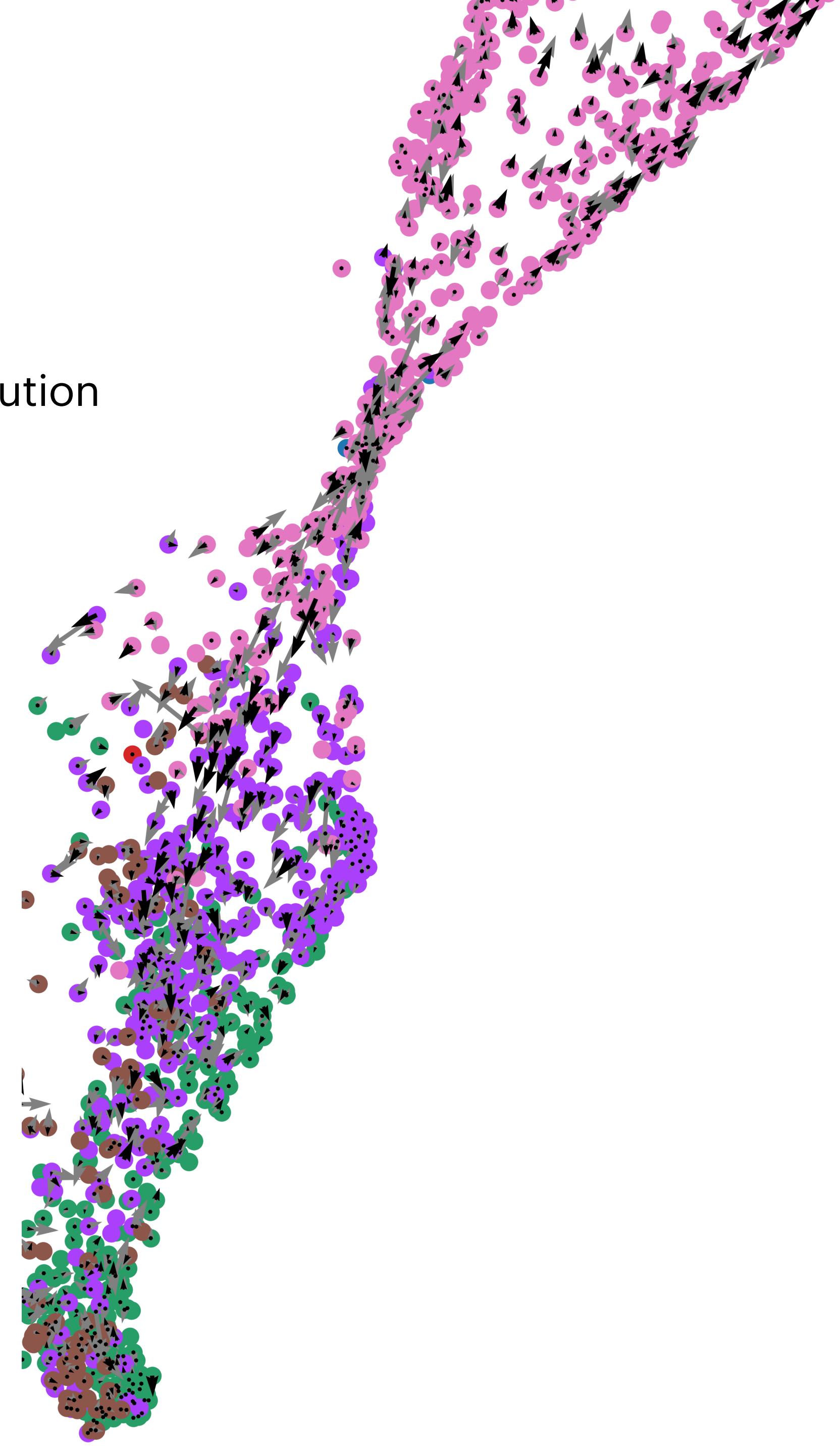
- Unfaithful to:
 - Velocity vectors norm
 - Cell-to-cell variety of velocities (stochasticity and cell plasticity)
 - General artefacts of out of distribution non-linear projection
 - Use PCA when possible

eco-velo upgrade

- No error-prone gene-wise parameter fitting
- Applicable for u-s as well as unlabeled-labeled- mRNA
- Fixed time-scale (labelling time)
- Joint embedding of current old-new cell states —> also resolves out-of-distribution projection problems



- Preserve velocity heterogeneity and quantify the diffusion component
- Compare to velocities post-inflammation



Summary: Dynamical inference & latent spaces

- Some good/bad latent spaces
- Geometrical PT and response PT
- Reliable velocity estimation
- Parallel methods PT, velocity and OT can be used for validation

Data Integration

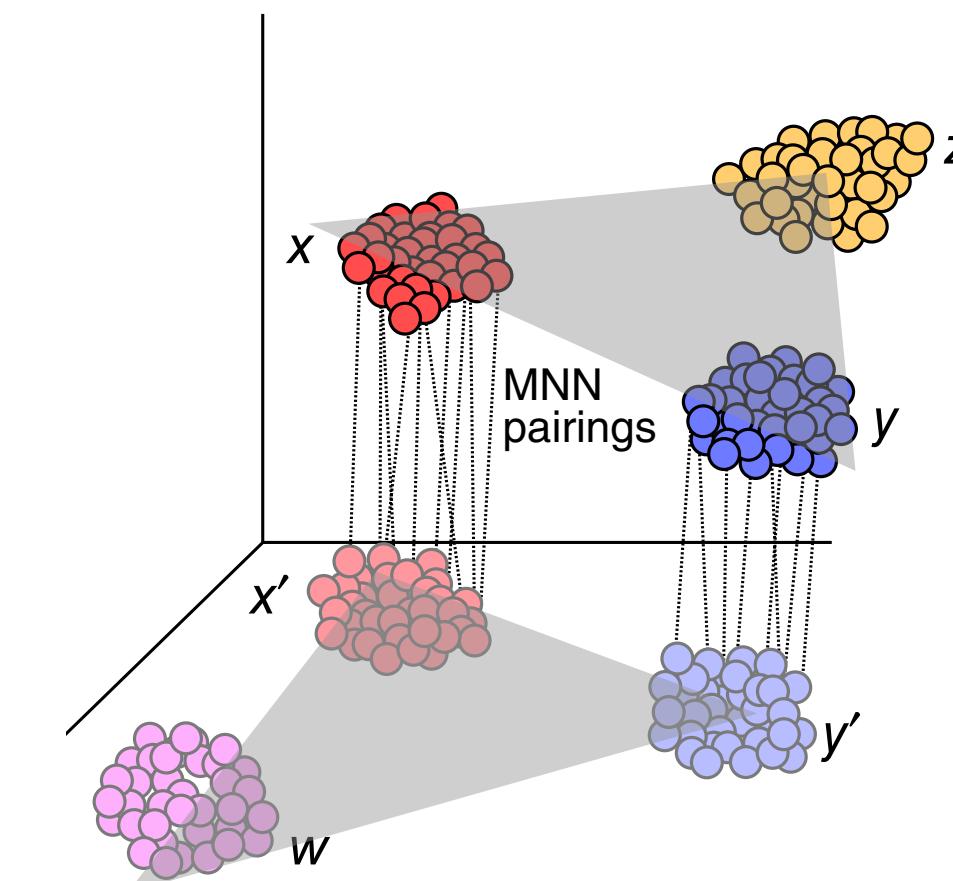


Comparing patients, time points, etc.

