Statistical Methods for High-dimensional Biology


# Single-cell Genomics: Advanced topics

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

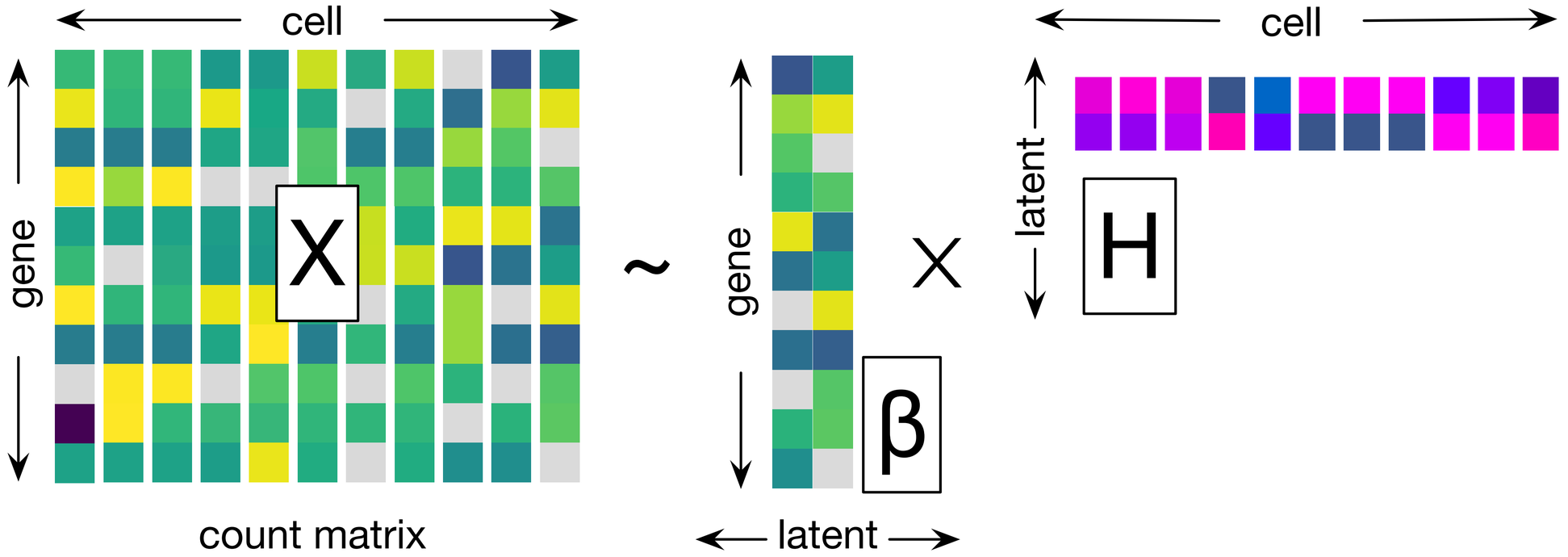
# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

Intensive Math

Fast survey

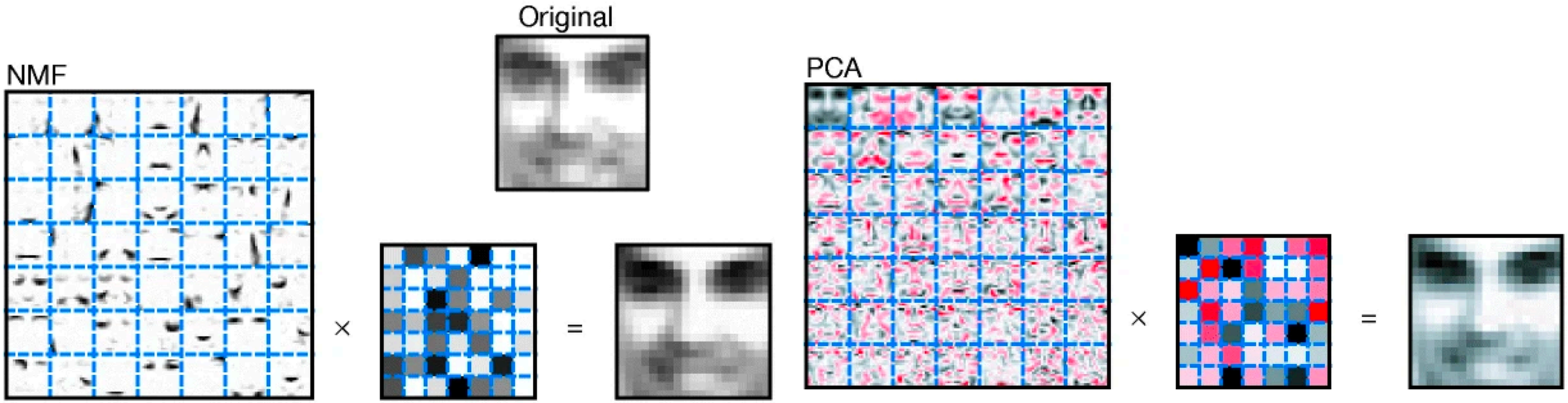
# A generative model of single-cell data?




$$\mathbb{E}[X] \approx f\left(H\beta\right)$$

Can we assign tens of thousands of cells to some hidden probability space ($H$)?

# Recall: Lect #12 about NMF

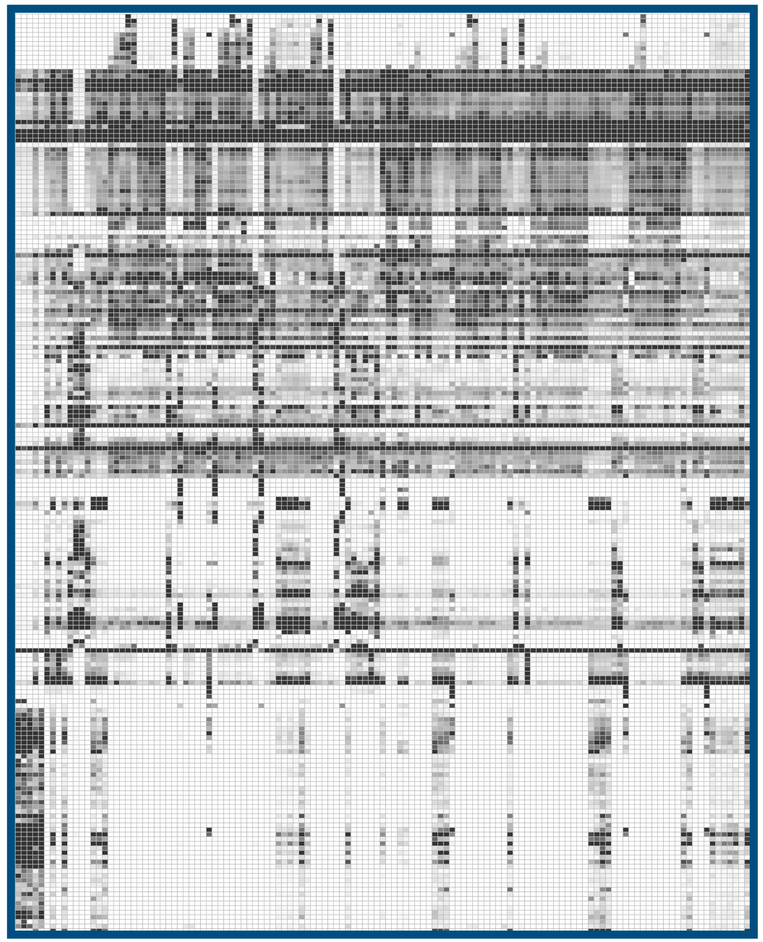



**Decomposition by "parts"**

**Positive + negative can cancel each other**

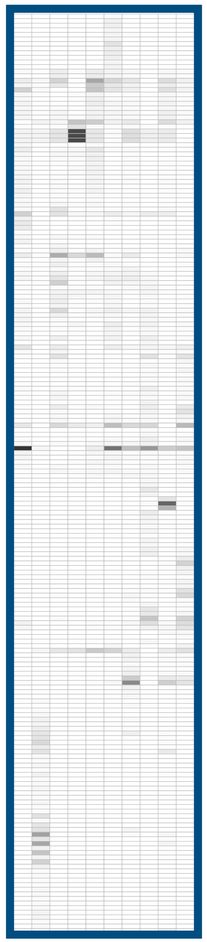
**Lee and Seung, *Nature* (1999)**

# NMF to factorize expression matrix



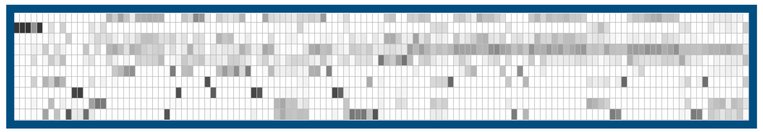
X

$\sim$



β

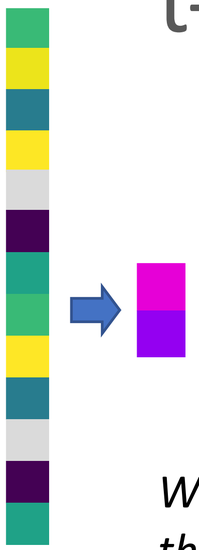
$\times$



H

# A model-based approach goes beyond visualization of the high-dim. scRNA-seq

t-SNE



D=35k obs. X



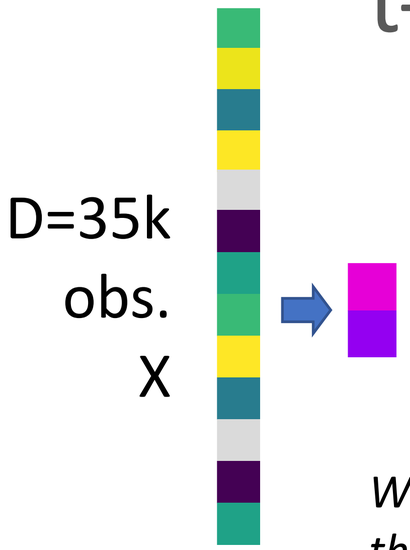
d=2, latent z

*Where are the model parameters?*

$$\min_{z} \quad D_{KL}\left(p_{ij}||q_{ij}\right) = \sum_{ij} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

# A model-based approach goes beyond visualization of the high-dim. scRNA-seq

## t-SNE

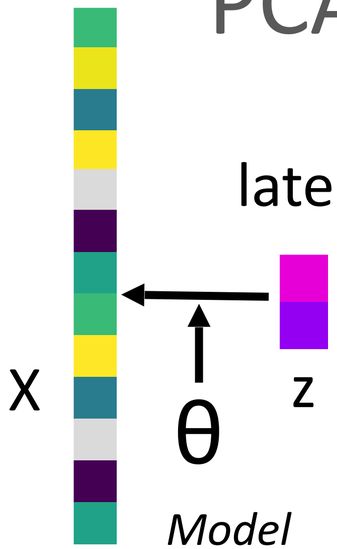



d=2,
latent
z

*Where are the model parameters?*

$$\min_{z} \quad D_{KL}\left(p_{ij} \,||\, q_{ij}\right) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

## PCA, MF




latent

z

θ

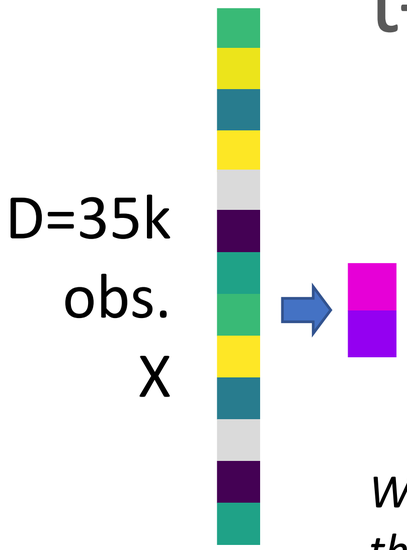
*Model parameters*

Can we generate/
estimate/impute
latent states
on unobserved data?

$$z \leftarrow f(x^*; \theta)$$

# A model-based approach goes beyond visualization of the high-dim. scRNA-seq

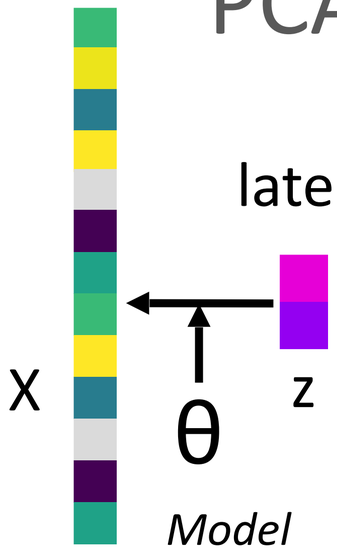
## t-SNE




*Where are the model parameters?*

$$\min_{z} \quad D_{KL}\left(p_{ij}\,||\,q_{ij}\right) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

## PCA, MF



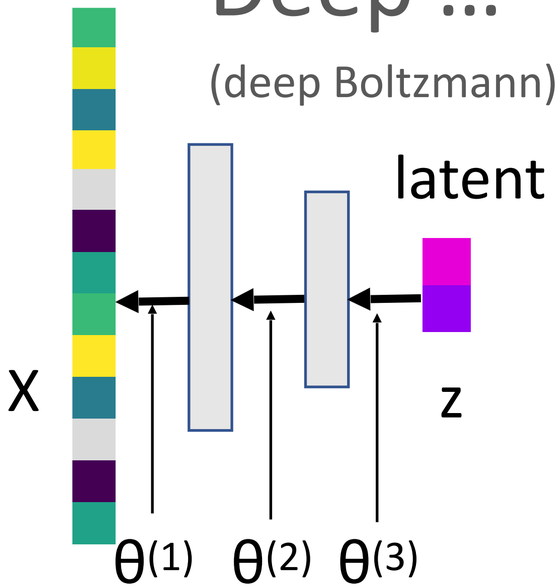

*Model parameters*

Can we generate/ estimate/impute latent states on unobserved data?

$$z \leftarrow f(x^*; \theta)$$

## Deep …

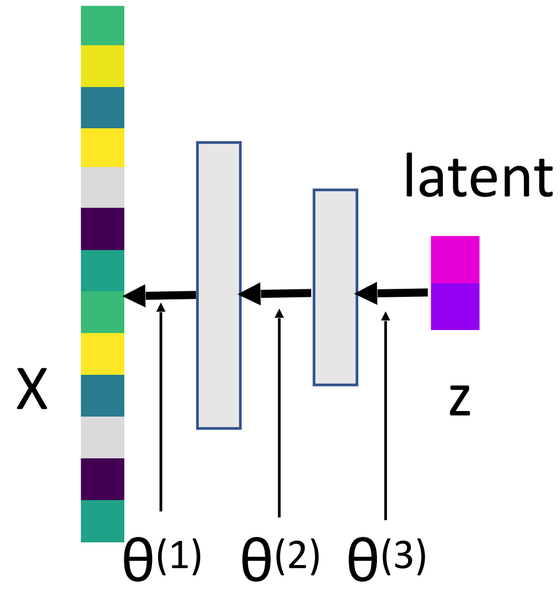
(deep Boltzmann)




Can we represent high-dimensional data using multiple functions?

$$z \leftarrow f(f(f(x^*; \theta^{(1)}); \theta^{(2)}); \theta^{(3)})$$

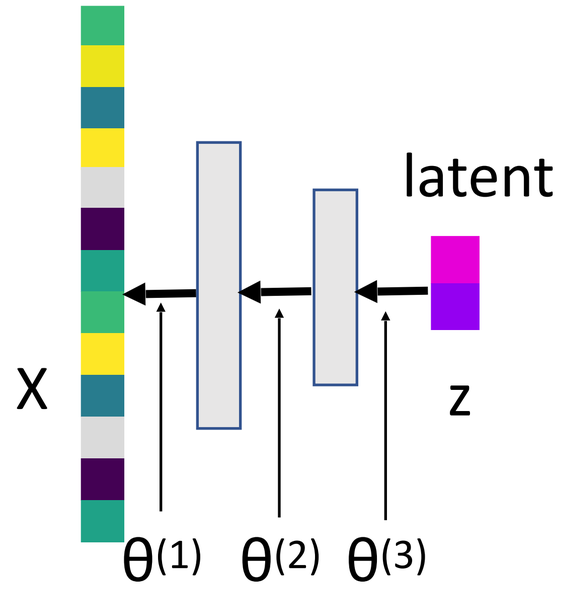
# How do we estimate the parameters of a deep latent variable model?



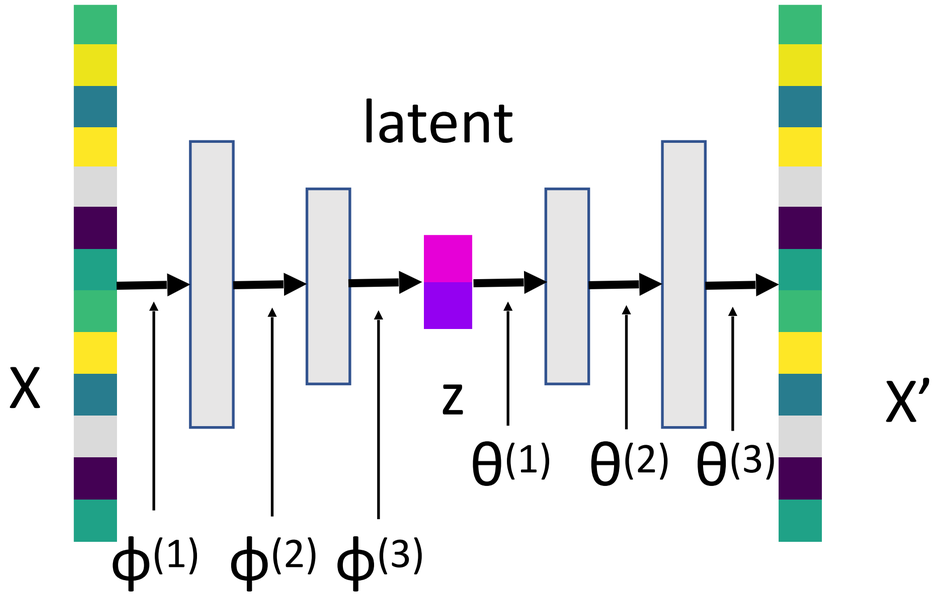

$$\int dZ\ p(Z)p(X \mid \theta, Z)$$

- E-step: Estimate/sample the latent Z
- M-step: Maximize the parameter θ

# How do we estimate the parameters of a deep latent variable model?




vs.




$$\int dZ\ p(Z)p(X \mid \theta, Z)$$

- E-step: Estimate/sample the latent Z
- M-step: Maximize the parameter θ

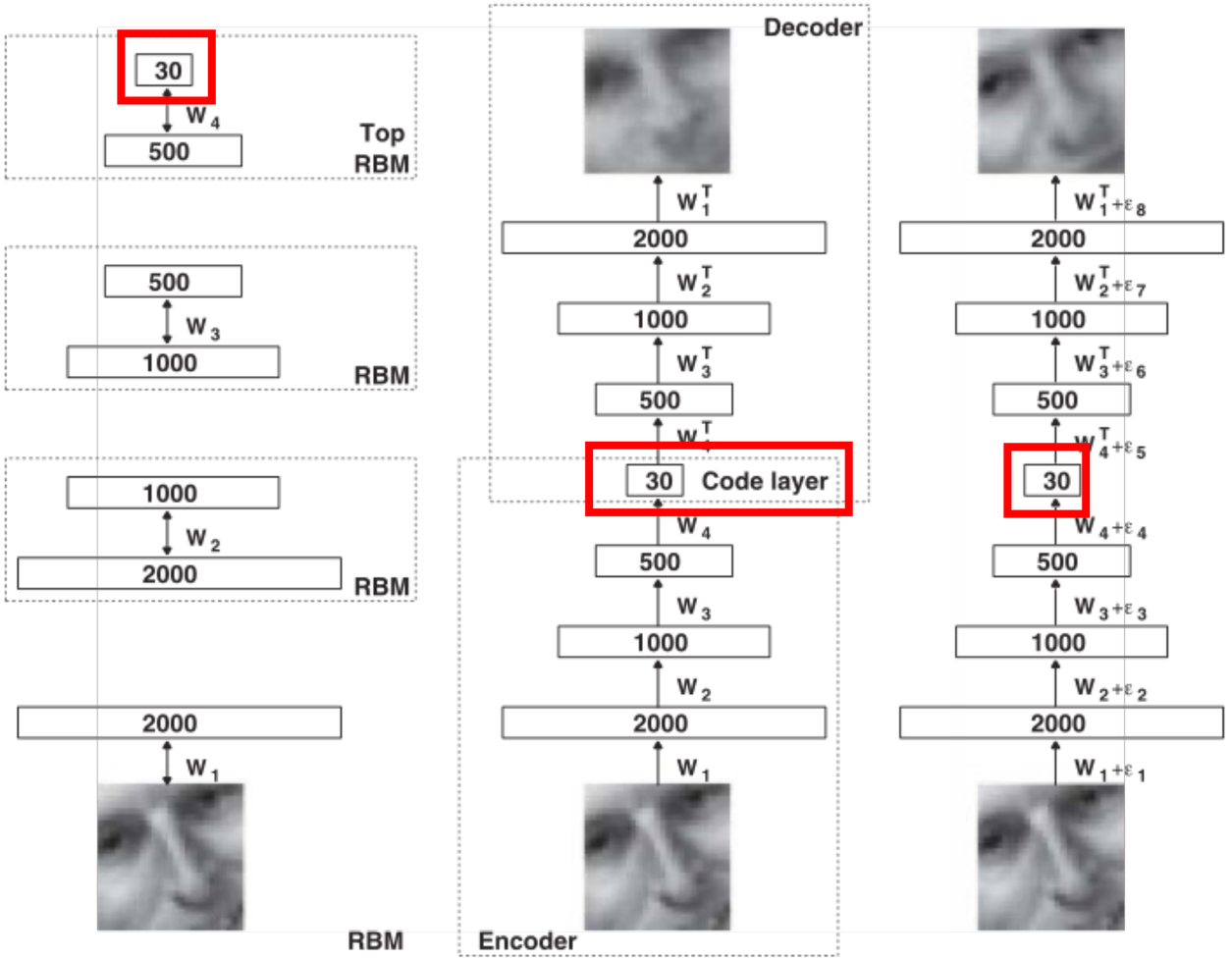
Goal: Make X and X' similar with respect to a data-generating model

No need to carry out an EM-type of algorithm
Just straight optimization of the parameters

# Digression

## Geoff Hinton's Deep Autoencoder model

Greedy layer-by-layer pretraining



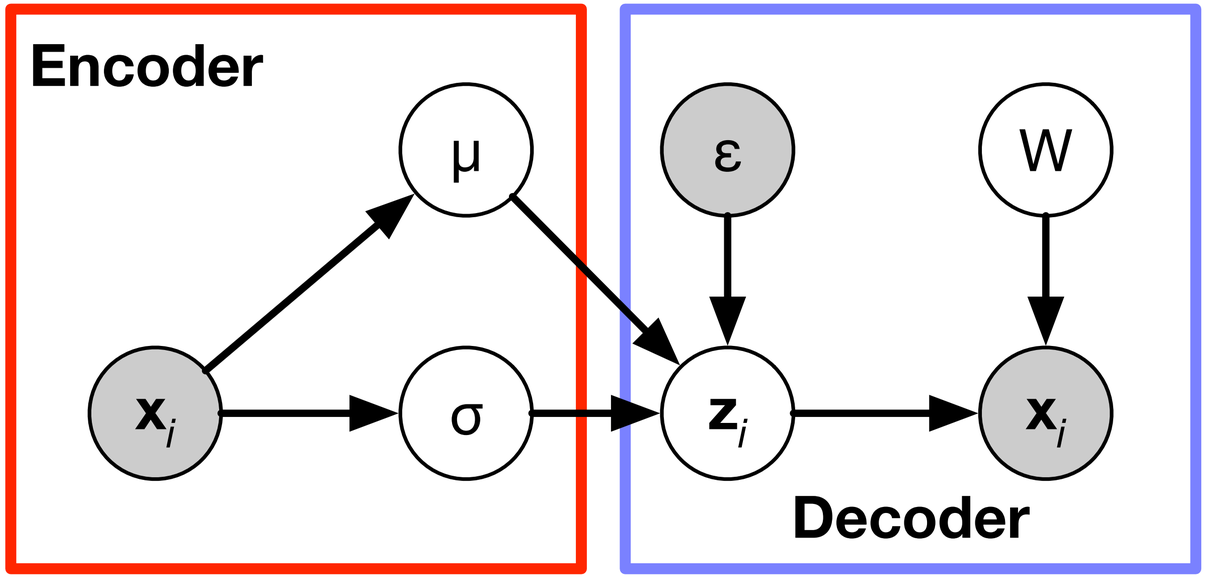

Fine-tuning by taking gradient steps

Hinton and Salakhutdinov, Science (2006)

# Why do we need unsupervised learning for single-cell RNA-seq data?

- Probabilistic interpretation of latent states
- Incomplete single-cell data, lots of drop-out measurements
- We can design generative model parameters as interpretable as possible!

# Variation autoencoder (VAE): a Bayesian inference framework for easy/scalable inference of latent variable model

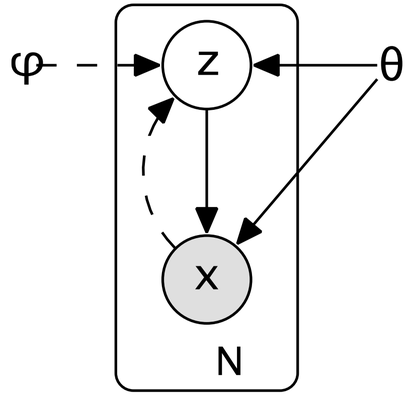



- Define relationships between variables (auto generative process)
- Usually, the decoder side captures our scientific hypothesis
- We can use an "auto-diff" algorithm (e.g., Facebook `torch` or Google `tensorflow`) to calculate gradients for the model parameters to optimize.

# Variational Inference, VI by Neural Net & SGD

True log-probability
(difficult due to integration
over all the latent variables)

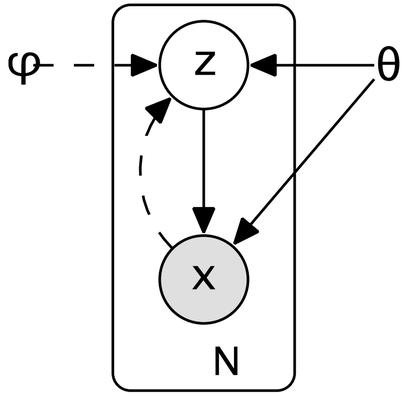
$$\ln \int dZ\, p(Z) p(X \mid \theta, Z)$$



# Variational Inference, VI by Neural Net & SGD

True log-probability
(difficult due to integration
over all the latent variables)

$$\ln \int dZ\, p(Z)p(X\,|\,\theta, Z)$$
$$= \ln \int dZ\, p(Z)p(X\,|\,\theta, Z)\frac{q(Z)}{q(Z)}$$

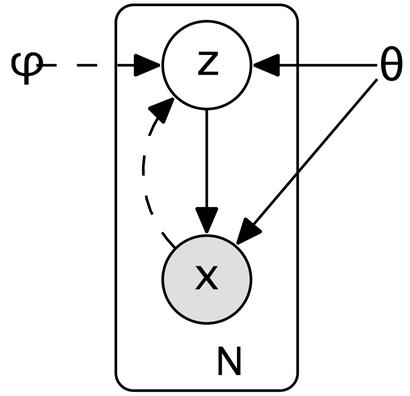


# Variational Inference, VI by Neural Net & SGD

True log-probability
(difficult due to integration
over all the latent variables)

$$\ln \int dZ\, p(Z)p(X\,|\,\theta, Z)$$
$$= \ln \int dZ\, p(Z)p(X\,|\,\theta, Z)\frac{q(Z)}{q(Z)}$$
$$\geq \int dZ \ln\left(\frac{p(Z)p(X\,|\,\theta, Z)}{q(Z)}\right) q(Z)$$

**Jensen's inequality**




What is your choice of q(Z)?

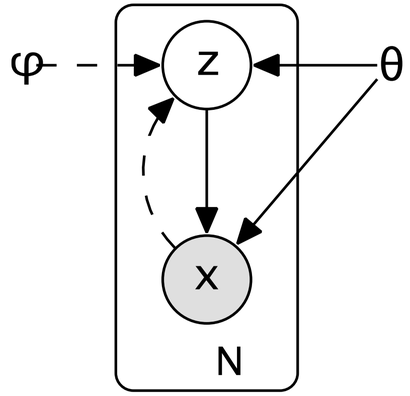
$q(Z) = \prod_d q(z_d)$

# Variational Inference, VI by Neural Net & SGD

True log-probability
(difficult due to integration
over all the latent variables)

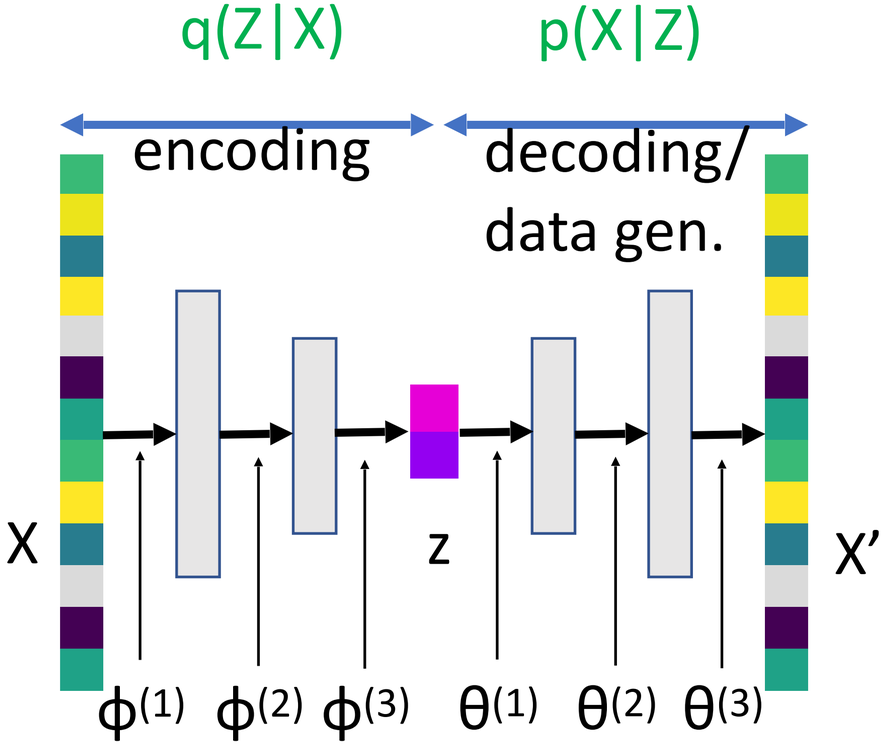
$$\ln \int dZ\, p(Z)p(X\,|\,\theta, Z)$$
$$= \ln \int dZ\, p(Z)p(X\,|\,\theta, Z)\frac{q(Z)}{q(Z)}$$
$$\geq \int dZ \ln\left(\frac{p(Z)p(X\,|\,\theta, Z)}{q(Z)}\right) q(Z)$$
$$= \mathbb{E}_q\left(\ln p(X\,|\,\theta, Z)\right)$$
$$+ \mathbb{E}_q\left(\ln p(Z)/q(Z)\right)$$

Lower-bound tractable
if variational q(Z) makes the function
easier to take average (E).



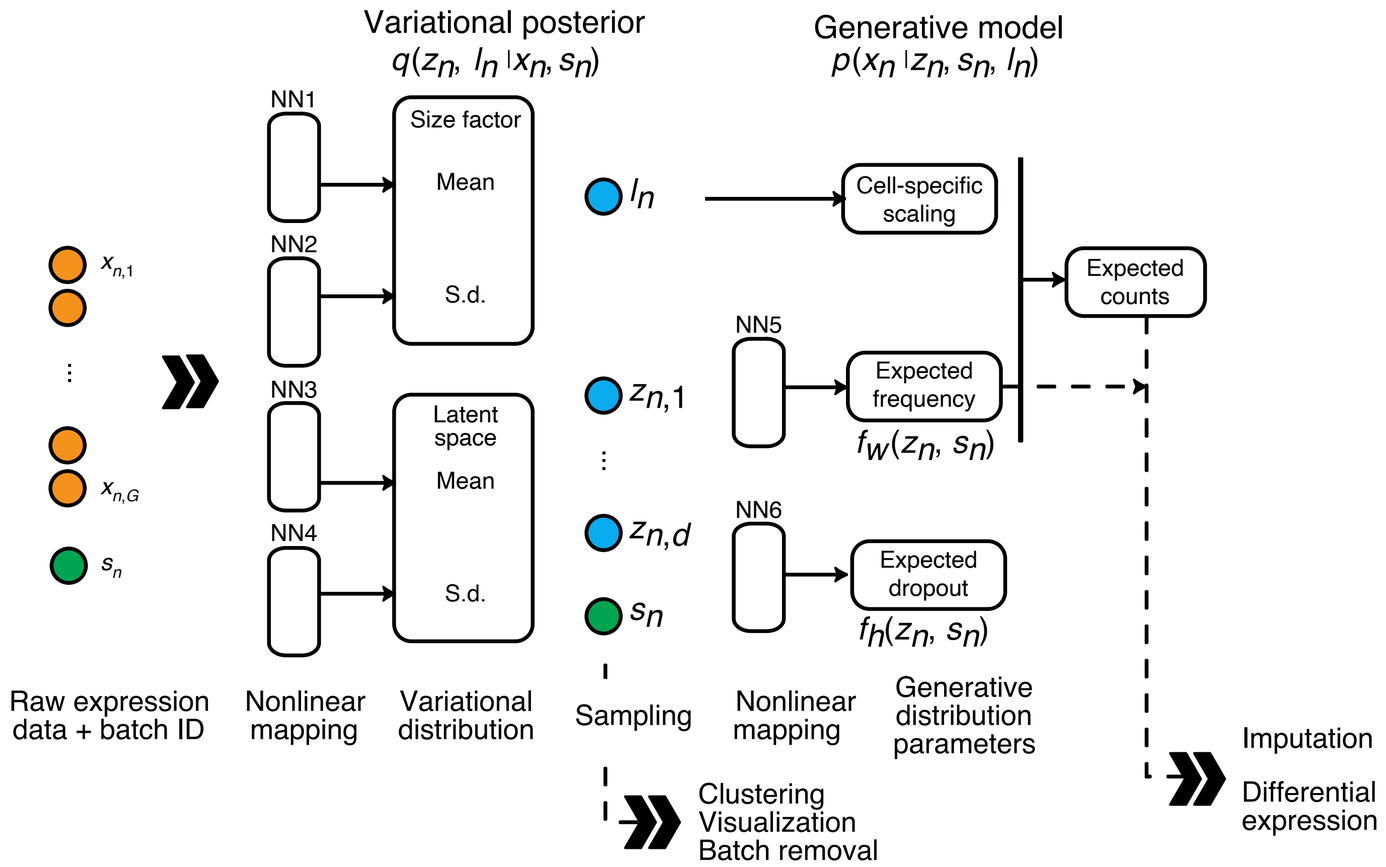

What is your choice of q(Z)?

What if taking expectation w.r.t. q(Z) is hard? **Sample z**




Kingma & Welling (2013)

# Deep generative modeling for single-cell transcriptomics

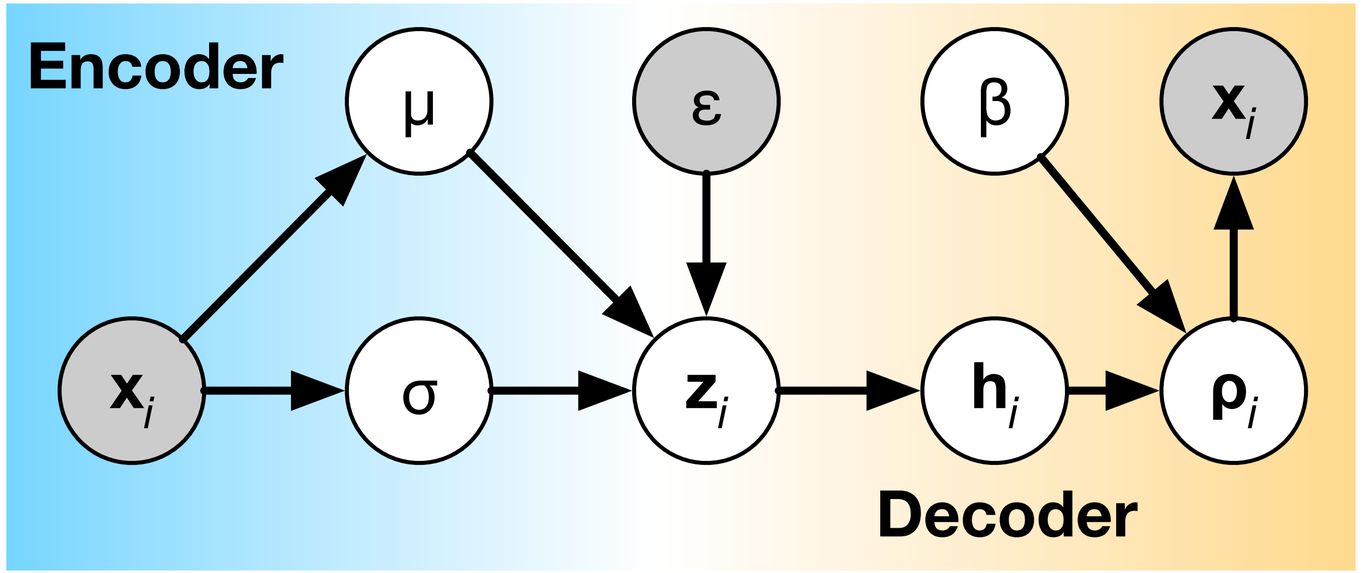



Generative model: zero-inflated negative binomial distribution

Lopez, .., Jordan, Yosef, *Nature Methods* (2018)

# Multinomial topic modelling for (incomplete) single-cell expression data

Can we simply model scRNA-seq counts by multinomial distribution?




- $X_{ig}$: gene expression of a gene $g$ in a single cell $i$
- $H_{ik}$: latent topic proportion of a cell $i$ to a topic $k$
- $\beta_{kg}$: topic $k$-specific gene probability

# Multinomial topic modelling for (incomplete) single-cell expression data

Can we simply model scRNA-seq counts by multinomial distribution?

Likelihood model:

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{g=1}^{\text{genes}} \left( \sum_{k} H_{ik} \beta_{kg} \right)^{X_{ij}}$$

- $X_{ig}$: gene expression of a gene $g$ in a single cell $i$
- $H_{ik}$: latent topic proportion of a cell $i$ to a topic $k$
- $\beta_{kg}$: topic $k$-specific gene probability

# Multinomial topic modelling for (incomplete) single-cell expression data

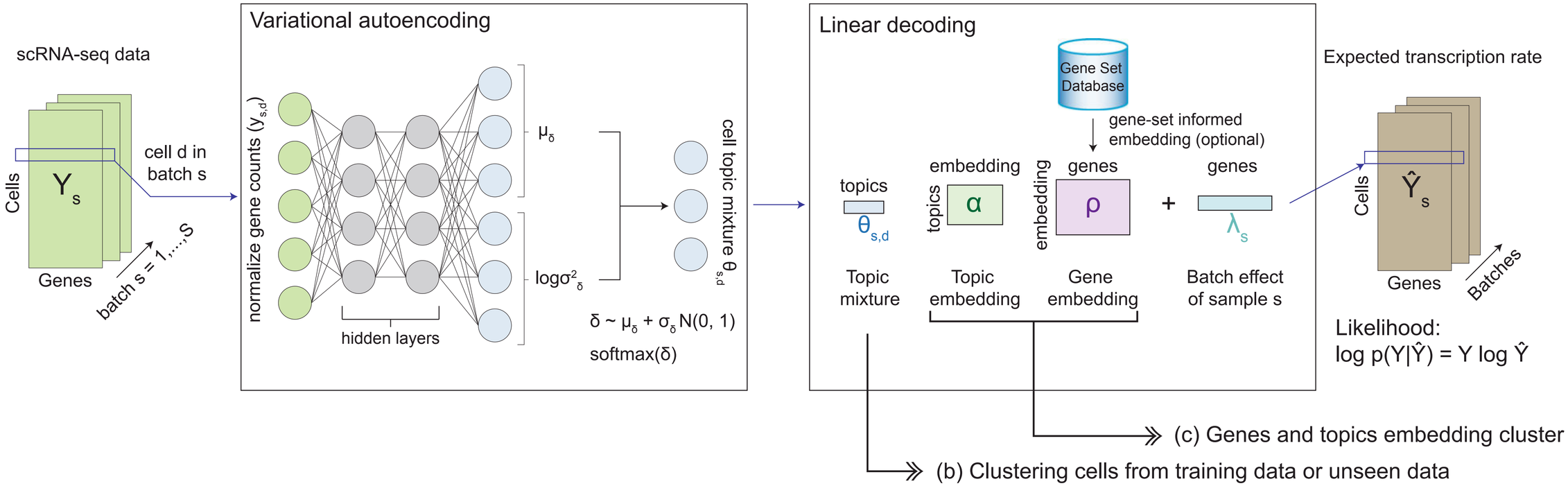
Can we simply model scRNA-seq counts by multinomial distribution?

Likelihood model:

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{g=1}^{\text{genes}} \left( \underbrace{\sum_{k} H_{ik}\beta_{kg}}_{\text{a gene } g\text{'s probability in a cell } \iota \equiv \rho_{ig}} \right)^{X_{ij}}$$

- $X_{ig}$: gene expression of a gene $g$ in a single cell $i$
- $H_{ik}$: latent topic proportion of a cell $i$ to a topic $k$
- $\beta_{kg}$: topic $k$-specific gene probability

# Single-cell Embedded Topic Model




We can factorize $\beta = \alpha\rho$.

Zhao, Cai, .., Li, *Nature Comm.* (2021)

# Topic Modelling: Comparison between document vs. single-cell

We think of a cell as a document, which is $\approx$ a bag of words, or $\approx$ a bag short mRNA reads.

| variables | in document topic model | in single cell ETM |
|---|---|---|
| $D$ | Total number of documents (corpus) | Total number of cells |
| $d$ | Document index | Cell index |
| $N_d$ | Number of words in a document $d$ | Number of read counts in a cell $d$ |
| $j$ | Word index, $j \in [N_d]$ | Read index |
| $K$ | Total number of topics | Total number of cell type topics |
| $k$ | Topic index, $k \in [K]$ | Cell topic index |
| $V$ | Size of vocabulary | Total number of genes |
| $v$ | Vocabulary index $v \in [V]$ | Gene index |
| $W_{dj}^v$ | Indicator for a word to vocabulary $\in \{0, 1\}$ | Indicator for a read to a gene $\in \{0, 1\}$ |
| $X_{dv}$ | Vocabulary $v$ occurrence in a document $d$ | Gene expression of a gene $v$ in a cell $d \in [0, N_d]$ |

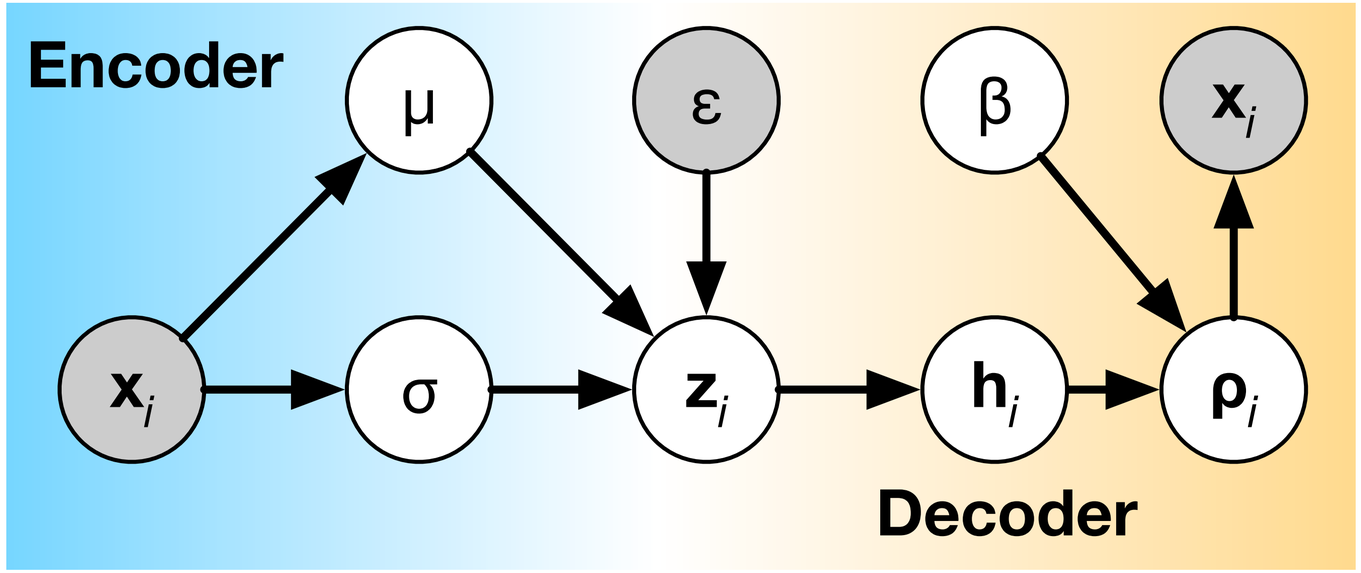
$W_{dj}^v = 1$ if and only if a word $j$ in a document $d$ takes $v$-th word in the vocabulary; otherwise, $W_{dj}^v = 0$.

# Single-cell Embedded topic model's latent states and model parameters

| variables | in document topic model | in single cell ETM |
|---|---|---|
| $Z_{dj}^k$ | Indicator for assigning a word to a topic $k$ | Indicator for assigning a read to a topic $k$ |
| $H_{dk}$ | Hidden state $k$ of a document $d$ | Hidden state $k$ of a cell $d$ |
| $\beta_{kv}$ | topic $k$-specific vocabulary $v$ frequency | topic $k$-specific, a gene $v$'s exression |

- In Latent Dirichlet Allocation: $\sum_{k=1}^{K} H_{dk} = 1$ and $H_{dk} > 0$, and $\mathbf{h}_d \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$ *a priori*. Approximately, we have $\hat{H}_{dk} = \sum_j^{N_d} Z_{dj}^k / N_d$.
- In Embedded Topic model, $H_{dk}$ with the simplex constraints; $H_{dk} = \exp(\delta_{dk}) / \sum_{k'} \exp(\delta_{dk'})$ where $\delta_{dk} \sim \mathcal{N}(0, 1)$ *a priori*.
- Additional constraints: $\beta_{kv} > 0$ and $\sum_v \beta_{kv} = 1$, meaning that only a handful of vocabulary $v$ contribute to a topic $k$.

# Multinomial topic modelling for (incomplete) single-cell expression data

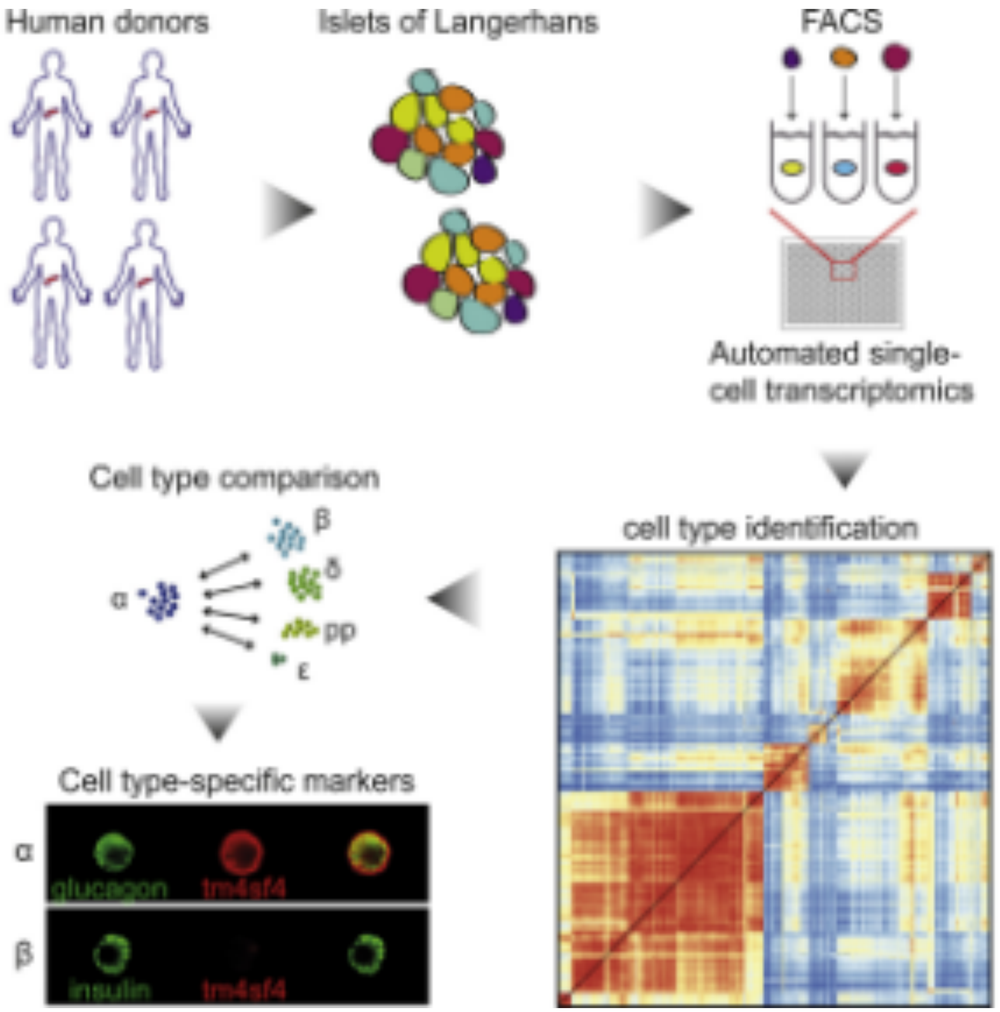



Probability of gene $g$ in a cell $i$:

$$\rho_{ig} = \sum_{k \in \text{topics}} H_{ik} \beta_{kg}$$

By **not** normalizing the probability of each cell, we do not worry about modelling sequencing depths.