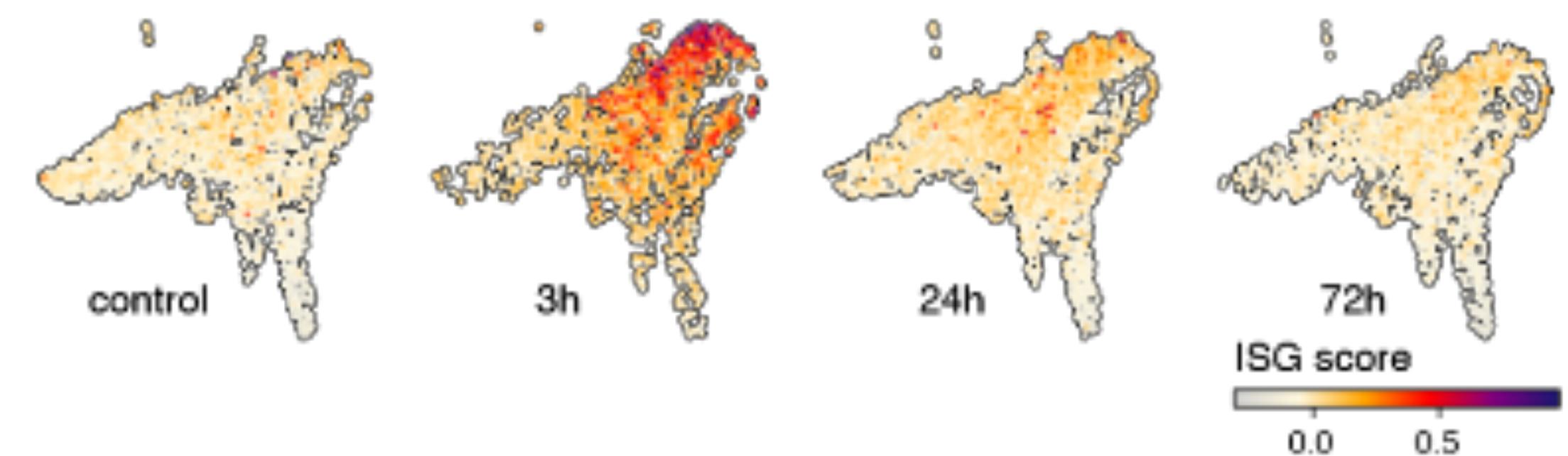
Lineage tracing

Data integration: Visualise different data views on a shared manifold

- Projection (linear, Nystroem)
- Haghverdi et al. Nature biotechnology 2018, **MNN**
- Butler et al. Nature biotechnology 2018, CCA + time warping (**Seurat** v2) →
- CCA + MNN (Seurat v3) →
- RPCA + MNN (Seurat v4-5)



- Hie et al. Nature biotechnology 2019, **Scanorama**
- **Harmony**, Korsunsky et al. Nature Methods 2019
- Deep learning approaches e.g. scVI Lopez et al. Nature Methods 2019



Compound-SNE: Comparative Alignment of t-SNEs

- Minimum intervention with the embedding algorithm
- Unlike data integration, respect dataset specific manifold structures
- Not exactly the same feature set among the datasets
- Comparing multiple time points, patients, modalities

Soft alignment in two steps

- Initialisation with Procrustes transformation in PCA space (Primary)
- Coupling forces between same clusters (Full)
- ~ modified cost function

Compound-SNE aligning different time points of HSPCs response to inflammation

- Respect dataset specific manifold structures
- Facilitates visual comparison
- Areas, densities, distances should not be over-interpretted!



Bioinformatics, 2024, 40(7), btae471
https://doi.org/10.1093/bioinformatics/btae471
Advance Access Publication Date: 25 July 2024
Applications Note

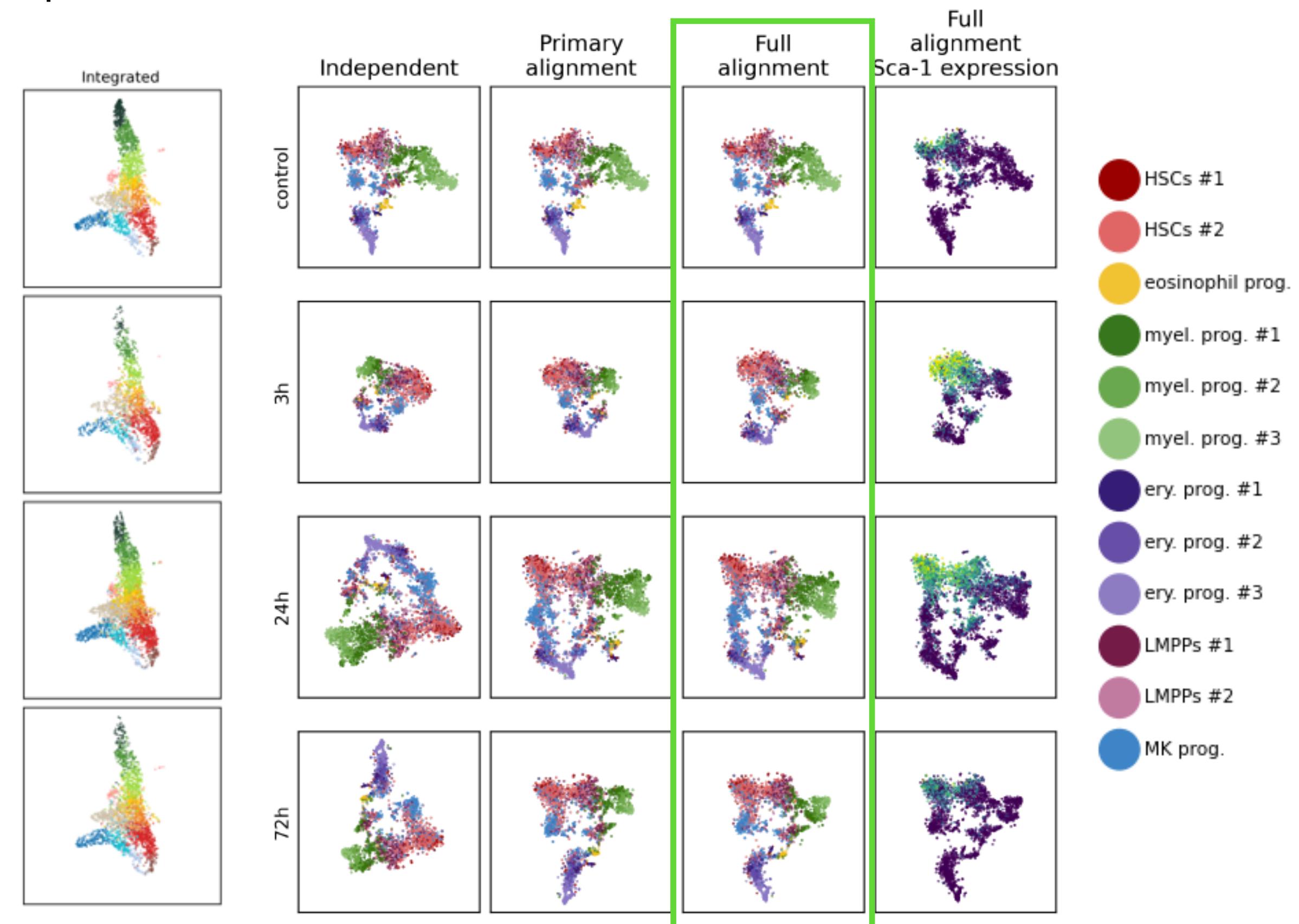
Gene expression

Download Compound-SNE!

Compound-SNE: comparative alignment of t-SNEs for multiple single-cell omics data visualization

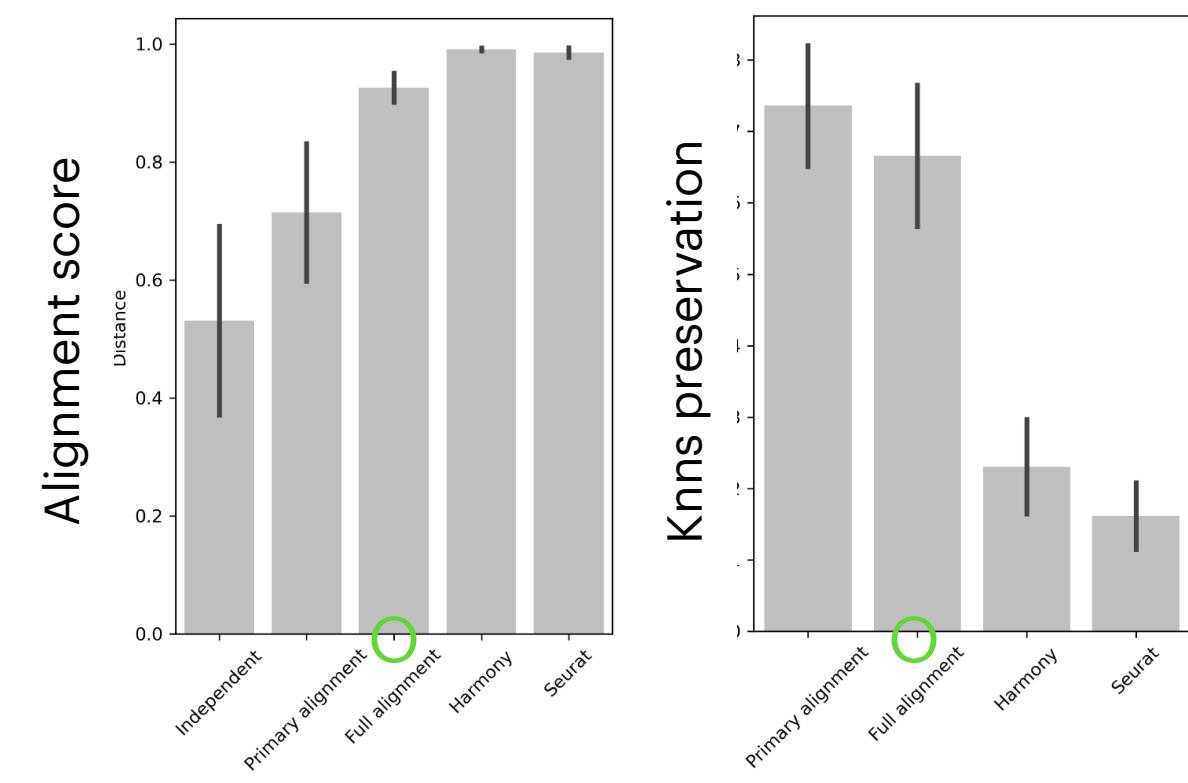
Colin G. Cess¹ and Laleh Haghverdi ^{1,*}