# Example: single-cell RNA-seq data of human pancreatic cells



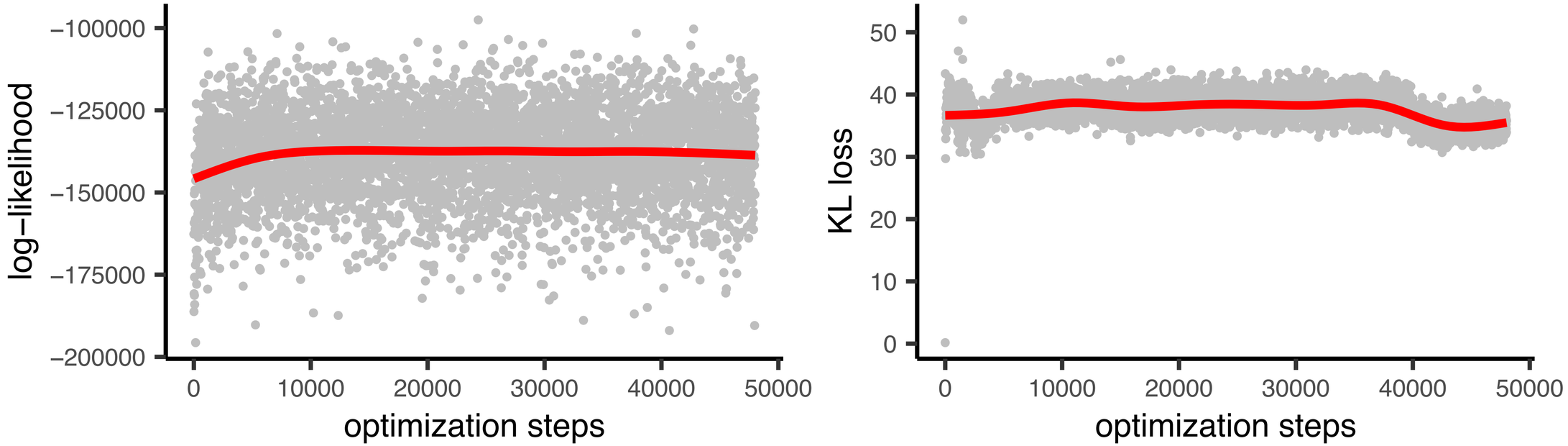

Muraro *et al. Cell Systems* (2016)

We will use scRNA-seq data (GEO accession: GSE85241) as a working example.

- genes/features/rows: 19,140
- cells/columns: 3,072
- non-zero elements: 12,442,034
- $\sim$ 21 % non-zero

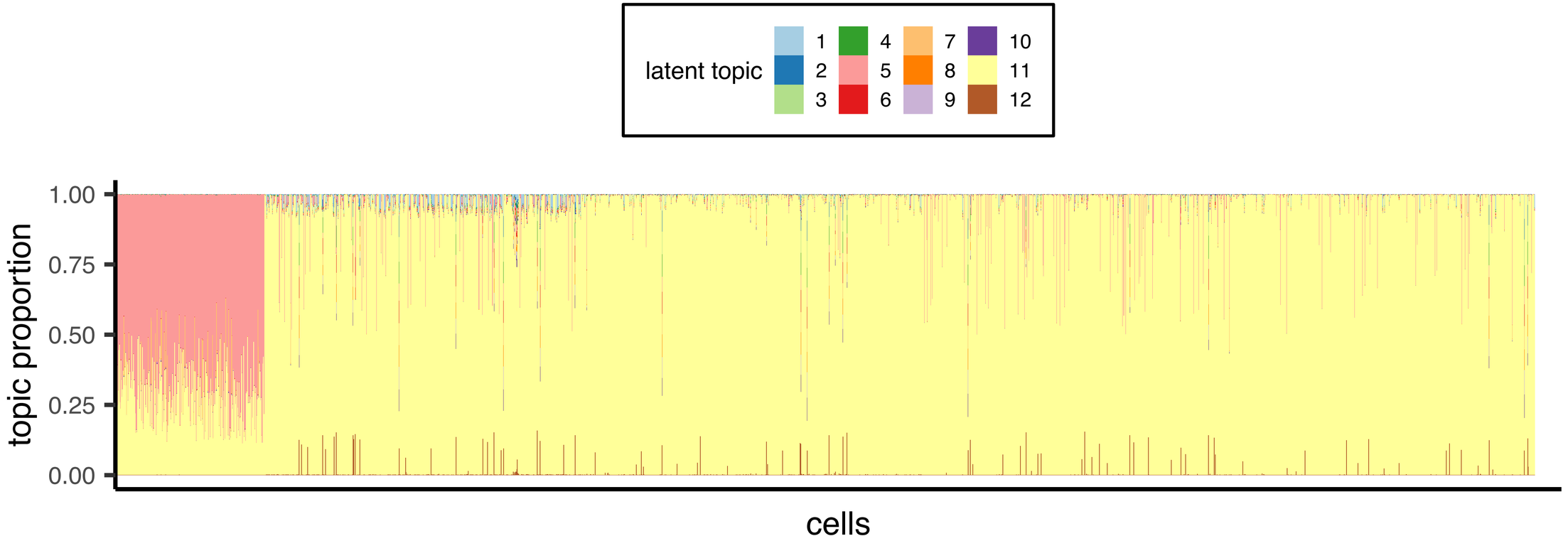
# Variational inference $\approx$ maximum likelihood regularized by a KL-divergence term

$$\underbrace{\mathbb{E}[\log p(\mathbf{x}|\mathbf{h}(\mathbf{z}))]}_{\text{expected data likelihood}} - \underbrace{\mathbb{E}[\log q(\mathbf{z})/p(\mathbf{z})]}_{\text{KL loss}}$$



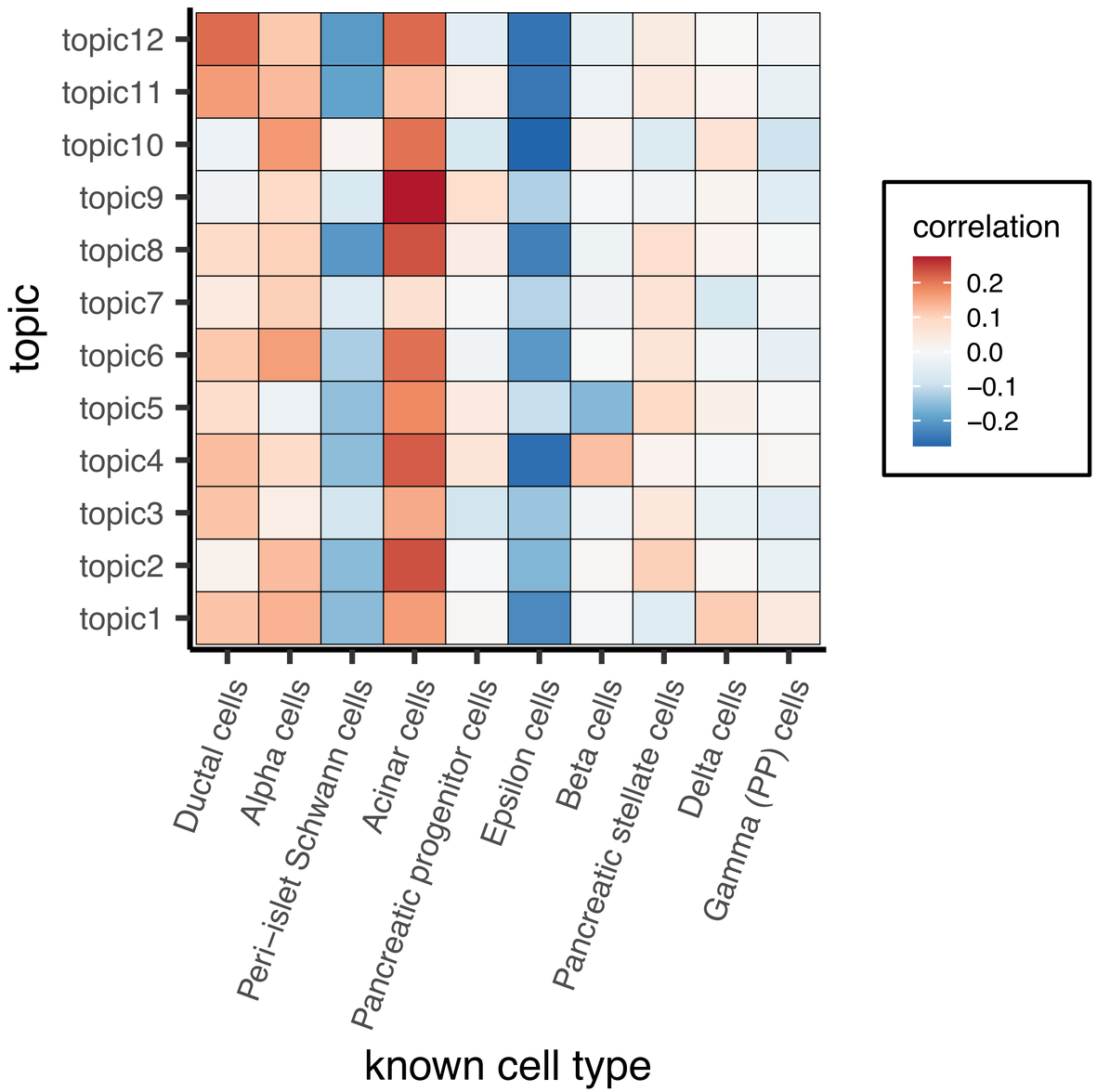

We may need to train longer than usual... (don't be fooled by log-likelihood)

# ETM learning just started ... (hidden states $\mathbf{h}$)




There is no obvious pattern... yet

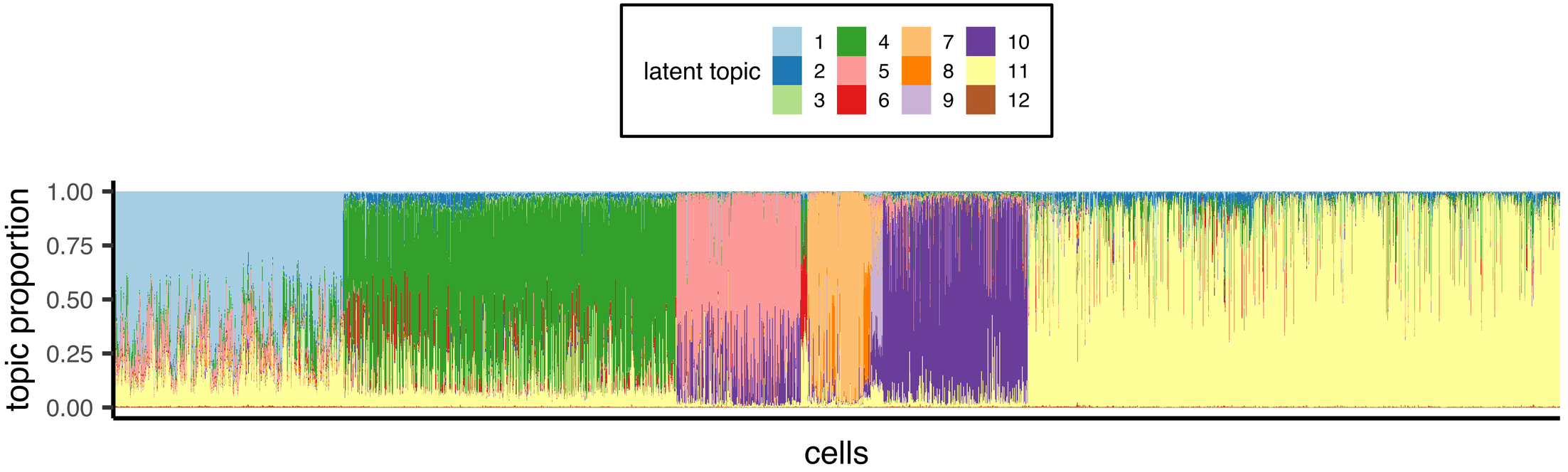
# ETM learning just started ... (weight parameters $\beta$)




- We can correlate each topic-specific gene $\times 1$ weight vector, $\beta_k$, with known cell type-specific marker genes
- No obvious concepts emerged yet, not so specific correlation patterns, yet...

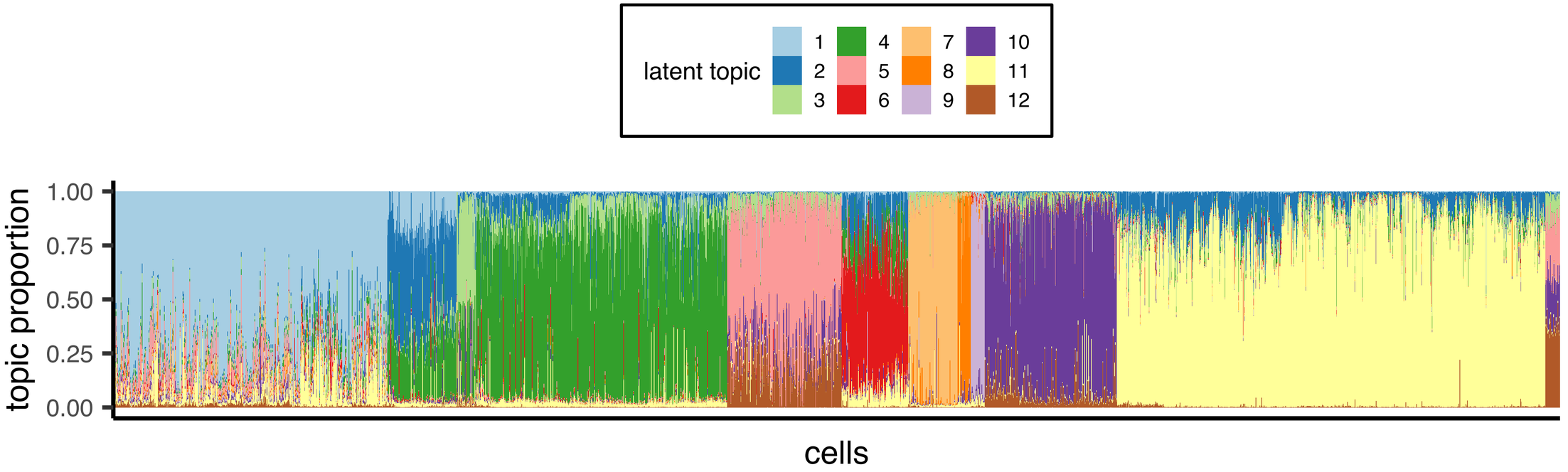
# If we keep on training ETM (hidden states) ..

epoch = 270



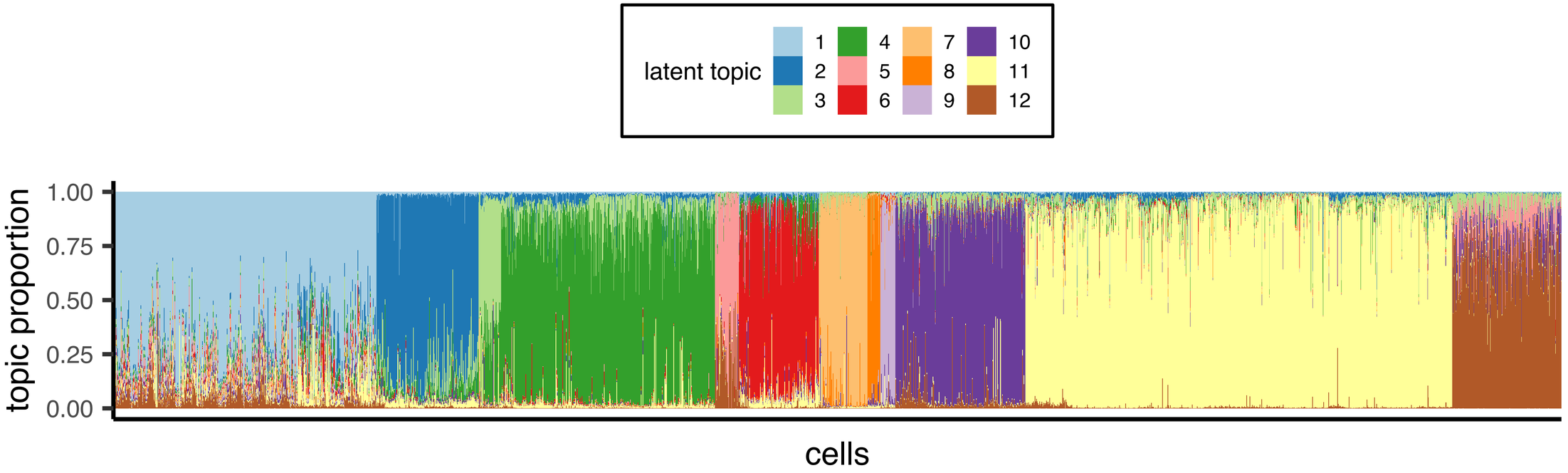
# If we keep on training ETM (hidden states) ..

epoch = 570

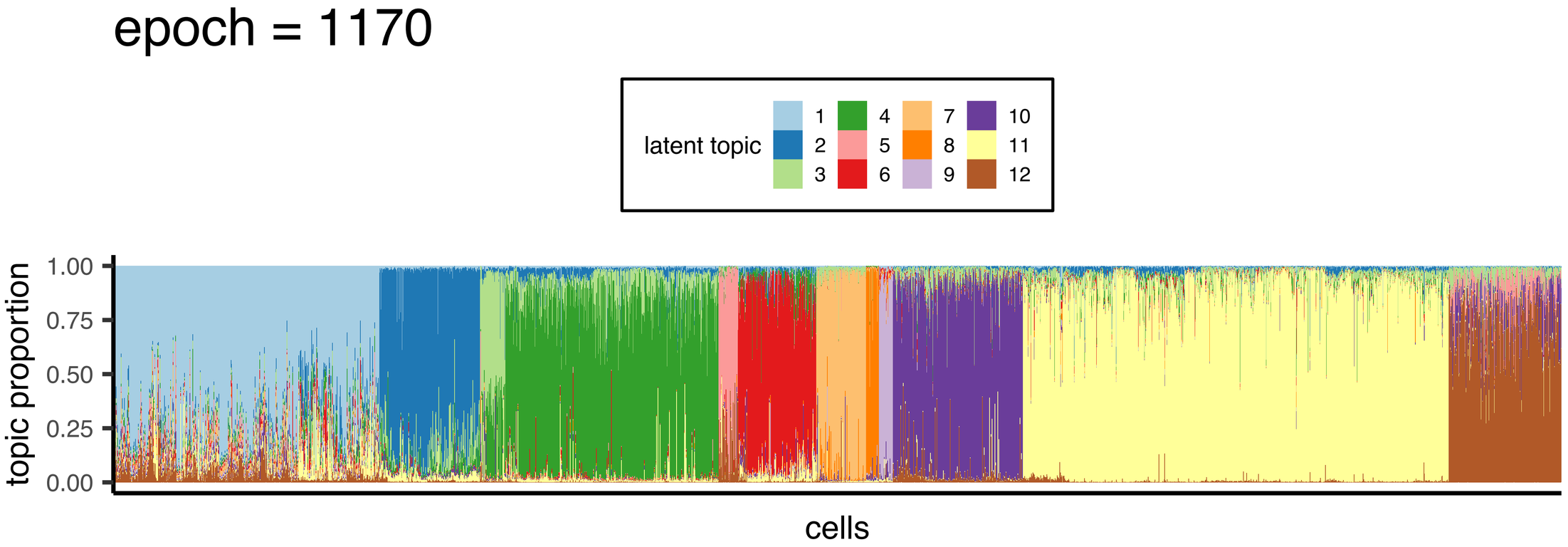


# If we keep on training ETM (hidden states) ..

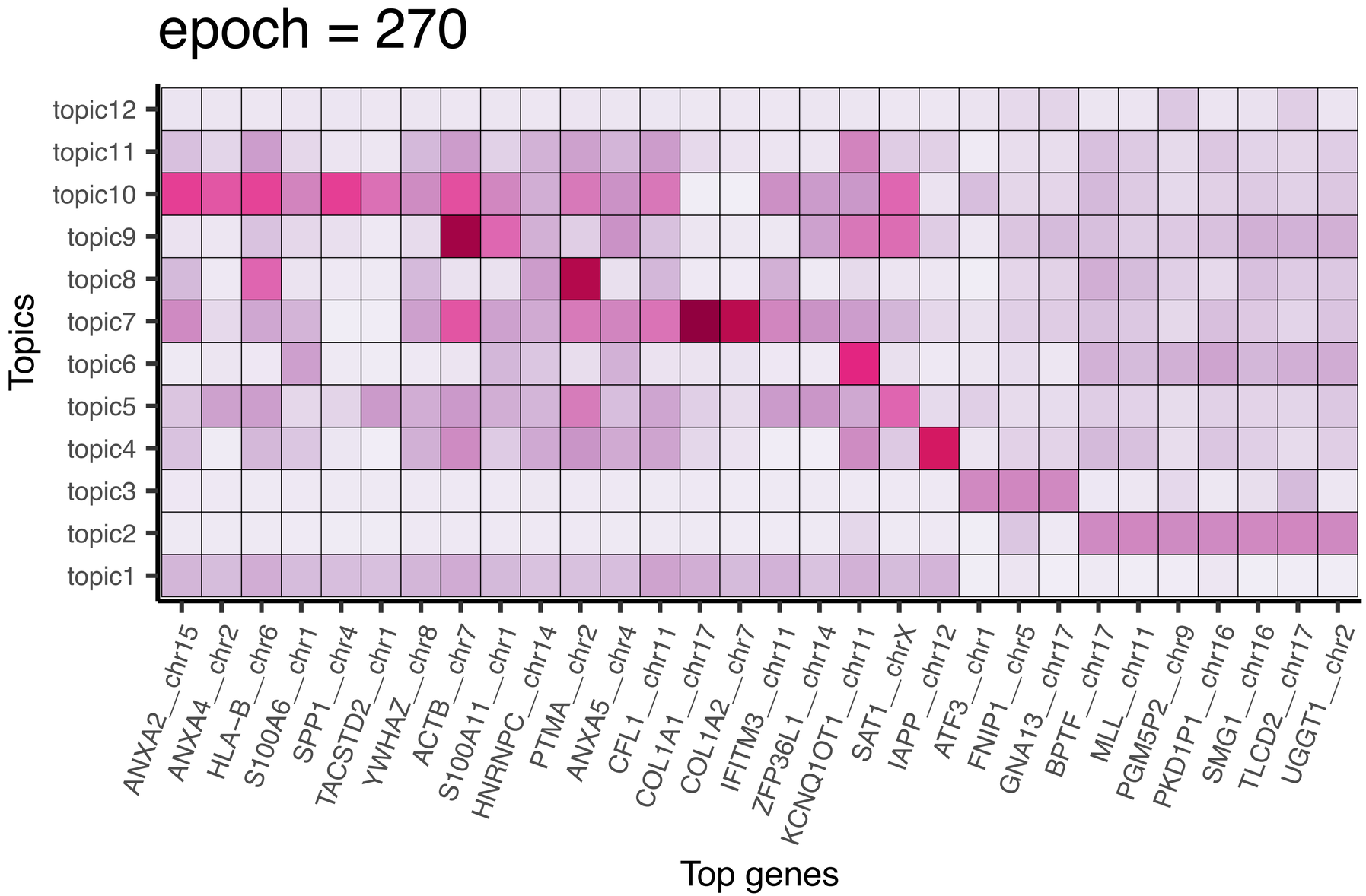
epoch = 870



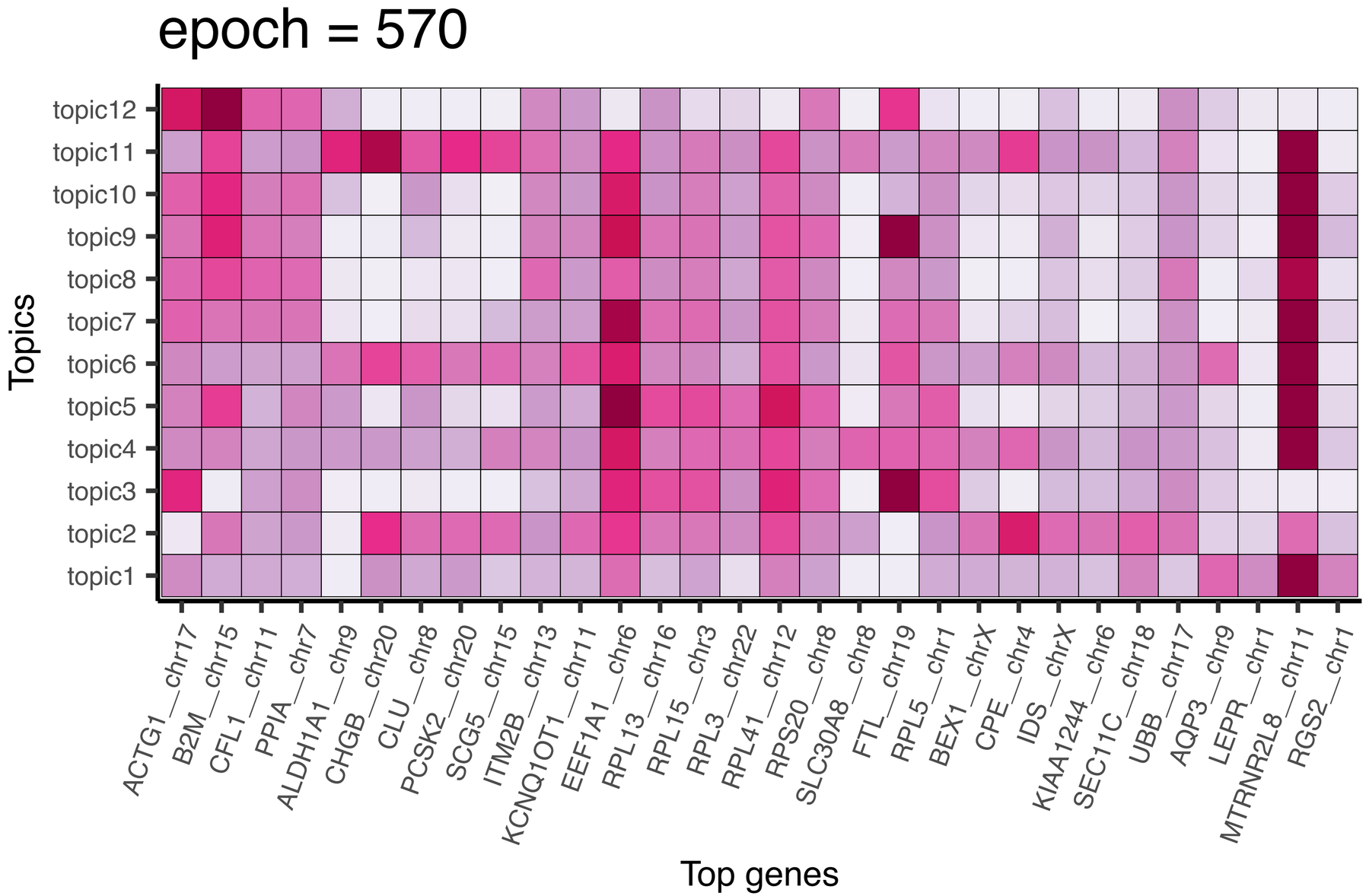
# If we keep on training ETM (hidden states) ..



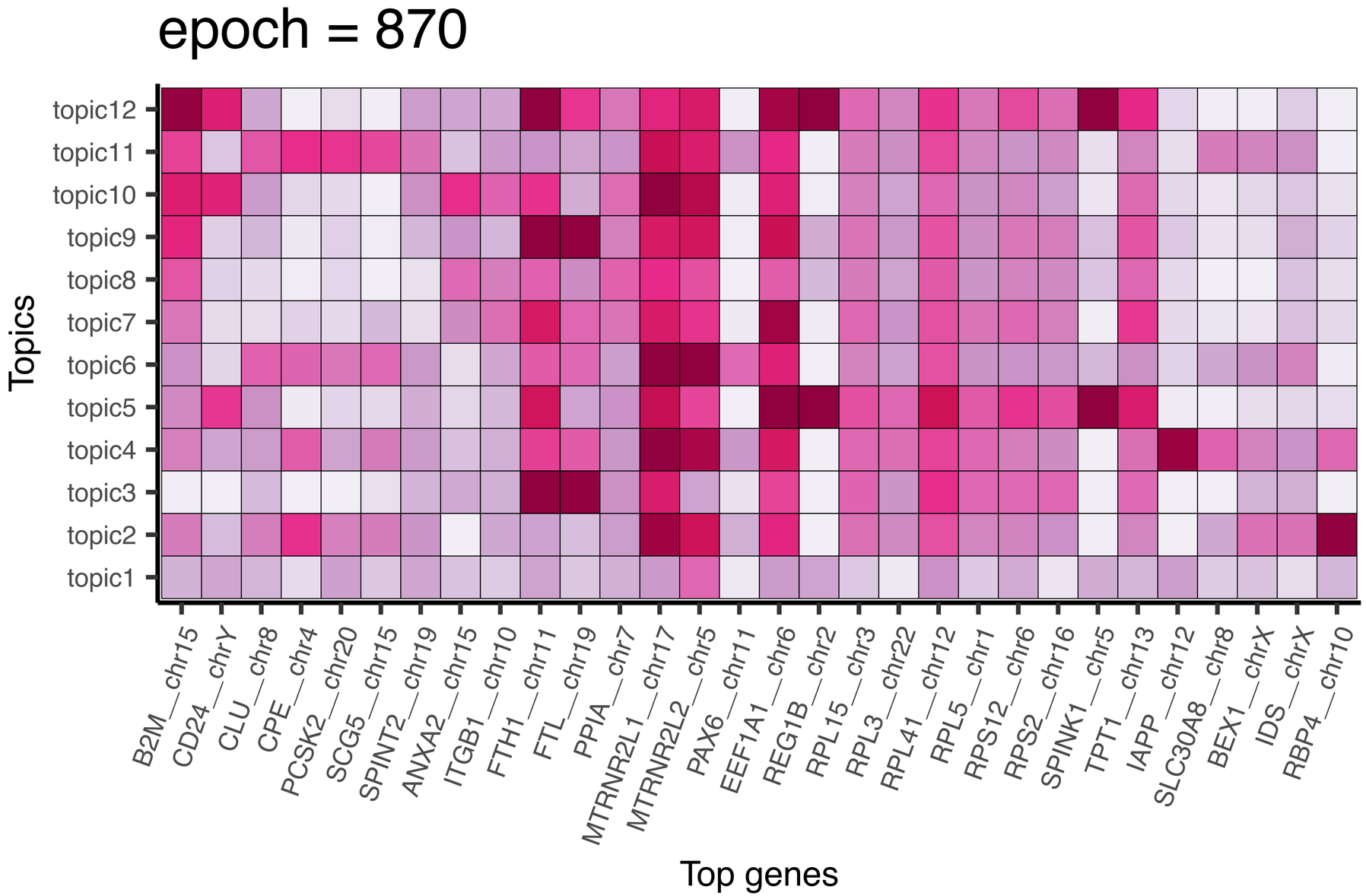
# If we keep on training ETM (weight parameters) ...



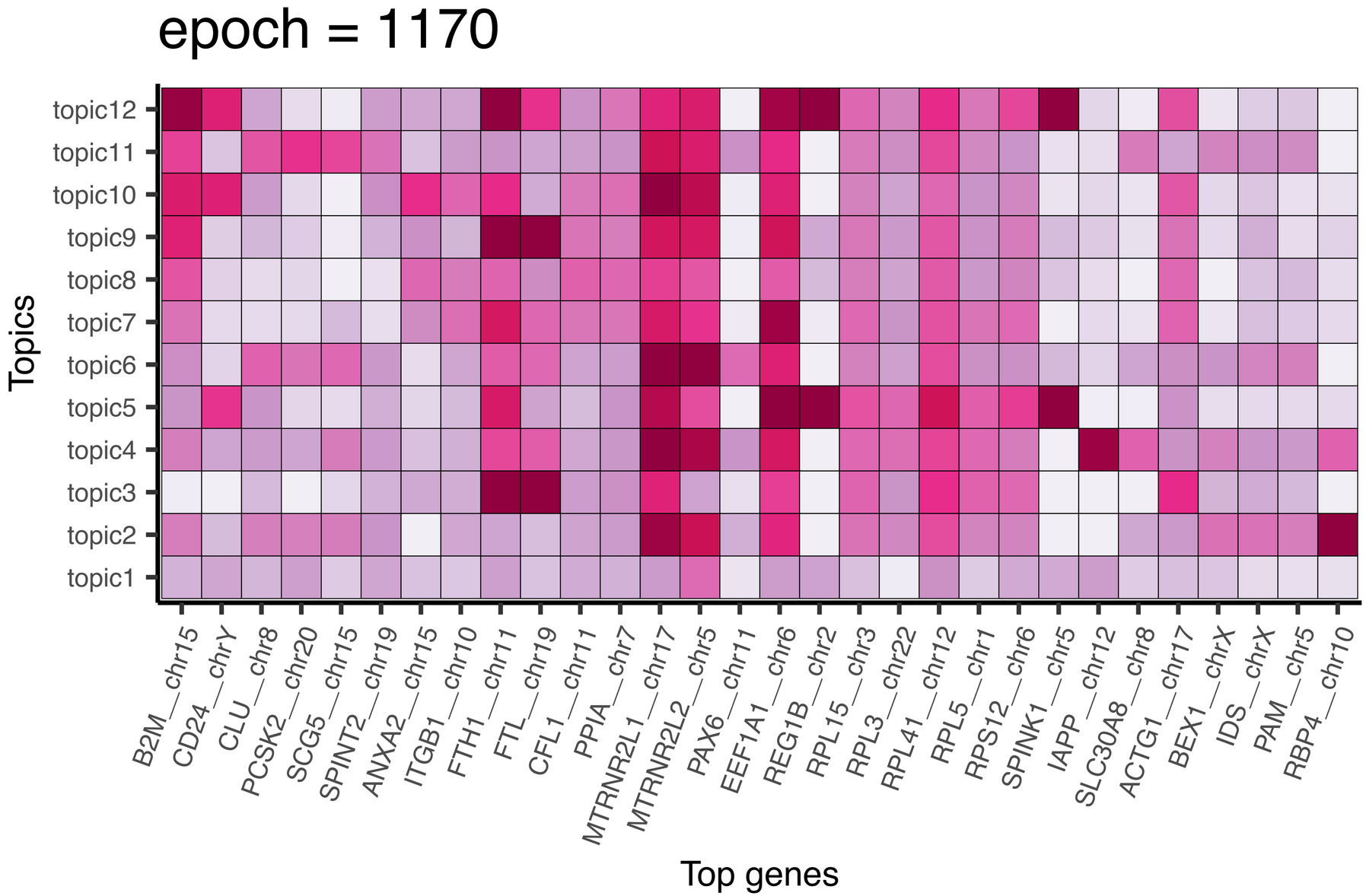
# If we keep on training ETM (weight parameters) ...



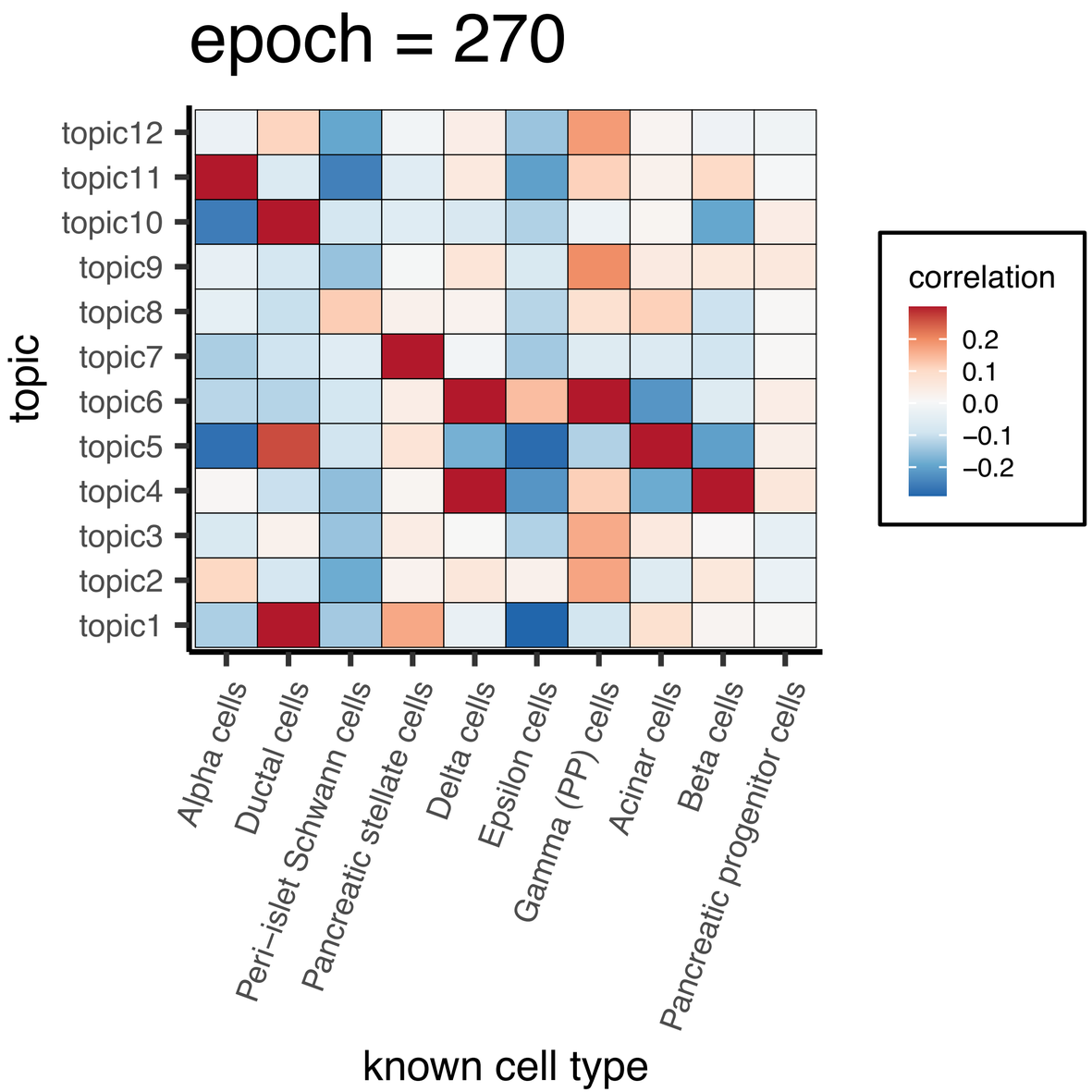
# If we keep on training ETM (weight parameters) ...



# If we keep on training ETM (weight parameters) ...

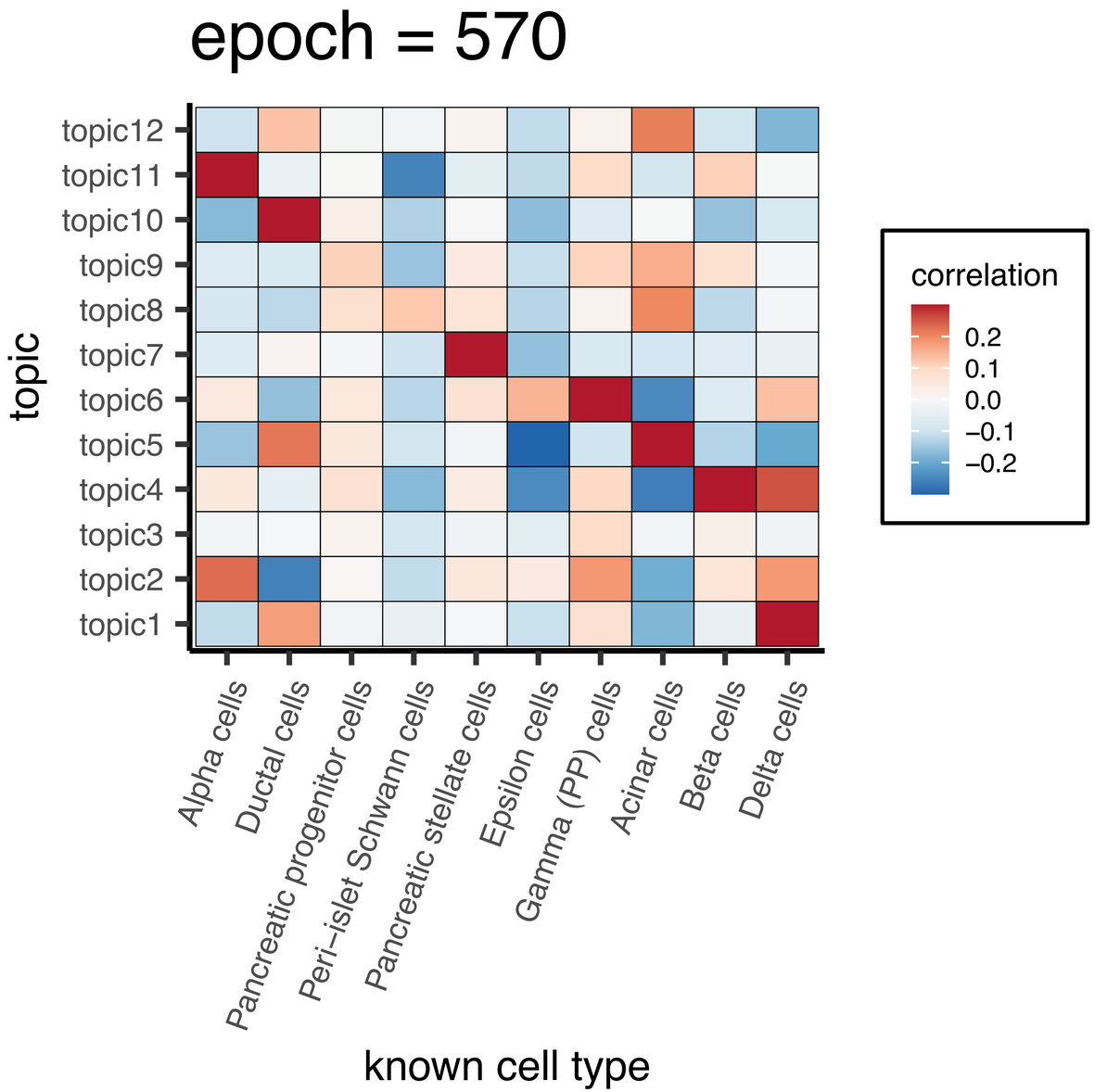


# If we keep on training ETM (weight parameters) ...



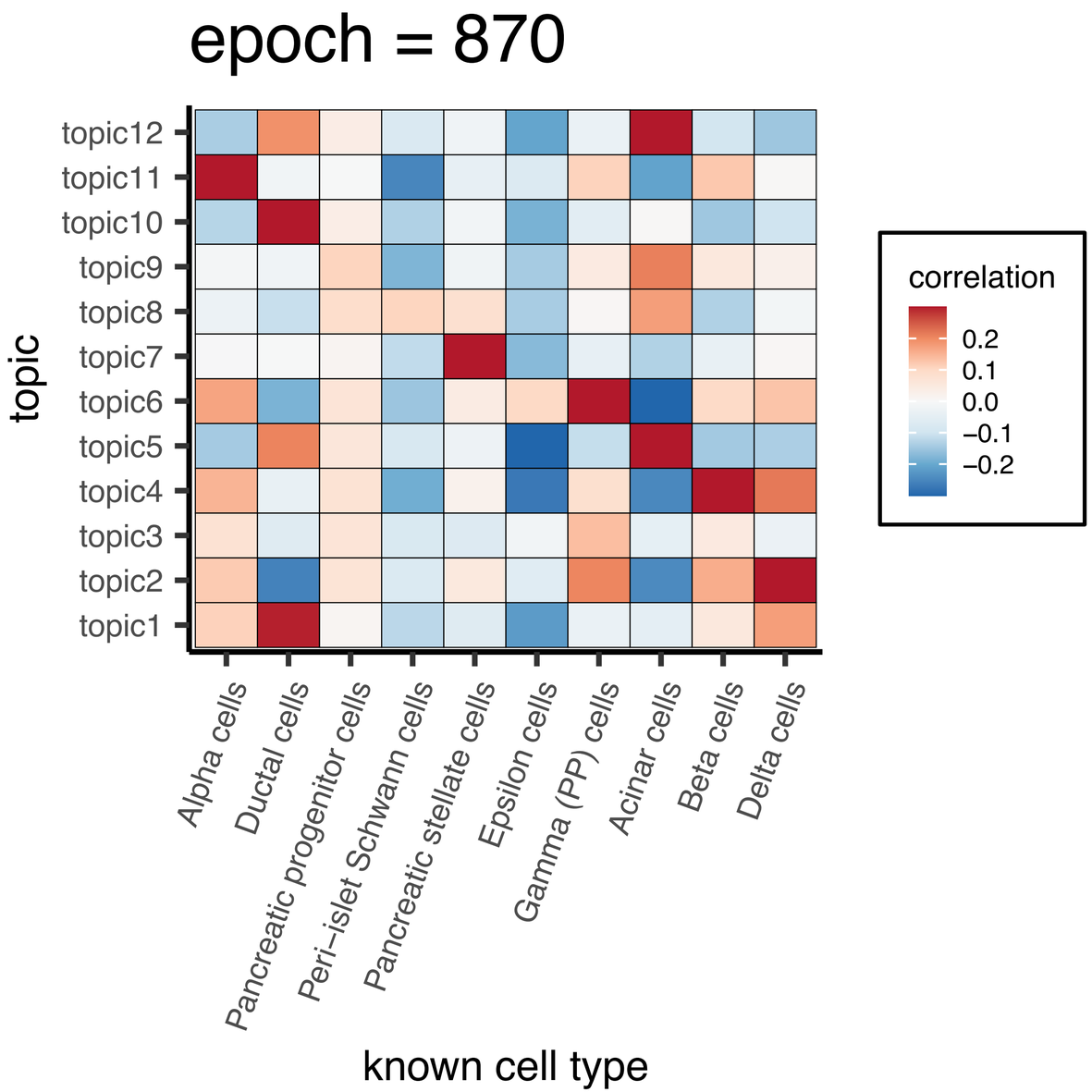

- We can correlate each topic-specific gene $\times 1$ weight vector, $\beta_k$, with known cell type-specific marker genes
- We retrieved marker gene information of known cell types from PangaloDB