¹Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (BIMSB-MDC), Berlin 10115, Germany



$$d_i = (Y_{r,center_i} - Y_{s,center_i})^2$$

$$L_{total} = L_{tsne} + \sum_{i=1}^K \lambda_i d_i$$



t-distributed Stochastic Neighbours Embedding

Van der Maaten, "Visualizing data using t-SNE.", 2008

Van Der Maaten, Laurens. "Accelerating t-SNE using tree-based algorithms." 2014

Original (high-dim) space X

$$p_{ij} = \frac{1}{z_X} \cdot \exp\left(\frac{-||\mathbf{X}_i - \mathbf{X}_j||^2}{2\sigma_i^2}\right)$$

$$\mathbf{P} \leftarrow \frac{\mathbf{P} + \mathbf{P}'}{2n}$$

embedding (e.g. 2D) space Y

$$q_{ij} = \frac{1}{z_Y} \cdot \frac{1}{1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2}$$

$$C = KL(\mathbf{P} || \mathbf{Q}) = \sum_{i \neq j} p_{ij} \frac{\log(p_{ij})}{\log(q_{ij})}$$

$$\frac{\partial C}{\partial \mathbf{Y}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{Y}_i - \mathbf{Y}_j)(1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2)^{-1}$$

$$\begin{aligned} \frac{\partial C}{\partial \mathbf{Y}_i} &= 4 \left(\sum_j p_{ij} q_{ij} z_Y (\mathbf{Y}_i - \mathbf{Y}_j) - \sum_j q_{ij}^2 z_Y (\mathbf{Y}_i - \mathbf{Y}_j) \right) \\ &= 4 \left(\underbrace{\sum_j p_{ij} \frac{\mathbf{Y}_i - \mathbf{Y}_j}{1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2}}_{\text{Repulsive}} - \underbrace{\sum_j \frac{1}{z_Y} \cdot \frac{\mathbf{Y}_i - \mathbf{Y}_j}{(1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2)^2}}_{\text{Attractive}} \right) \end{aligned}$$

Repulsive

Attractive

t-SNE

- Initialise with $\mathbf{Y}^{(0)}$
- $$\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathbf{Y}_i} + \alpha(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)})$$

Learning rate Momentum

(Move Y in the direction of the largest gradient)
- The final solution for Y depends on the initialisation \mathbf{Y}^0

Extracting multi-modal information from scRNA-seq

- Big data concept
- Sound of the keyboard → text string, room temperature, etc.
- Correspondence between feature sets
- More efficiently using the available data
- Cost efficient experiments
- Study of rare samples
- scRNA-seq as the most available data modality
- Noisy, uncertain scRNA-seq SNV for clonal tracing?
 - Germline variants
 - Allele frequency drop-out
 - Cell type specific expression
 - Sequencing and alignment errors
 - RNA-edits



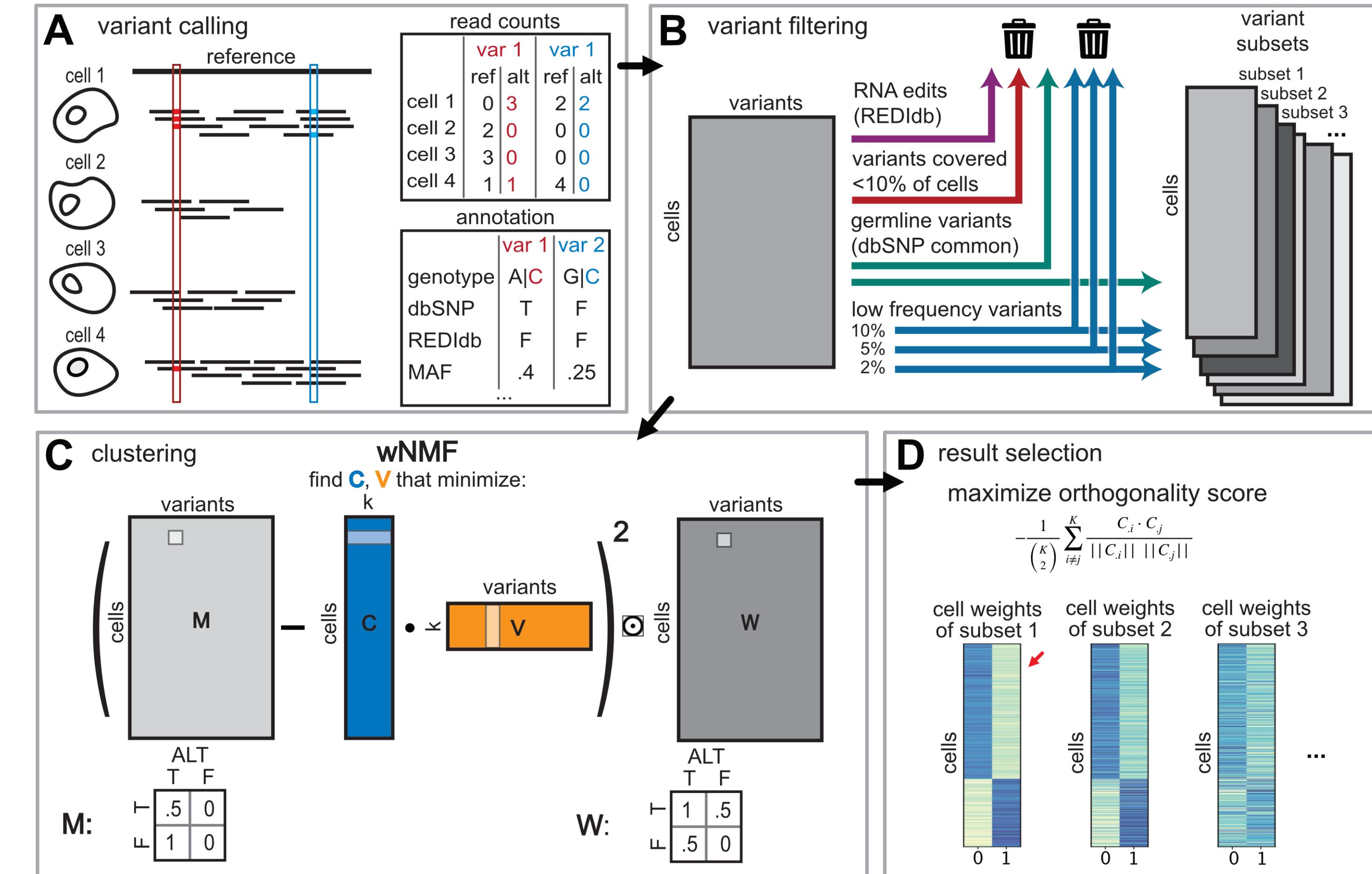
CCLONE: Clonal tracing from scRNA-seq variant calls

Gene expression

Identifying cancer cells from calling single-nucleotide variants in scRNA-seq data