# If we keep on training ETM (weight parameters) ...



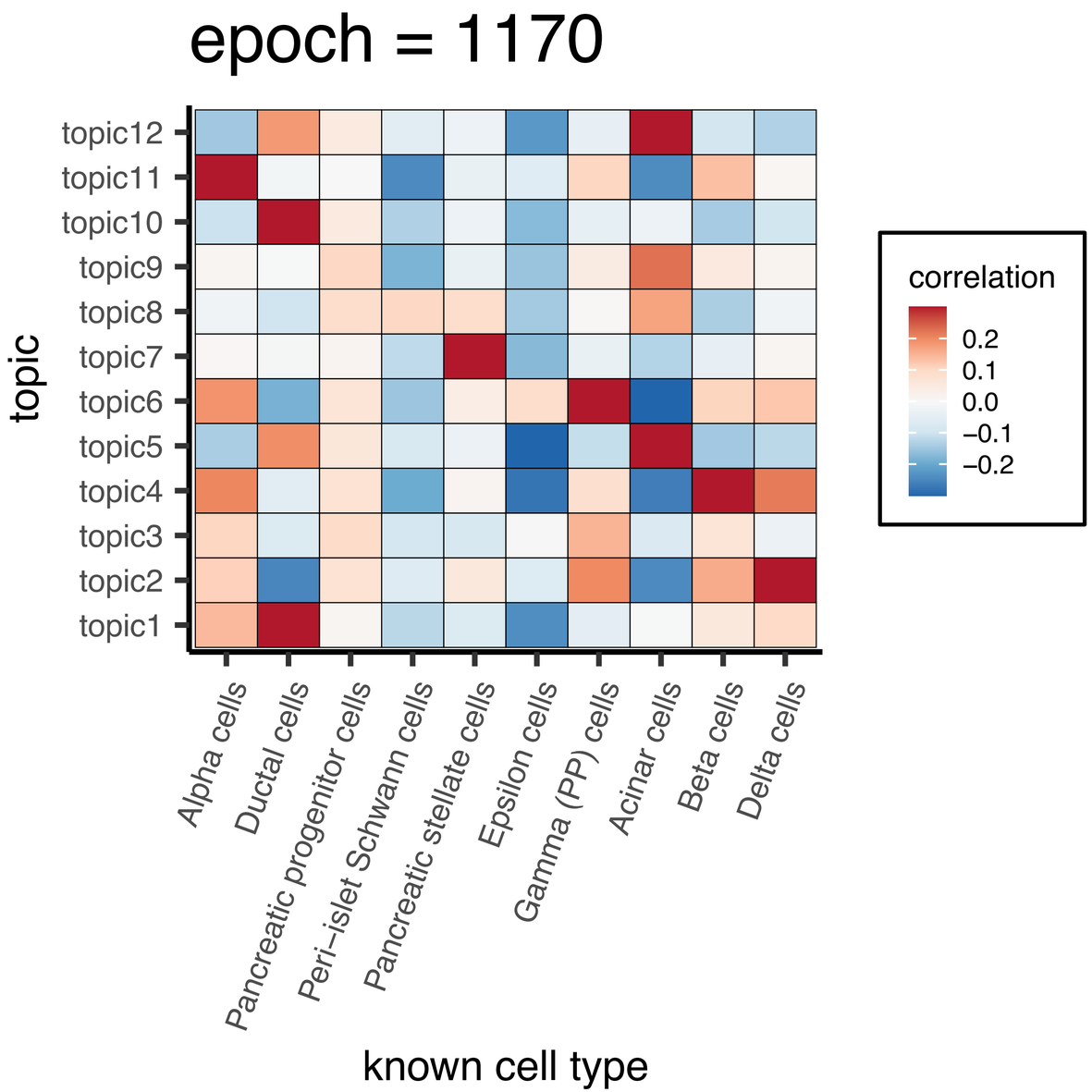

- We can correlate each topic-specific gene $\times 1$ weight vector, $\beta_k$, with known cell type-specific marker genes
- We retrieved marker gene information of known cell types from PangaloDB

# If we keep on training ETM (weight parameters) ...



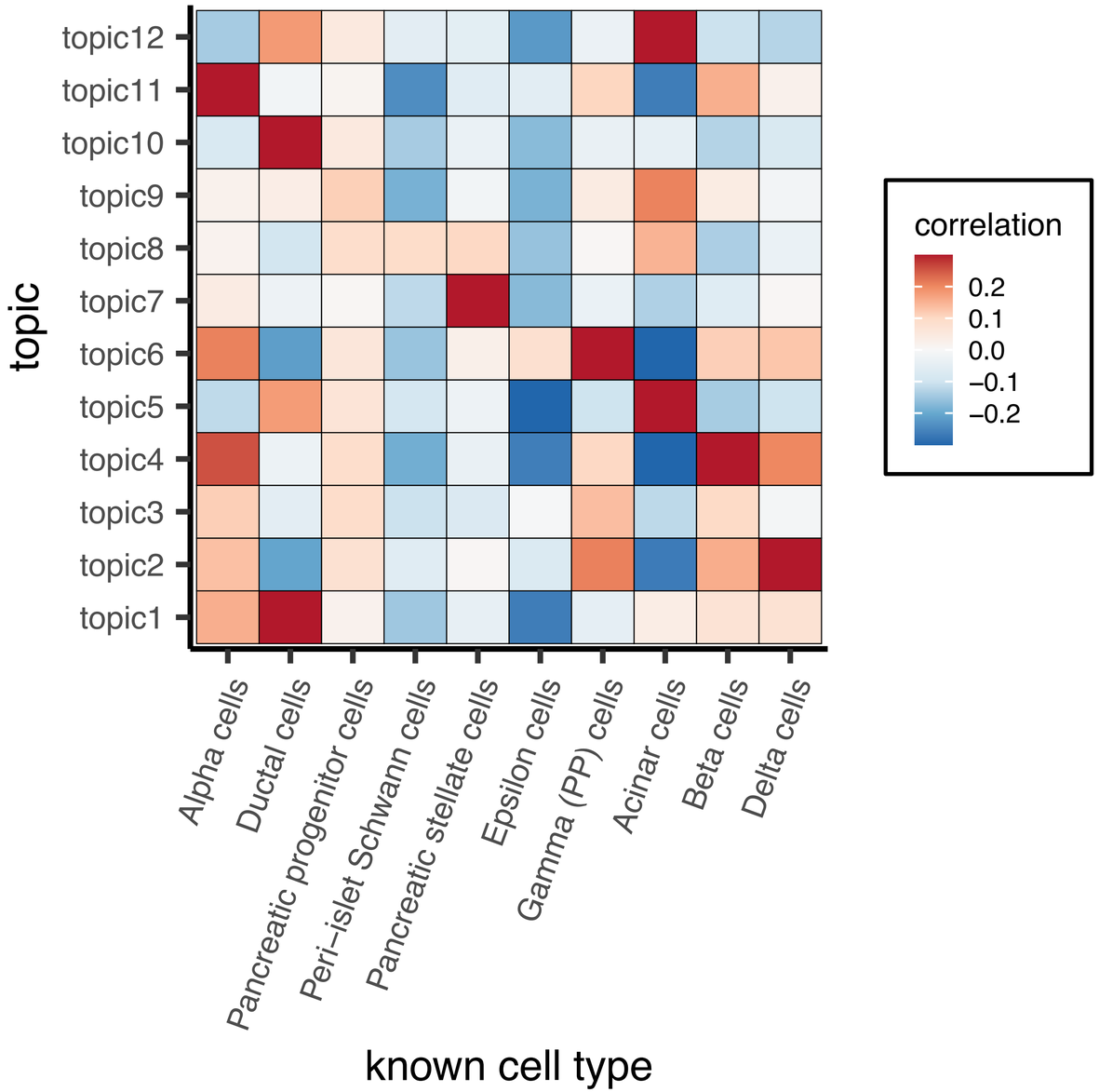

- We can correlate each topic-specific gene $\times 1$ weight vector, $\beta_k$, with known cell type-specific marker genes
- We retrieved marker gene information of known cell types from PangaloDB

# If we keep on training ETM (weight parameters) ...



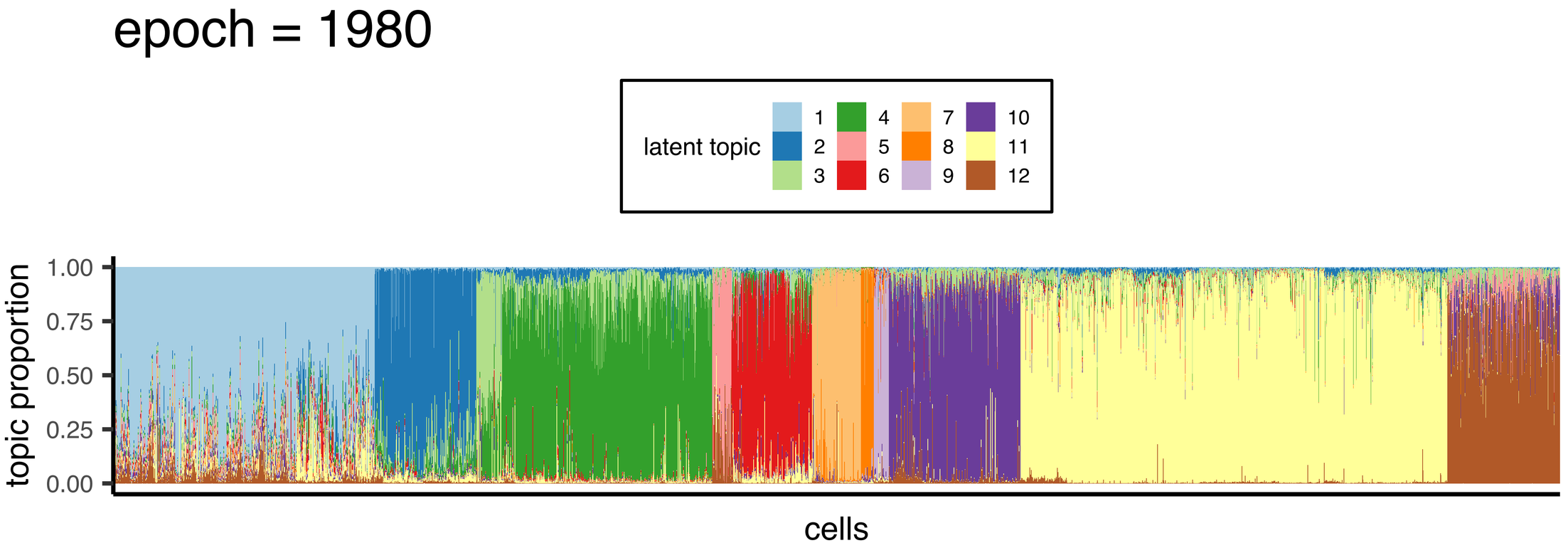

- We can correlate each topic-specific gene $\times 1$ weight vector, $\beta_k$, with known cell type-specific marker genes
- We retrieved marker gene information of known cell types from PangaloDB

# After enough training steps...

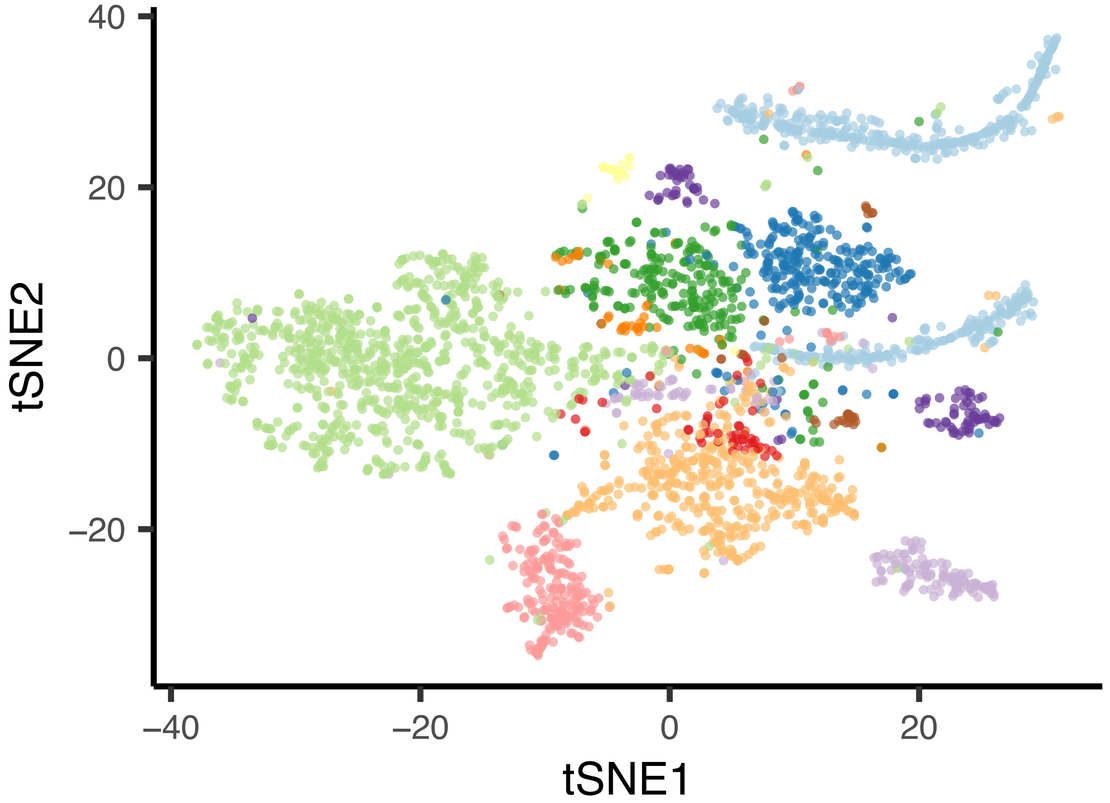



- We have recapitulated most known Pancreatic cell types in our single-cell analysis

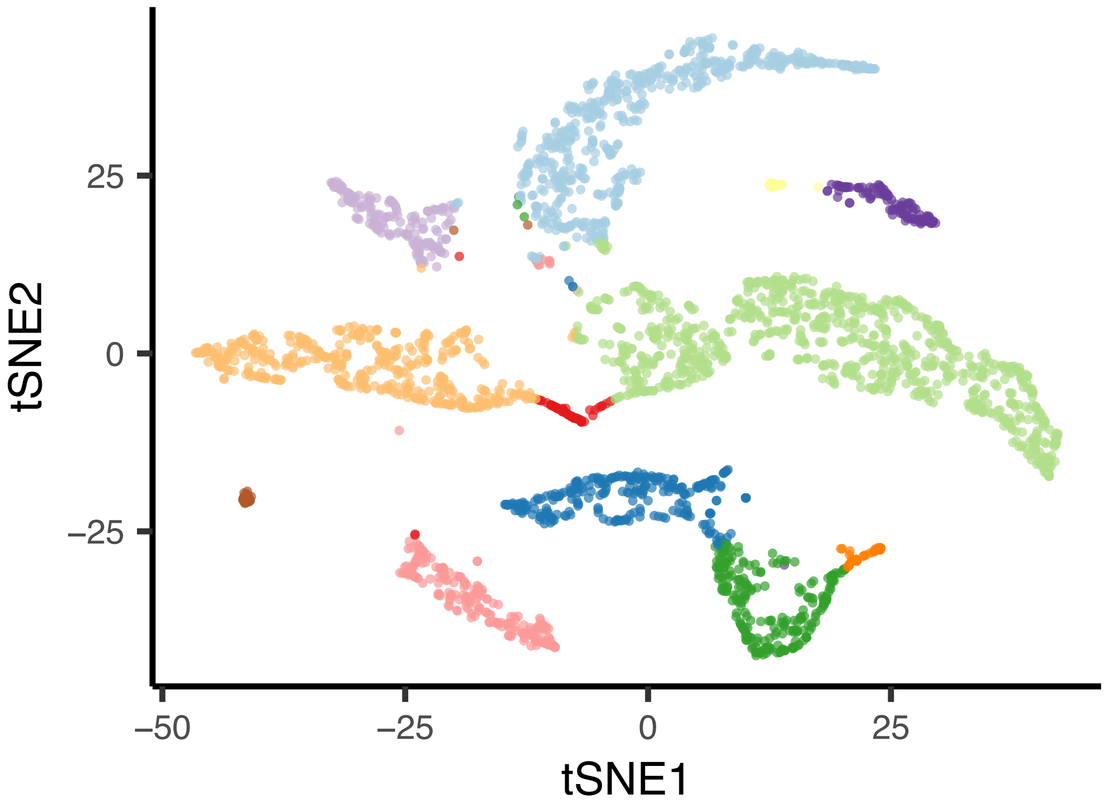
# Single-cell ETM effectively learns cellular admixture model



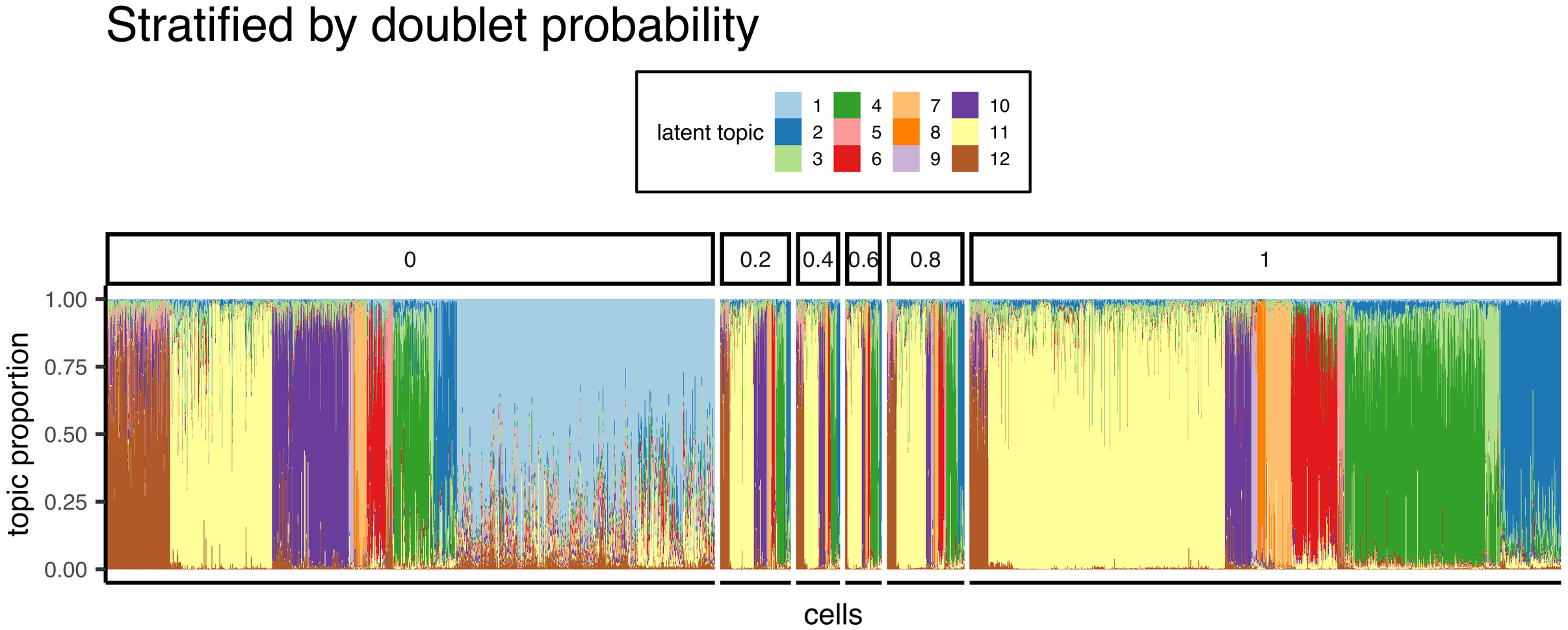
## using top 50 PCs




## tSNE on the latent topic space



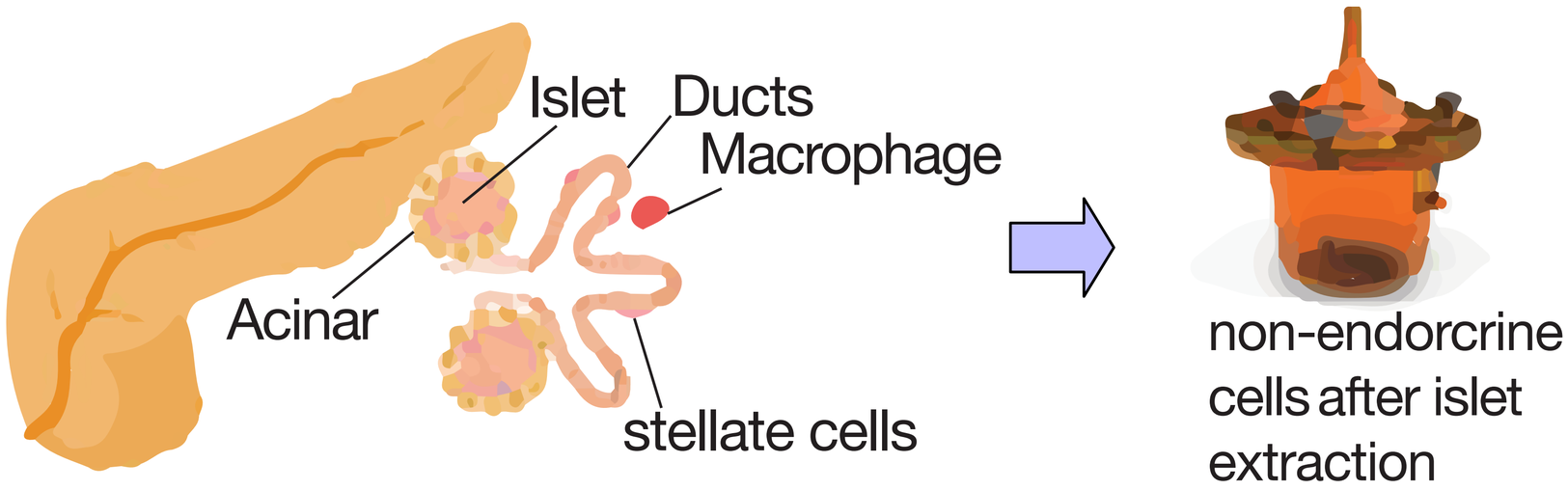
# Wait, what about the doublets?




What do you think?

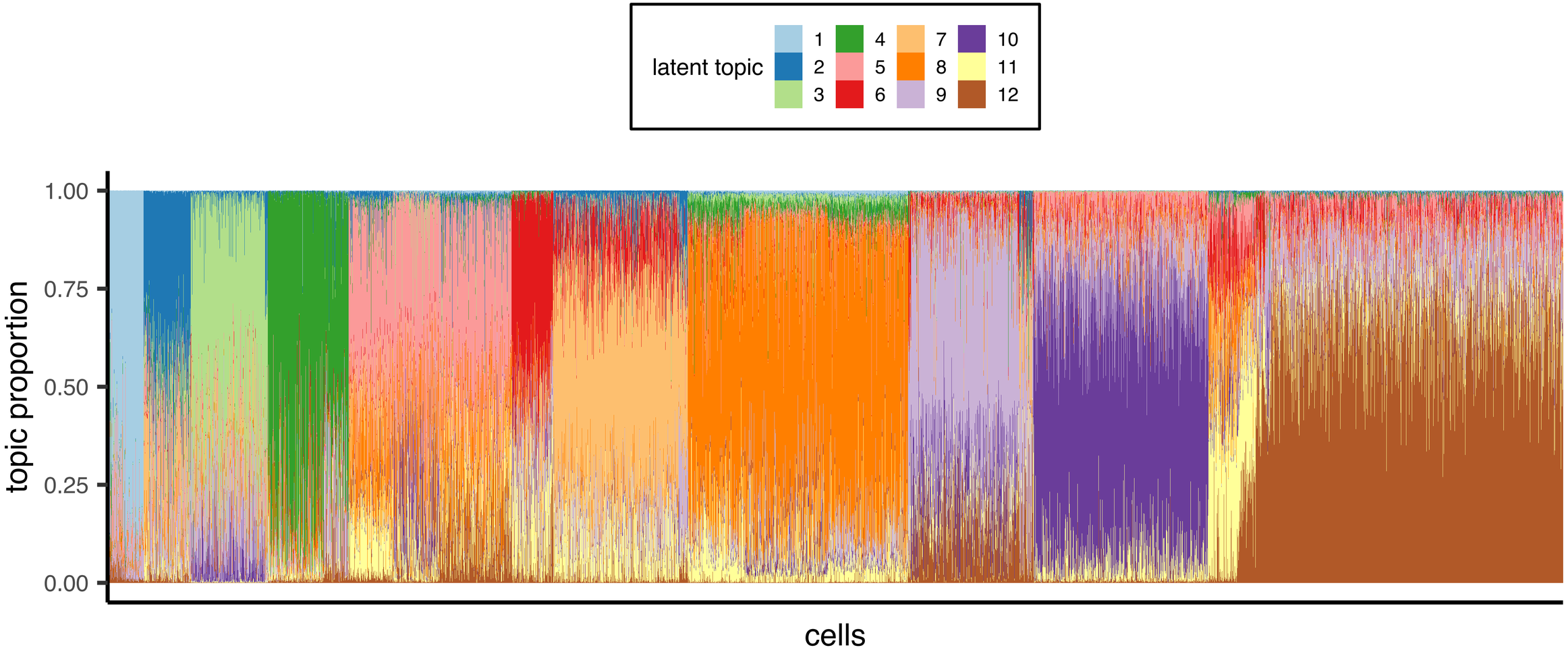
# A latent topic model robustly capture cell states, avoiding batch effects

Single-cell RNA-seq data from three donors (three batches)

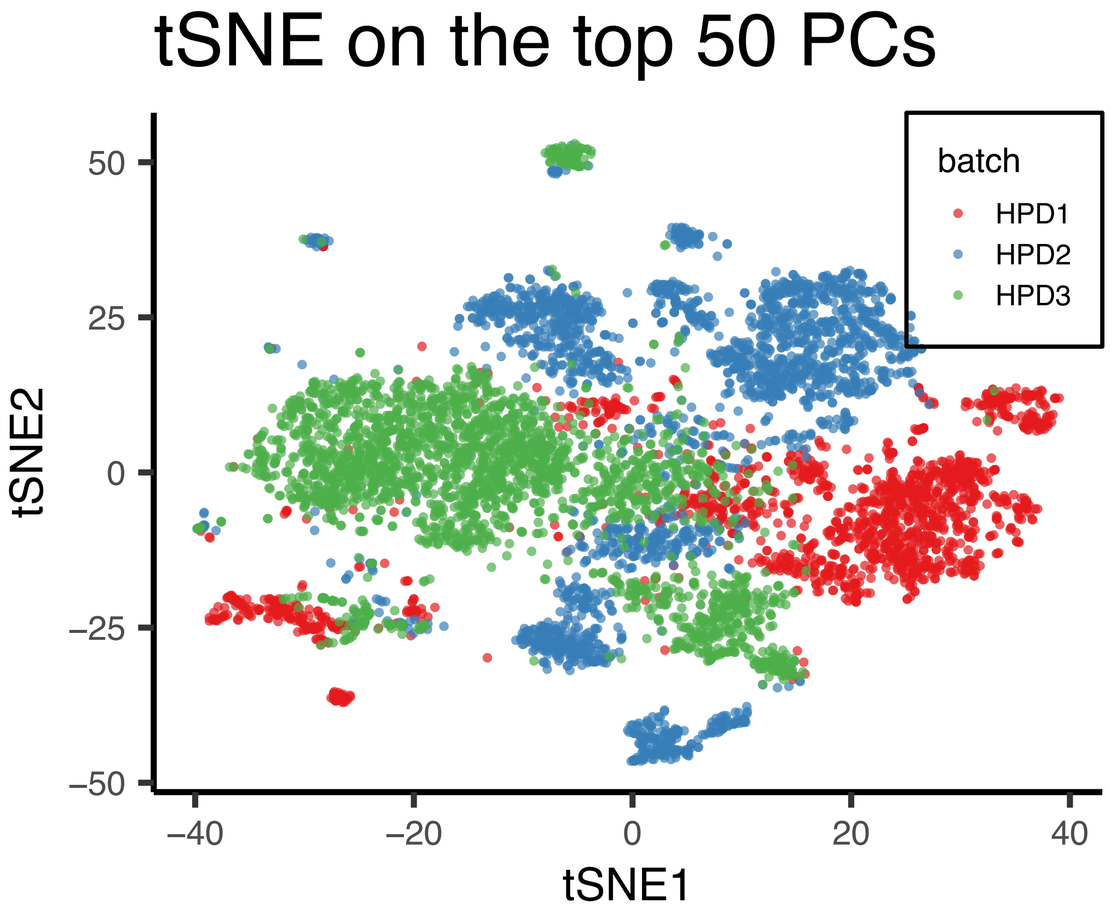


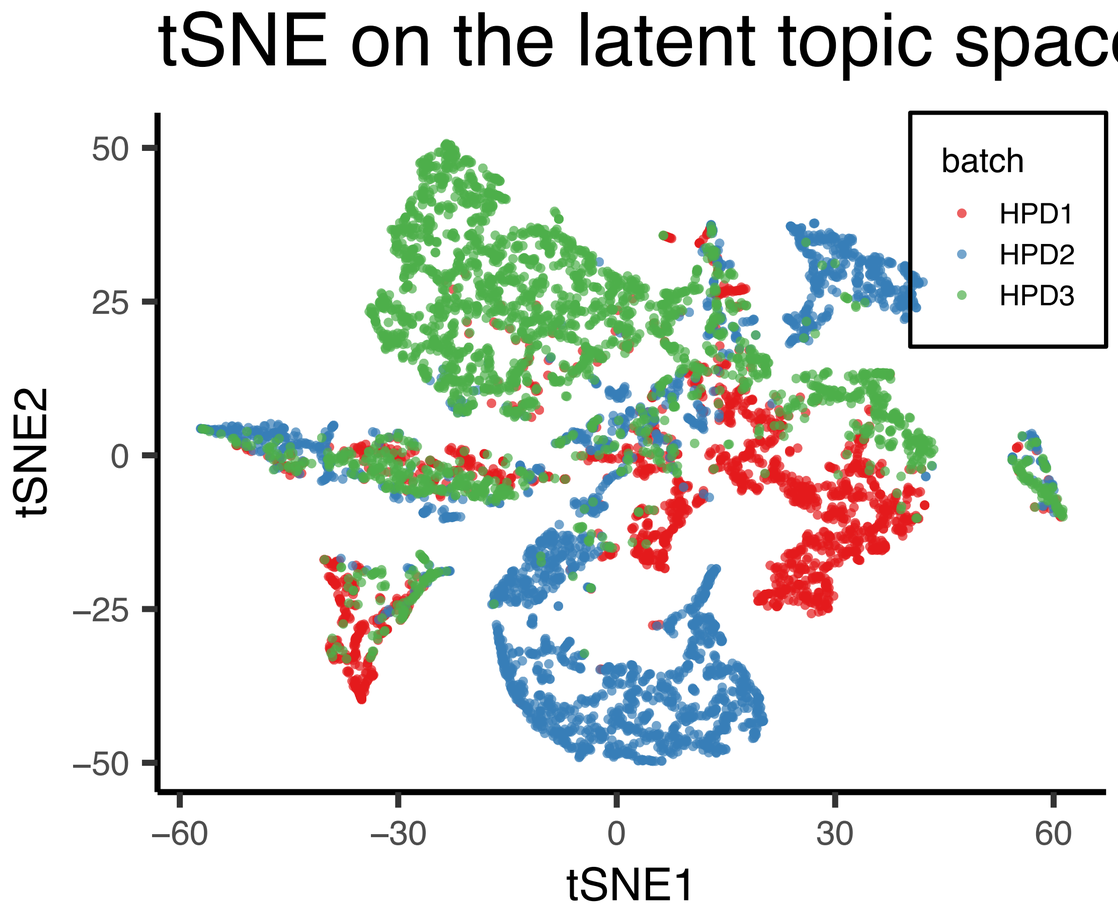

*Goal*: Topic space for 6,873 cells shared across multiple donors (batches).

Qadir, *et al.*, *PNAS* (2020)

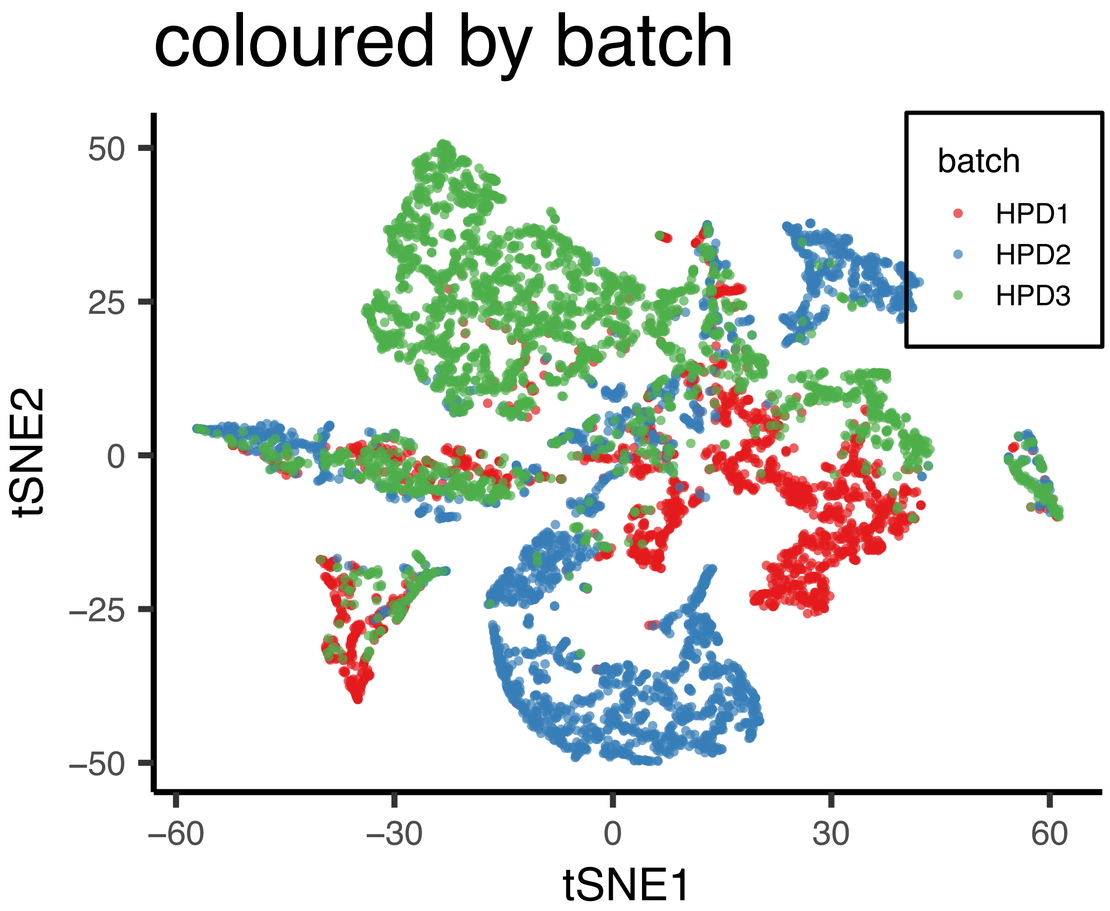

latent topic
1
2
3
4
5
6
7
8
9
10
11
12
1.00
0.75
0.50
0.25
0.00
topic proportion
cells

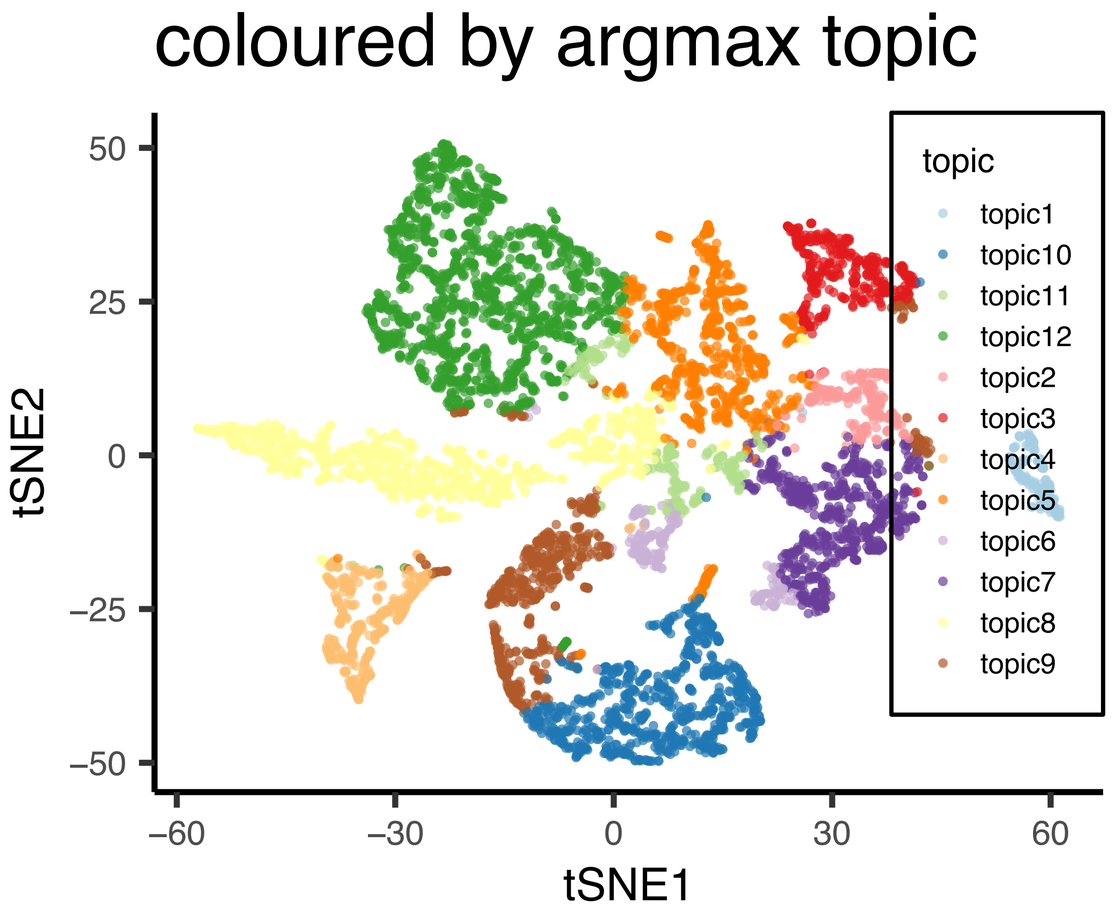
# Multiple batches mingle well in latent topic space!



# Multiple batches mingle well in latent topic space!

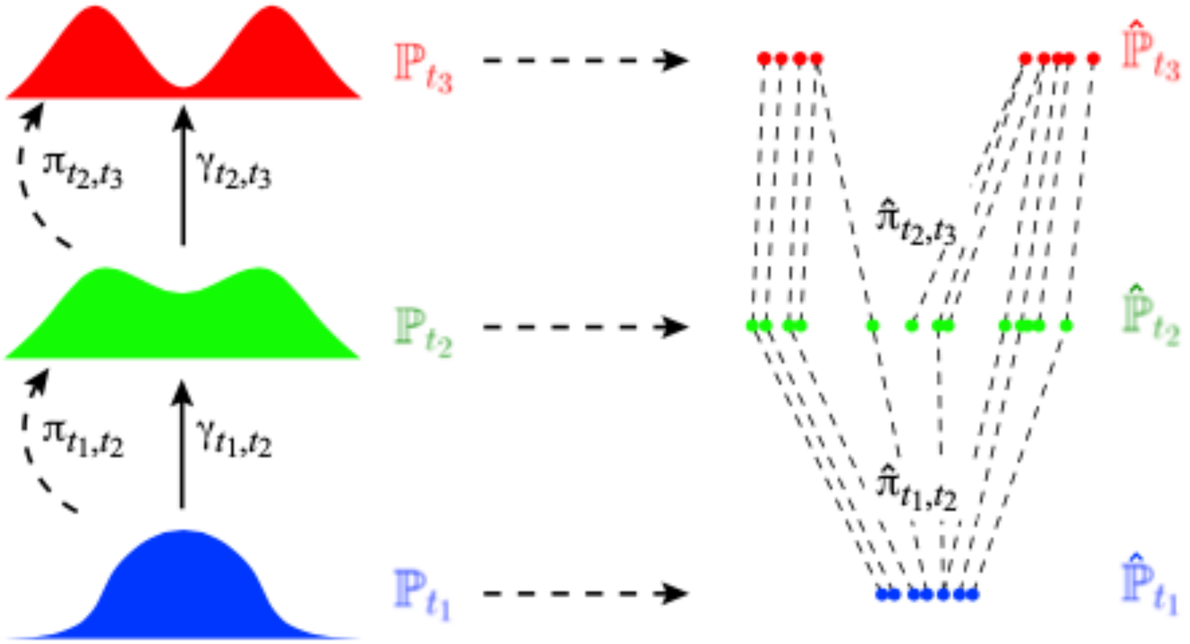


# Today’s lecture: Single-cell Part 2

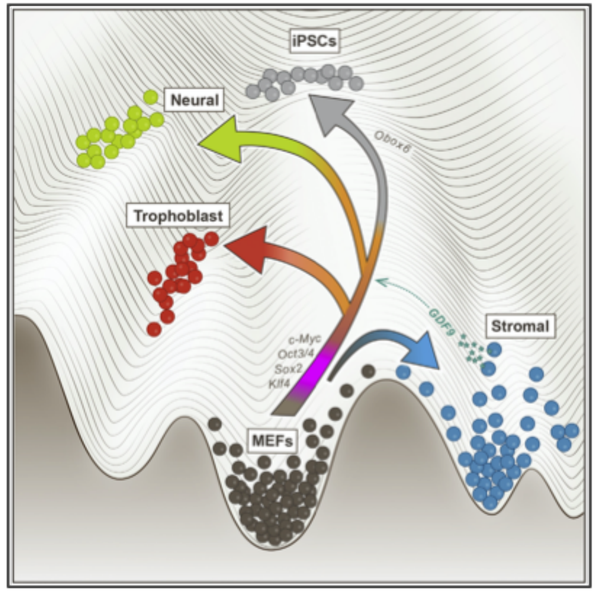
- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

# Optimal Transport to interrogate developmental trajectories from time-series scRNA-seq data




**Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming**

Graphical Abstract




Authors

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

Correspondence

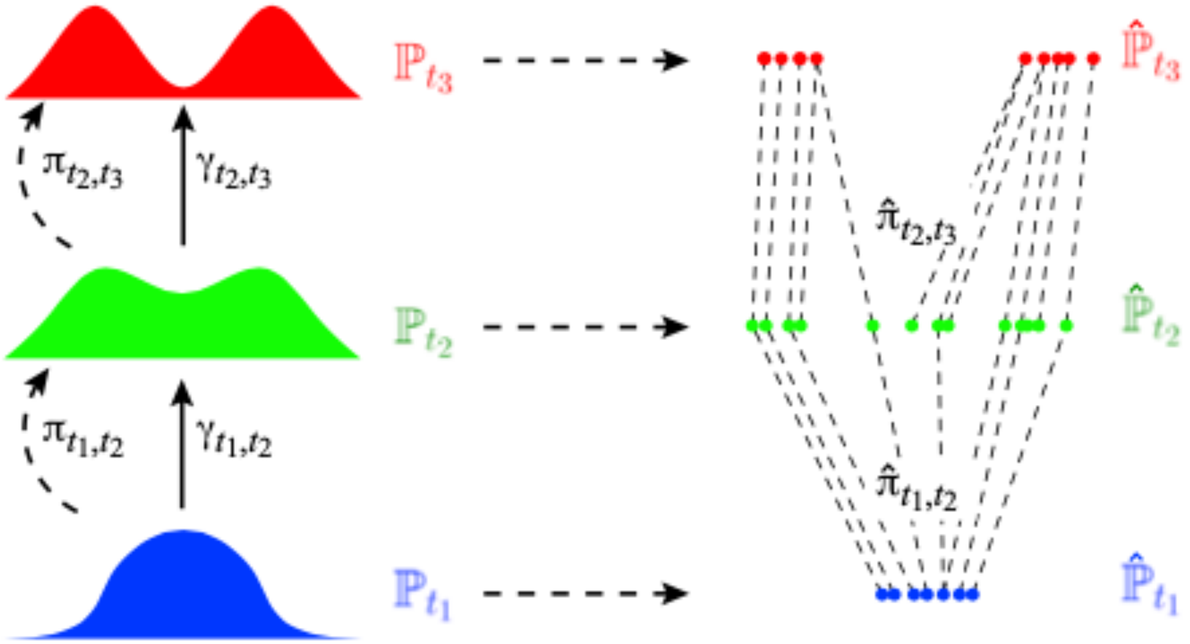
jianshu@broadinstitute.org (J.S.), aregev@broadinstitute.org (A.R.), lander@broadinstitute.org (E.S.L.)

In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.

Schiebinger *et al.* Cell (2019)

# Optimal Transport to couple cells between adjacent time points




**Input** (Coupling t1 and t2):

- Cost function (distance matrix) cells in t1 (x) cells in t2 (y)

  C(x,y)

- Marginal (empirical) distributions over G genes:

  Q(x) and P(y) on $\mathbb{R}^G$

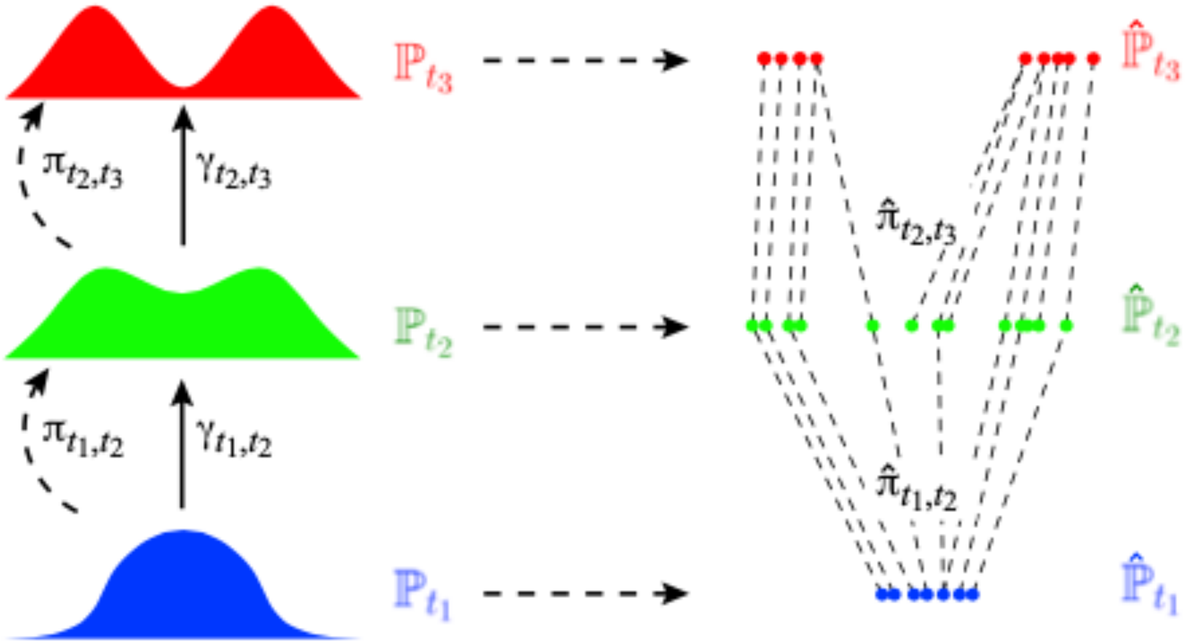
**Output**:

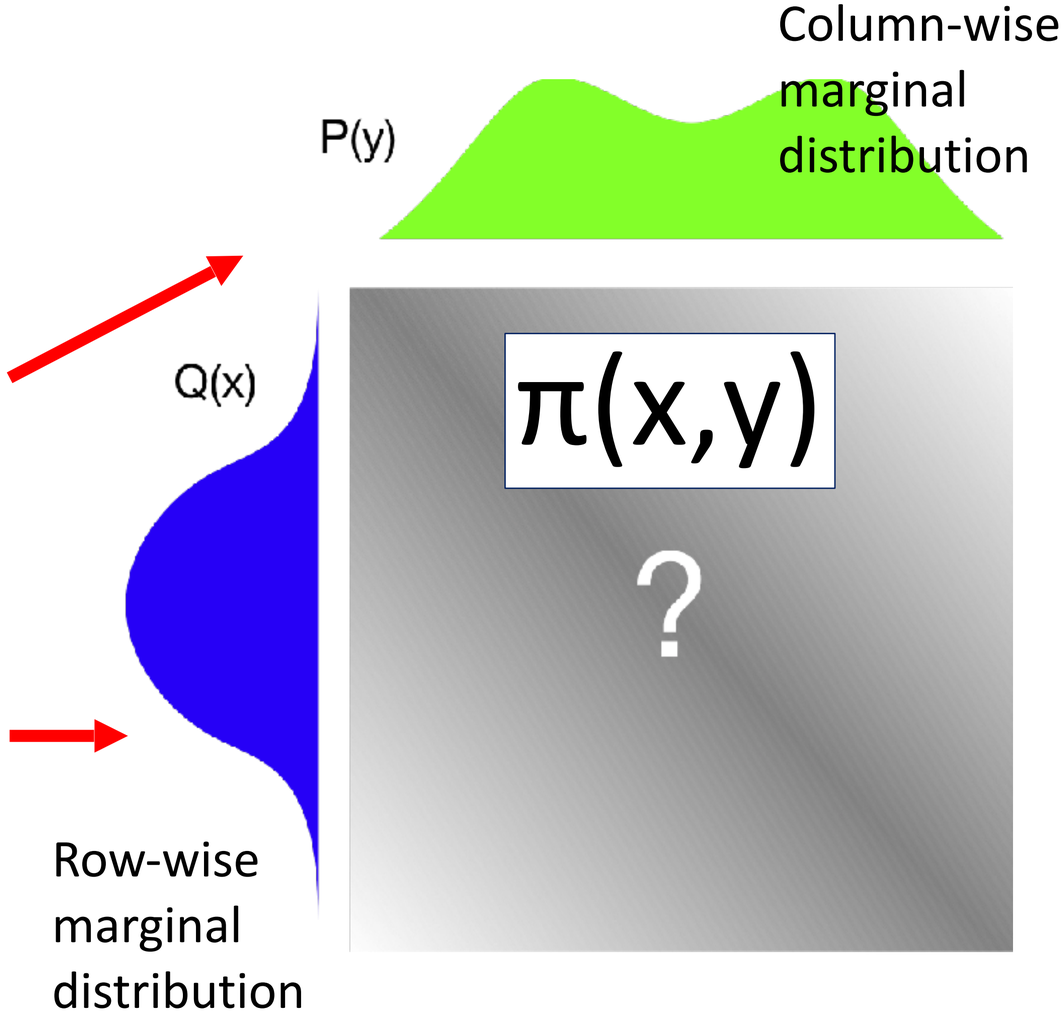
- Joint probability

  π(x,y)

Schiebinger *et al.* Cell (2019)

# OT: Two marginal probs → Joint probability




Assume: biology is optimal

# Solving OT by constrained optimization

**Objective function**

$$\min_{\pi} \sum_{x}\sum_{y} C(x,y)\pi(x,y) - \epsilon H(\pi)$$

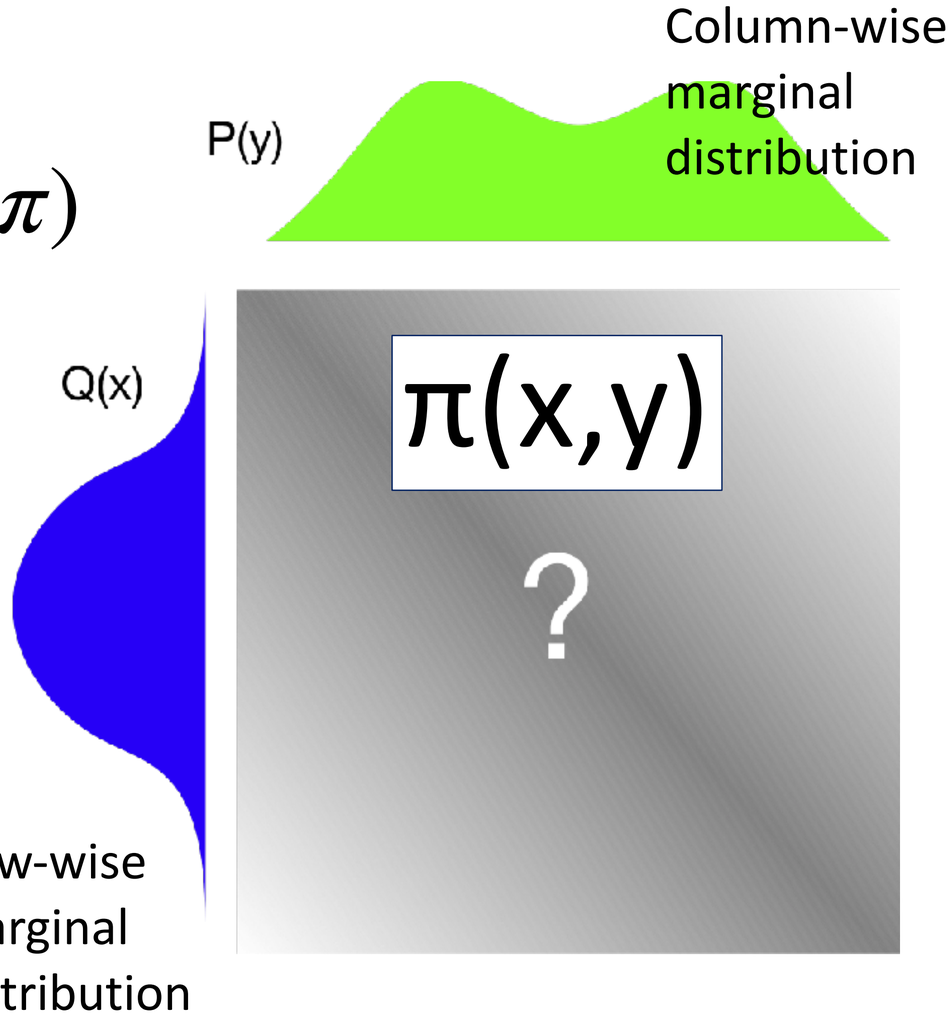
Entropy for smoothness & a unique soln.

**Constraints**

$$D_{KL}\left(\sum_{x} \pi(x,y) \,||\, P(y)\right) < s1$$

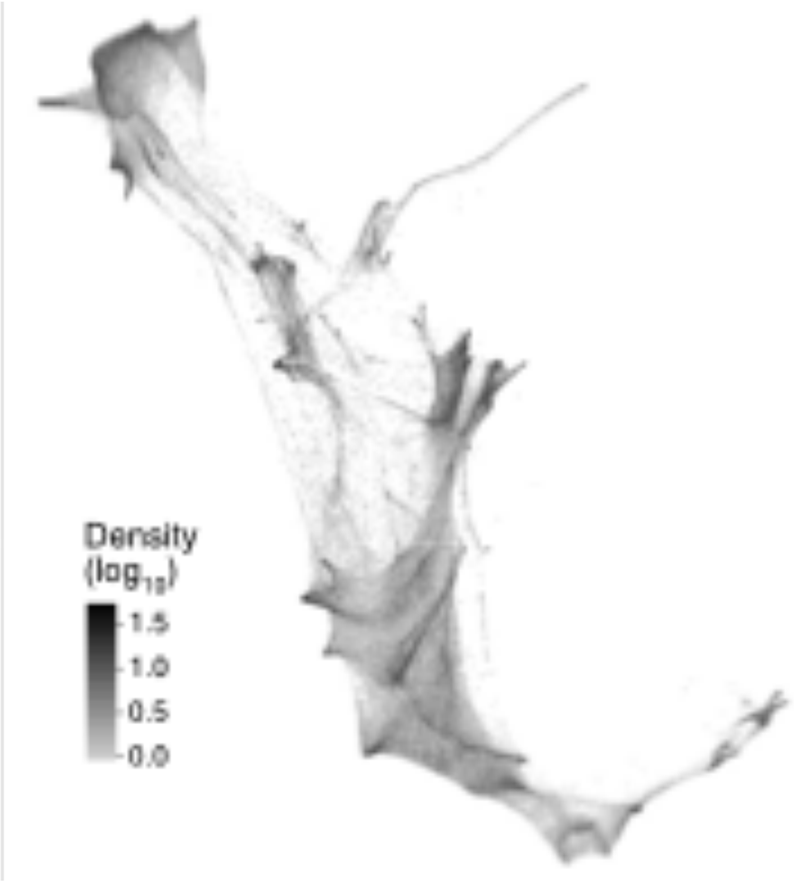
$$D_{KL}\left(\sum_{y} \pi(x,y) \,||\, Q(x)\right) < s2$$

Row and col-wise marginal distrib. should match

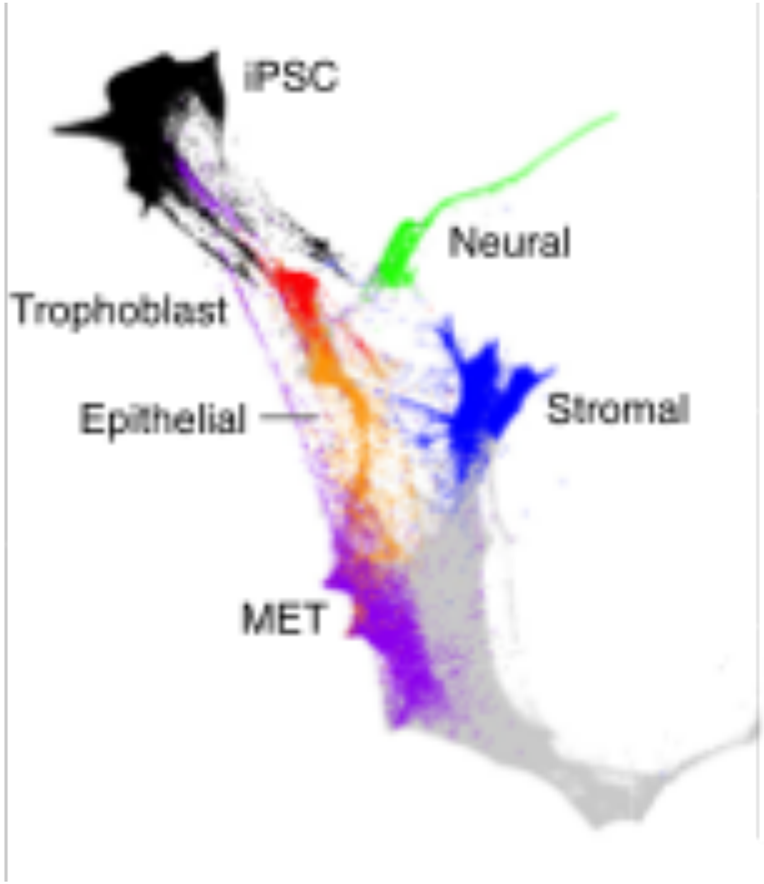


# OT to reconstruct a developmental trajectory

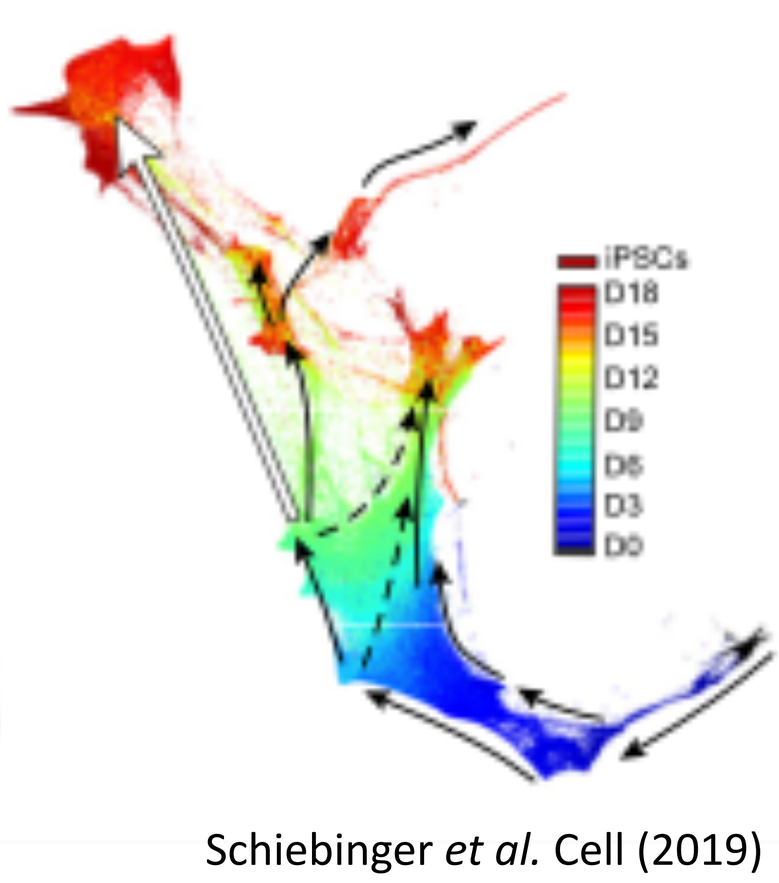
Observed marginal density




Cell type annotation




Trajectory inference



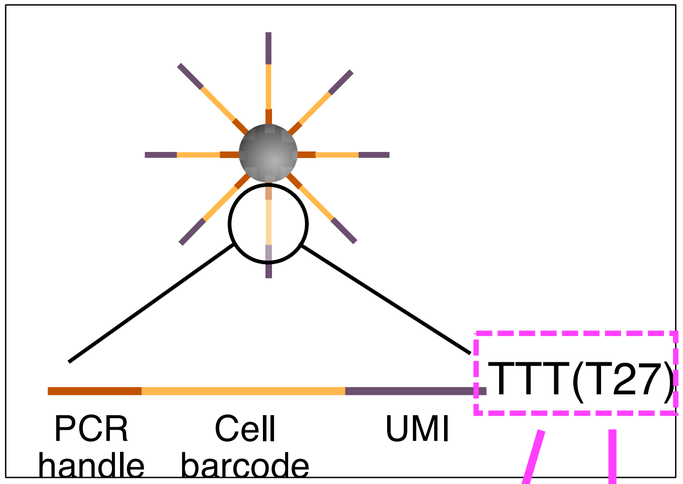

Schiebinger *et al.* Cell (2019)

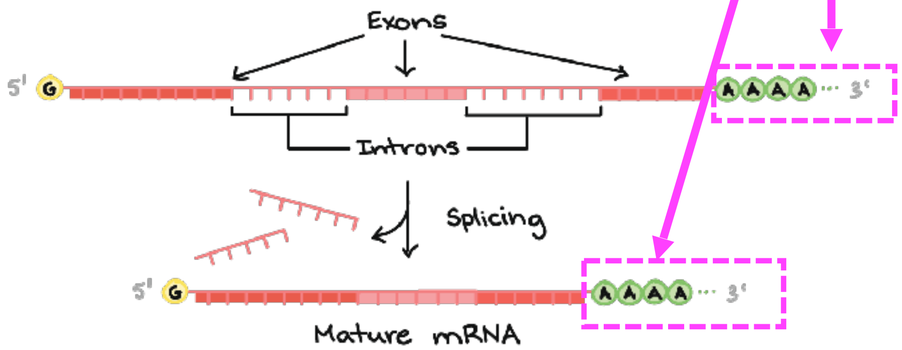
# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

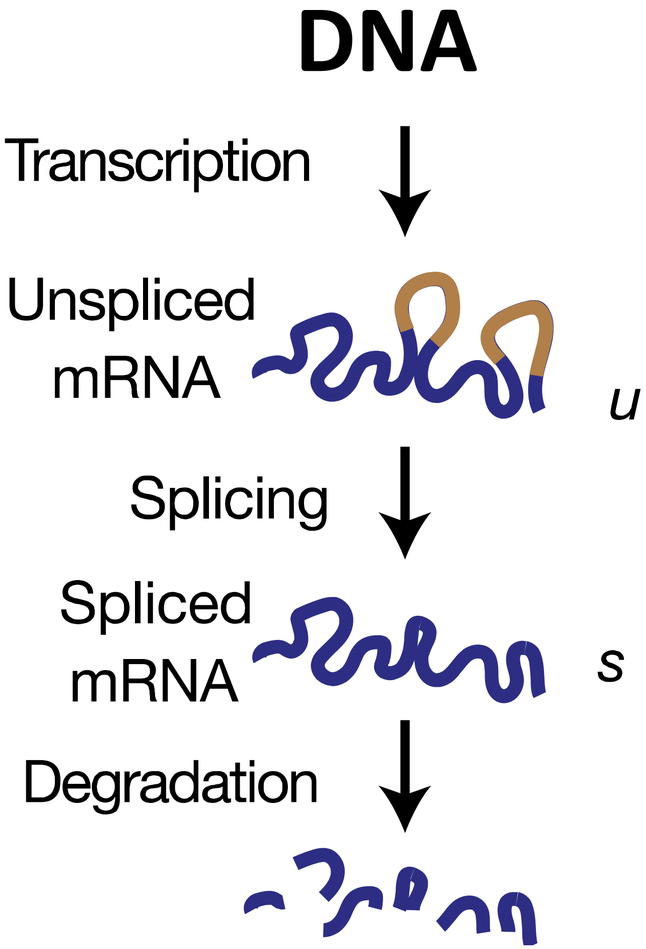
# Capturing the dynamics of gene expression regulation: pre mRNA → mature mRNA

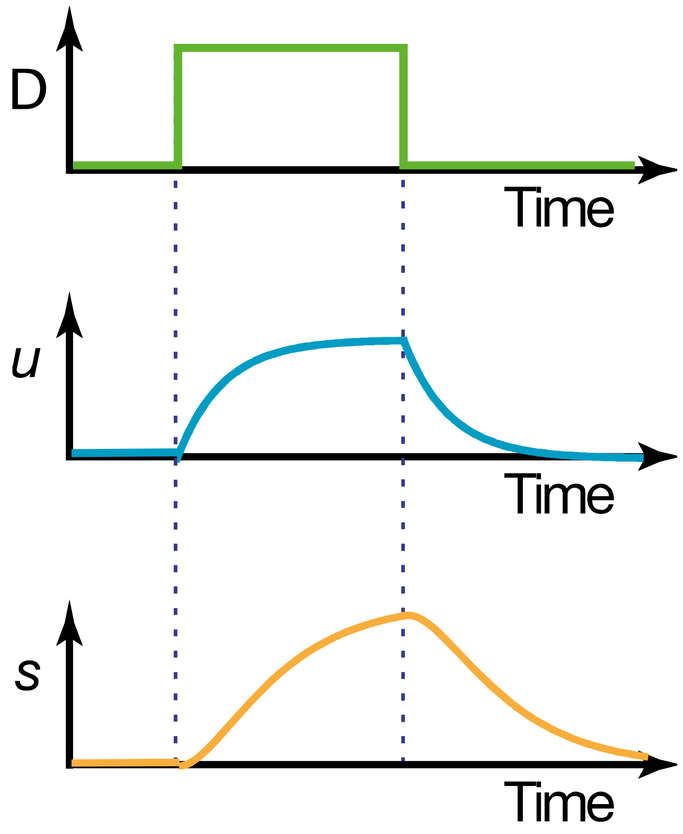
poly-A/T for both U & S



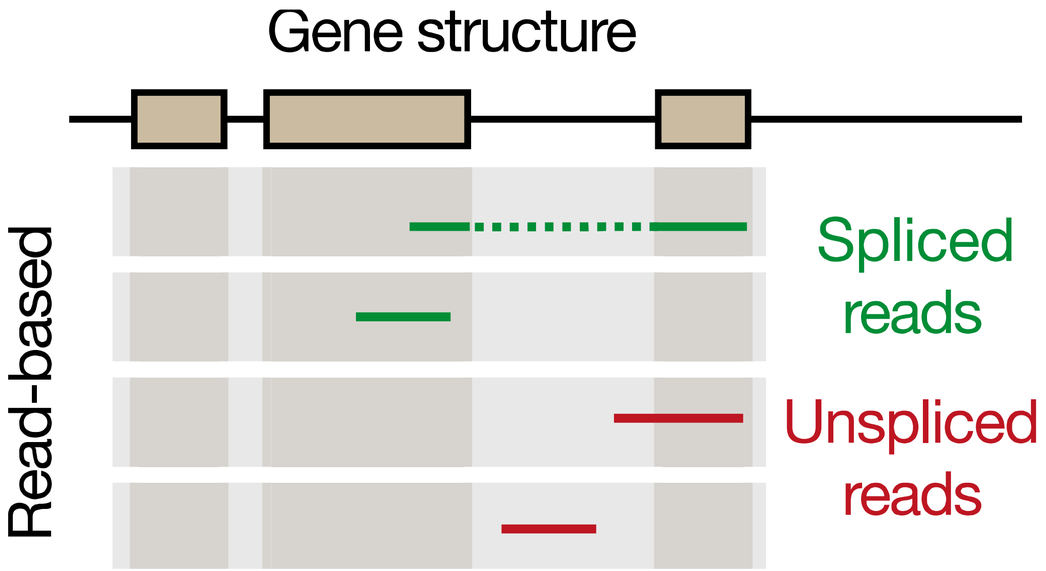

La Manno *et al*. Nature (2018)



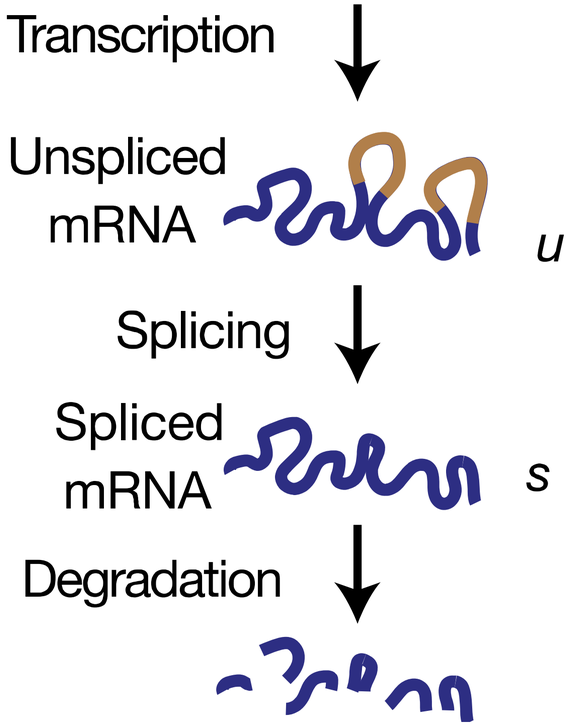

Goal: mRNA dynamics (ODEs)

# Estimate the state of each gene in each cell by counting the number of spliced/unspliced



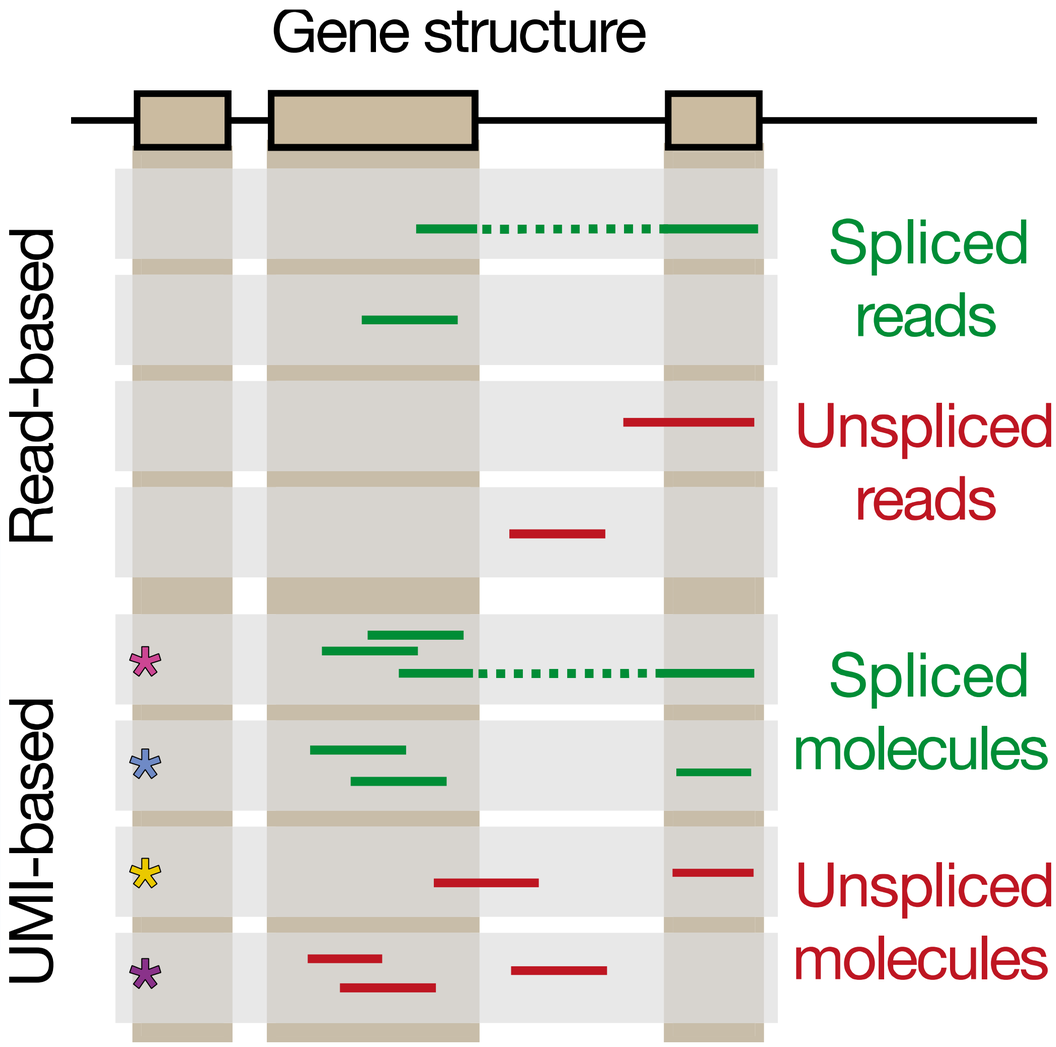

**Unspliced:** short reads landed on introns (gap) and the boundary between exon and intron.

**Spliced:** short reads, not including any intronic regions

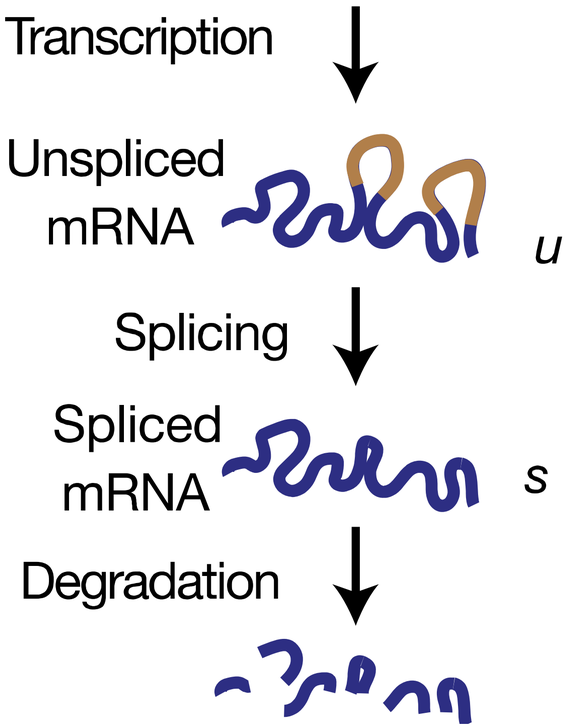



La Manno *et al*. Nature (2018)

# Estimate the state of each gene in each cell by counting the number of spliced/unspliced

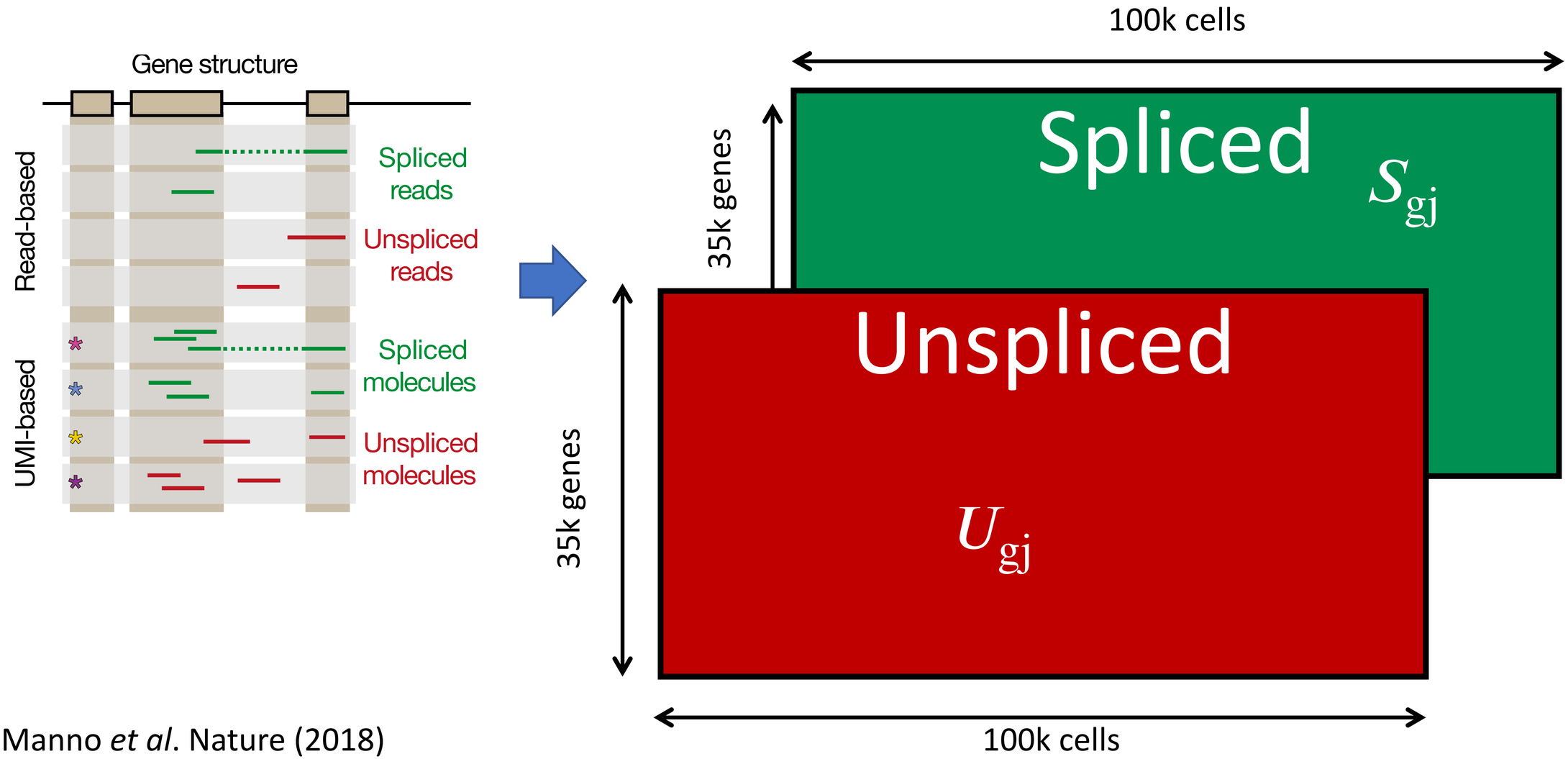



With UMI, we count the number of unique molecules (we could have multiple reads per UMI)




La Manno *et al*. Nature (2018)

# Two count matrices = spliced + unspliced




La Manno *et al*. Nature (2018)

# Ordinary Differential Equations capture the dynamics of splicing events

$$\frac{dU}{dt} = \alpha - \beta U(t)$$

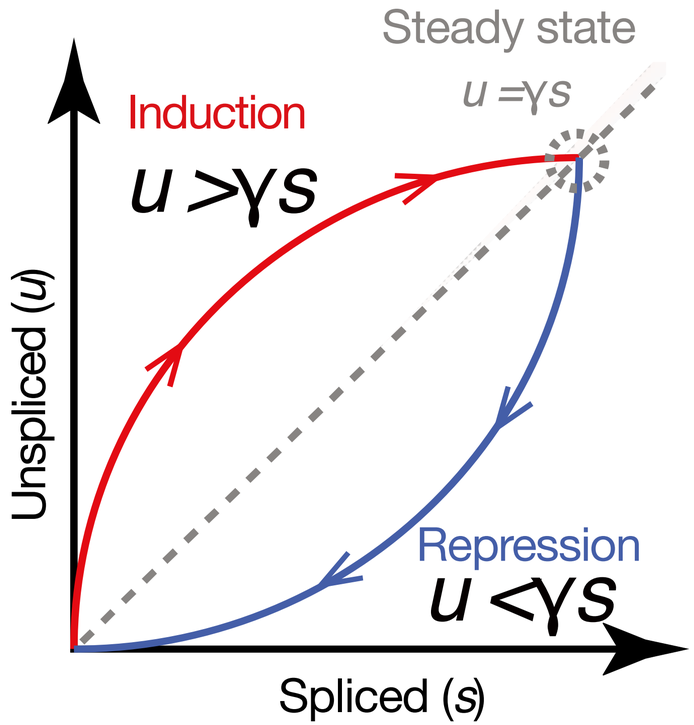
splicing rate

transcription initiation rate

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate

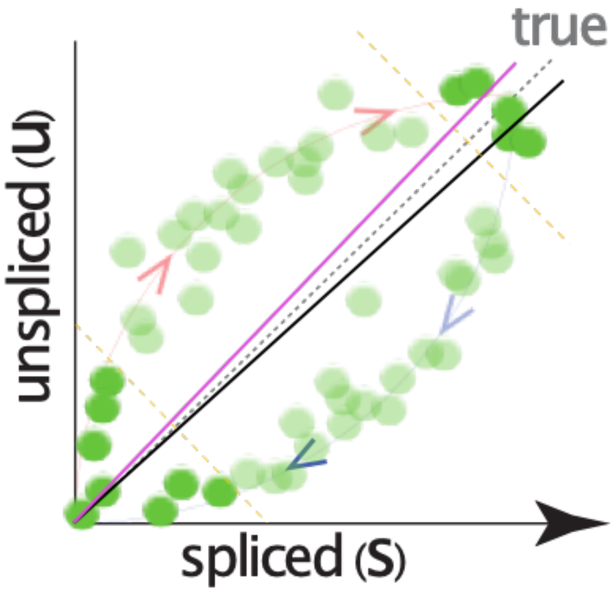
mRNA degradation




La Manno *et al*. Nature (2018)

# How do we estimate the model parameters?

If scRNA-seq profiled all the steady state for this gene (each point = cell)




A quick solution can be derived by least-square estimates

$$u_{gj} \sim s_{gj}\,\gamma_g$$

$$\nabla x_{gj} \leftarrow u_{gj} - \hat{\gamma}_g s_{gj}$$

$$x_{gj}(\Delta t) \leftarrow x_{gj} + \nabla x_{gj}\Delta t$$

La Manno *et al*. Nature (2018)

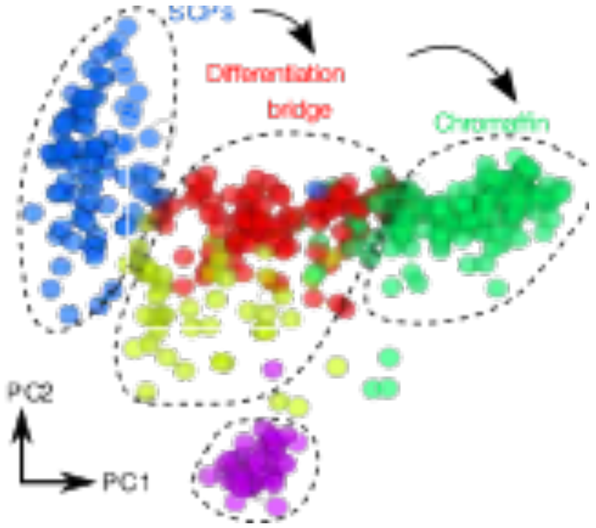
# Velocity → Vector field

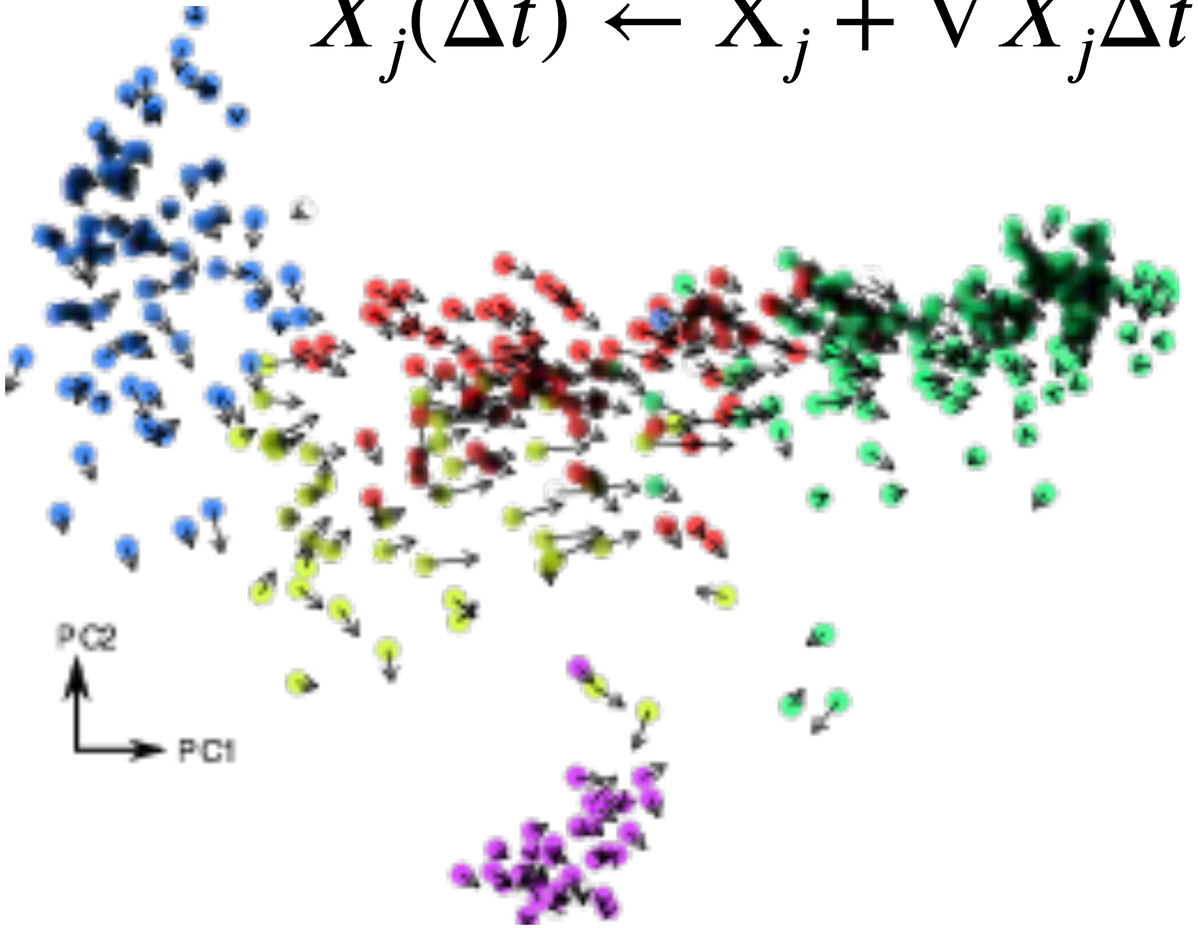
Estimate γ for the aggregate U and S

$$\nabla X_j \leftarrow U_j - \hat{\gamma} S_j$$

$$X_j(\Delta t) \leftarrow X_j + \nabla X_j \Delta t$$

$U_j \sim S_j \gamma$




La Manno *et al*. Nature (2018)

# We can estimate the model parameters in a better way…

transcription initiation rate — **splicing rate**

$$\frac{dU}{dt} = \alpha - \beta U(t)$$

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

**splicing rate** — mRNA degradation

Analytical solution:

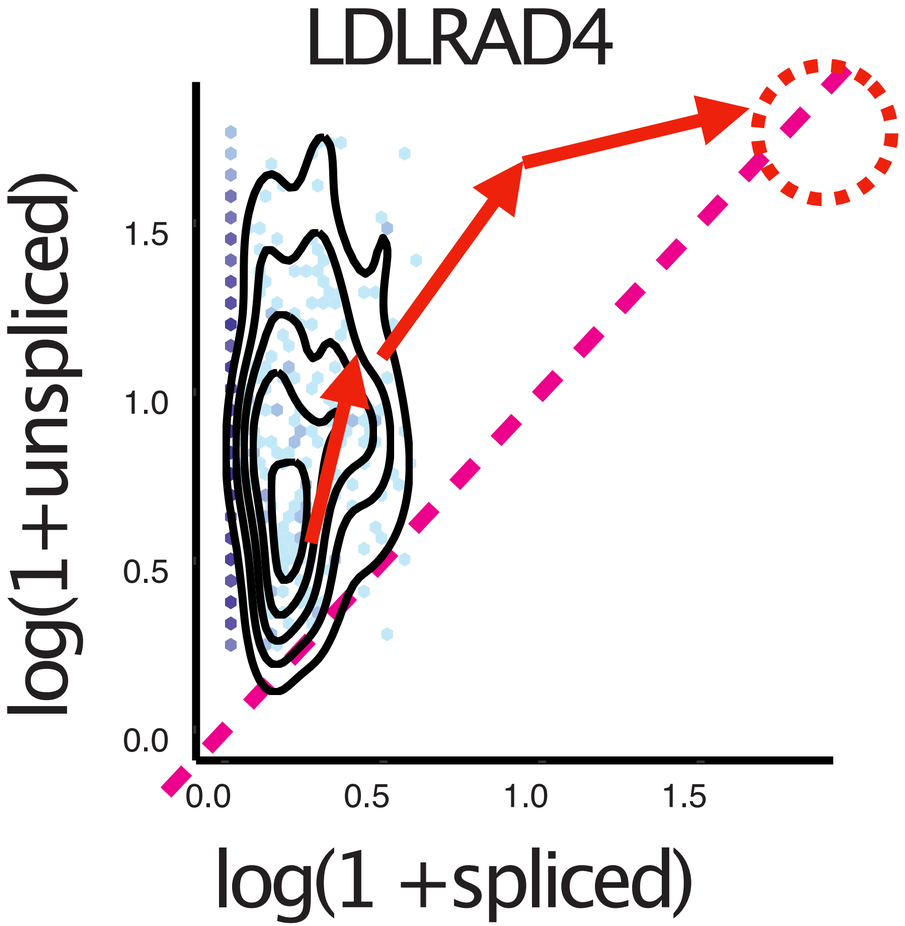
$$U(t) = U_0 e^{-\beta t} + (\alpha/\beta)(1 - e^{-\beta t})$$

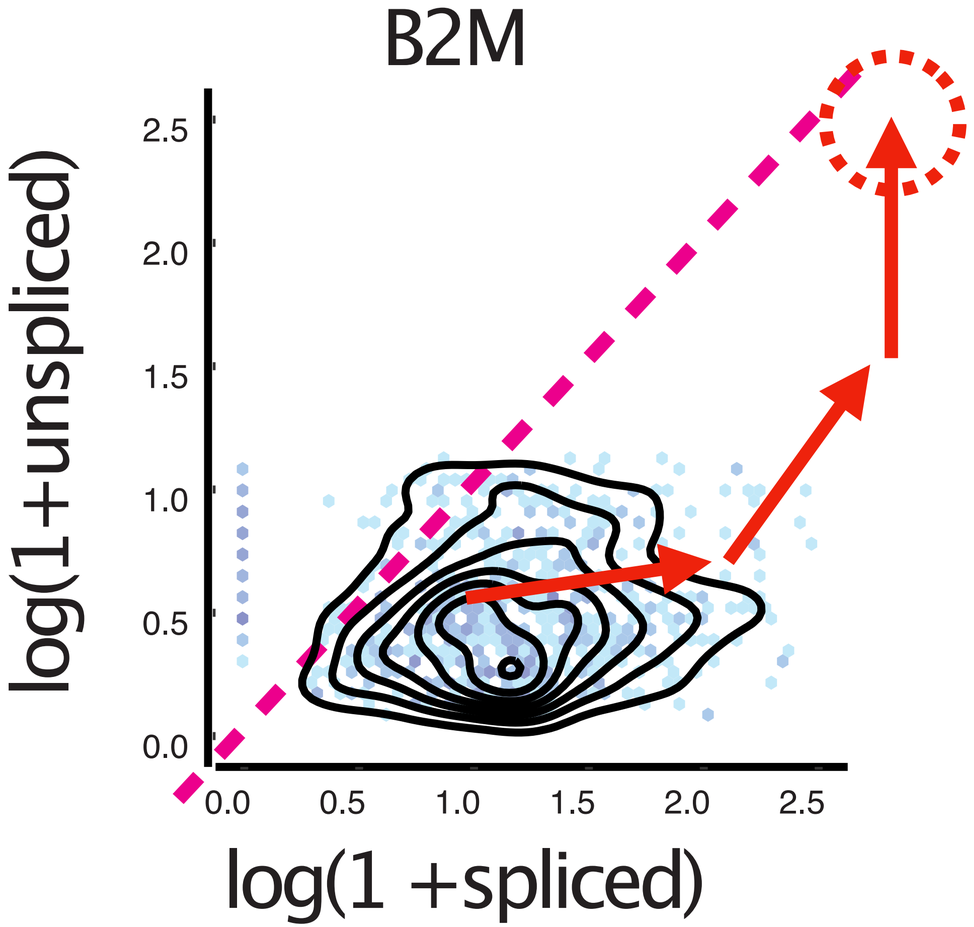
$$S(t) = S_0 e^{-\gamma t} + \frac{\alpha - \beta U_0}{\gamma - \beta}(e^{-\gamma t} - e^{-\beta t})$$

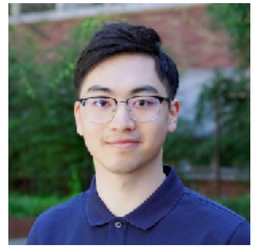
EM-algorithm

- E-step: Estimate latent time for each data point
- M-step: Optimize the ODE parameters

# However, in reality, we don't see full dynamics

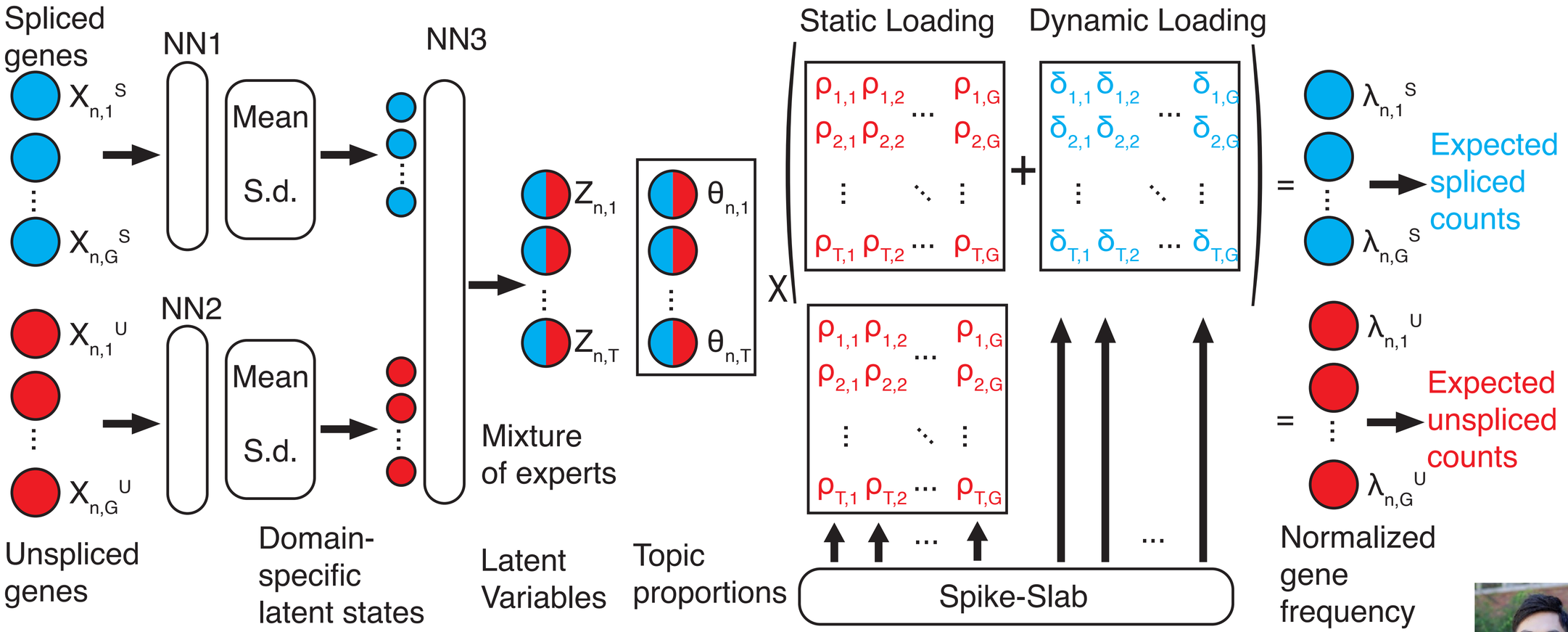


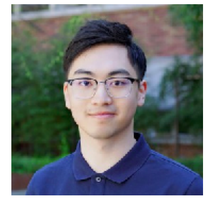



Zhang .. Park, *in preparation*

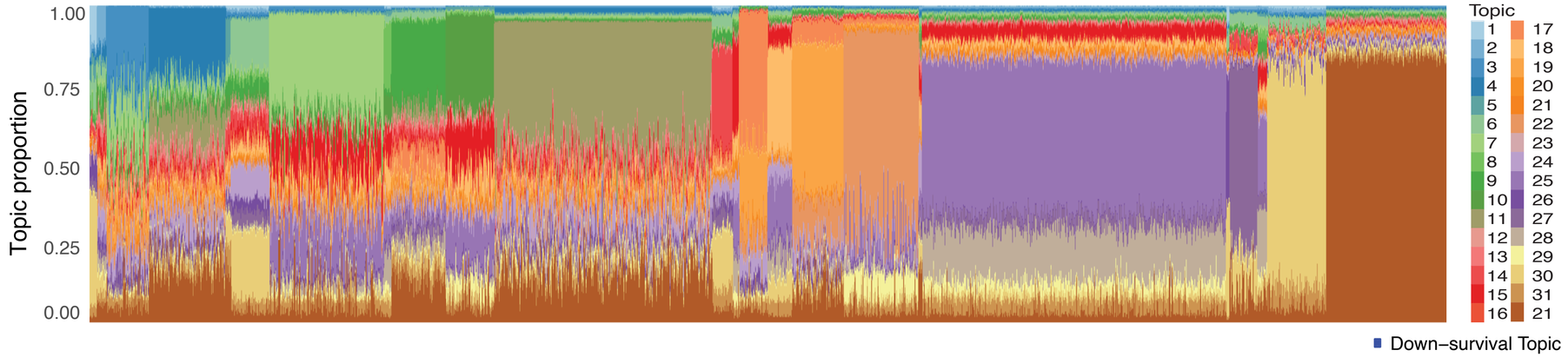
# Δ topic model to capture the difference between the spliced and unspliced counts