Valérie Marot-Lassauzaie , Sergi Beneyto-Calabuig ^{3,4}, Benedikt Obermayer ⁵, Lars Velten ^{3,4}, Dieter Beule ^{5,6}, Laleh Haghverdi ^{1,*}

- NMF for noisy data
- Weighted NMF for allelic or **cell type dependent** dropout and heteroplasity (e.g. $W_{ij}=0$ if position j not expresses in cell i)
- SNVs from (high coverage) scRNA-seq data can be used for clonal inference
- Identifying somatic mutations in scRNA-seq and scATAC-seq:
 - Scomatic Muyas et al. Nat. Biotechnol. 2024 (comparing variant distribution in tumour and healthy cells)
 - Monopogene Dou et al. Nat. Biotechnol. 2024 (using linkage disequilibrium for distinguishing germline variants)

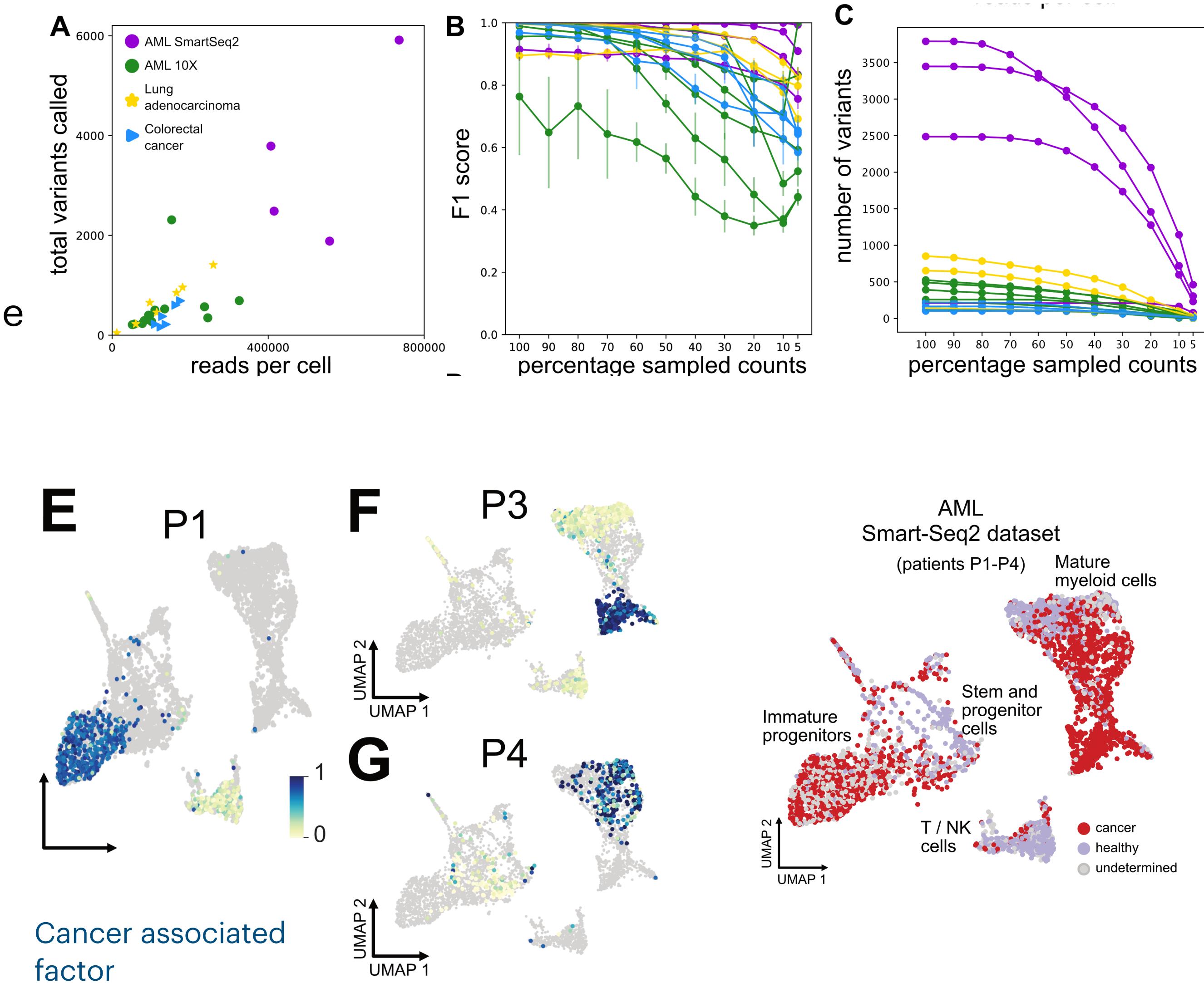
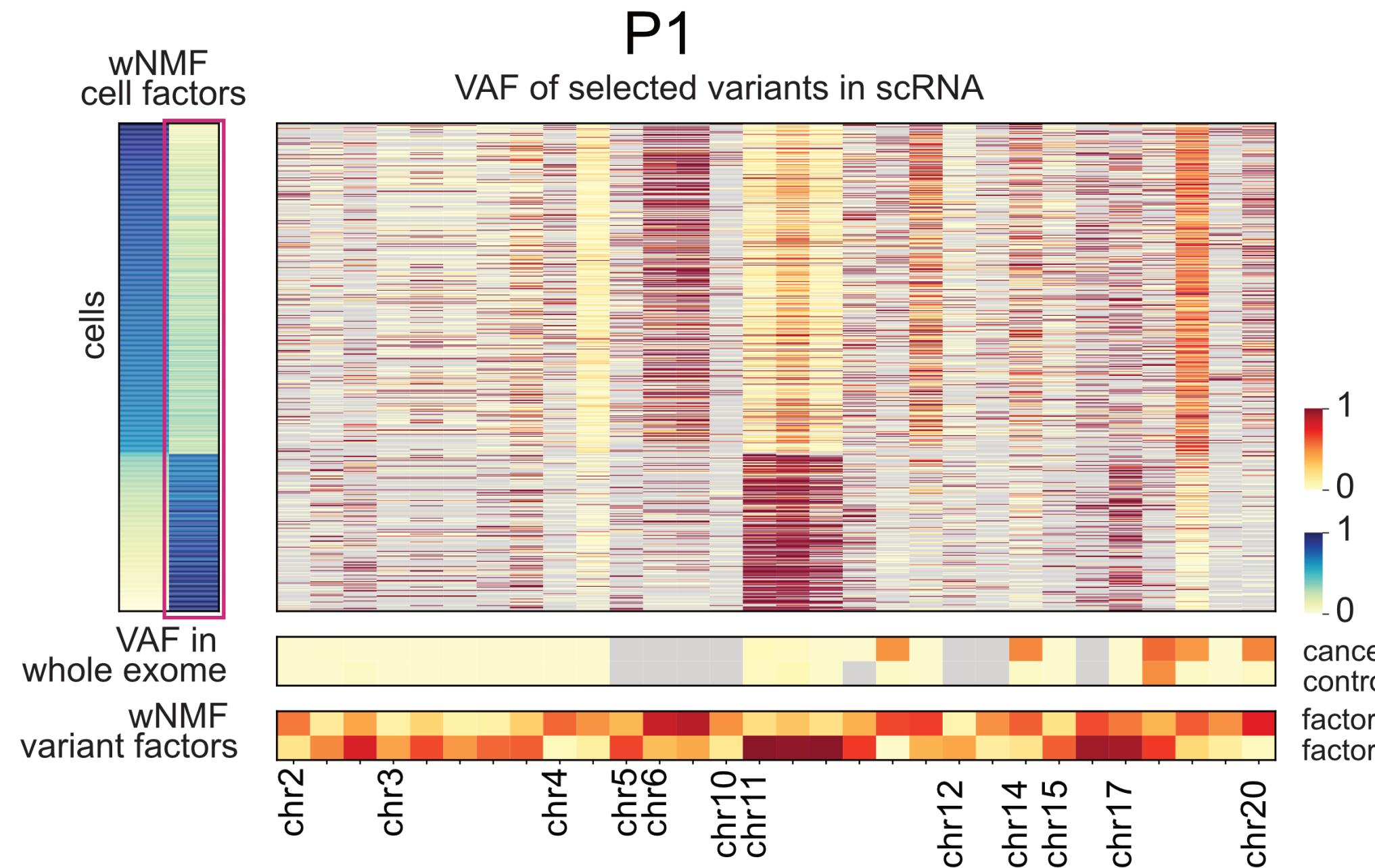


- M: Variant Allele Frequency
- W: Certainty of observation

- $C_{i,j}$: i th column of C
- K : # of factors

CCLONE: Cancer Cell Labelling On Noisy Expression

- Used for comparison as ground-truth: targeted sc-sequencing
- Higher success rate for higher mutation burden and sequencing coverage
- Interpretable factors and associated variants
- Interestingly some SNVs were not detected in WE (due to small population size, capture efficiency, different allelic expression in cancer vs. health...)



Summary: Data integration

- Hard integration
- Soft integration (Compound-SNE)
- Example of multi-modal information in scRNA-seq
 - Gene expression
 - Velocity (Eco-velo)
 - Clonal structure (CCLONE)

Summary: Dynamical inference & latent spaces

- Some good/bad latent spaces
- Geometrical PT and response PT
- Reliable velocity estimation
- Parallel methods PT, velocity and OT can be used for validation

My group:

Valerie Marot-Lassauzaie

Colin Cess

Thank You!

Roshni Biswas

Shashank Tiwari

Melanie Fattohi

Carla Mölbert

Yiftach Kolb

Zengya Yang



Berlin Institute
for Medical
Systems Biology



Bundesministerium
für Bildung
und Forschung



Main collaborators:

Lars Velten (CRG Barcelona)

Dieter Beule (Charite Berlin)

Simon Haas (BIH Berlin)

Marieke Essers (DKFZ Heidelberg)

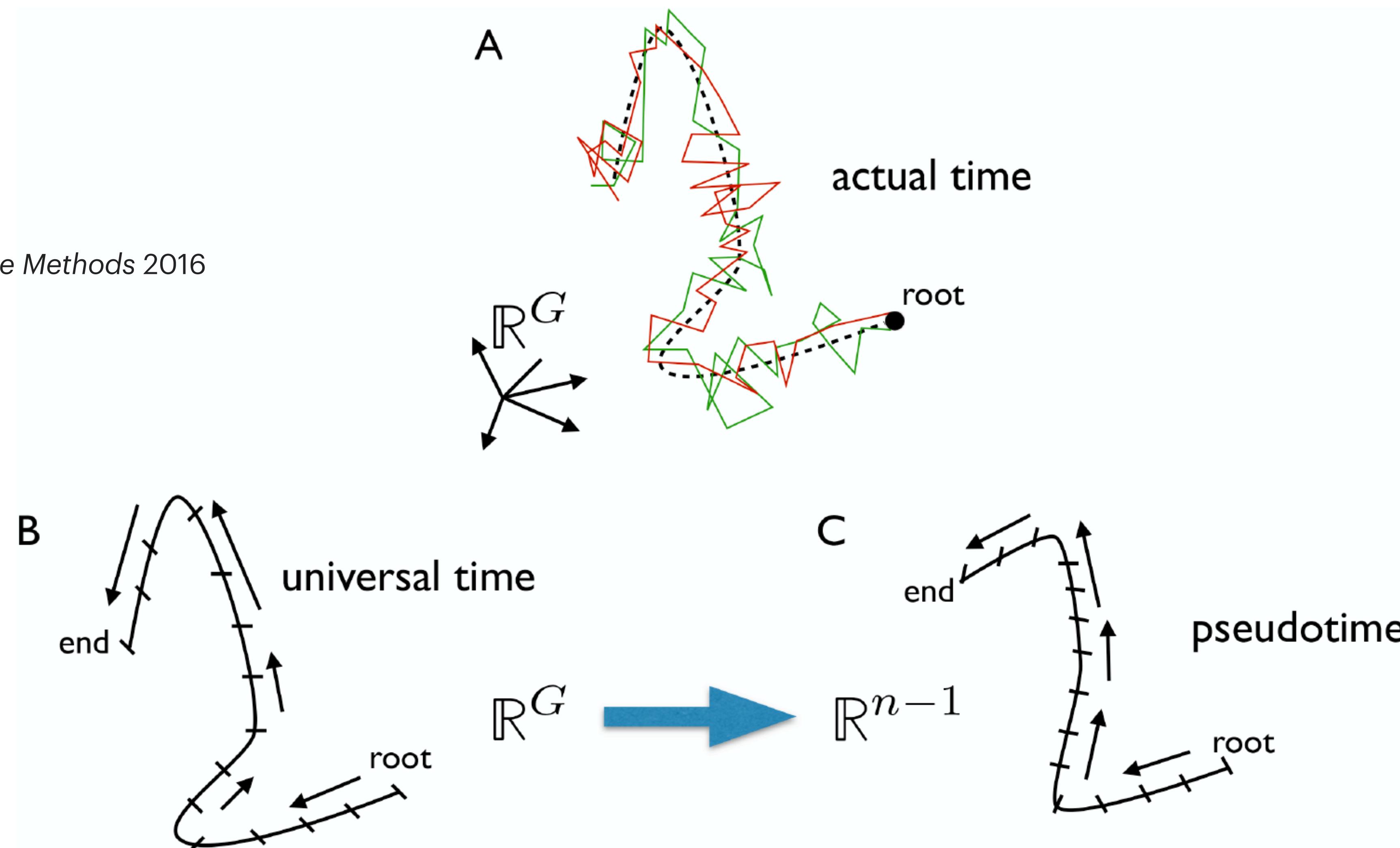
Leif Ludwig (BIH Berlin)



Dissecting and Reengineering the Regulatory Genome

How pseudotime is defined

Haghverdi et al. *Nature Methods* 2016
(Supplement)

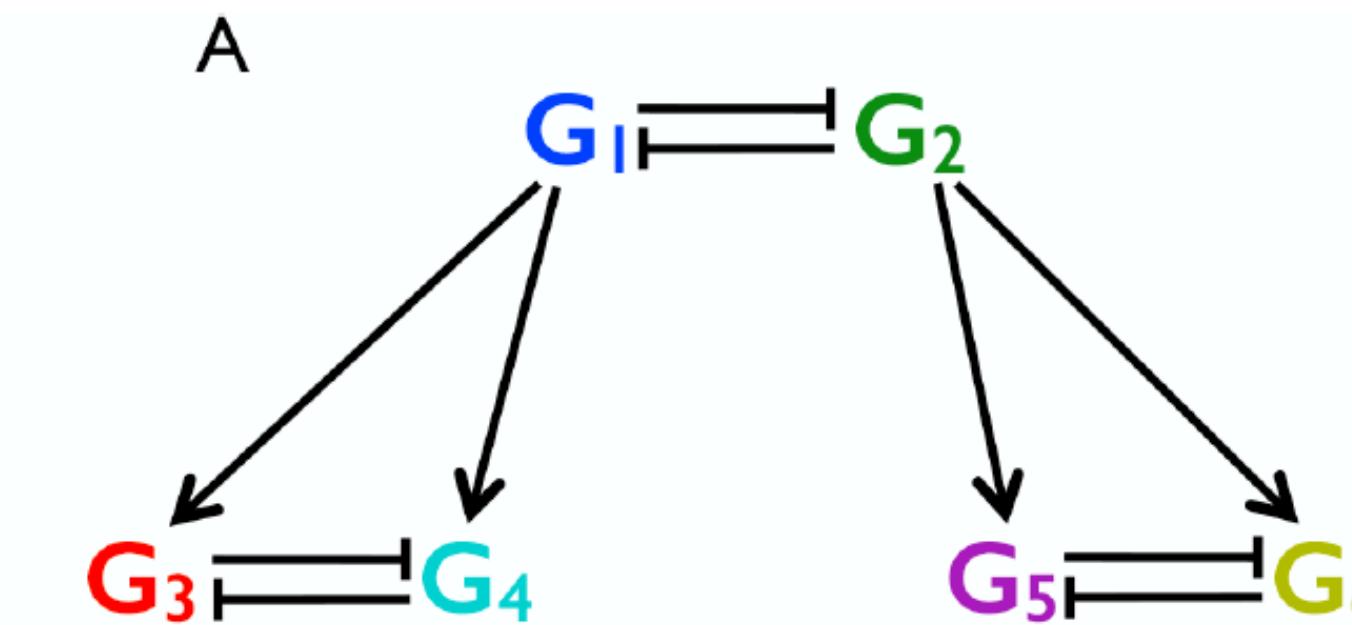


Supplementary Figure N7: A) Two (red and green) actual time single cell trajectories in gene expression space(\mathbb{R}^G). Each jump on a trajectory happens in an (equidistant) unit of actual time. B) Universal time is defined as arc length on the data manifold starting from the root. This manifold $C \subset \mathbb{R}^G$ remains the same for several single cell trajectories, as well as for snapshot samples of single cells. C) Pseudotime (in