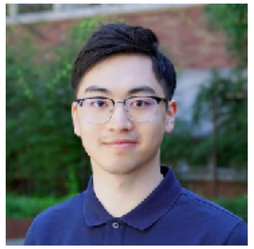


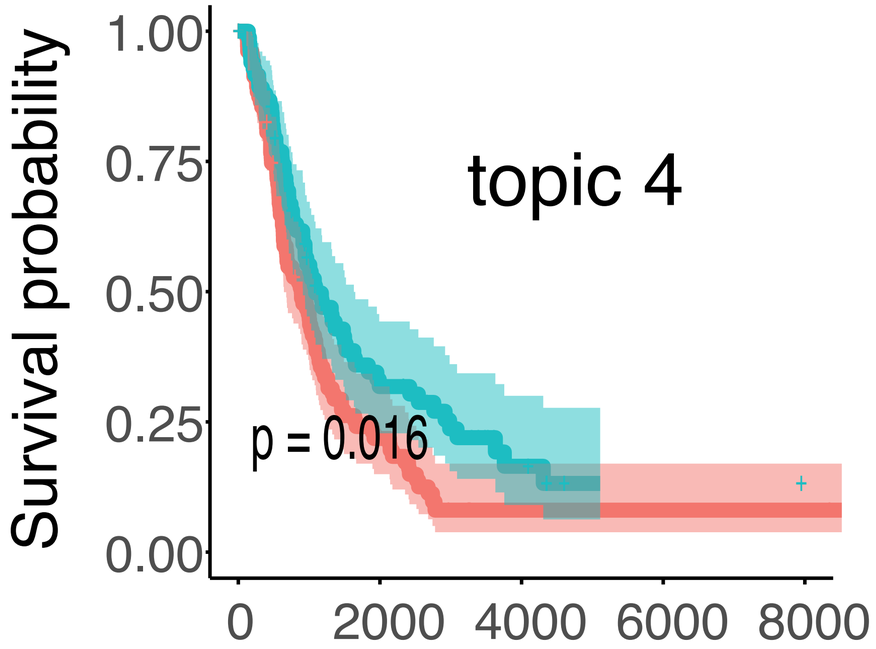



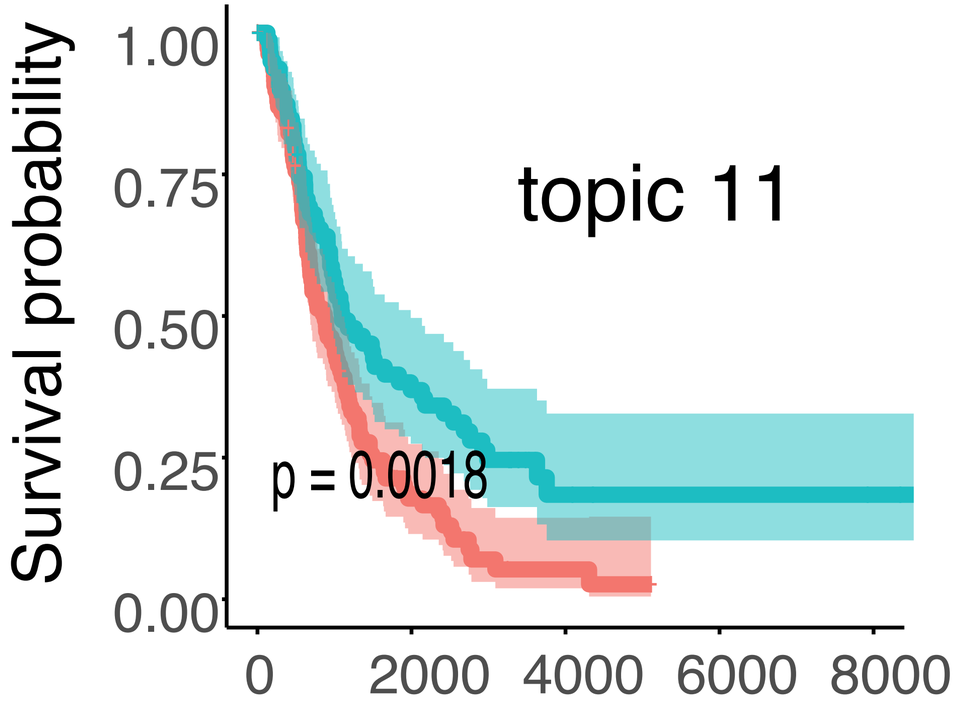
**Zhang .. Park, *in preparation***

# Δtopic captures cancer-progression specific transcriptional dynamics





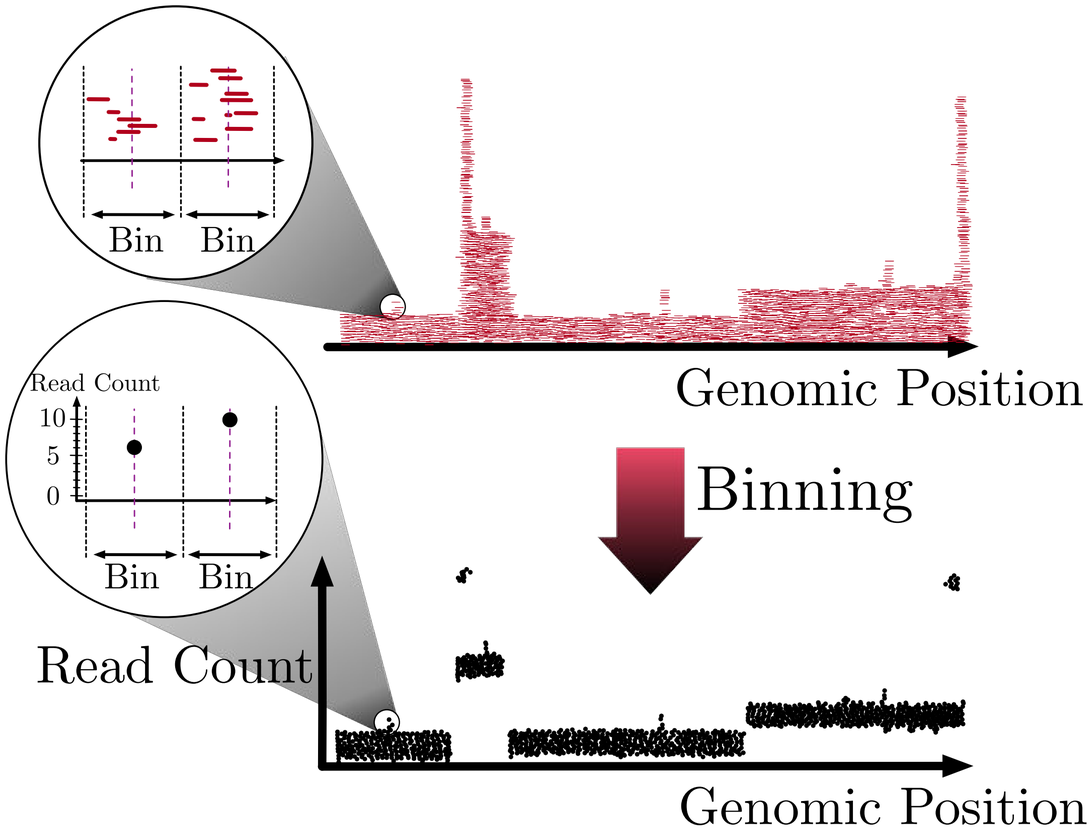



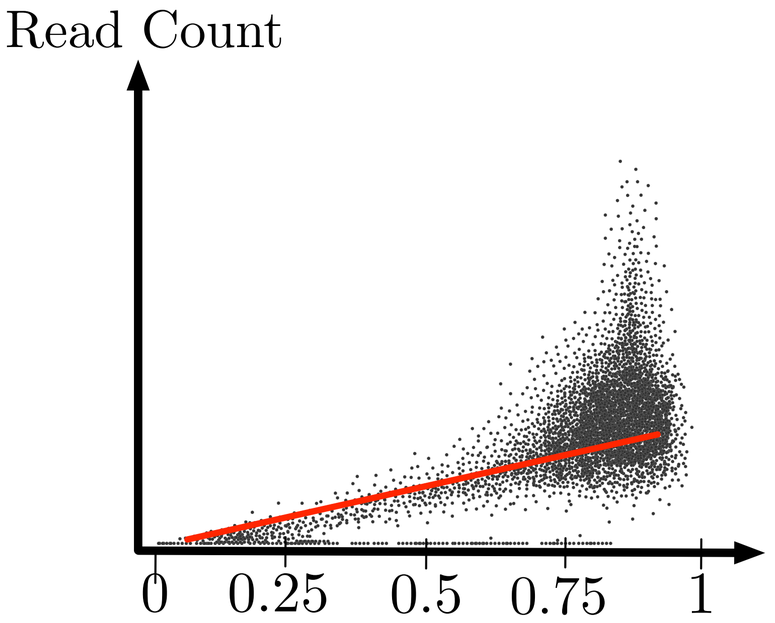

**Zhang .. Park, *in preparation***

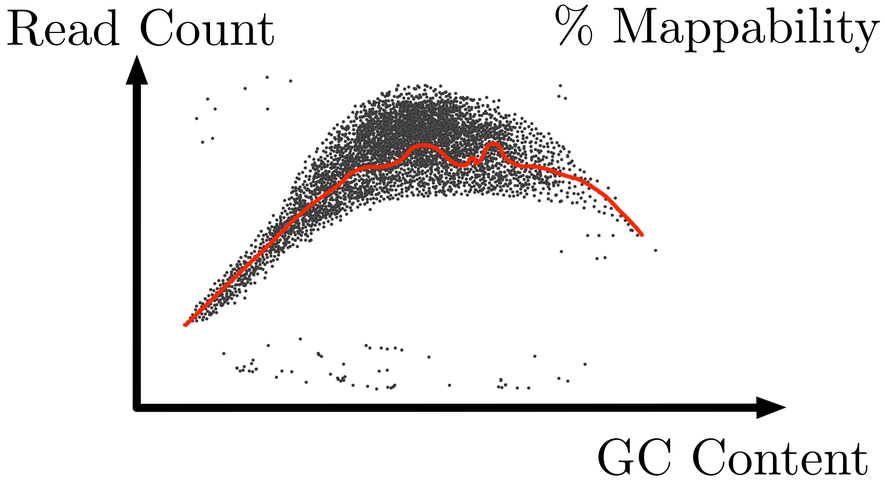
# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

# Copy number variation in single-cell DNA- data

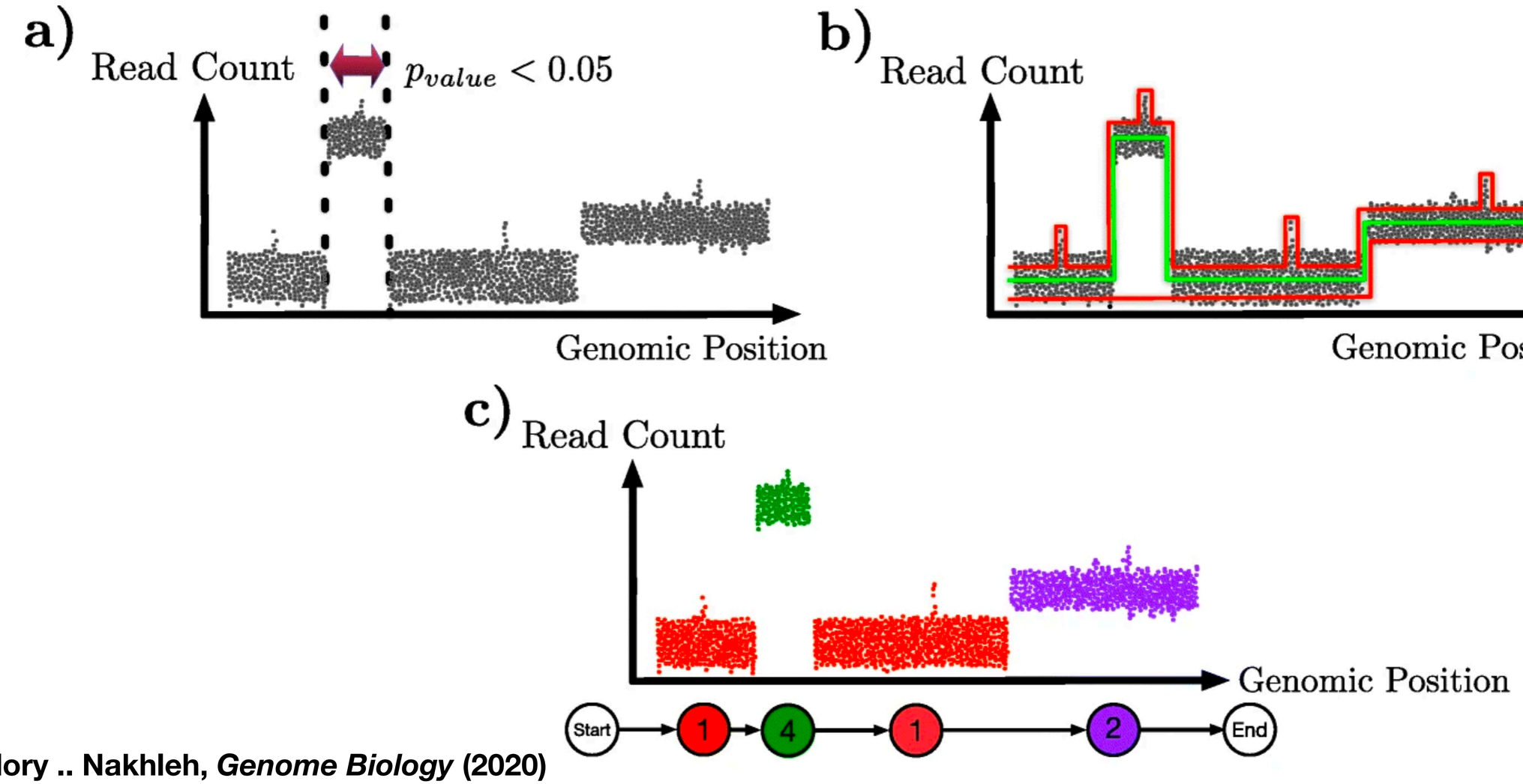



**Mallory .. Nakhleh, *Genome Biology* (2020)**

# Copy number variation in single-cell DNA- data




Mallory .. Nakhleh, *Genome Biology* (2020)

# Copy number profiles implicate clonal lineage

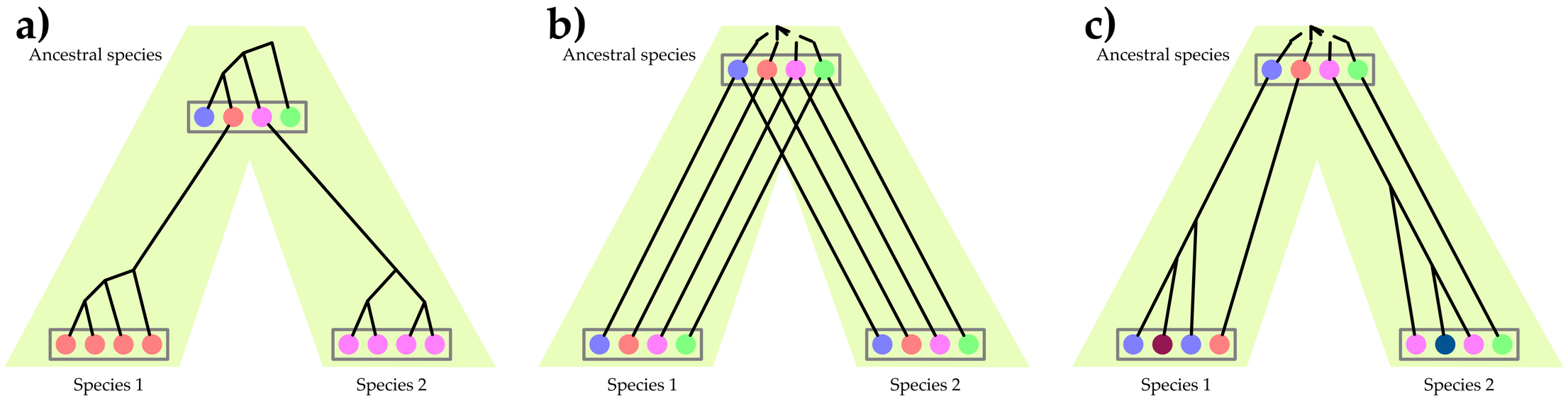
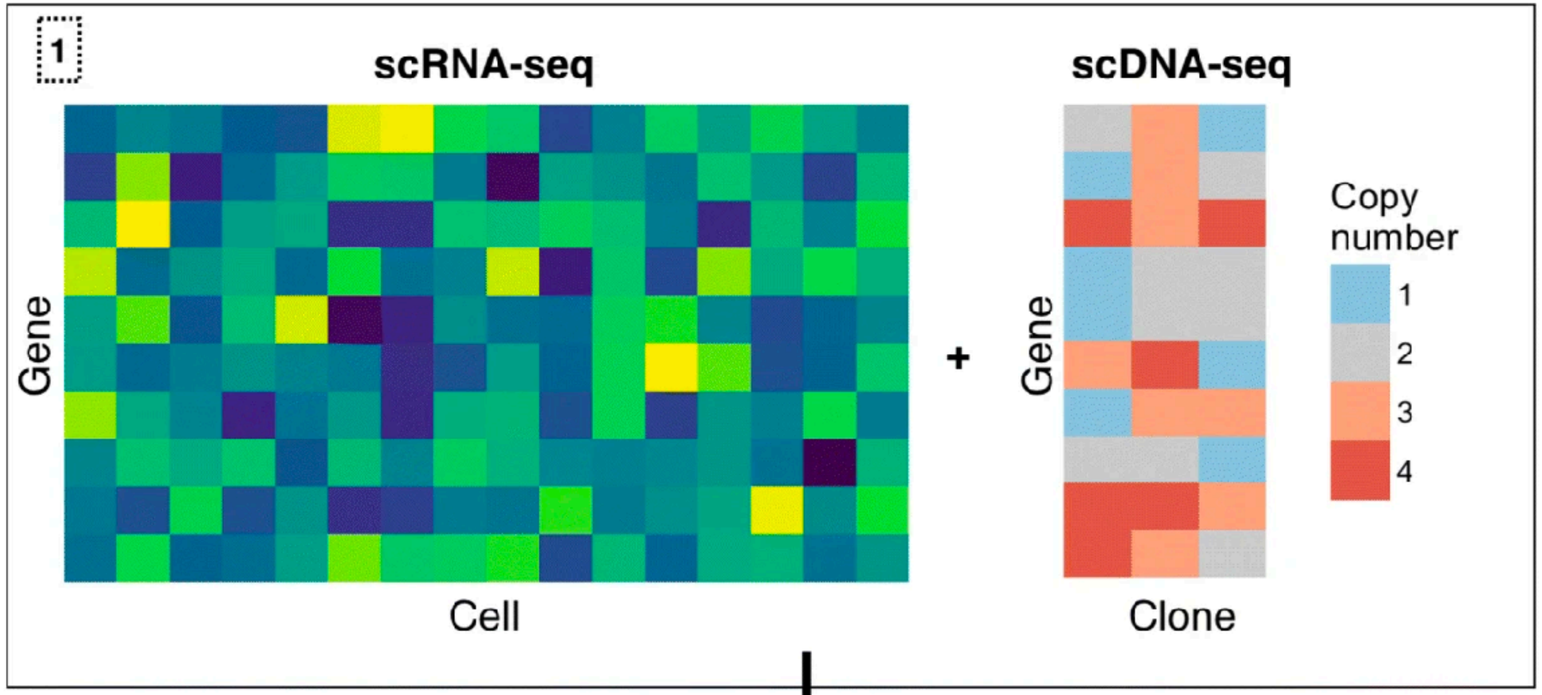
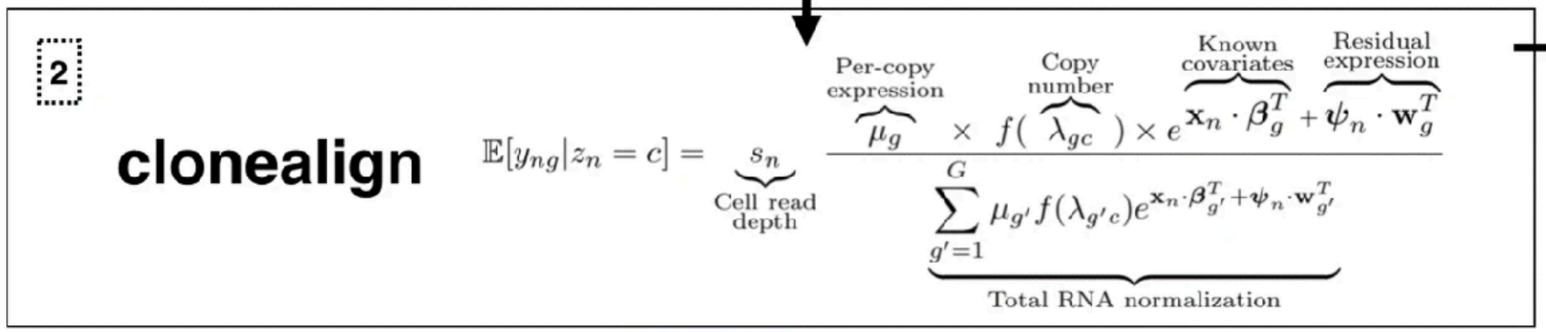


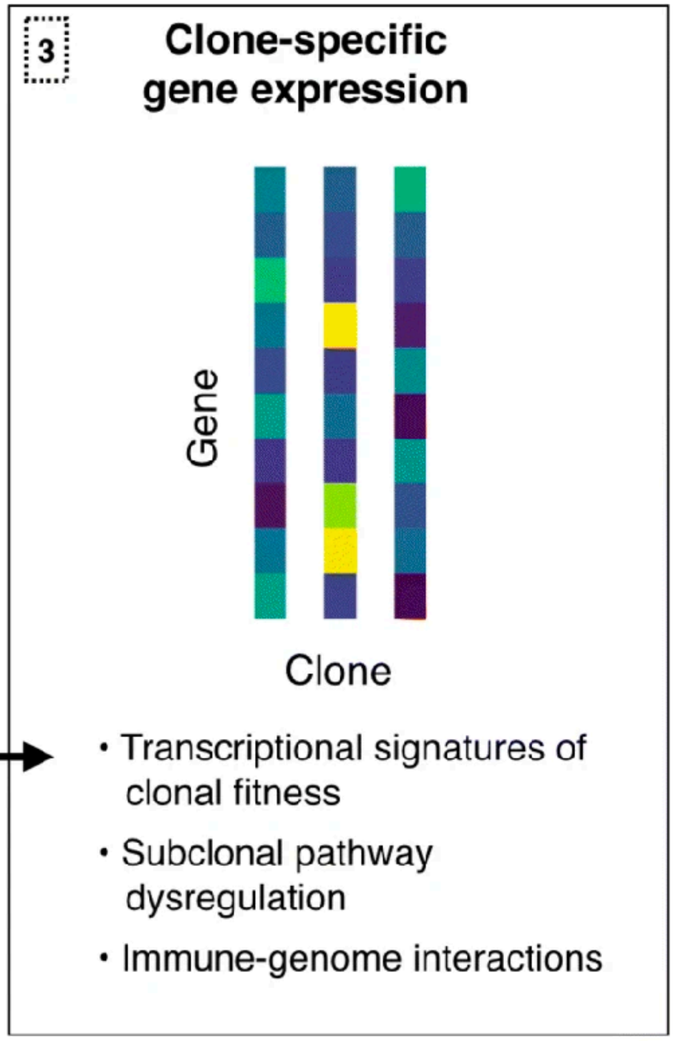

**Fig. 3** Three modes of evolution of multigene families. **a** Concerted evolution. **b** Divergent evolution. **c** Evolution by birth and death process. (Reproduced from [95])

**Mallory .. Nakhleh, *Genome Biology* (2020)**

# How do we match cells across scRNA- and scDNA-seq data?

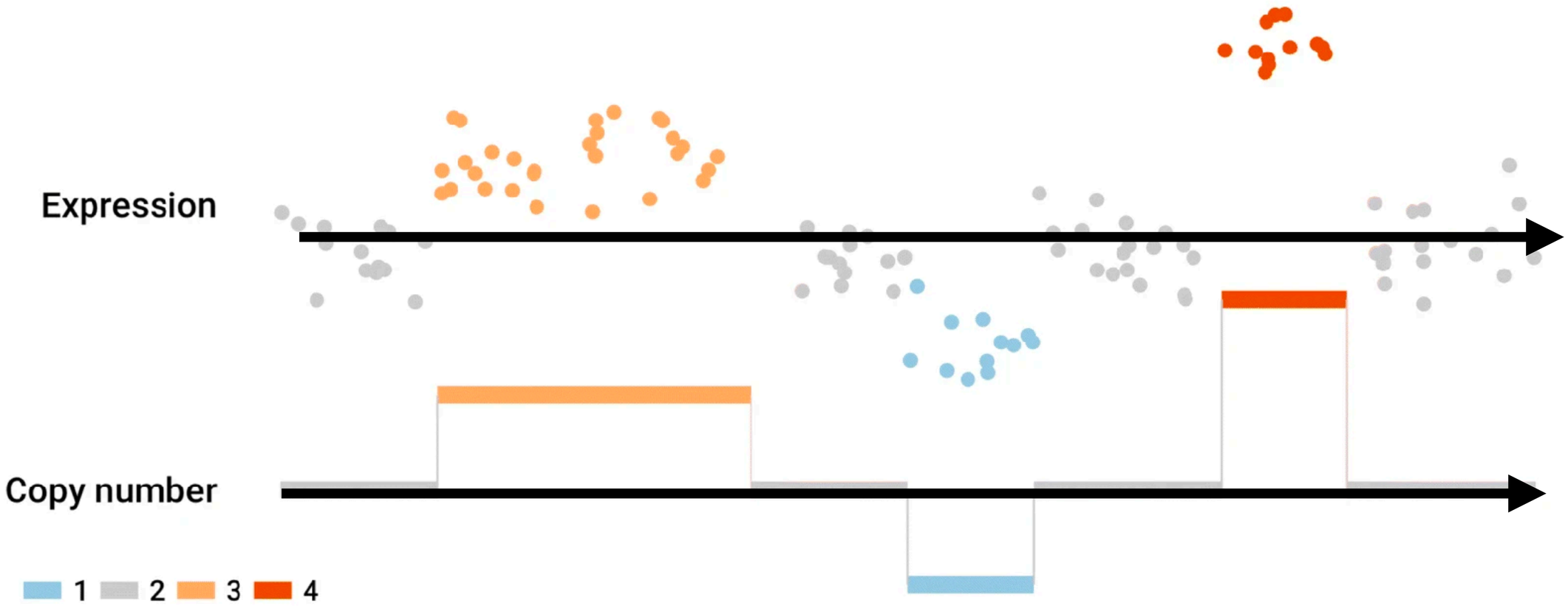


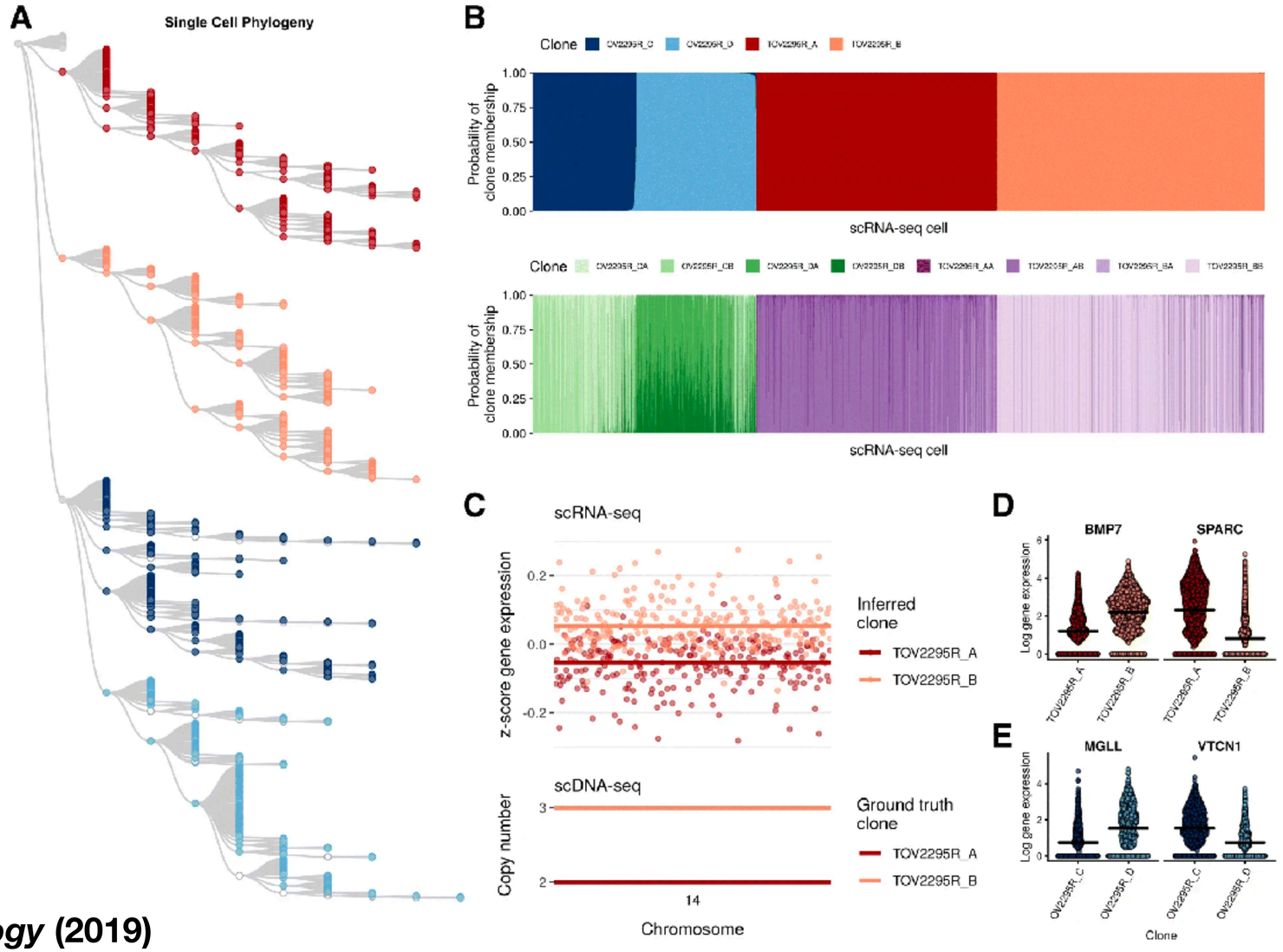

**Campbell .. Shah, *Genome Biology* (2019)**

# Idea: expression high ~ copy number high



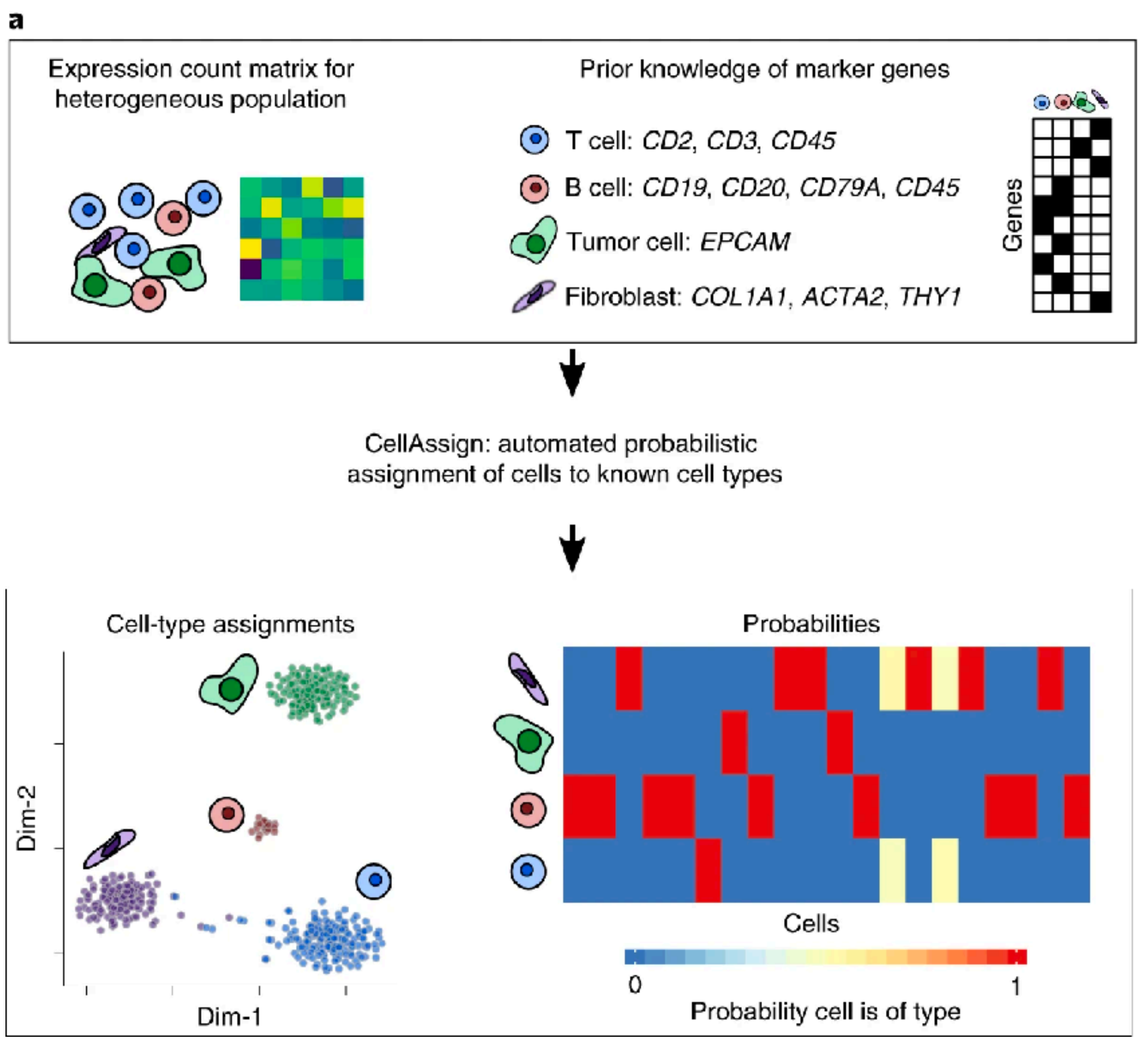

Campbell .. Shah, *Genome Biology* (2019)

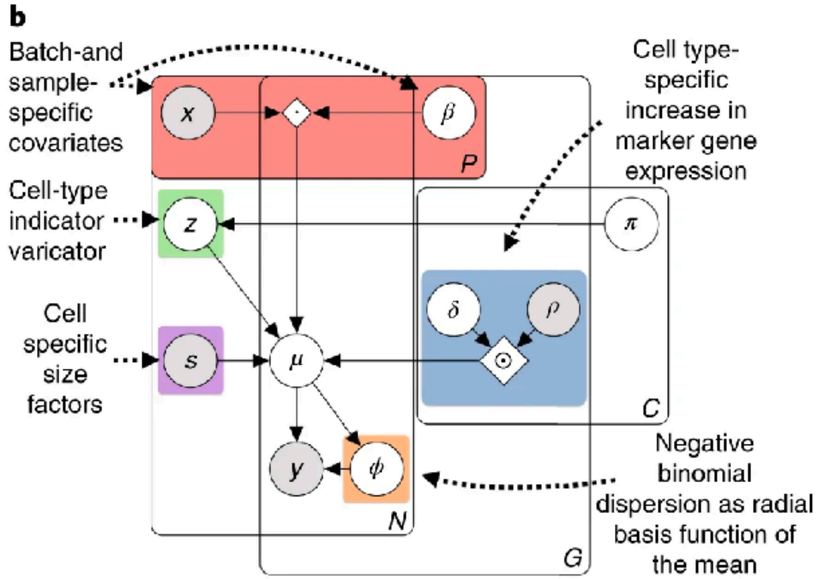
# Assign clonal information to each cell




Campbell .. Shah, *Genome Biology* (2019)

# Using marker gene annotations, we can also assign cell type identity without expensive modelling



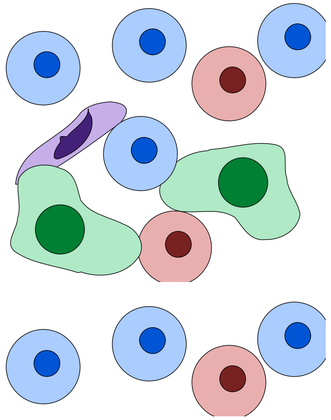

c

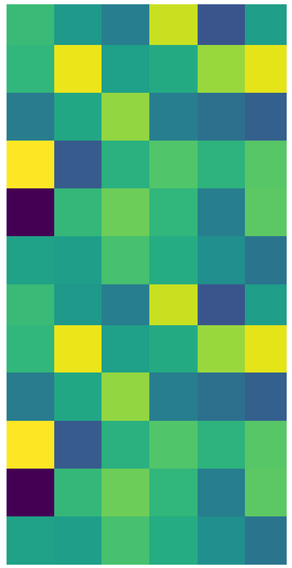
| Variable | Distribution | Description |
|---|---|---|
| $y_{ng}$ | Negative bionomial | Single-cell count |
| $s_n$ | None | Cell size factor |
| $z_n$ | Categorical | Cell type indicator |
| $\mu_{ngc}$ | Deterministic f$^n$ | Modeled average expression |
| $\phi_{ngc}$ | Deterministic f$^n$ | Negative binomial dispersion |
| $\delta_{gc}$ | log-normal | Marker overexpression |
| $\rho_{gc}$ | None | Marker/cell type matrix |
| $x_{np}$ | None | Covariates (batch or sample) |
| $\beta_{pg}$ | Gaussian | Covariate coefficients |
| $a, b$ | None | Dispersion basis coefficients |
| $\pi_c$ | Dirichlet | Prior probability of cell type |

Zhang .. Shah, *Nature Methods* (2019)

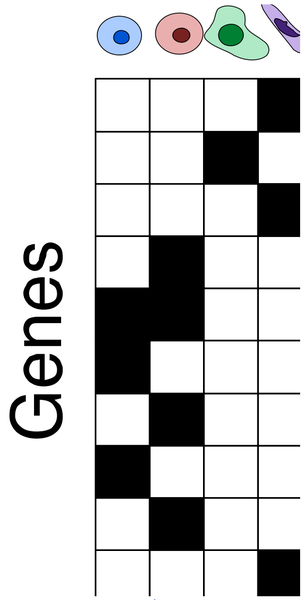
# Direct cell type annotation:
# If our goal is to annotate a cell type to each cell

single-cell RNA-seq data matrix
(with distinctive cell types)





Marker gene to known cell type
membership matrix (0/1)




T cell: *CD2*, *CD3*, *CD45*

B cell: *CD19*, *CD20*, *CD79A*, *CD45*

Tumor cell: *EPCAM*

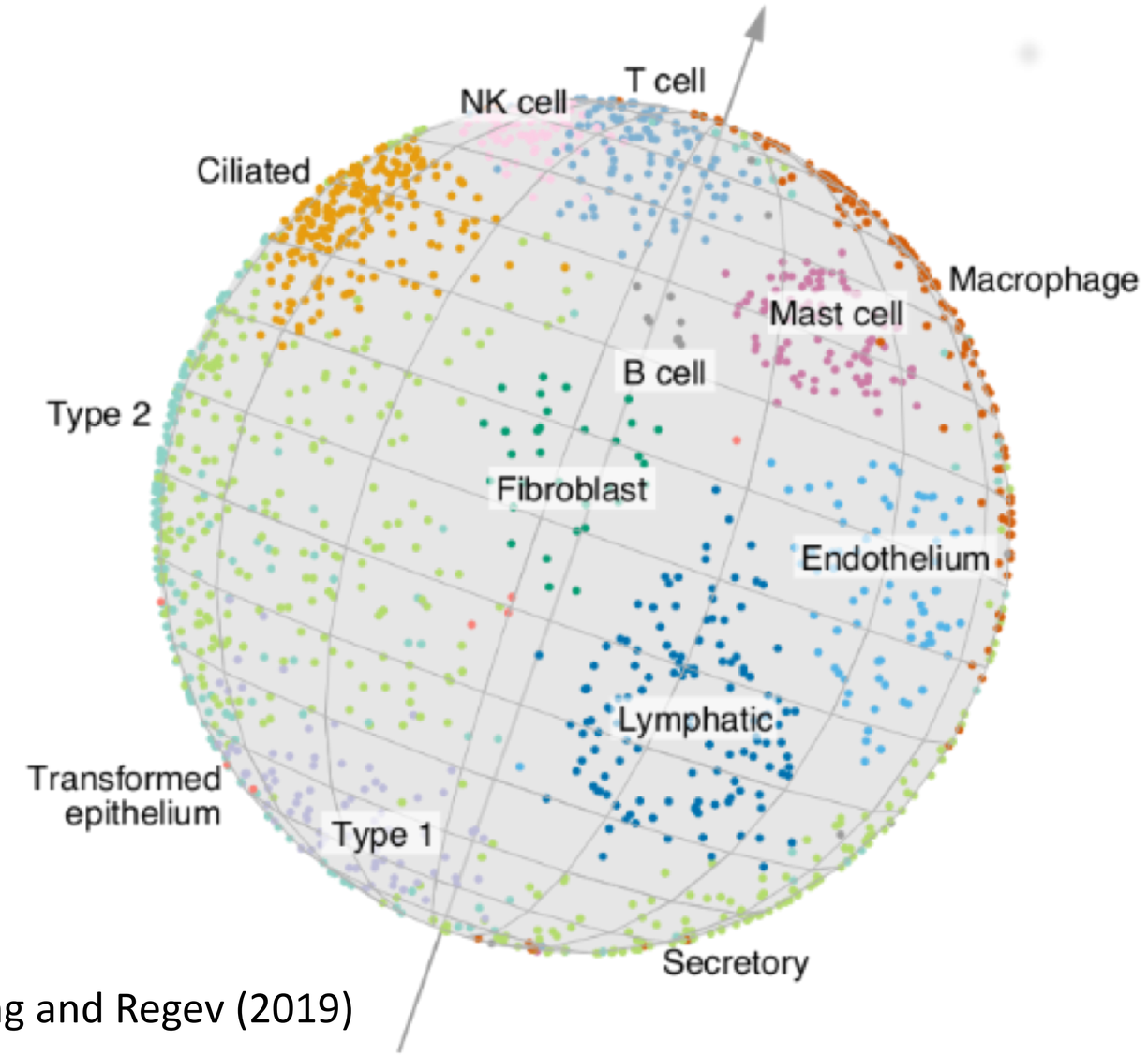
Fibroblast: *COL1A1*, *ACTA2*, *THY1*

Zhang *et al.*
Nature Methods (2019)

Check column-wise similarity
on the limited set of genes/features

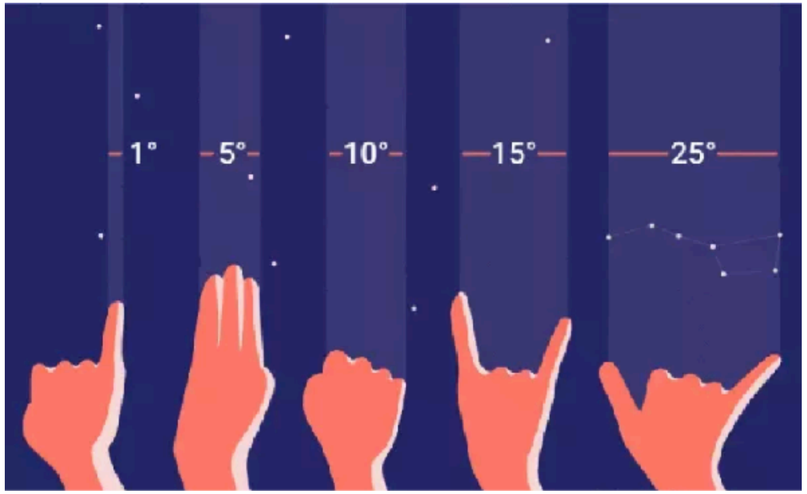
What's your robust metric?

# Angular distance and von Mises-Fisher



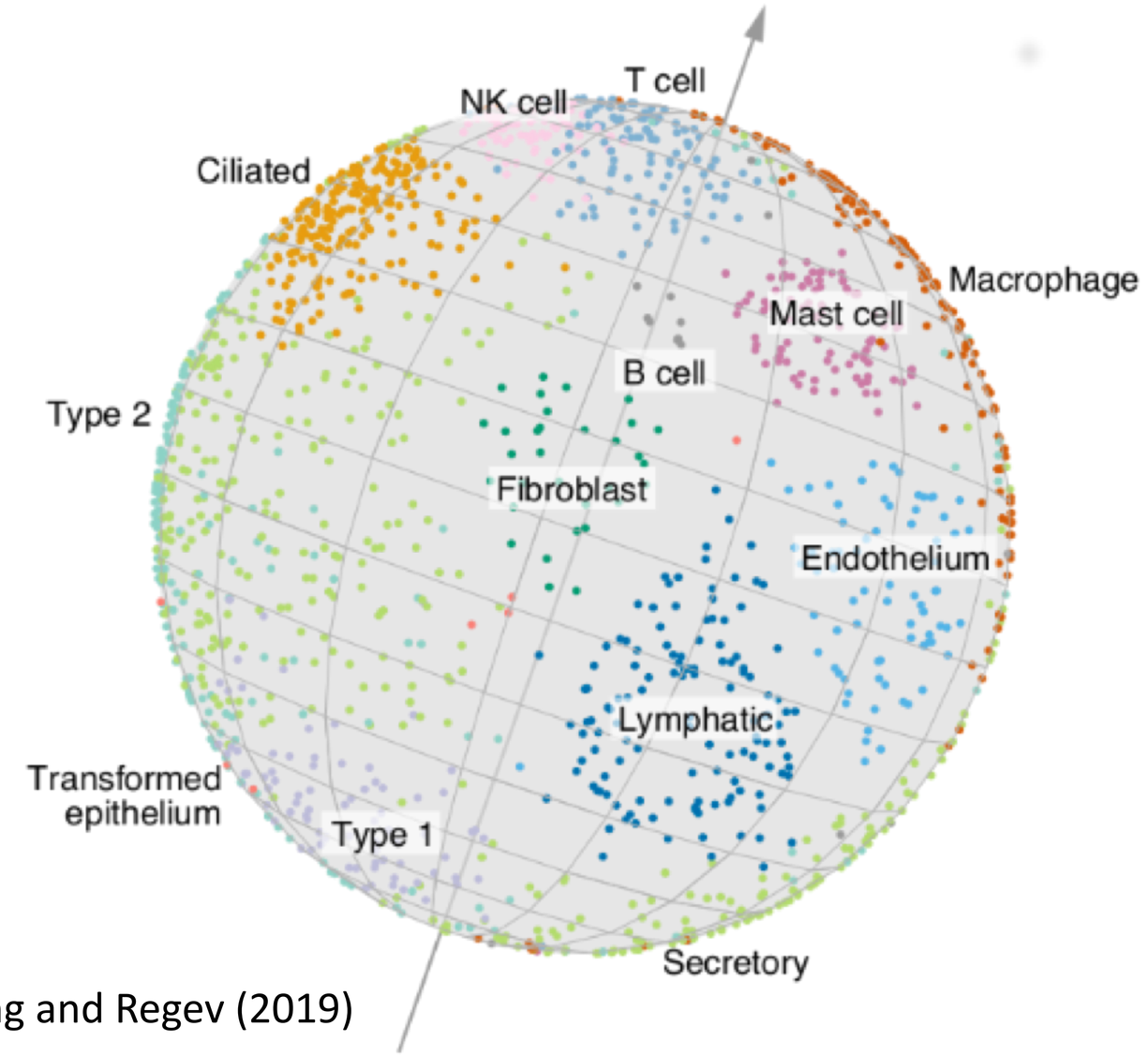

Ding and Regev (2019)

What if we project single-cell gene vectors onto hyper-sphere?