Actual time, universal time, pseudotime

"Diffusion pseudotime robustly reconstructs lineage branching"
Haghverdi et al. Nature methods 2016 (Supplemental Note 1)

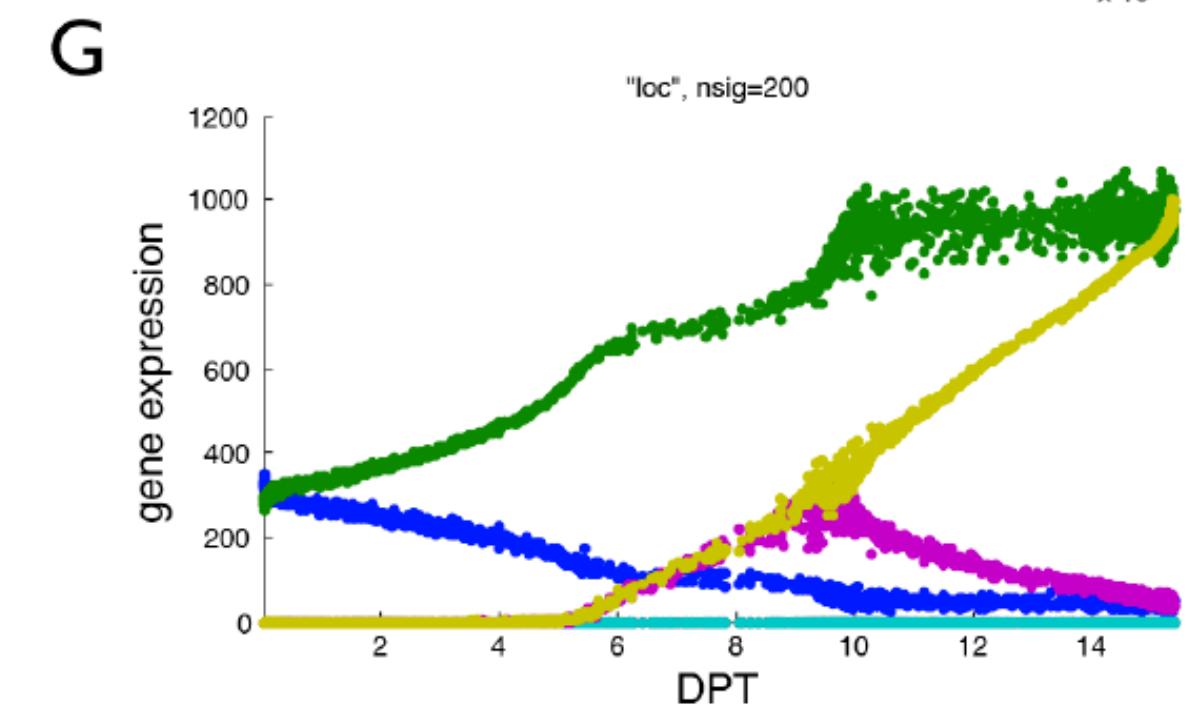
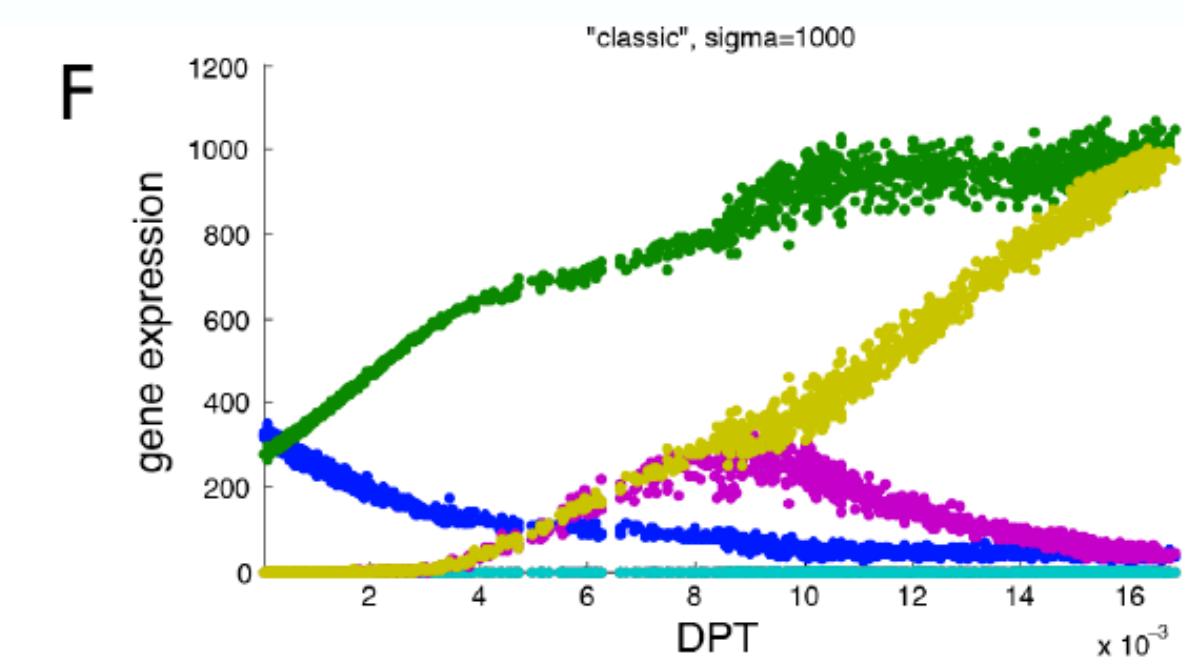
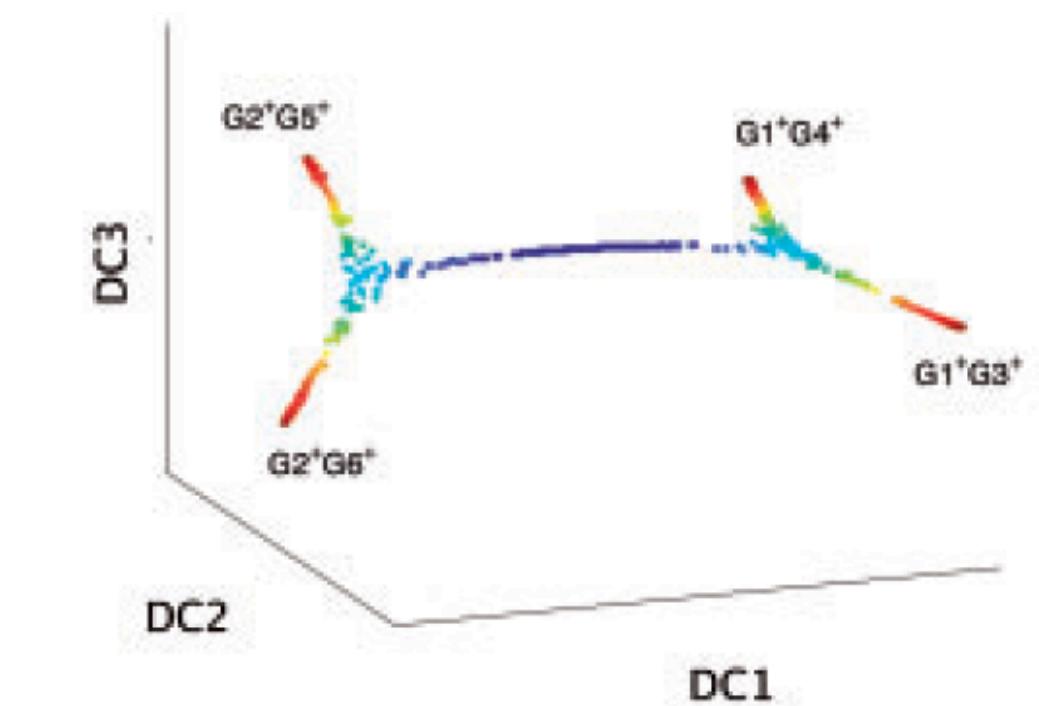
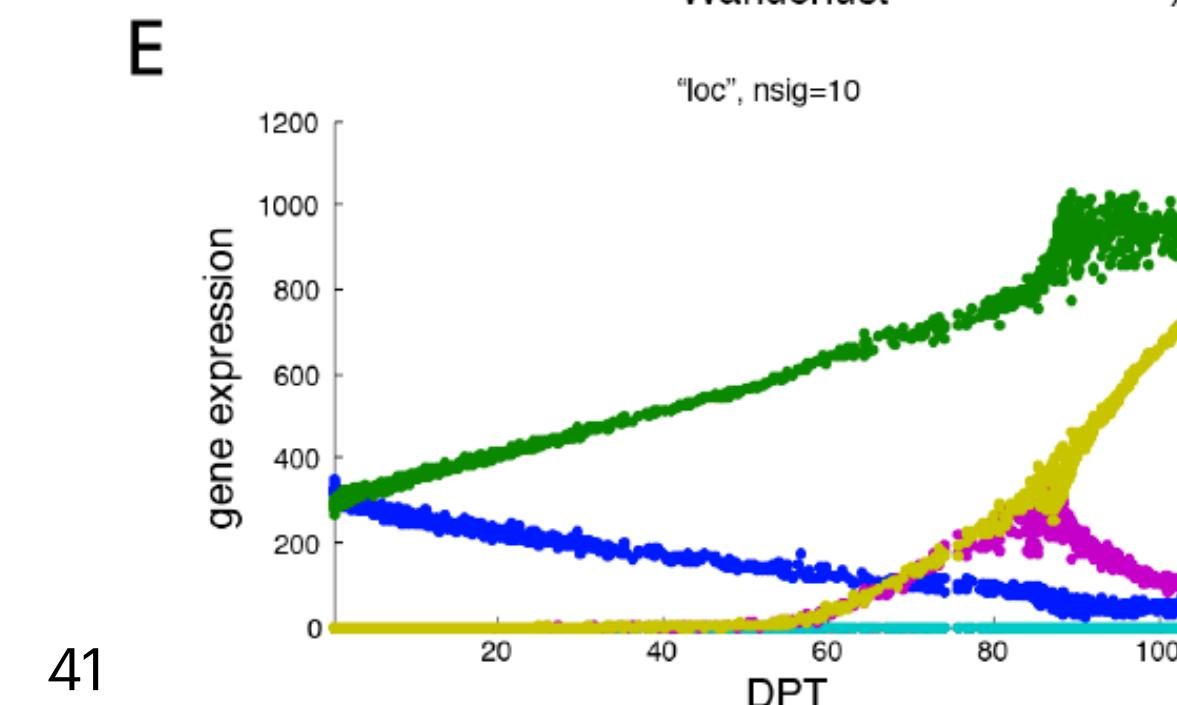