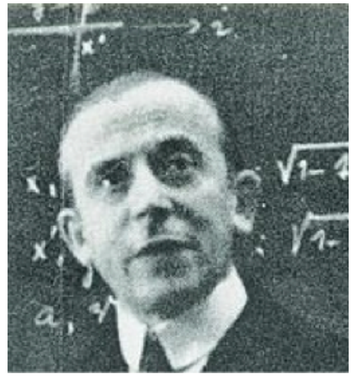



Angular distance between stars on Earth

# Angular distance and von Mises-Fisher




Ding and Regev (2019)



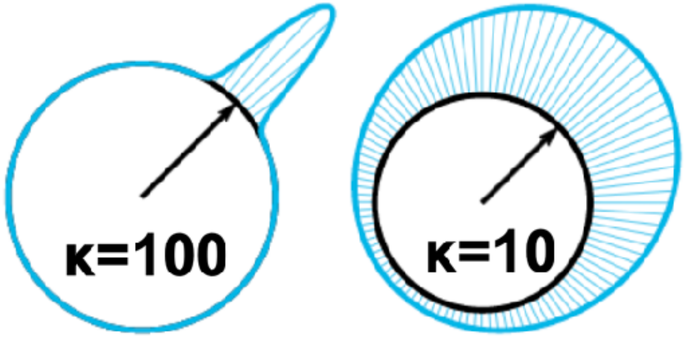
von Mises (p=2)

Fisher extended (p>2)

$$\exp\left\{ \kappa x_j^\top \mu_k \right\} C(\kappa)$$

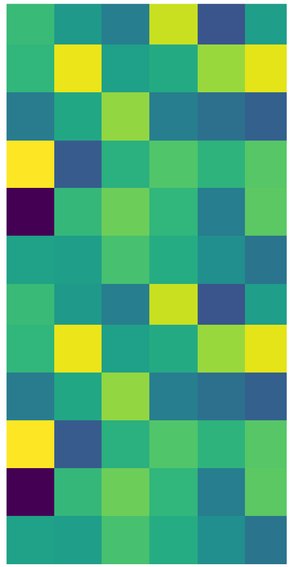
$\mu$: direction

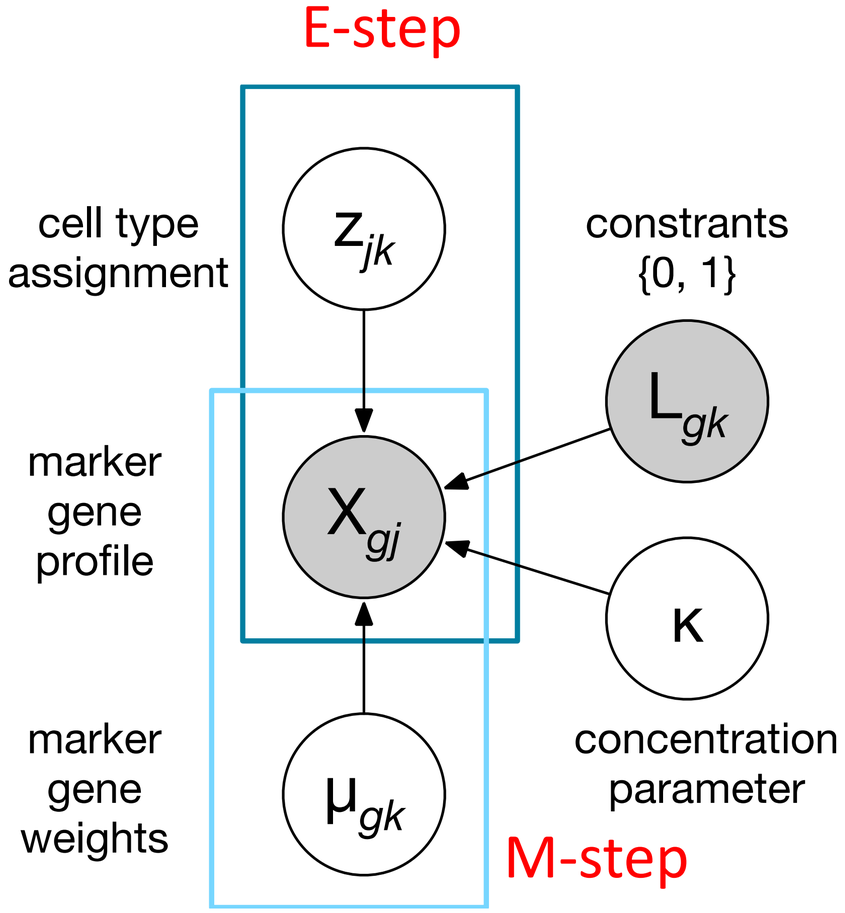
$\kappa$: concentration

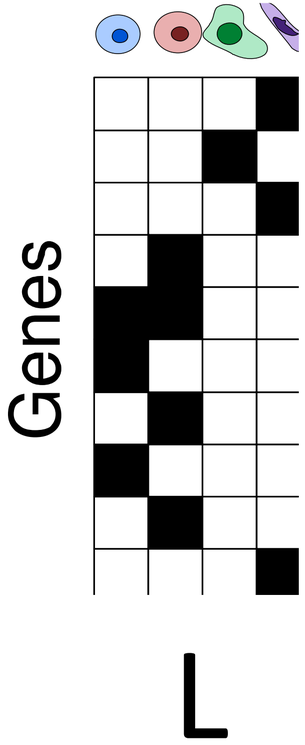


# Model a finite mixture of marker genes (with constraints: gene → cell type membership)

X






L

T cell: *CD2*, *CD3*, *CD45*

B cell: *CD19*, *CD20*, *CD79A*, *CD45*

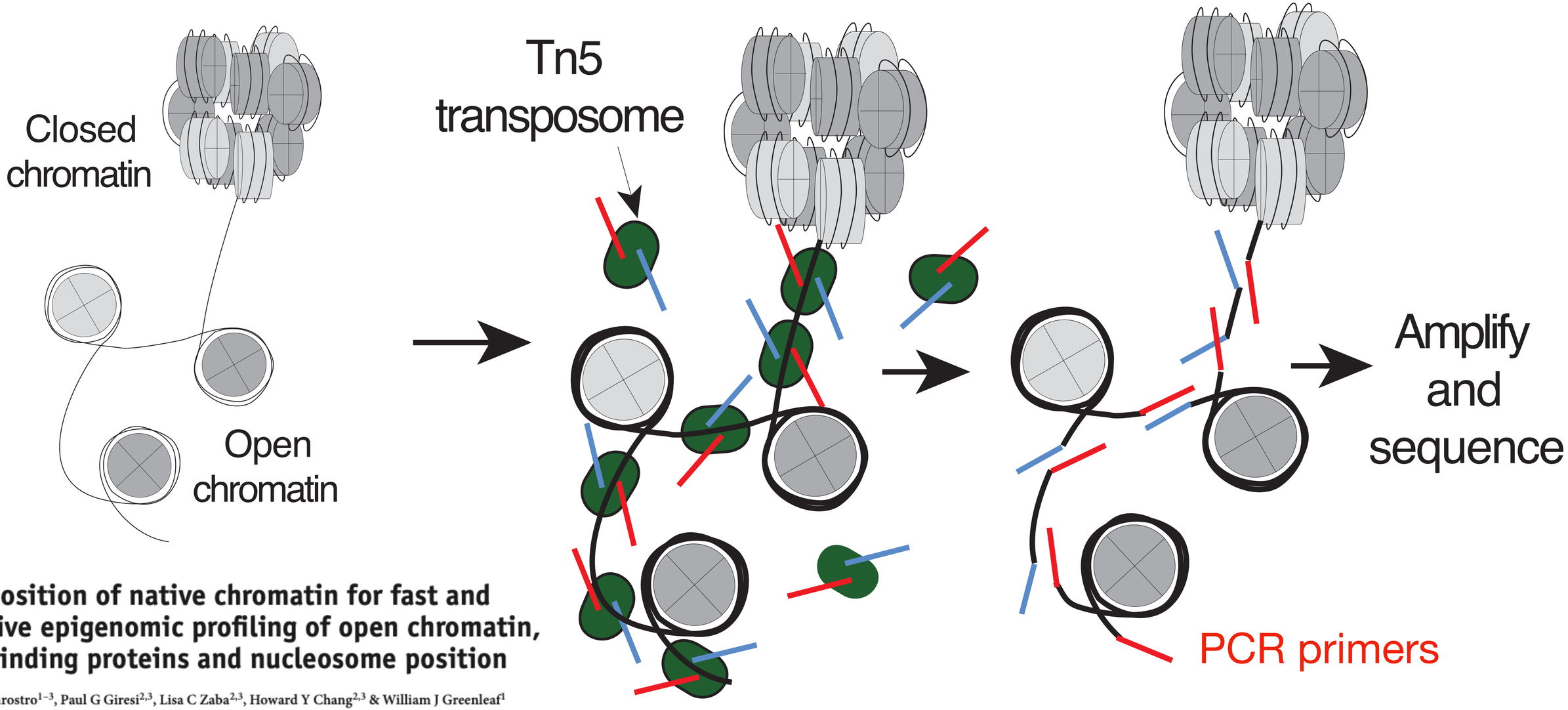
Tumor cell: *EPCAM*

Fibroblast: *COL1A1*, *ACTA2*, *THY1*

# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

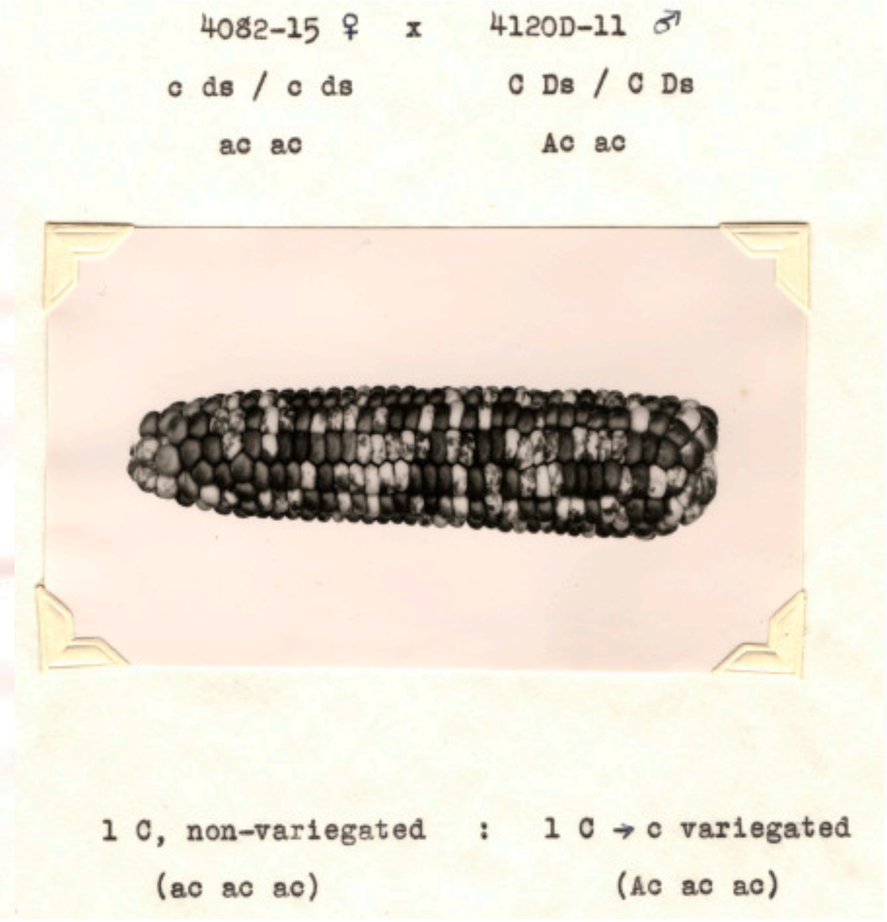
# ATAC-seq: How to profile single-cell epigenomics data?




**Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position**

Jason D Buenrostro[1–3], Paul G Giresi[2,3], Lisa C Zaba[2,3], Howard Y Chang[2,3] & William J Greenleaf[1]

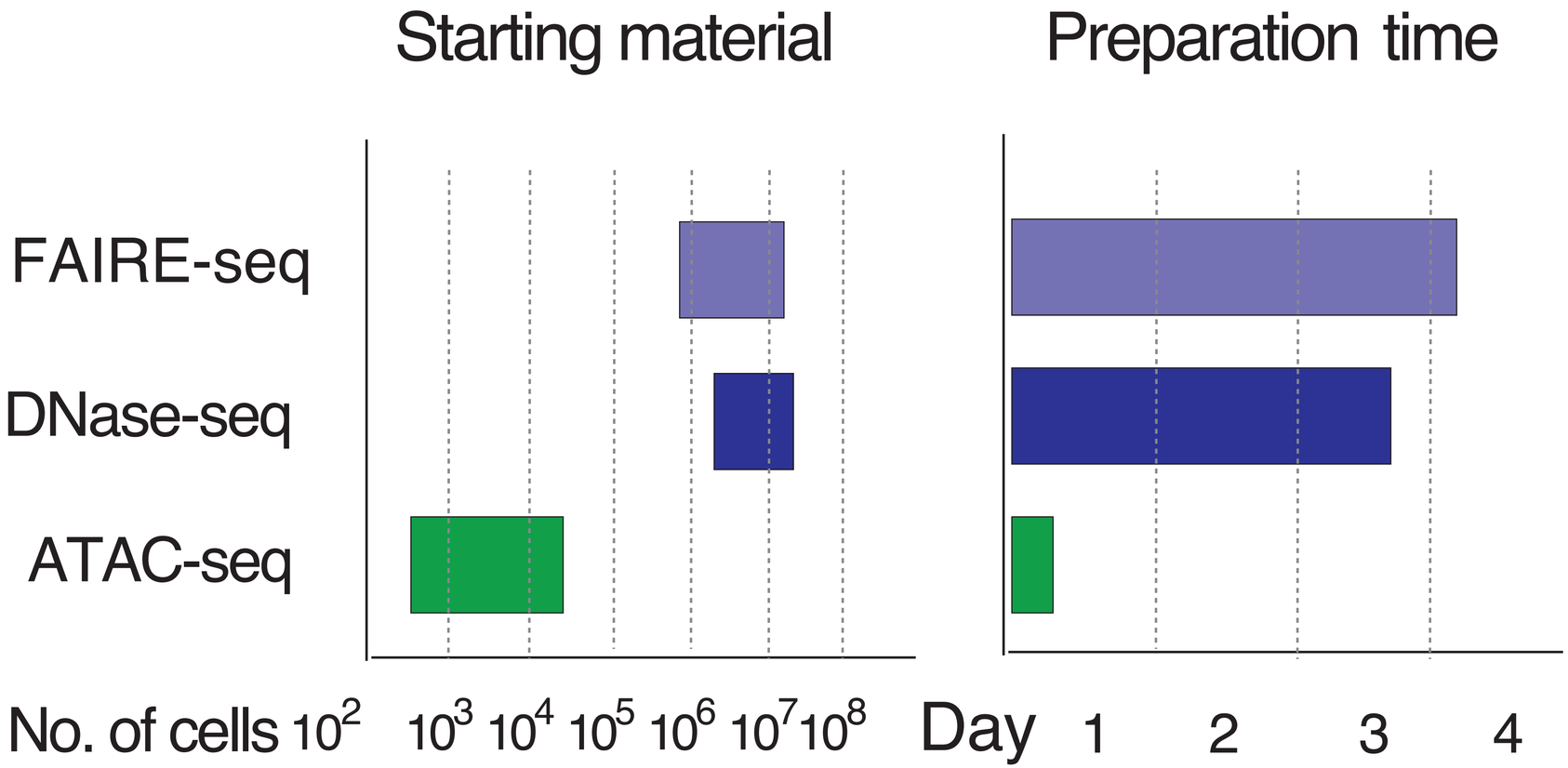
# Transposon: “jumping genes” + transposase




Barbra McClintock discovered genes could “jump”

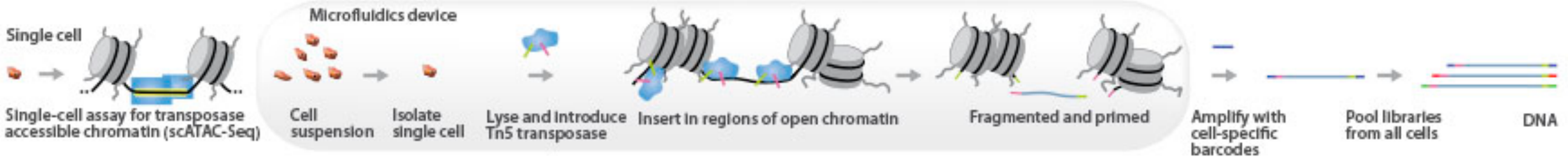
https://www.nobelprize.org/womenwhochangedscience/stories/barbara-mcclintock

# Why is ATAC-seq significant?




**Buenrostro *et al.* *Nature Methods* (2013)**

# Single-cell ATAC-seq



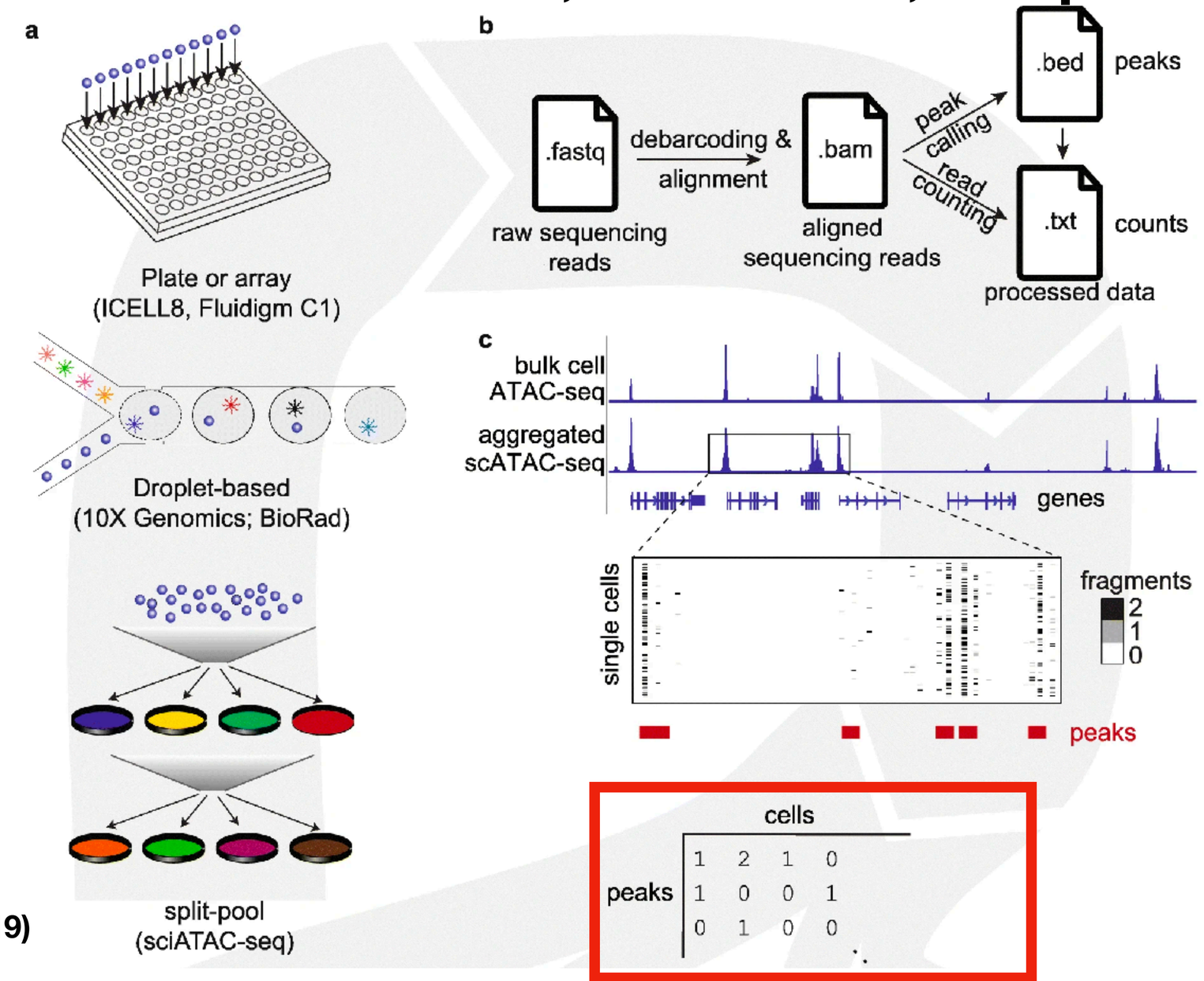

How many unique regions/reads are in a cell?

What are the dimensions? # features? # samples?

What are the min and max count per region and cell?

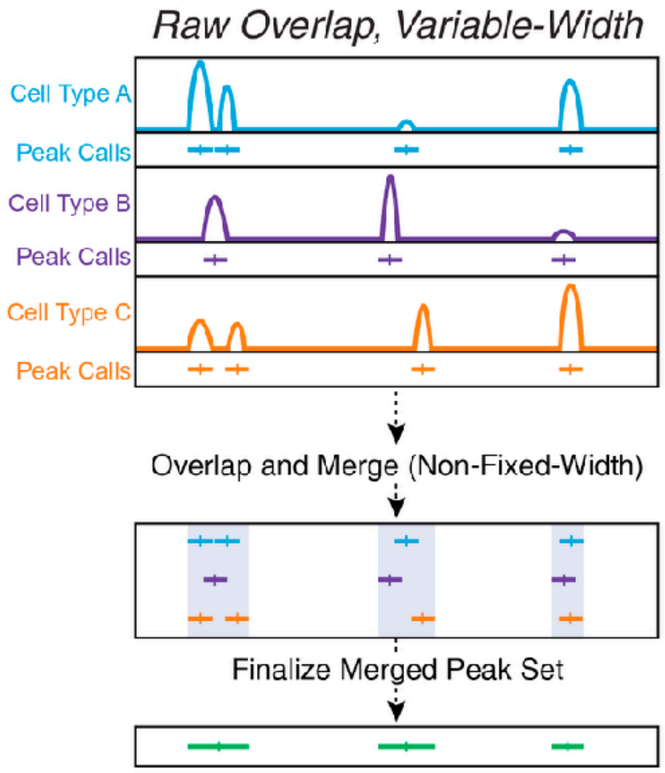
**https://www.illumina.com/techniques/popular-applications/epigenetics/atac-seq-chromatin-accessibility.html**

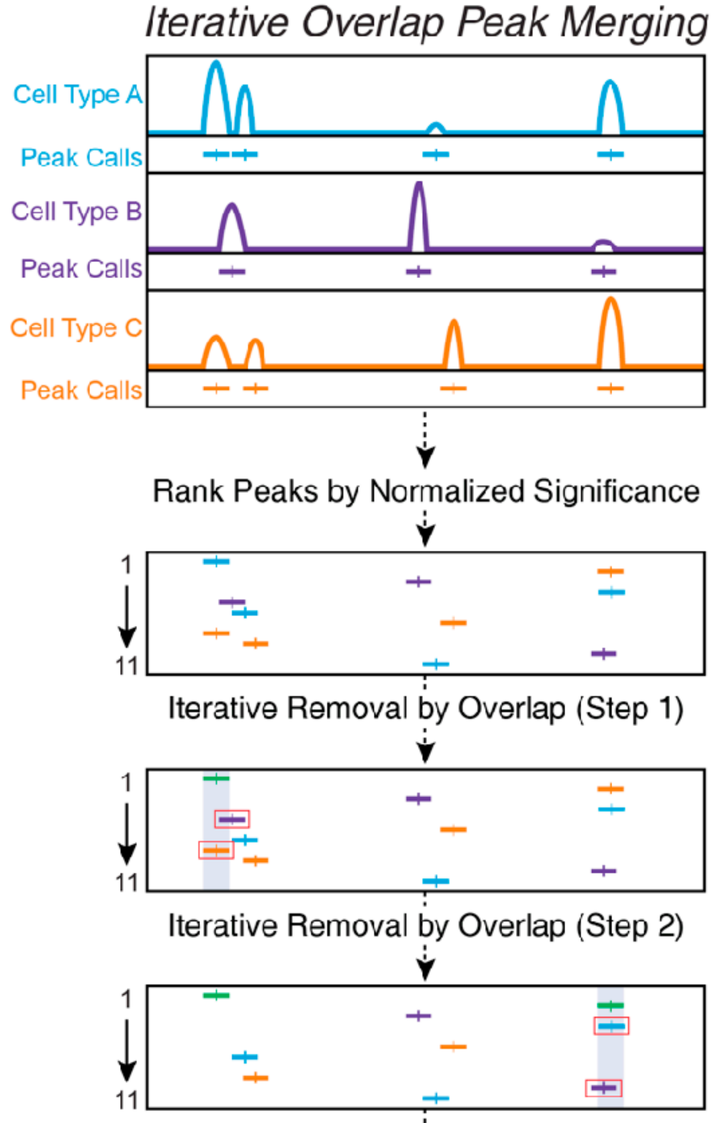
# The same theme: index cells, combine, sequence




Chen .. Pinello, *Genome Biology* (2019)

# ATAC-seq: How to process single-cell epigenomics data?



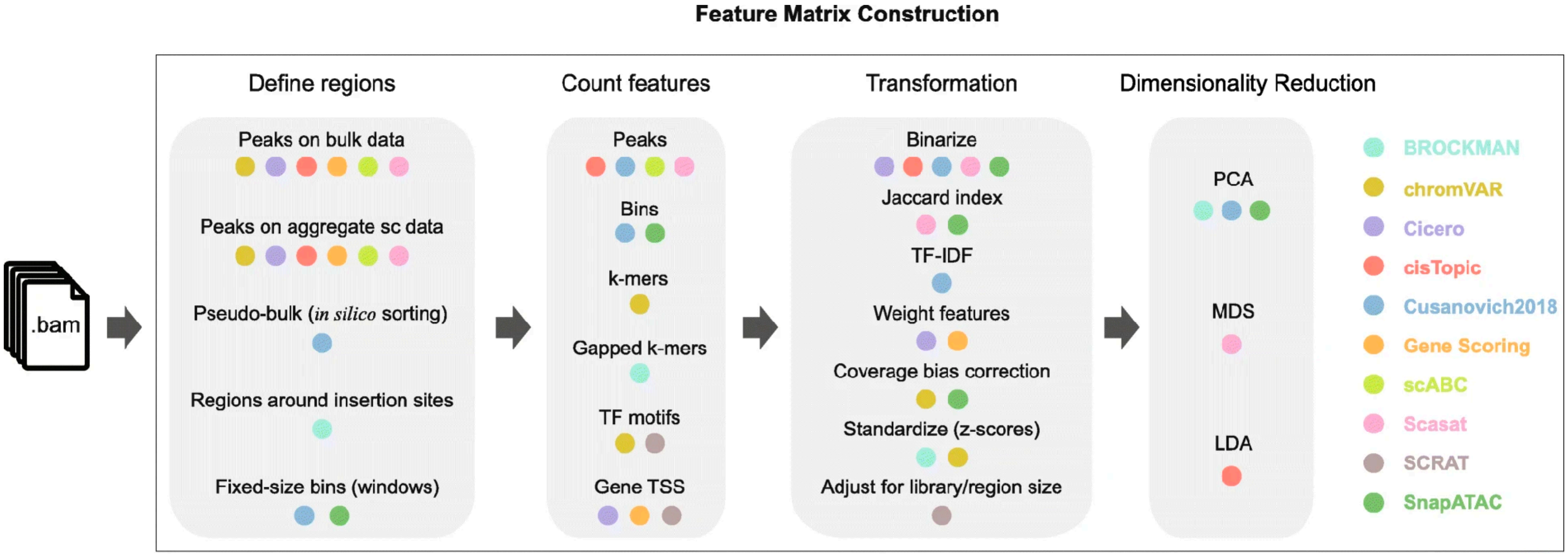




Why do we need to call “peaks?”

**A typical data point:**
**$chr:$start-$end <tab> intensity**

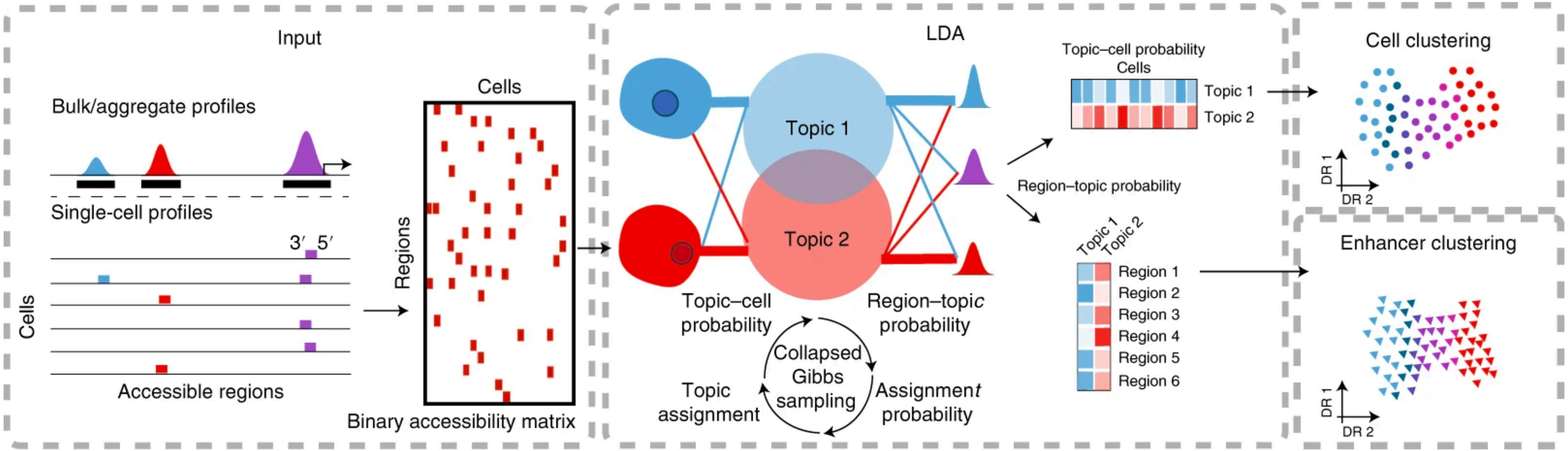
ArchR

# We treat “DNA regions” as features



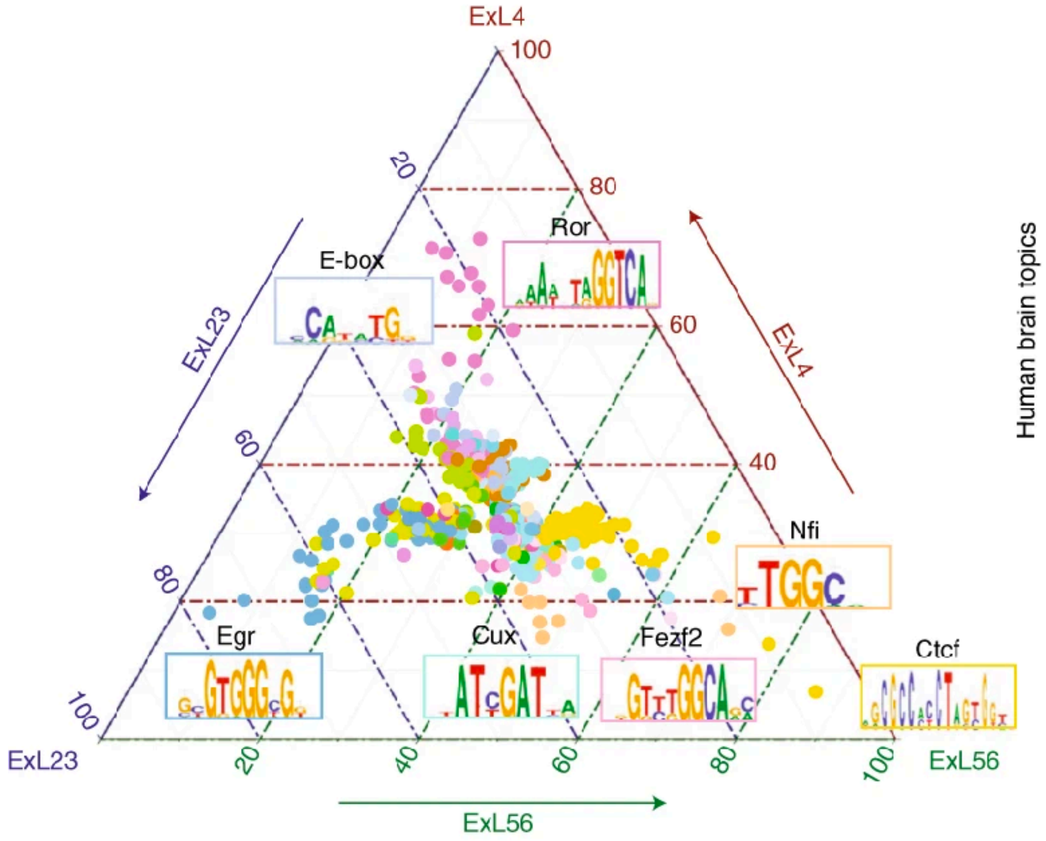

Chen .. Pinello, *Genome Biology* (2019)

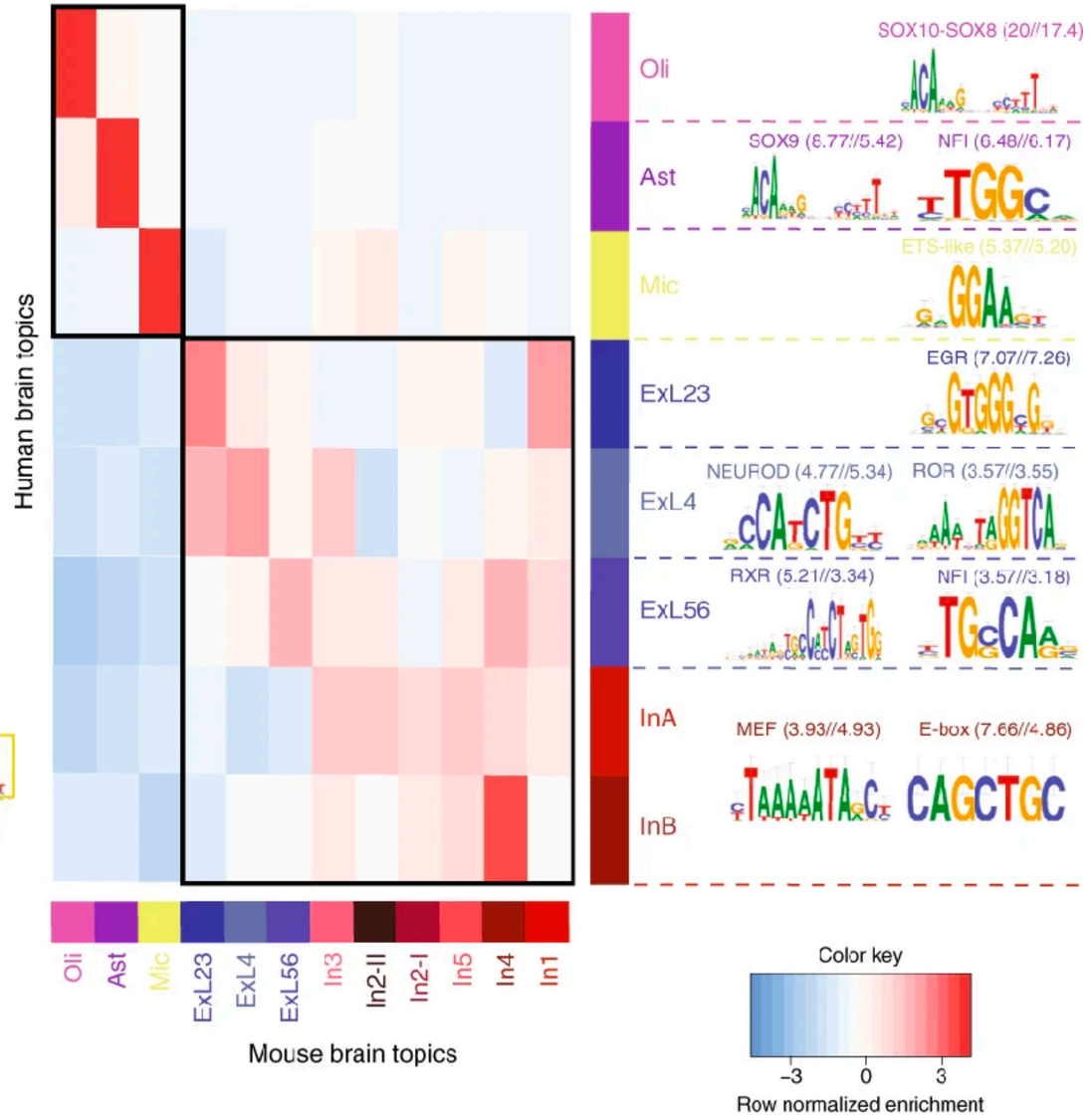
# We can treat “regions” as features




Gonzalez-Blas .. Aerts, *Nature Methods* (2019)

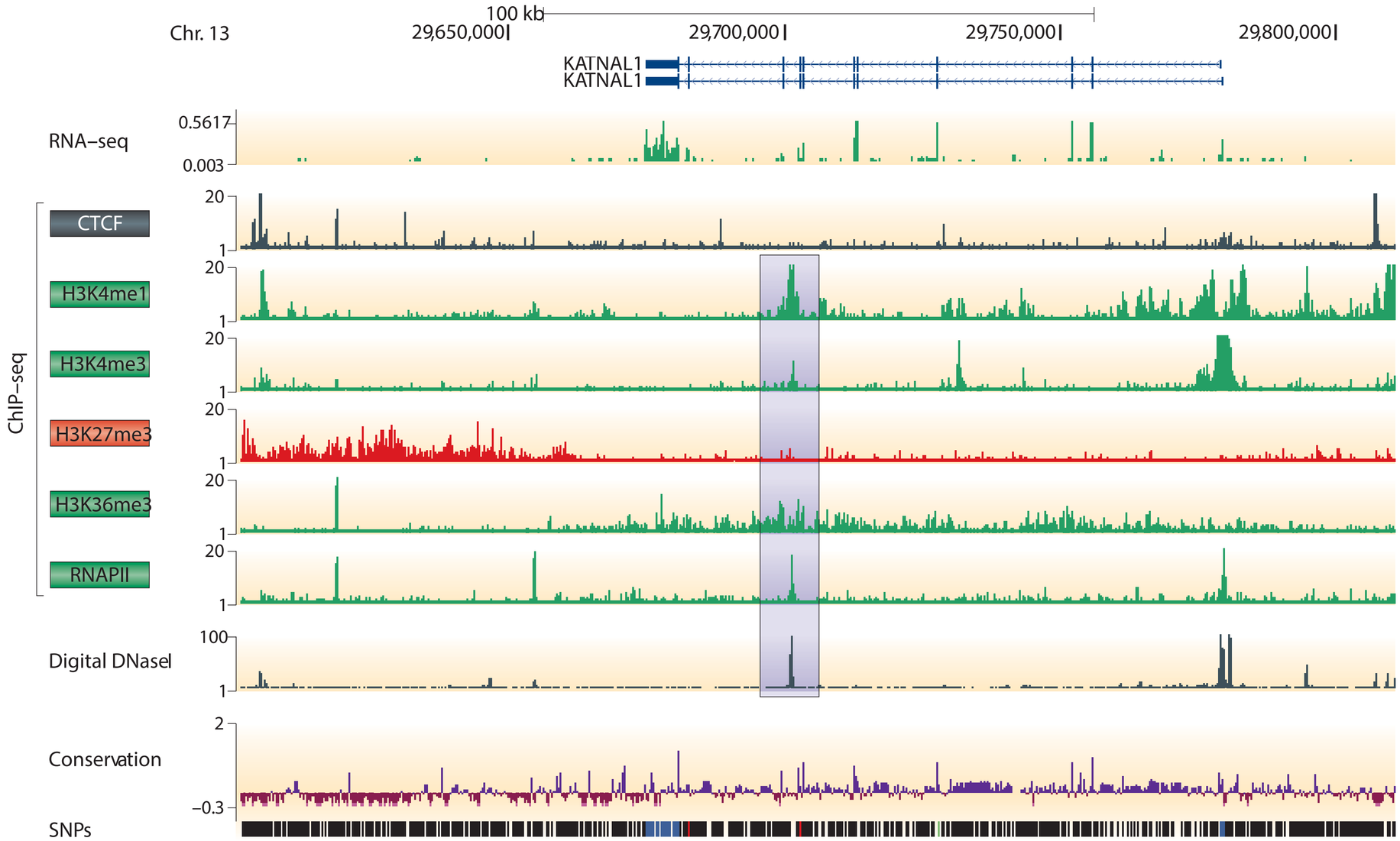
# Probabilistic topic models identify regulatory topics and cell types (and enriched motifs)



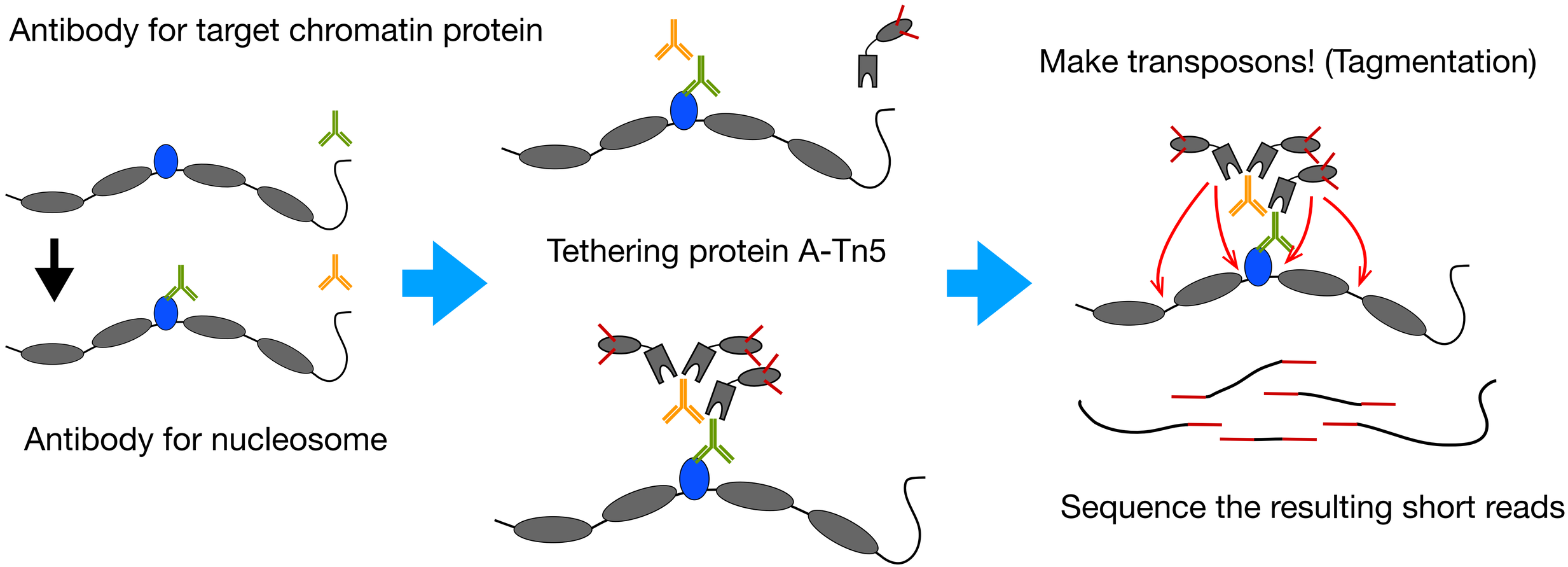

Gonzalez-Blas .. Aerts, *Nature Methods* (2019)

# An example of multiple ChIP-seq tracks



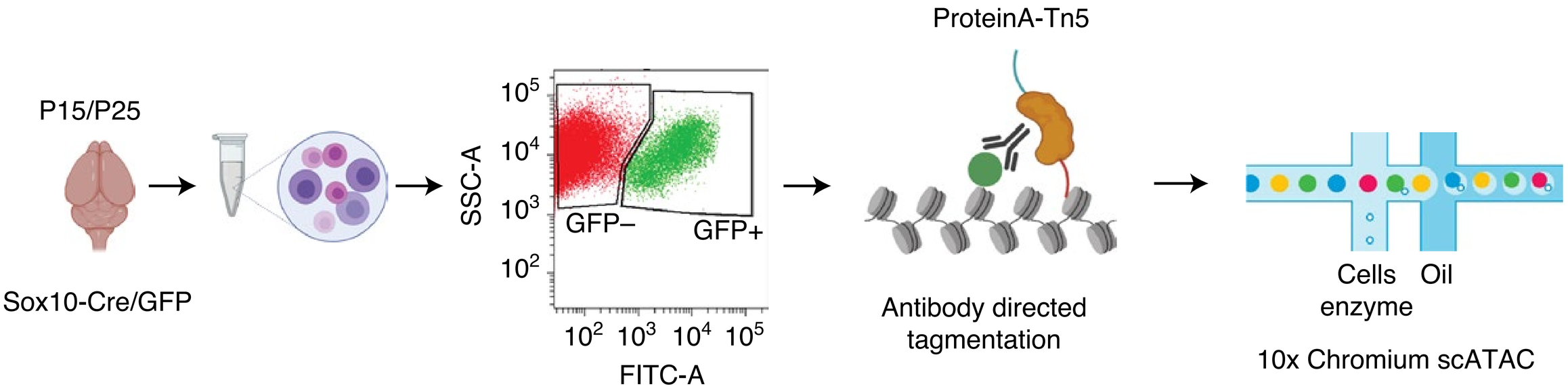

Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

# CUT & Tag = Cleavage Under Target + Tn5 Tagmentation



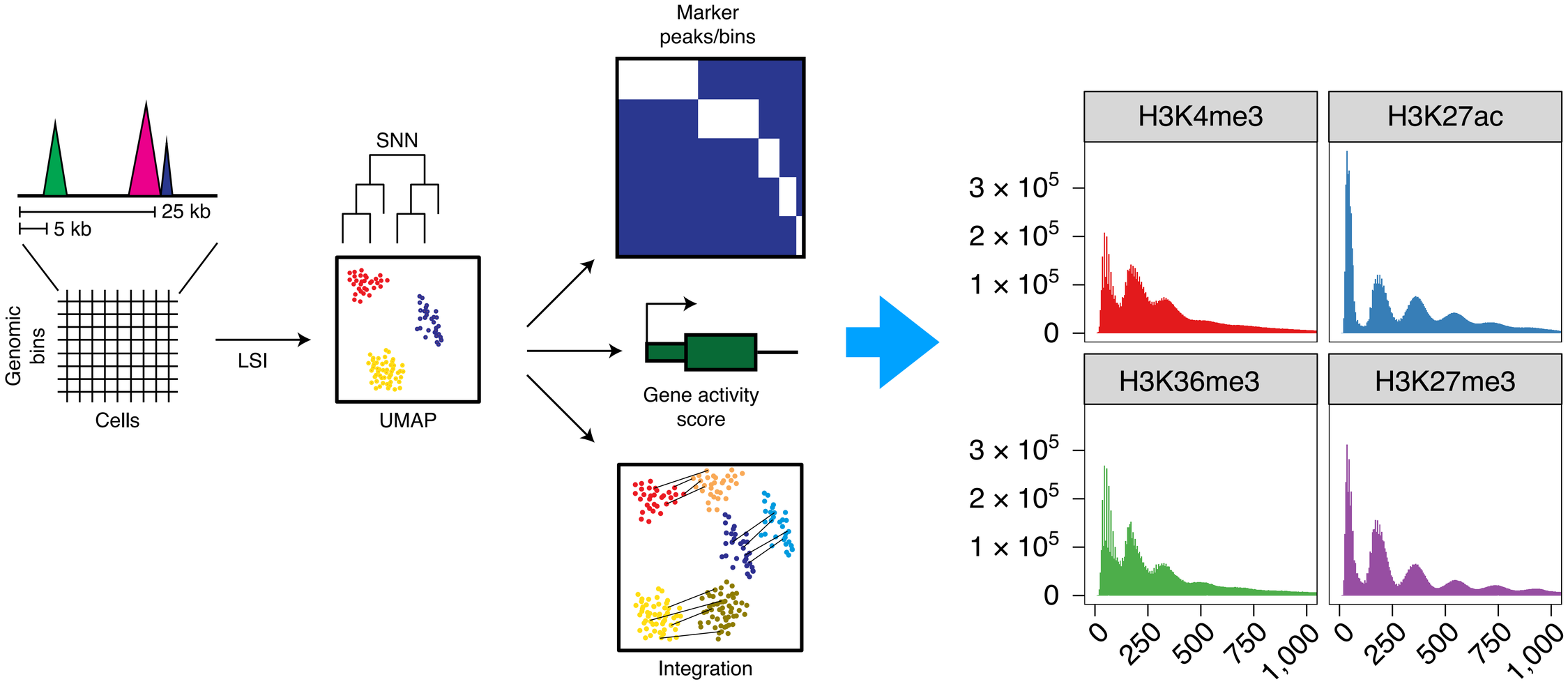

**Kaya-Okur .. Henikoff, *Nature Comm.* (2019)**

# Single-cell CUT & Tag




**Bartosovic .. Castelo-Branco, *Nature Biotech.* (2021)**

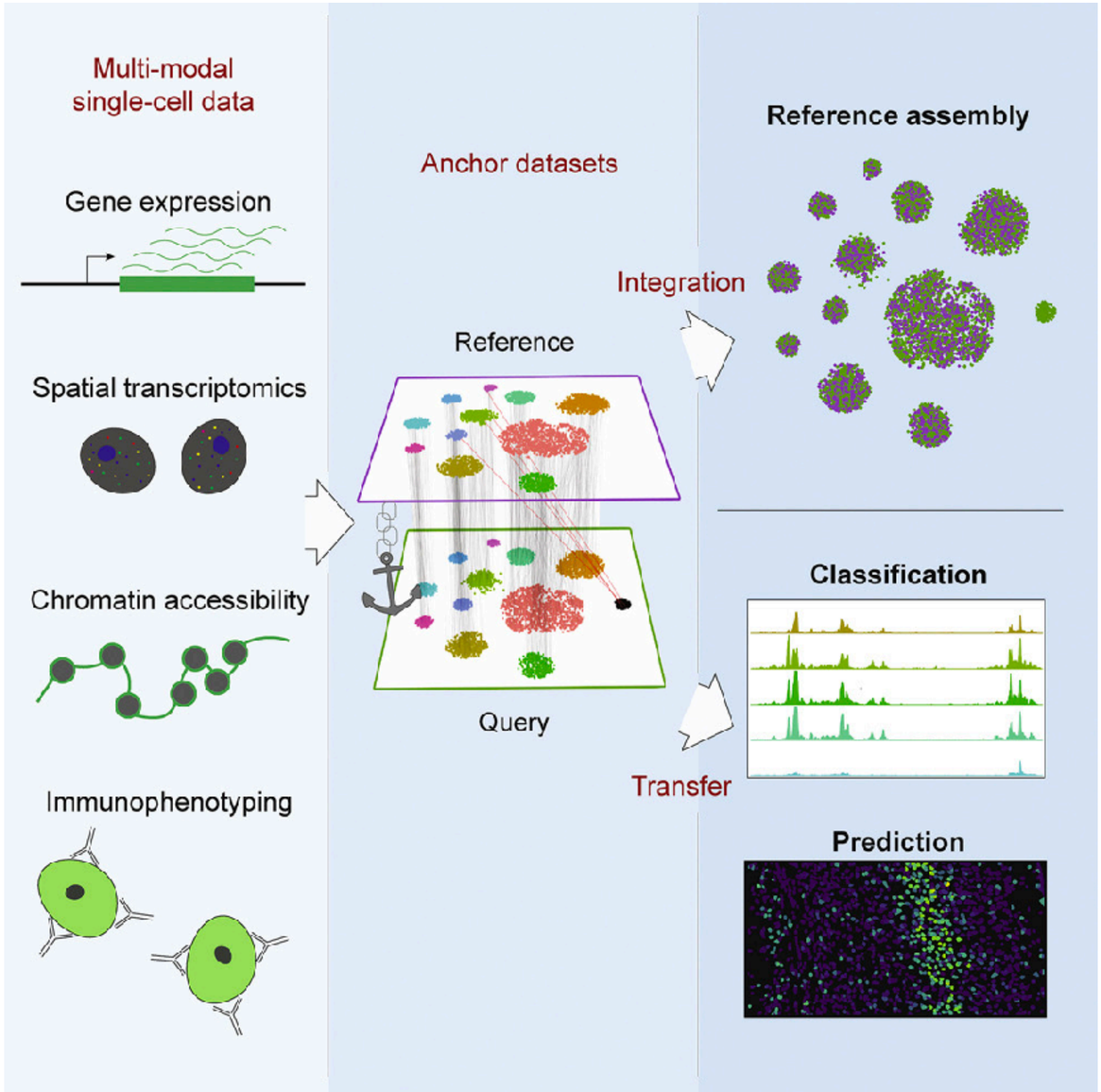
# Single-cell CUT&Tag followed by a typical scATAC-seq analysis



# Today’s lecture: Single-cell Part 2

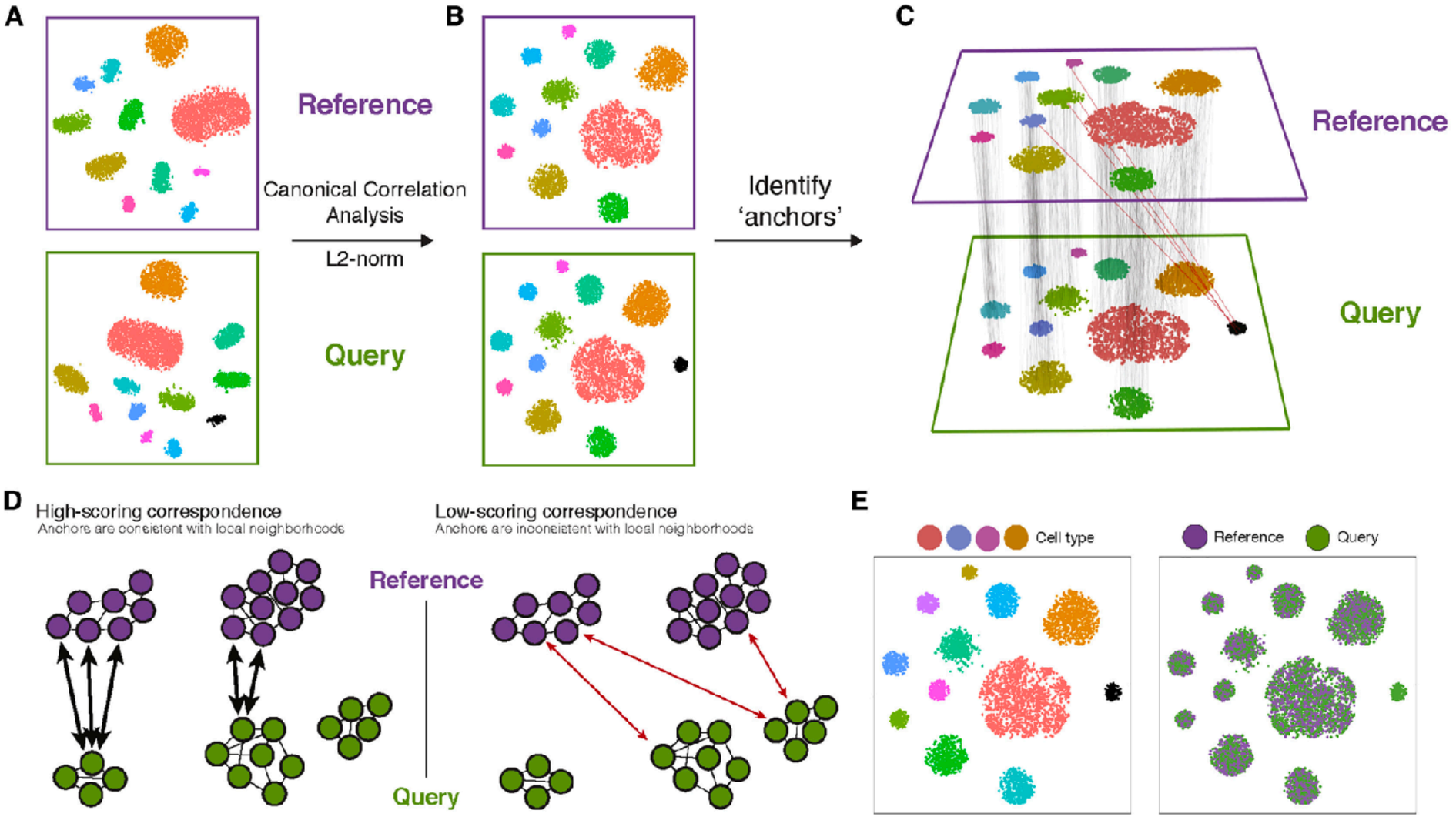
- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

# How can we do multimodal data integration?



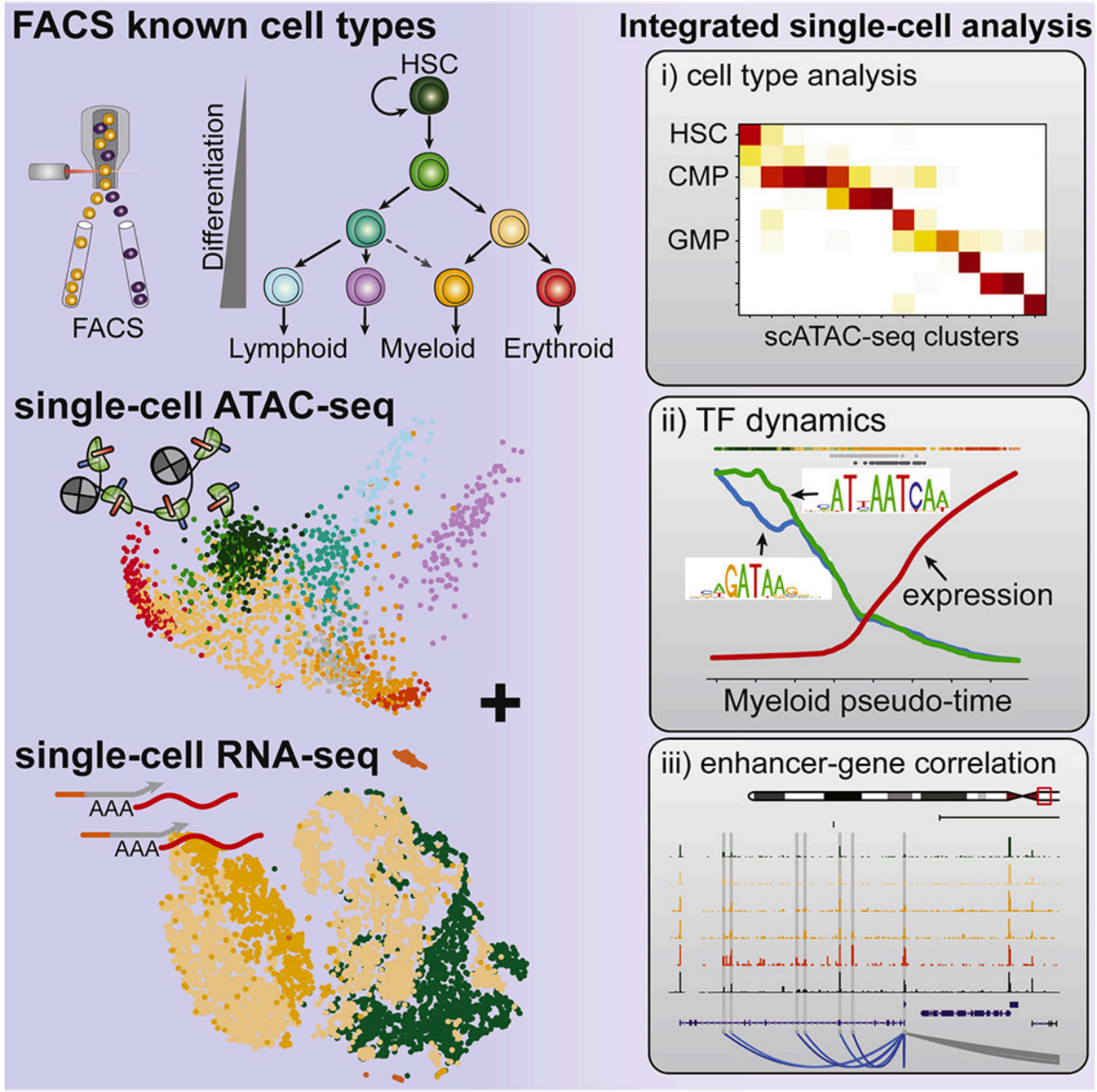

Stuart .. Satija, *Cell* (2019)

# How can we do multimodal data integration?



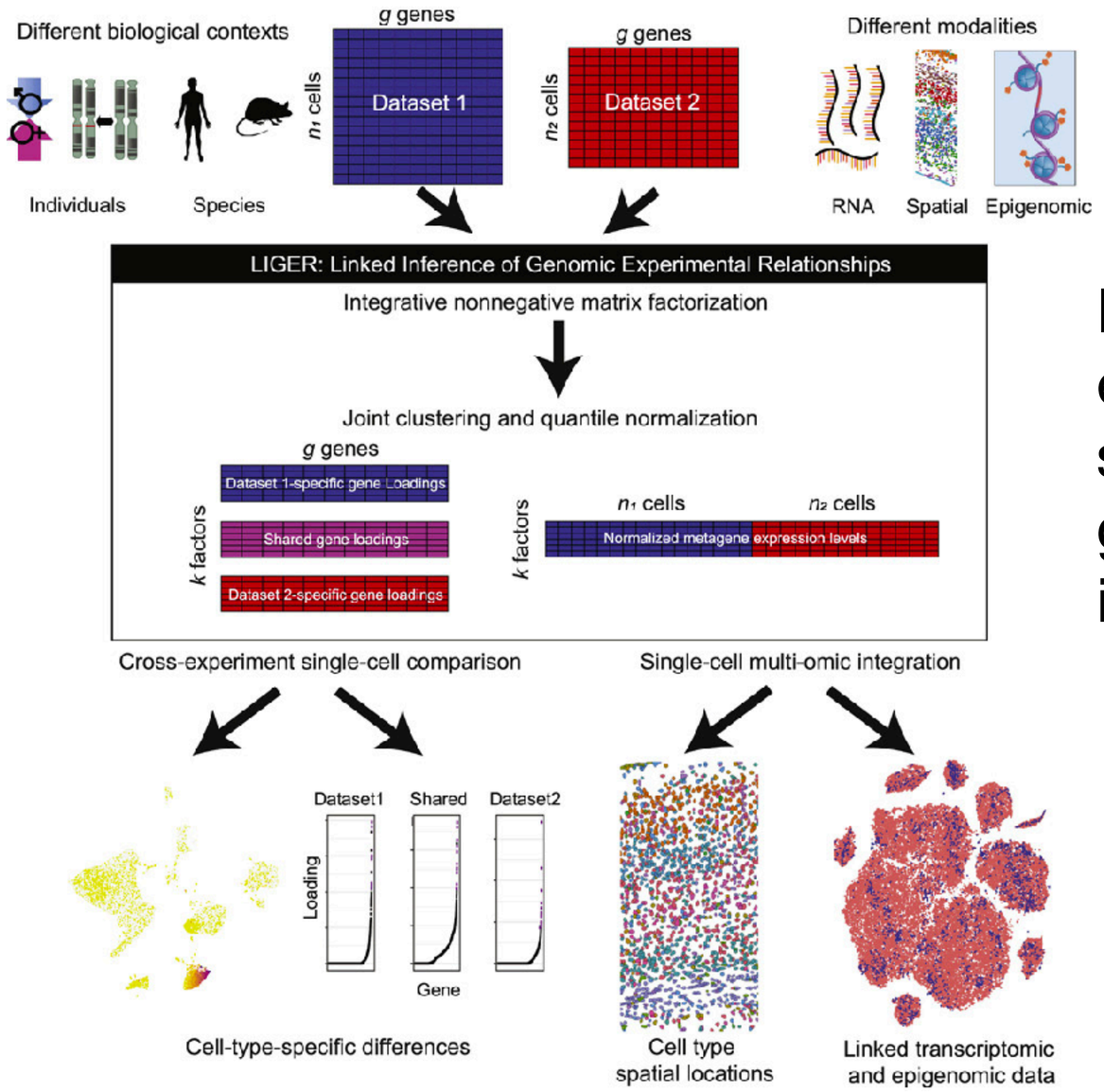

Stuart .. Satija, *Cell* (2019)

# Multi-omics data integration has become increasingly common.




Buenrostro .. Greenleaf, *Cell* (2018)

# A general principle: Gene-centric data integration

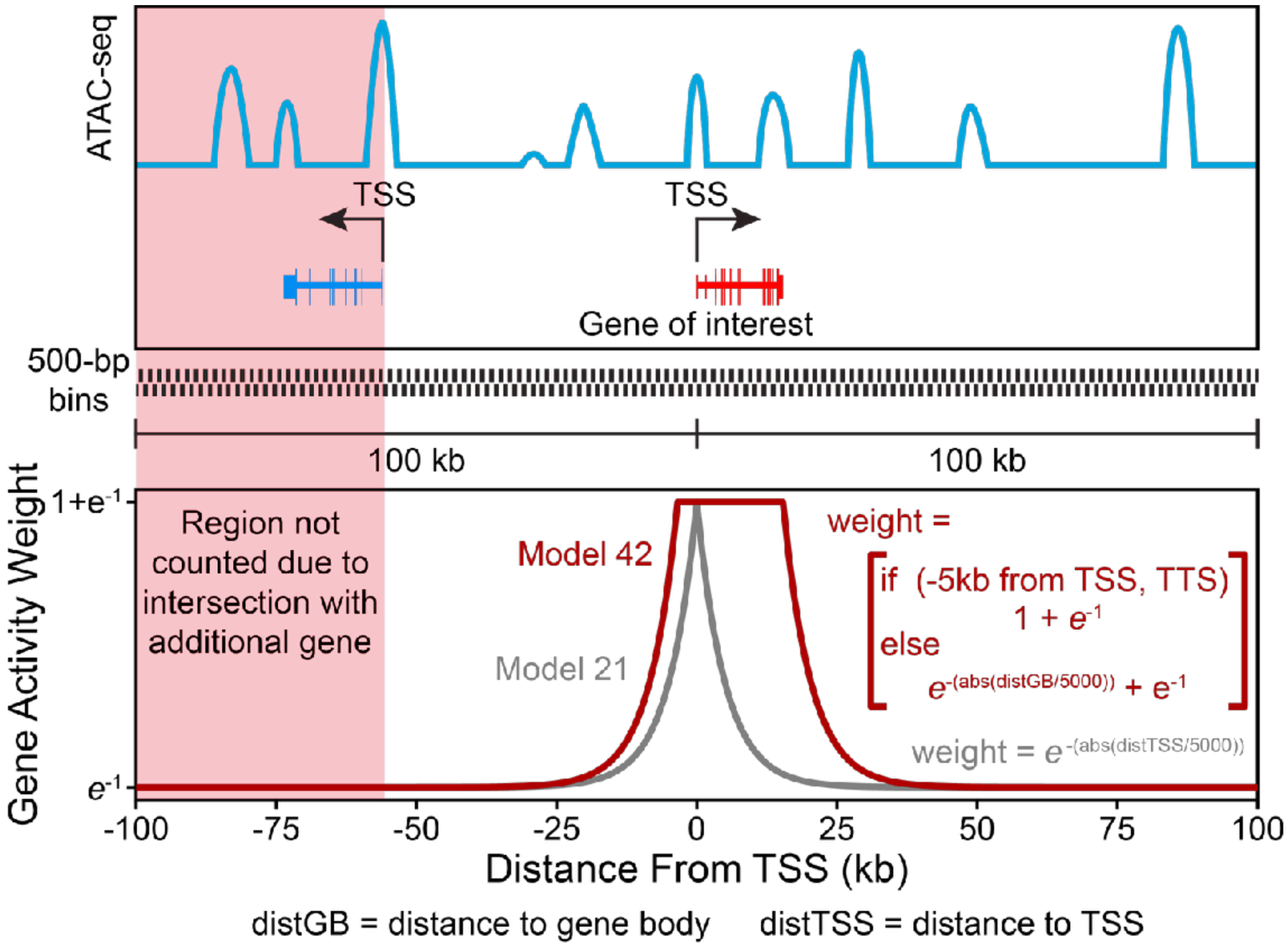



If two data sets encompass the same set of features (e.g., genes), data integration is straightforward.

**Welch .. Macosko, *Cell* (2019)**

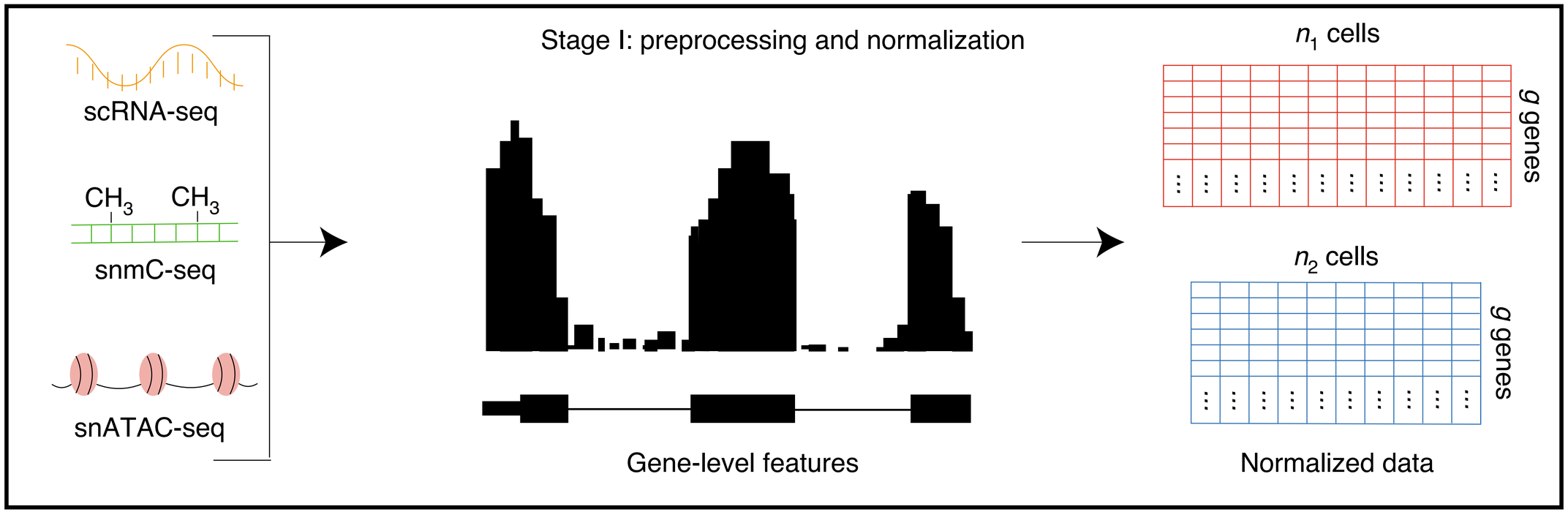
# A general principle: Gene-centric data integration

- Calculate “gene” scores to aggregate peaks into a gene-level feature.
- Straightforward joint analysis with scRNA-seq
- Why it makes sense? Rationale: most peaks correspond to promoter signals.



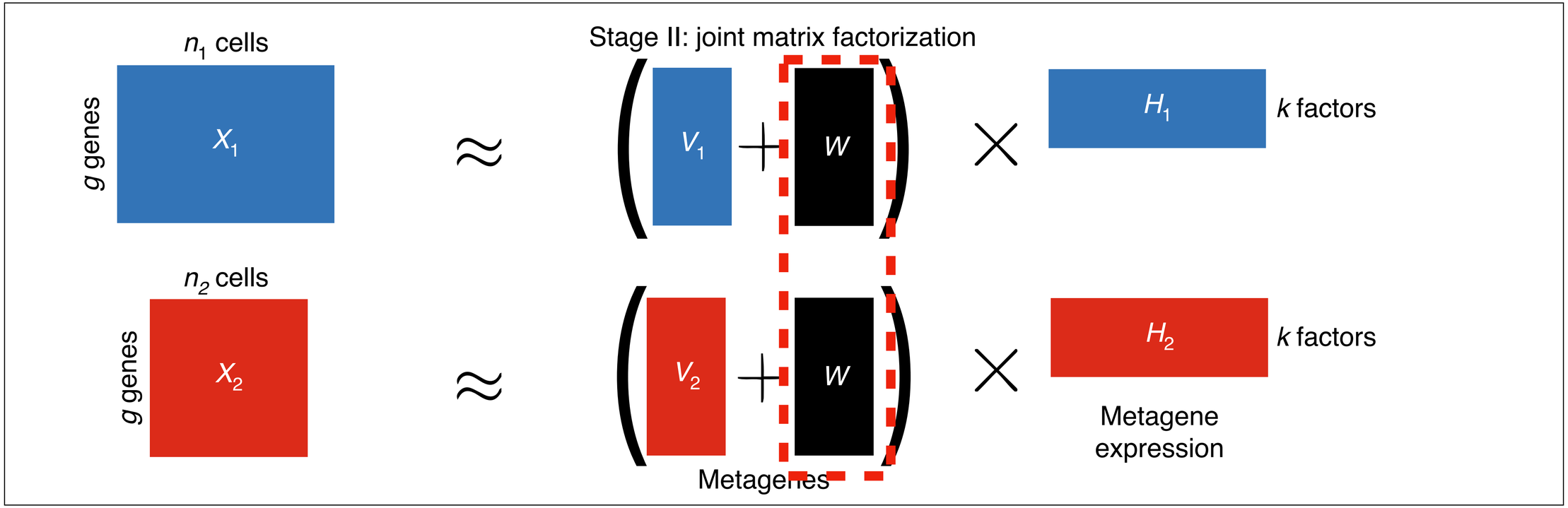

ArchR

# Gene-centric data integration



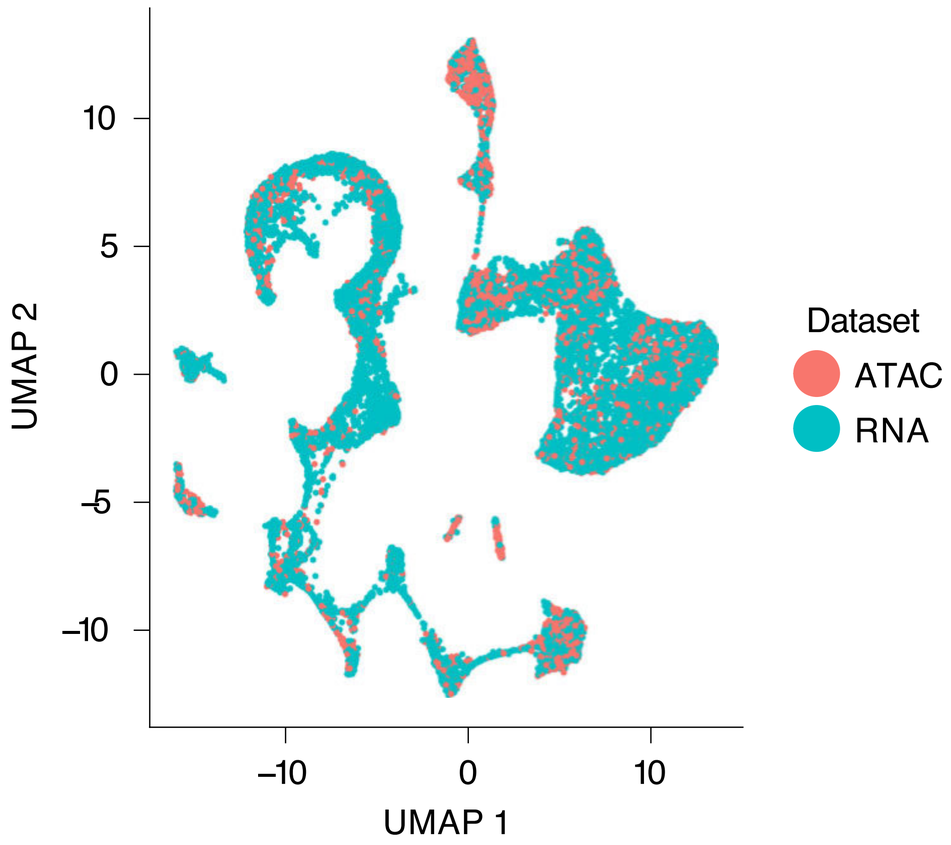

**Liu .. Welch, *Nature Protocols* (2020)**

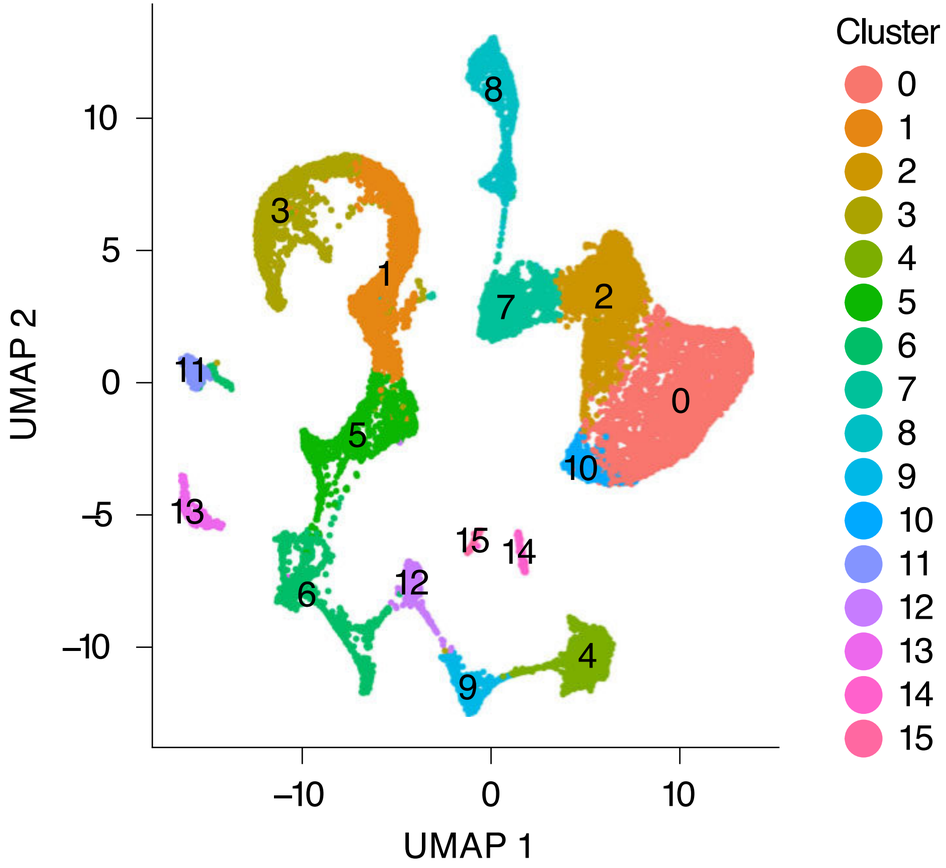
# Gene-centric data integration model




**Liu .. Welch, *Nature Protocols* (2020)**

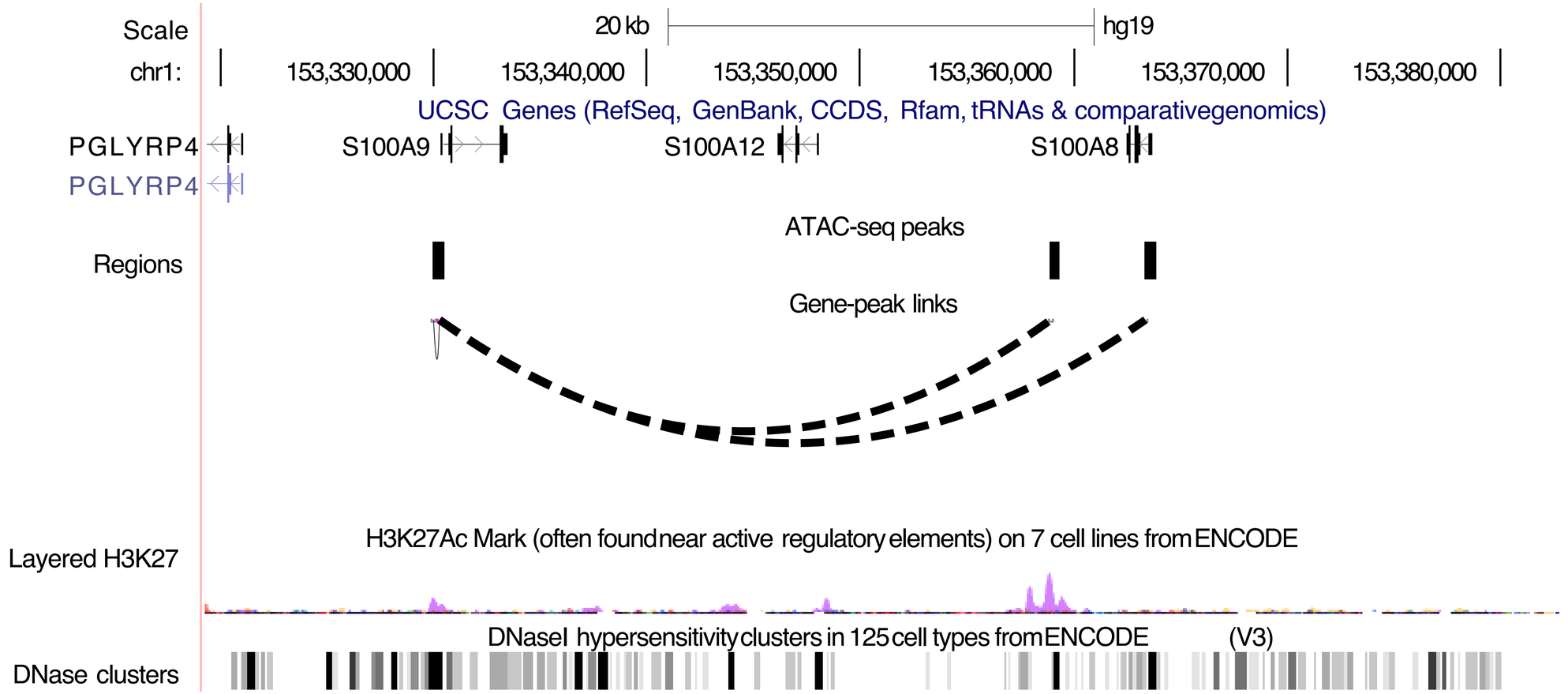
# After mapping genes and peaks in the same space… impute unobserved expressions/peaks



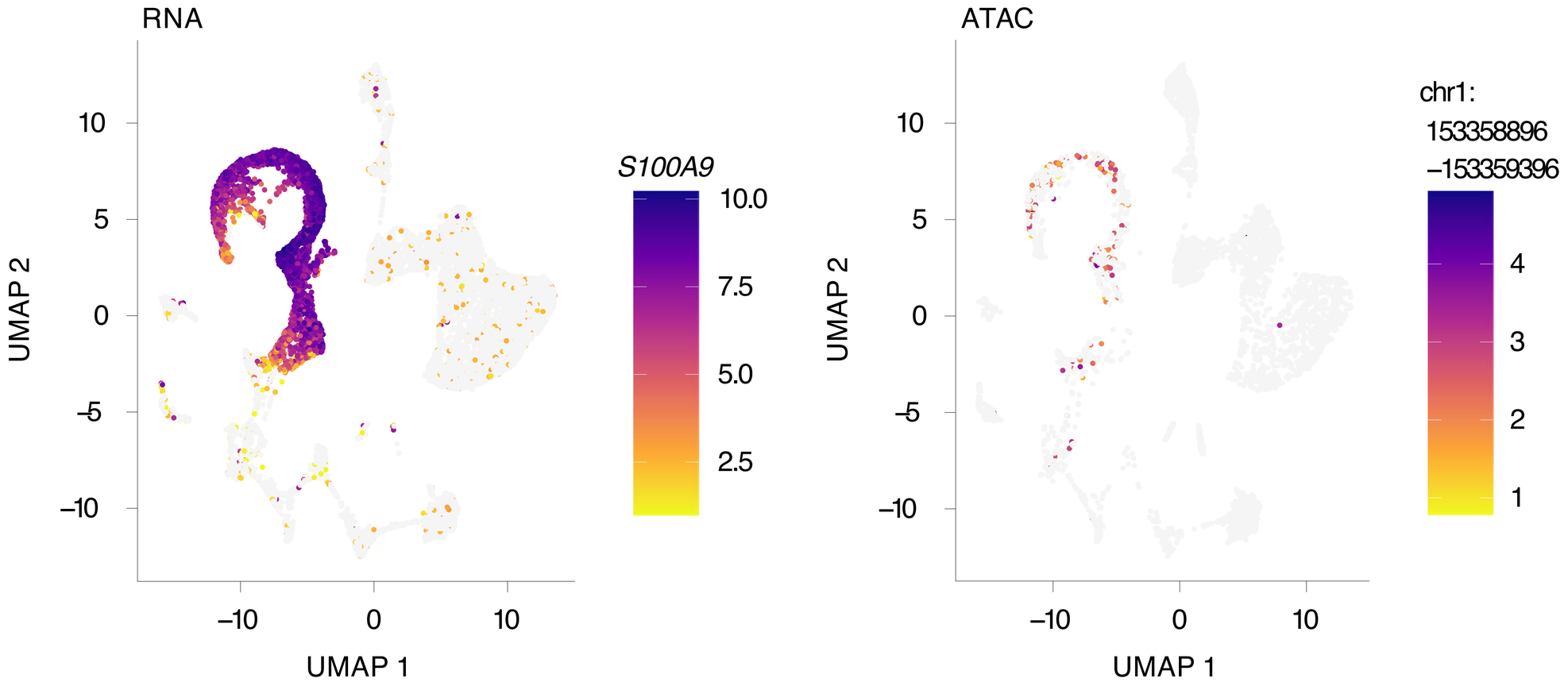

Liu .. Welch, *Nature Protocols* (2020)

# Measure correlation between the imputed genes and peaks



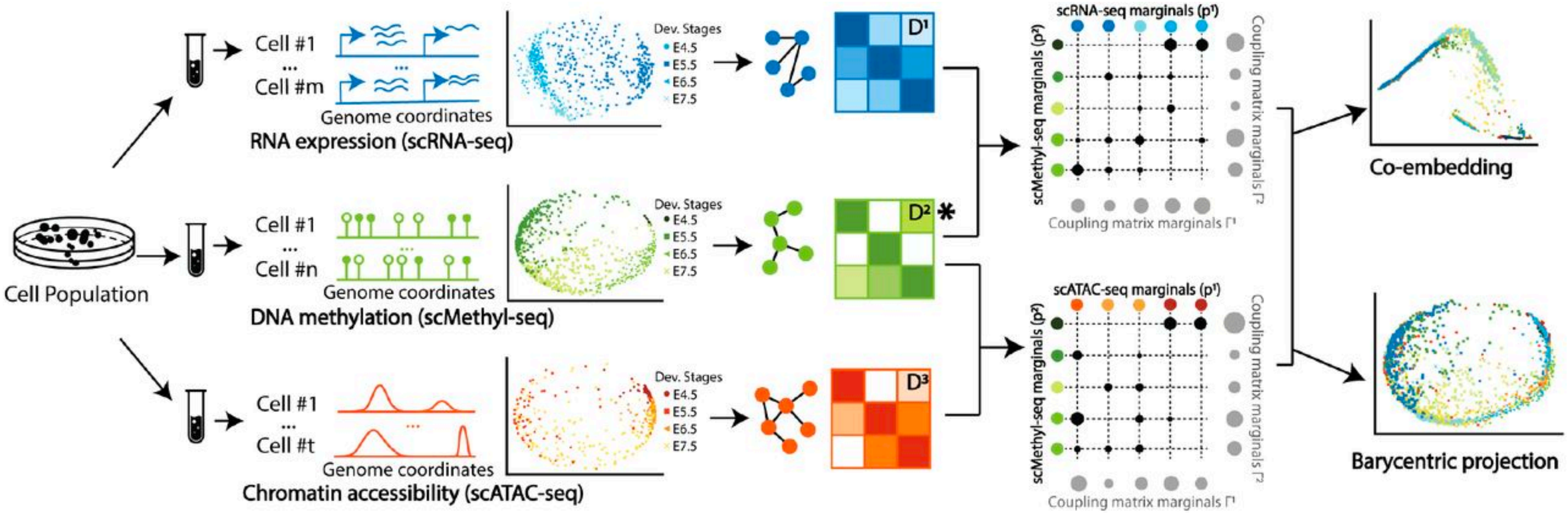

**Liu .. Welch, *Nature Protocols* (2020)**

# Measure correlation between the imputed genes and peaks




Liu .. Welch, *Nature Protocols* (2020)

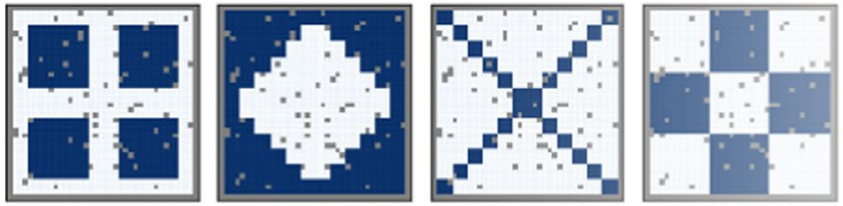
# We can do optimal transport to map one data type to the other data type



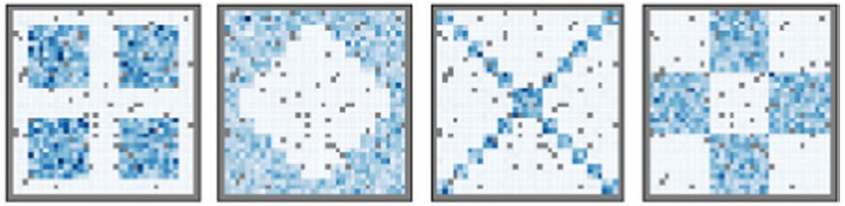
# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

# Spatial transcriptomic analysis demands another layer of modelling: dependency between neighbouring locations

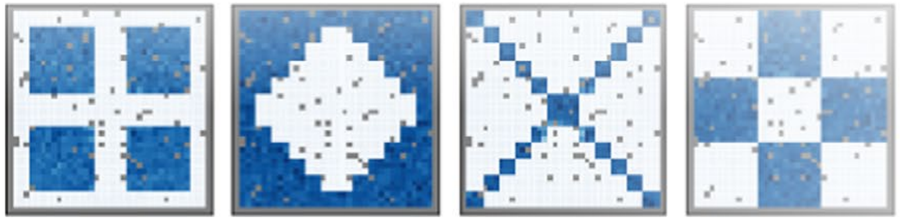


Ground truth

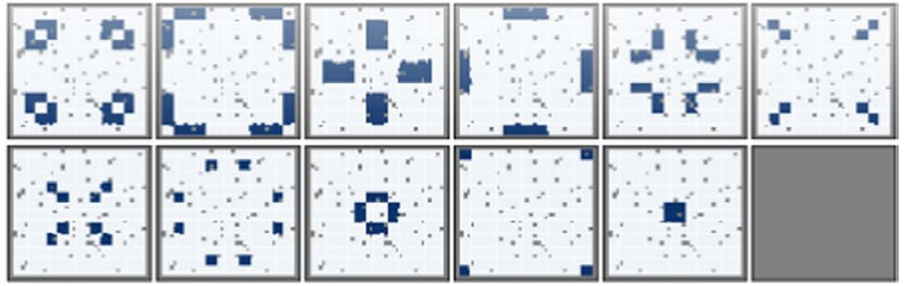


Stochastic simulation

NMF with dependency modelling

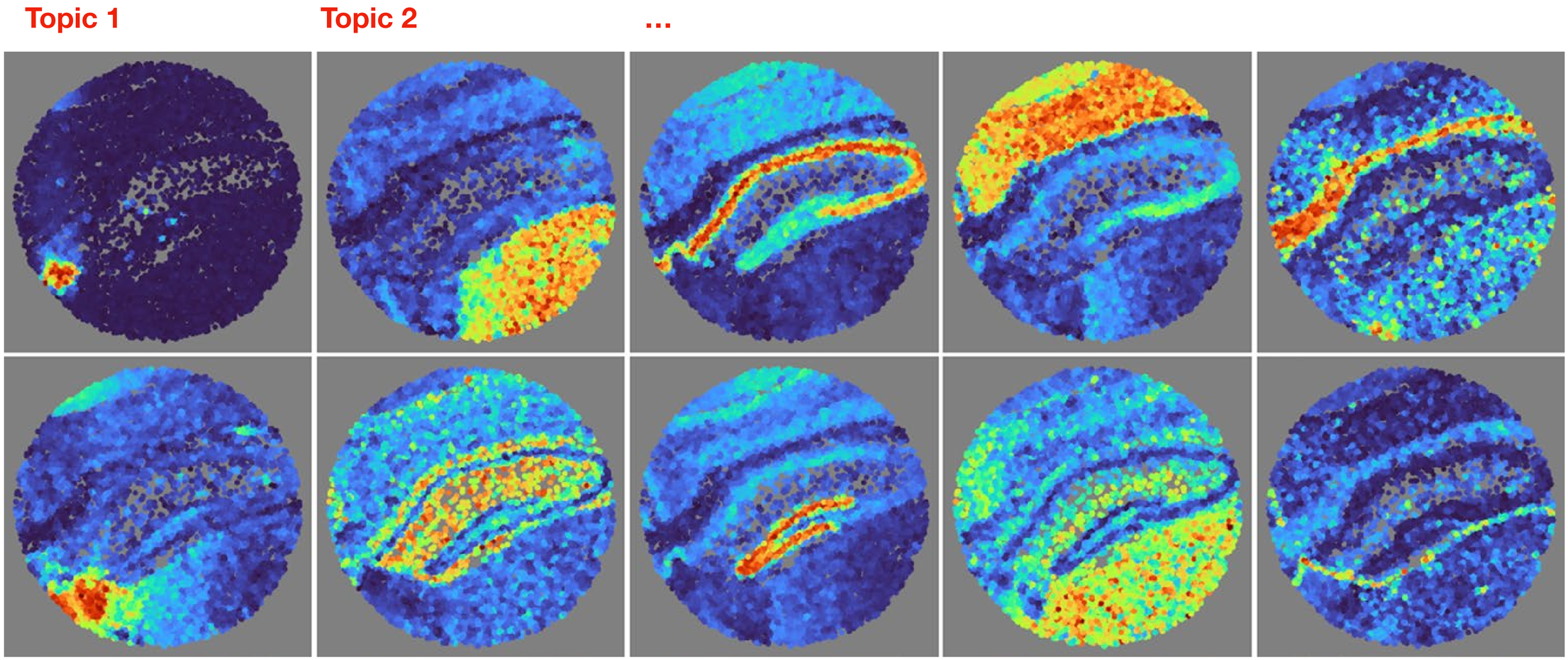


Clustering w/o dependency modelling



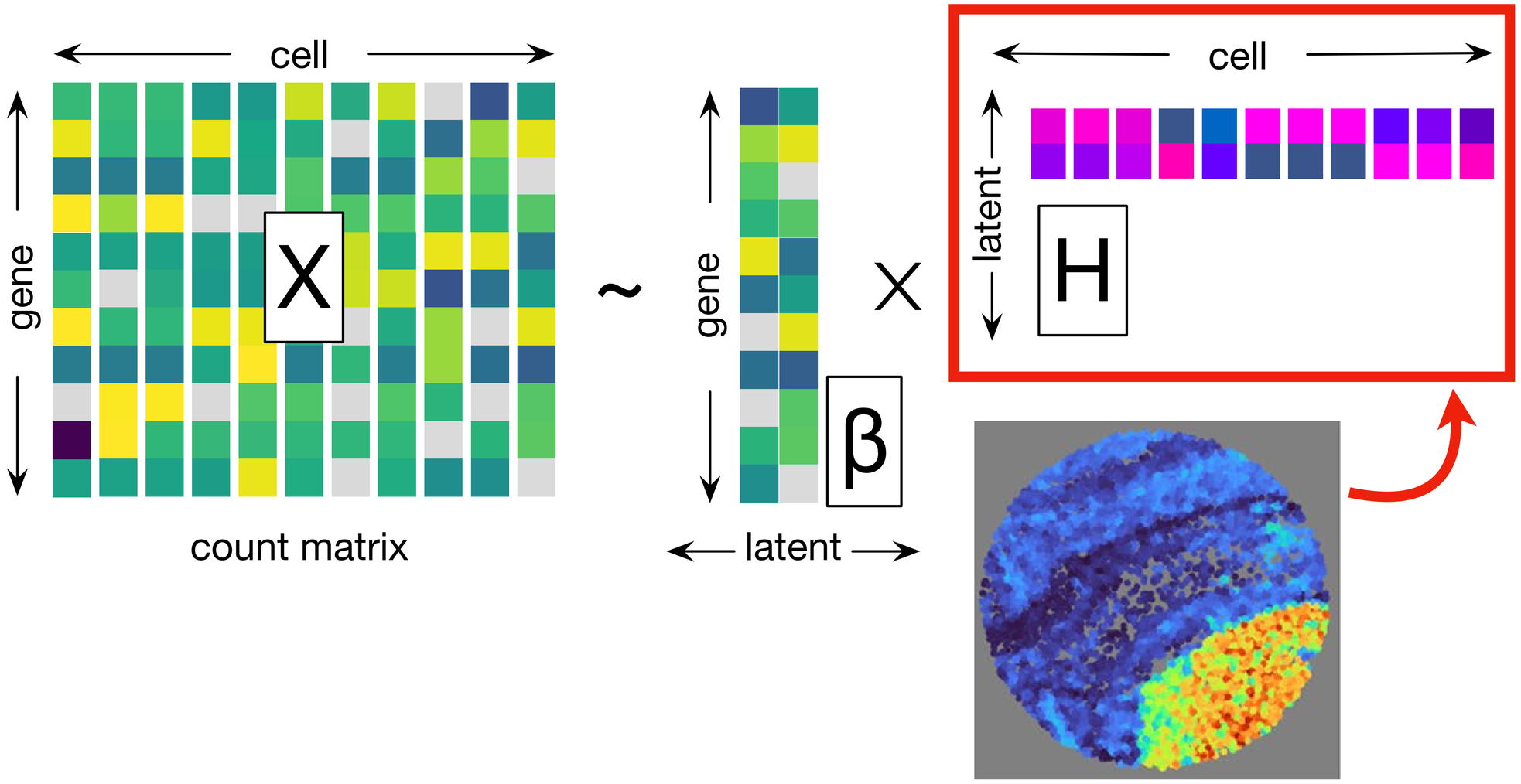
Townes and Engelhardt, *Nature Methods* (2022)