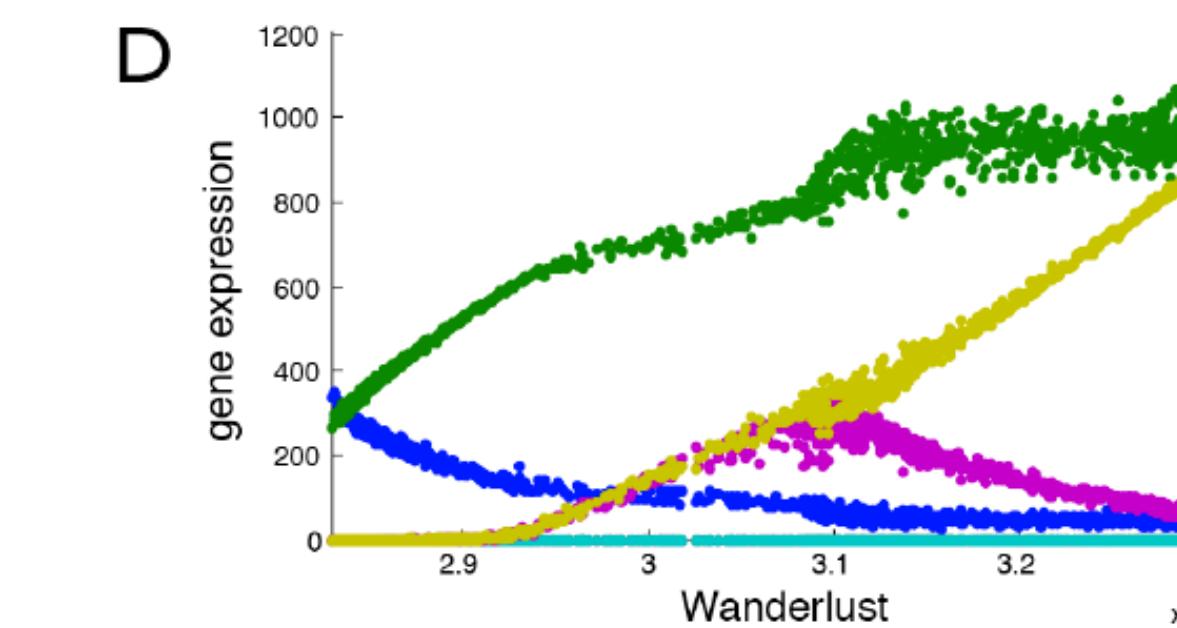
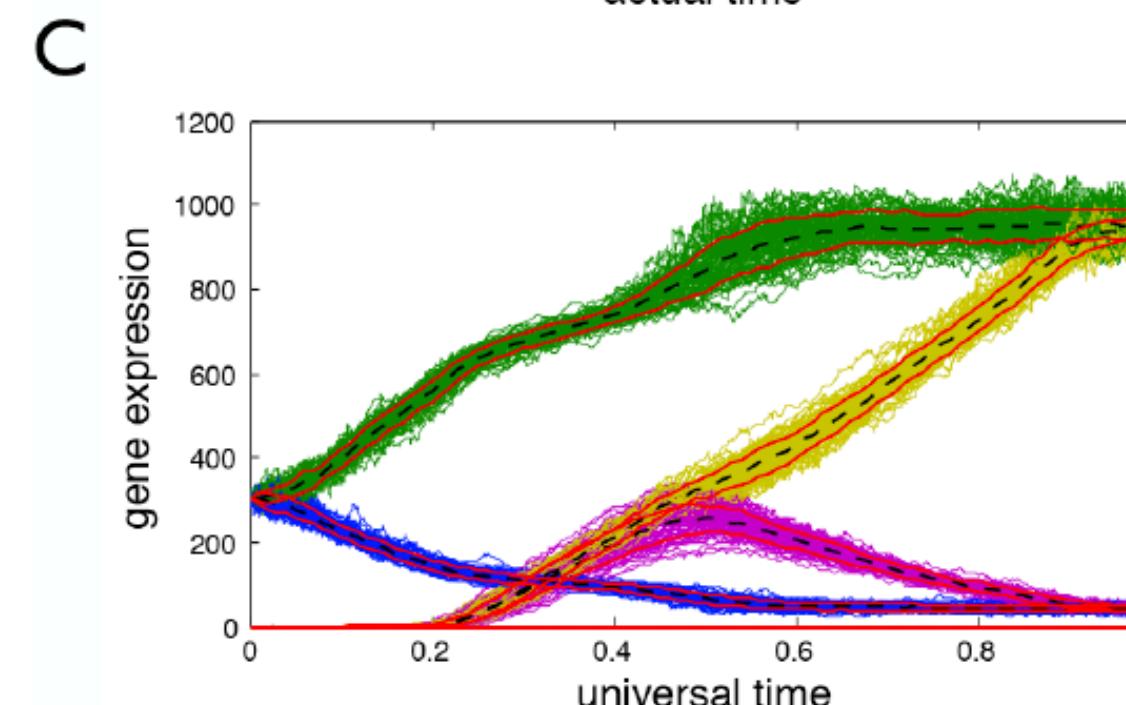
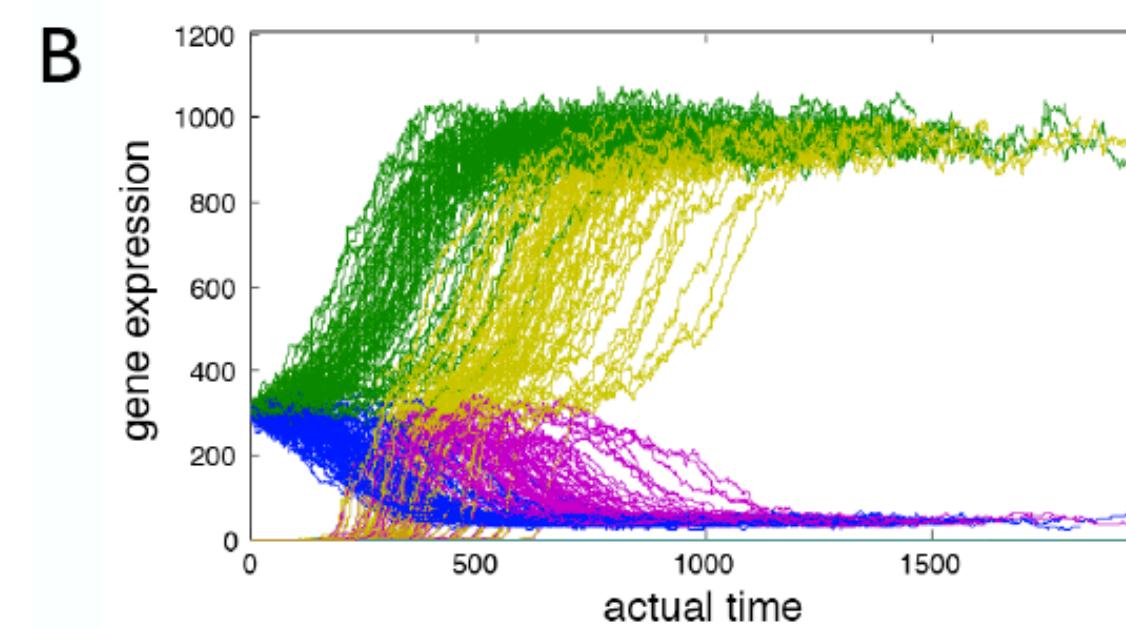
- Under assumption of equal time steps of sampling from an individual trajectory, as well as stationary distribution of cells sampled from multiple trajectories we have $1/\langle \text{density} \rangle \sim \langle \text{velocity} \rangle$



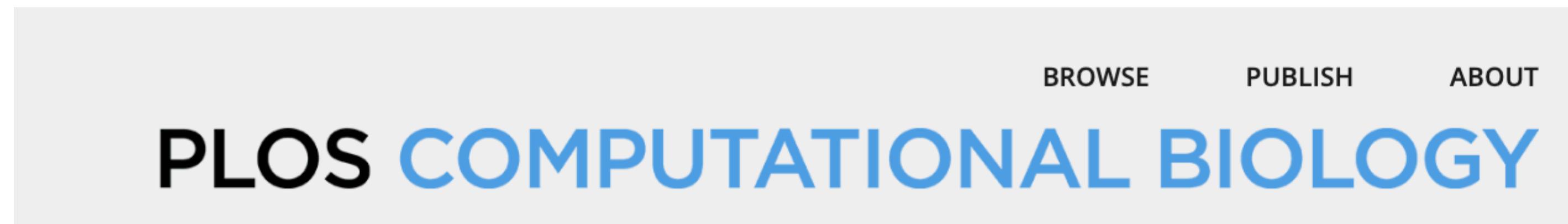
$$1/\rho(t) = v(t)$$

universal time

$$u(t) = \int_0^t \frac{1}{\rho(t')} dt$$



Cell state velocities



OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

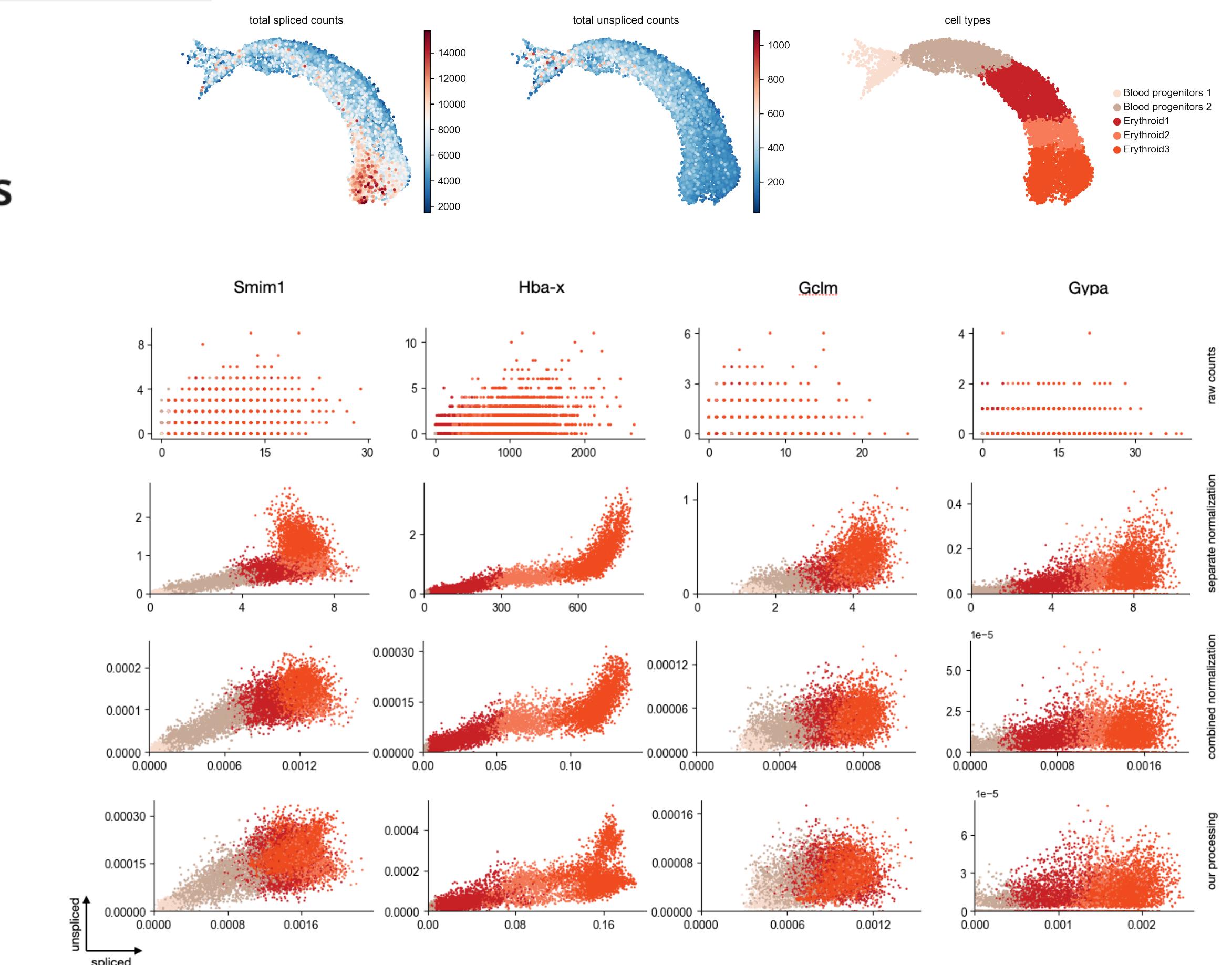
Towards reliable quantification of cell state velocities

Valérie Marot-Lassauzaie , Brigitte Joanne Bouman , Fearghal Declan Donaghy, Yasmin Demerdash, Marieke Alida Gertruda Essers, Laleh Haghverdi

Version 2 Published: September 28, 2022 • <https://doi.org/10.1371/journal.pcbi.1010031>

Problems in:

- Input data inaccuracies
- Preprocessing
- Parameter estimation, gene-wise curve fitting
, scaling factors
- Visualisation



Compound-SNE: Comparative Alignment of t-SNEs

- Minimum intervention with the embedding algorithm
- Unlike data integration, respect dataset specific manifold structures
- Not exactly the same feature set among the datasets
- Comparing multiple time points, patients, modalities

Soft alignment in two steps

- Initialisation with Procrustes transformation in PCA space (Primary)
- Coupling forces between same clusters (Full)

• ~ modified cost function

$$d_i = (Y_{r,center_i} - Y_{s,center_i})^2$$

$$L_{total} = L_{tsne} + \sum_{i=1}^K \lambda_i d_i$$