# Capture latent factors in space!




Townes and Engelhardt, *Nature Methods* (2022)

# The basic idea: incorporate cell-cell dependency based on physical locations



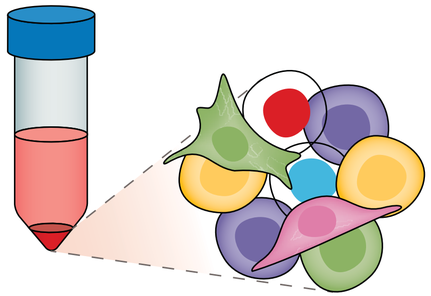
# Today’s lecture: Single-cell Part 2

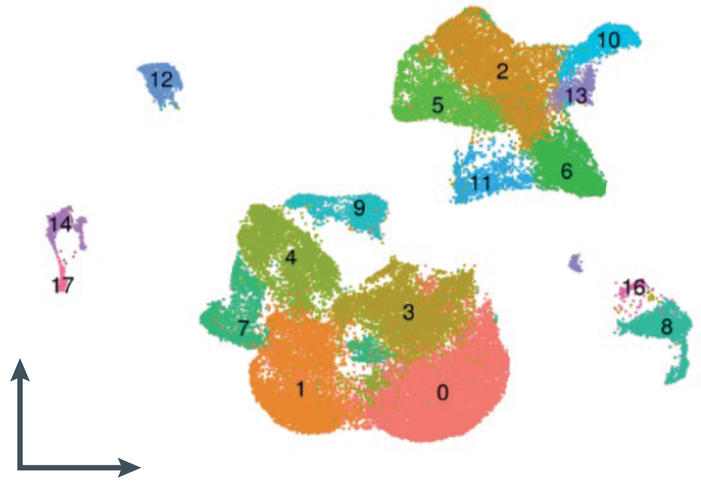
- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

# Overall work flow of Spatial Transcriptomics

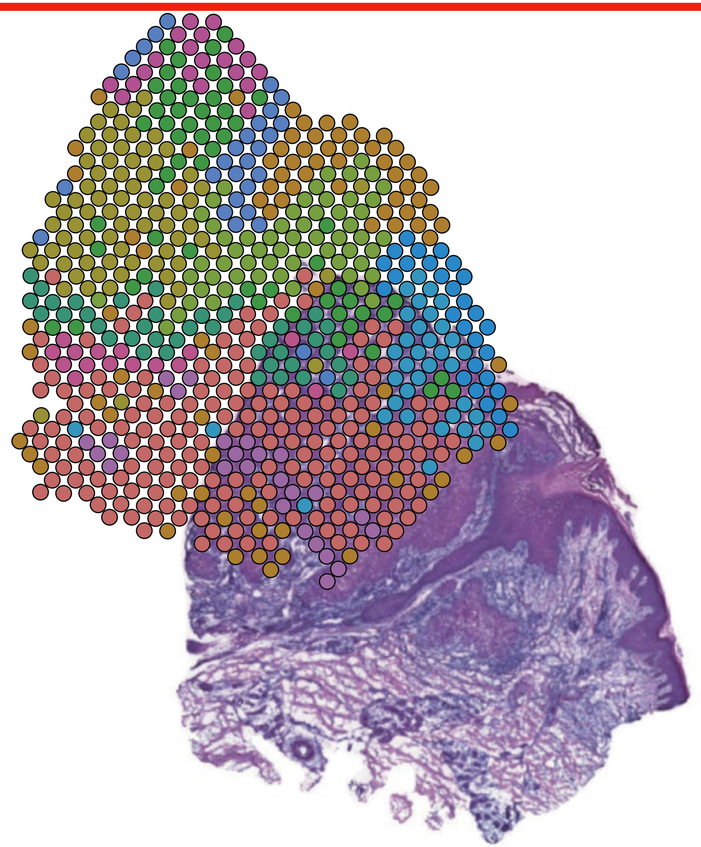
Identification of cell subpopulations

scRNA-seq






ID cell subtypes

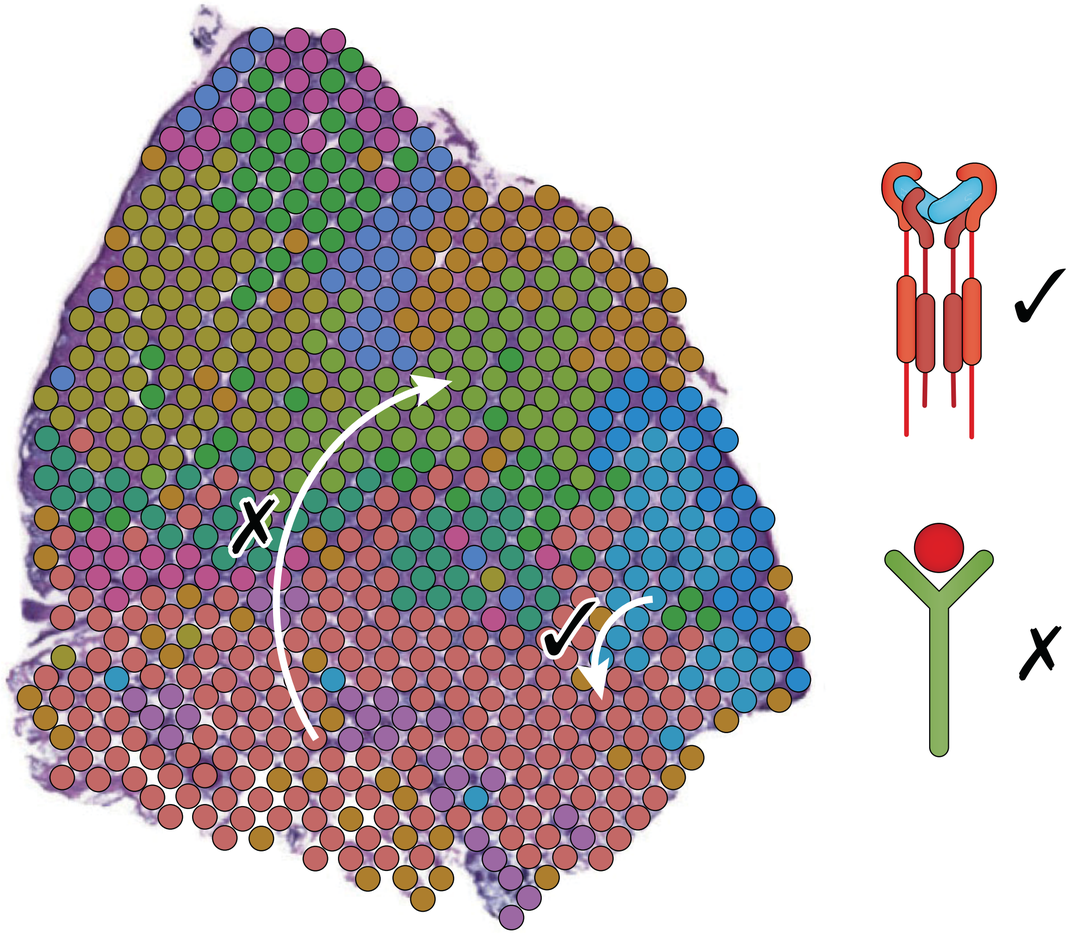


Spatial transcriptomics
(high-plex RNA imaging
or spatial barcoding)

**Today's focus**

**Longo *et al*. (2021)**

# Spatial transcriptomics → Intercellular communications

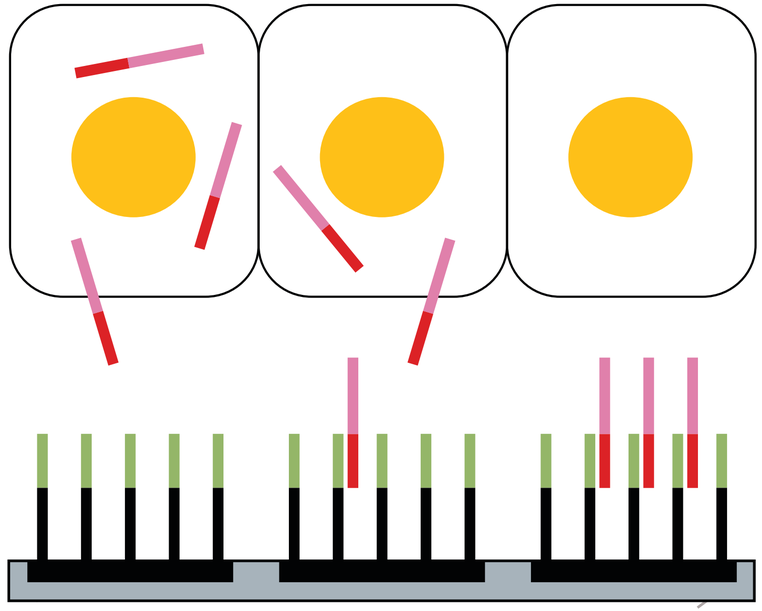


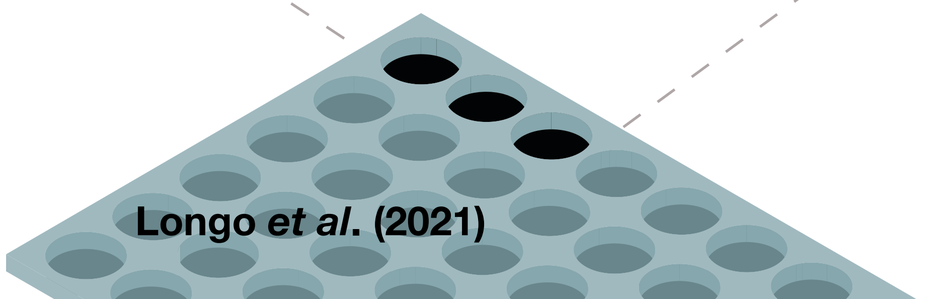
Spatially informed ligand–receptor algorithms

**Longo *et al*. (2021)**

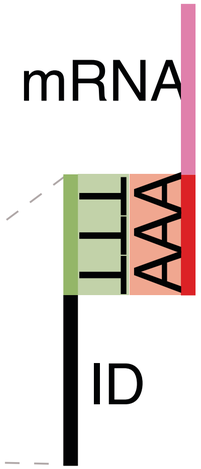
# How spatial barcoding works

**Each spot will have a unique ID**
**(ID = location barcode)**



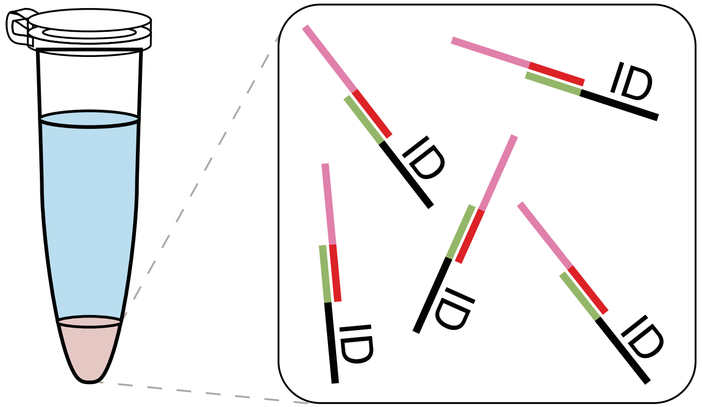


Longo *et al*. (2021)




**Poly Ts match with Poly-A tails**

**Micro-wells with many capture spots**

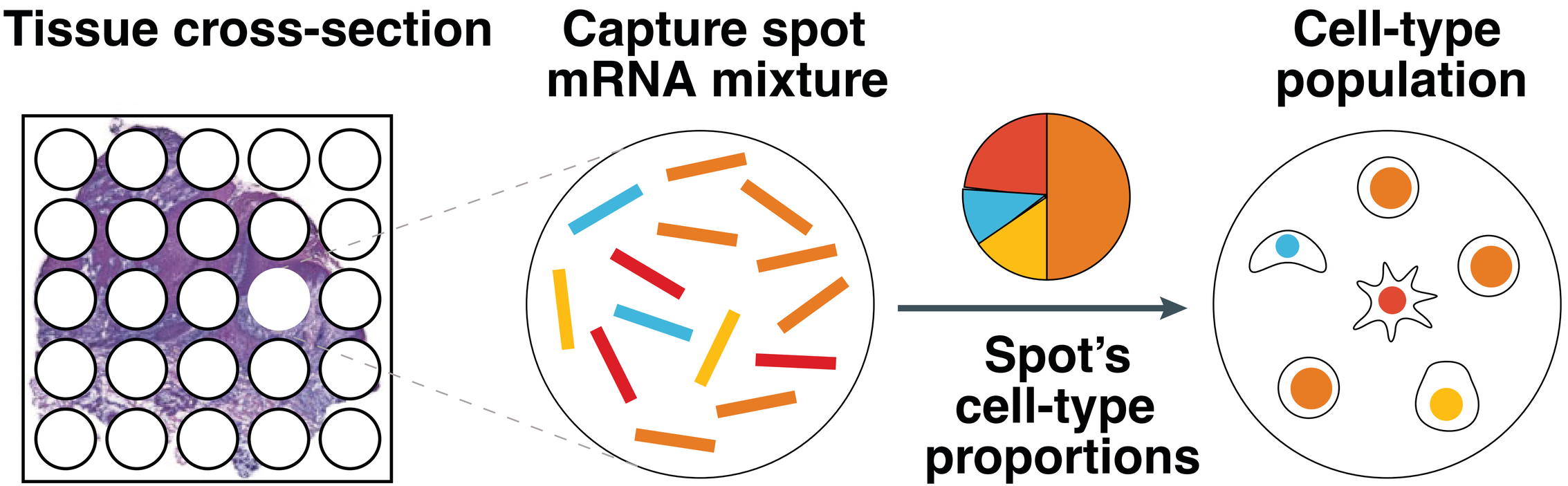



**Pooled & Seq.**

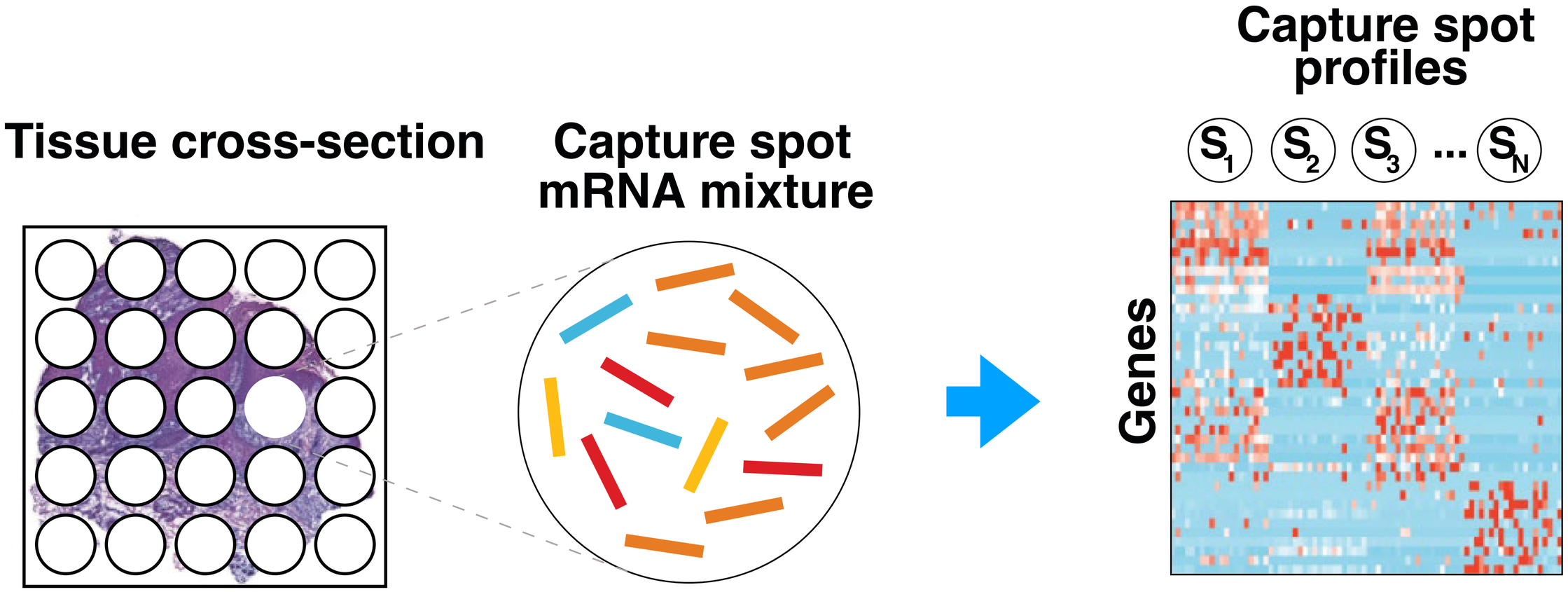
| | |
|---|---|
| ID1 | AATGG |
| ID2 | TACCC |
| ID3 | CCGGA |

…

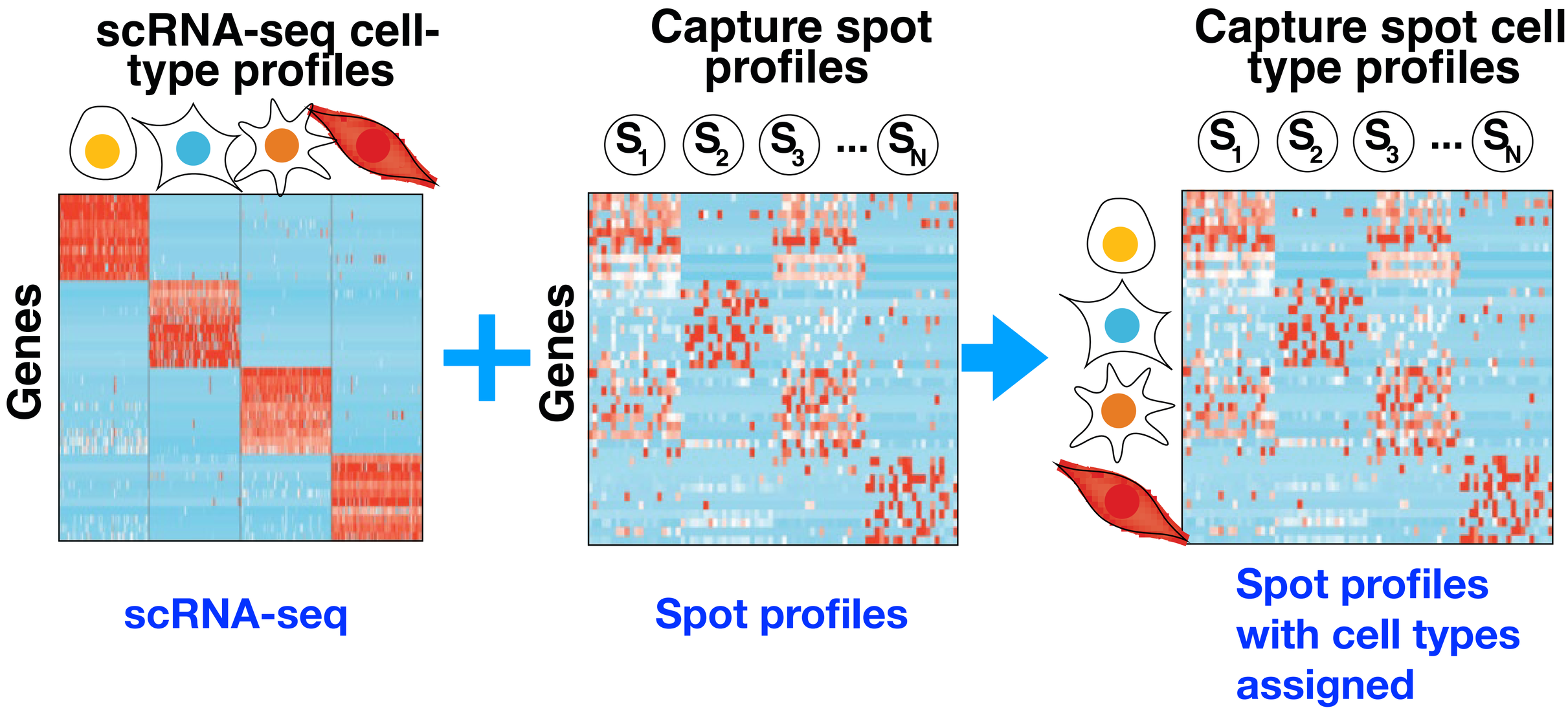
# A spot will contain a mixture of cell types/states, so we need to estimate the fraction



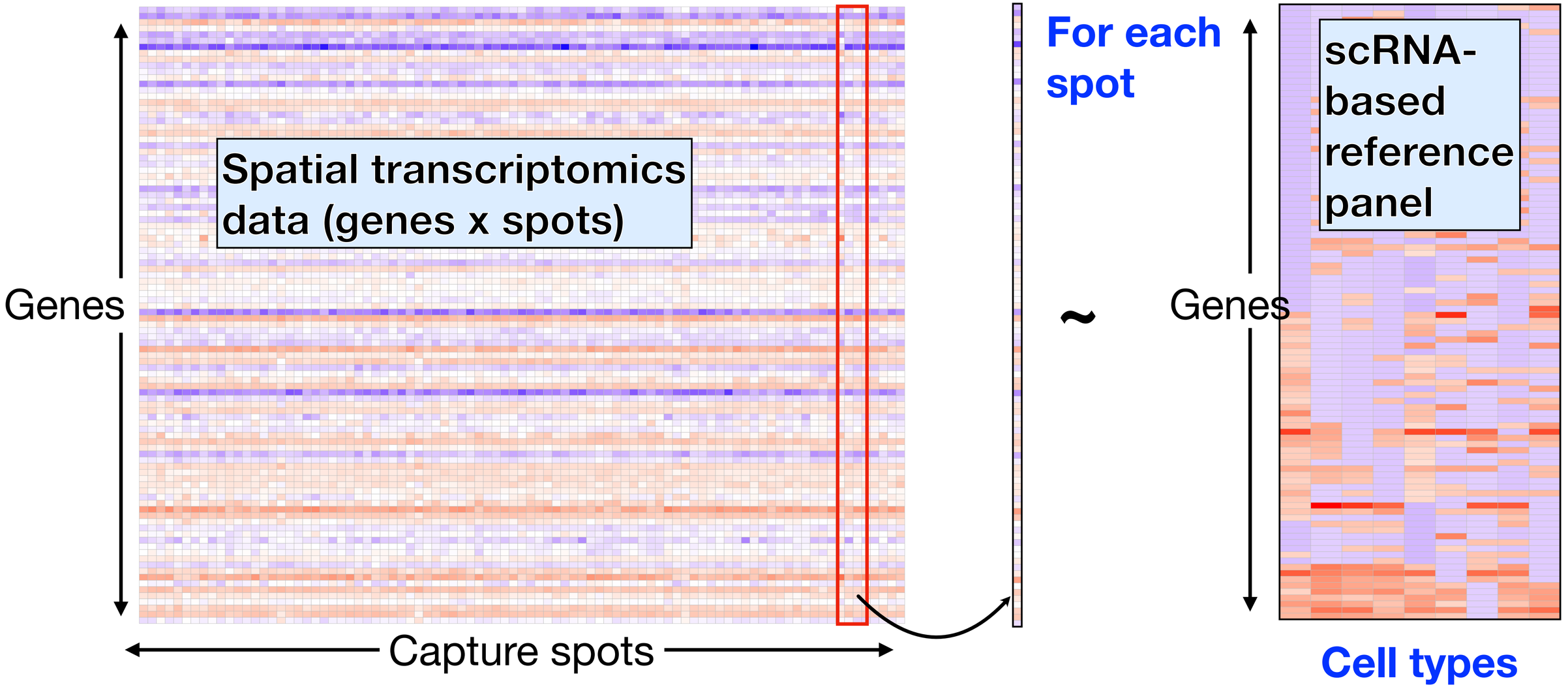
# Cells within each capture spot will be noisy, containing missing variables



# Joint analysis with high-quality scRNA-seq can improve the quality of spot profiles



# Cell type deconvolution brings cellular contexts to bulk RNA-seq data



# Negative Binomial GLM: directly modelling RNA-seq count

- $Y$: number of successfully "observed" reads in RNA-seq (~targeting)
- $r$: number of permitted "dropped" reads until $Y$ observed (~budget)
- $\rho$: success rate

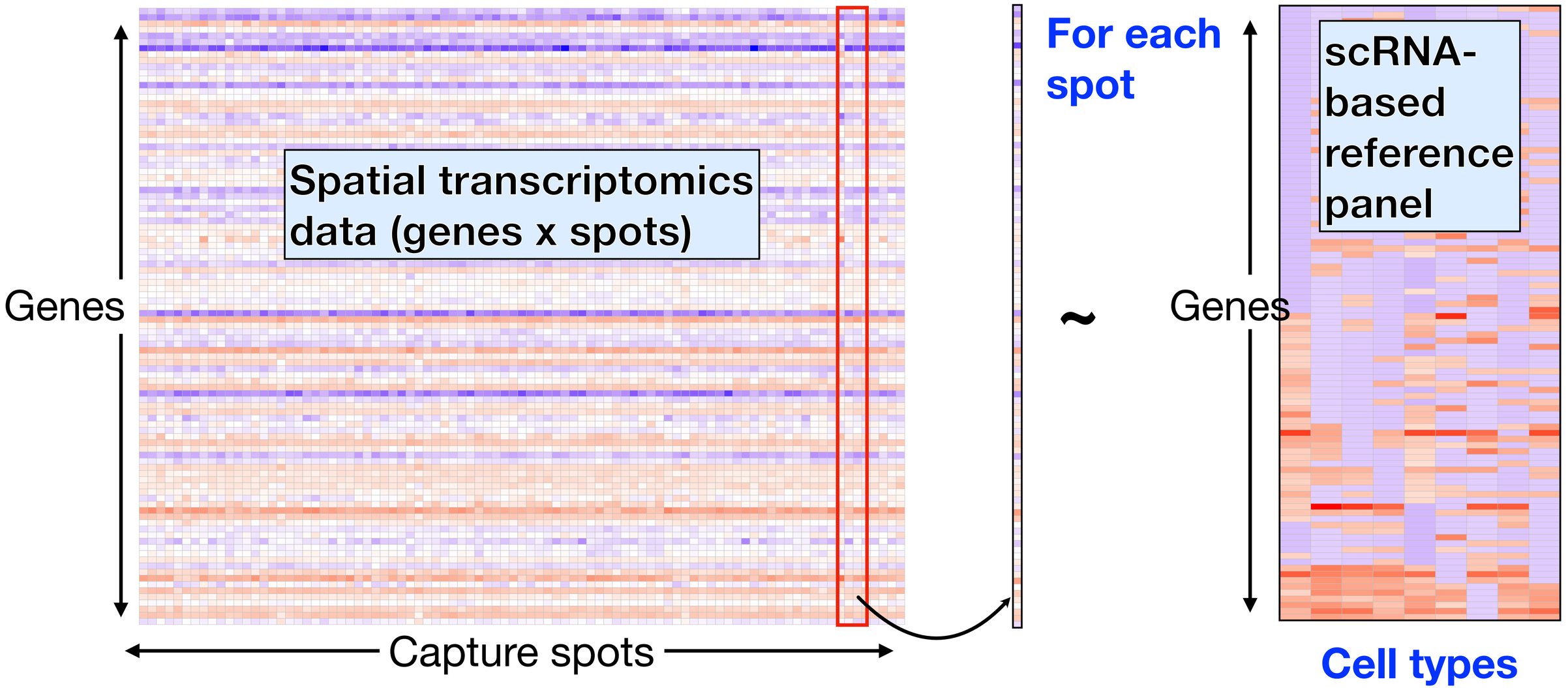
$$
\begin{aligned}
p(Y_i|\mu_i,\phi) &= \underbrace{\binom{Y_i+r-1}{Y_i}}_{\text{negative binomial}} \underbrace{\rho_i^{Y_i}}_{\text{success rate}} \underbrace{(1-\rho_i)^r}_{\text{drop rate}} \\
&= \text{NB}(Y_i|r=\phi^{-1},\rho_i=\mu_i/(\phi^{-1}+\mu_i)) \\
\text{or} &= \text{NB}(Y_i|\text{mean}=\mu_i,\text{overdispersion}=\phi)
\end{aligned}
$$

We can check:

mean: $\mathbb{E}[Y_i|r,\rho] = \rho r/(1-\rho) = \mu_i$

variance: $\mathbb{V}[Y_i|r,\rho] = \rho r/(1-\rho)^2 = \mu_i + \mu_i^2\phi$ (overdispersed mean-variance)

# Reference-based deconvolution for each spot

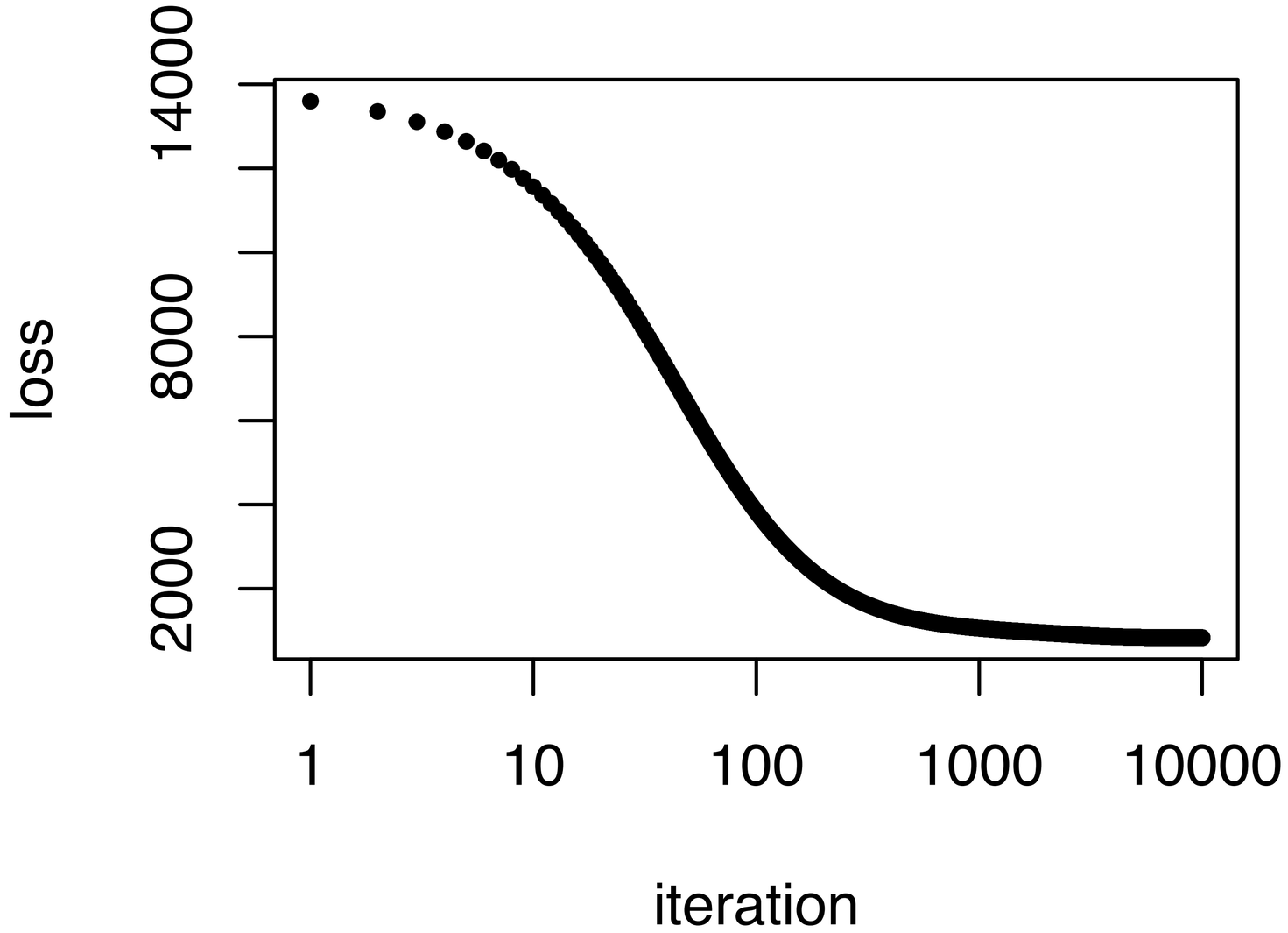


# Step 2: NB GLM for the cell type deconvolution

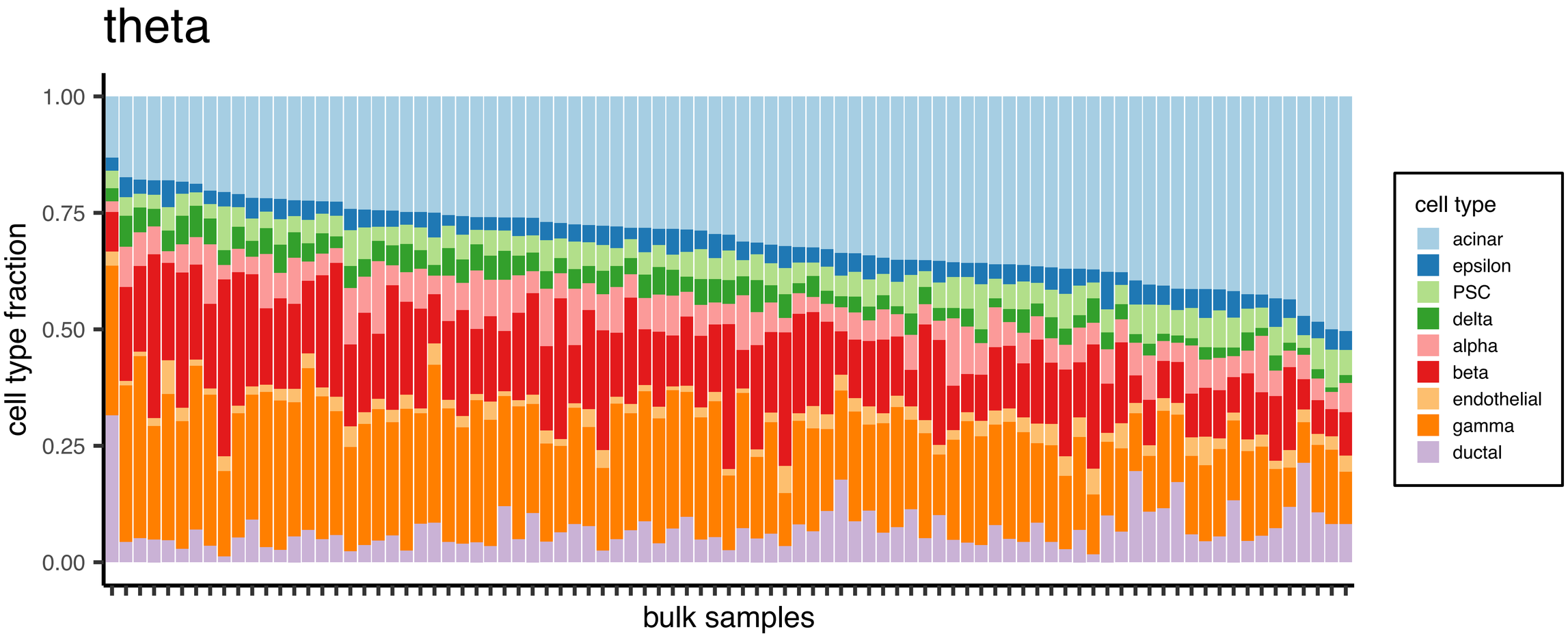
$$Y_{gi} \sim \text{NB}\left(\textbf{mean} = s_i \sum_t X_{gt}\theta_{ti}, \textbf{overdispersion} = \phi\right)$$

$$\theta_{ti} > 0, \sum_t \theta_{ti} = 1$$

# Step 3. Fit the model to the data



# Step 4. Show the cell type fraction estimates



# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**

nature cancer

TECHNICAL REPORT

https://doi.org/10.1038/s43018-022-00356-3

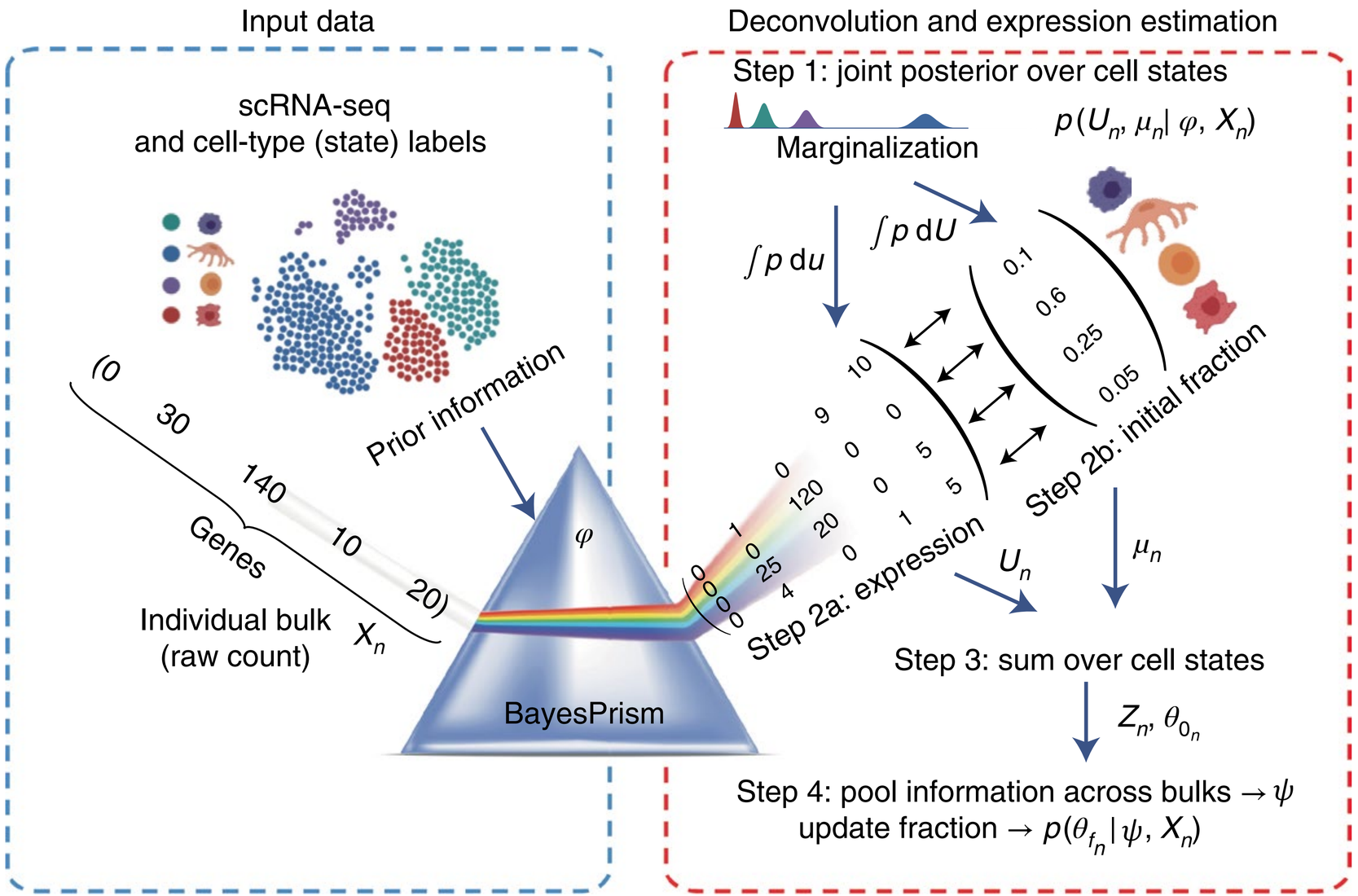
Check for updates

OPEN

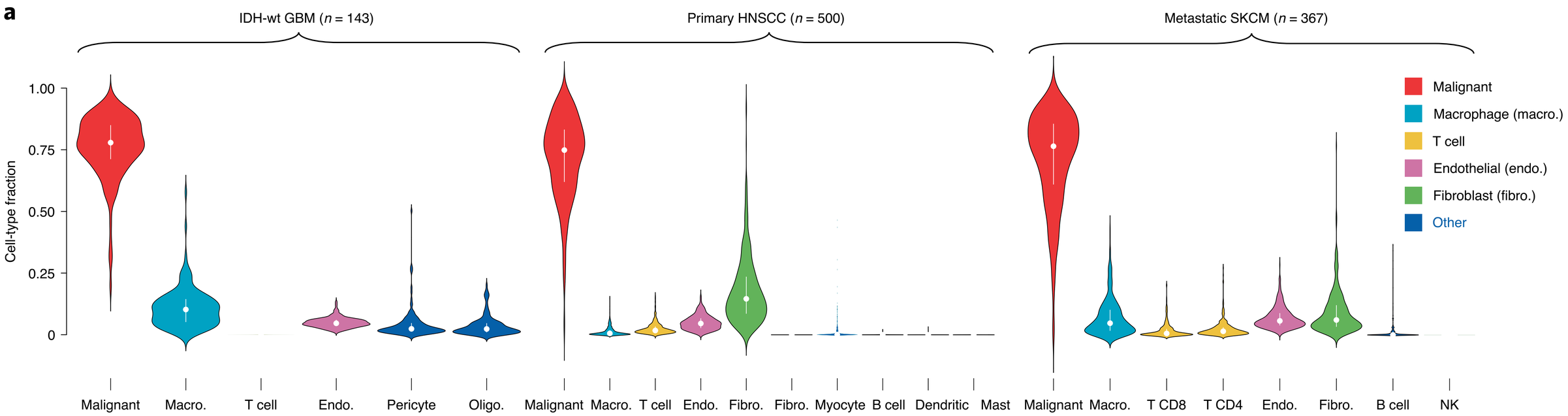
# Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology

Tinyi Chu[1,2,3] ✉, Zhong Wang[4], Dana Pe'er[3] and Charles G. Danko[1,5] ✉

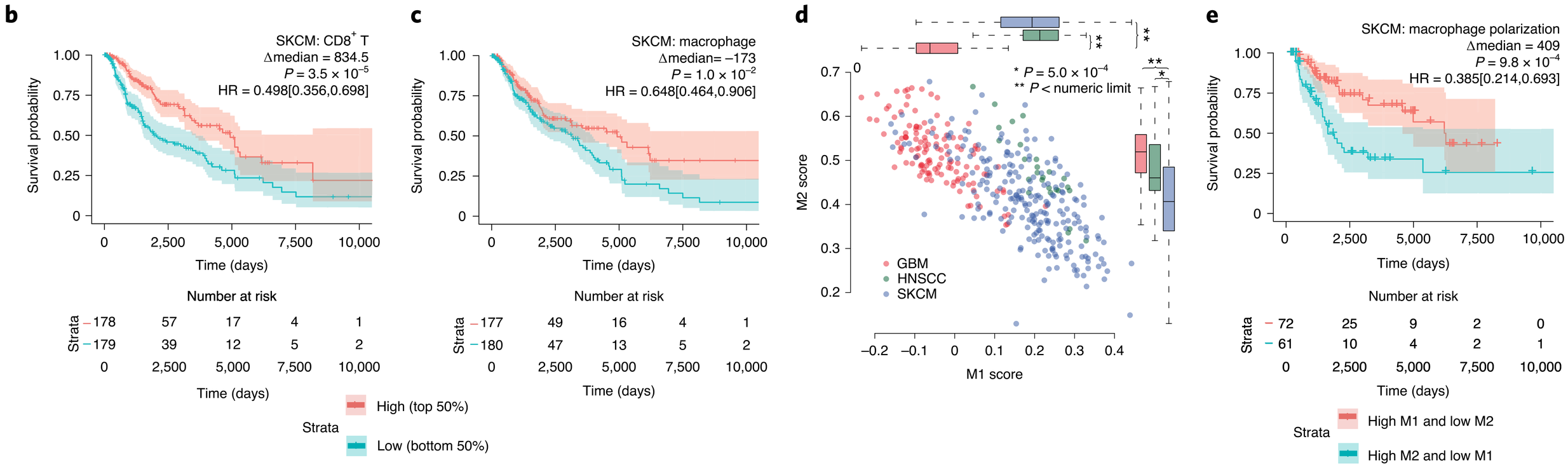
# BayesPRISM


Input data
scRNA-seq
and cell-type (state) labels
Prior information
(0 30 140 10 20)
Genes
Individual bulk
(raw count)
$X_n$
$\varphi$
BayesPrism
Deconvolution and expression estimation
Step 1: joint posterior over cell states
Marginalization
$p(U_n, \mu_n | \varphi, X_n)$
$\int p \, du$
$\int p \, dU$
0.1
0.6
0.25
0.05
Step 2b: initial fraction
Step 2a: expression
$U_n$
$\mu_n$
Step 3: sum over cell states
$Z_n, \theta_{0_n}$
Step 4: pool information across bulks → $\psi$
update fraction → $p(\theta_{f_n} | \psi, X_n)$

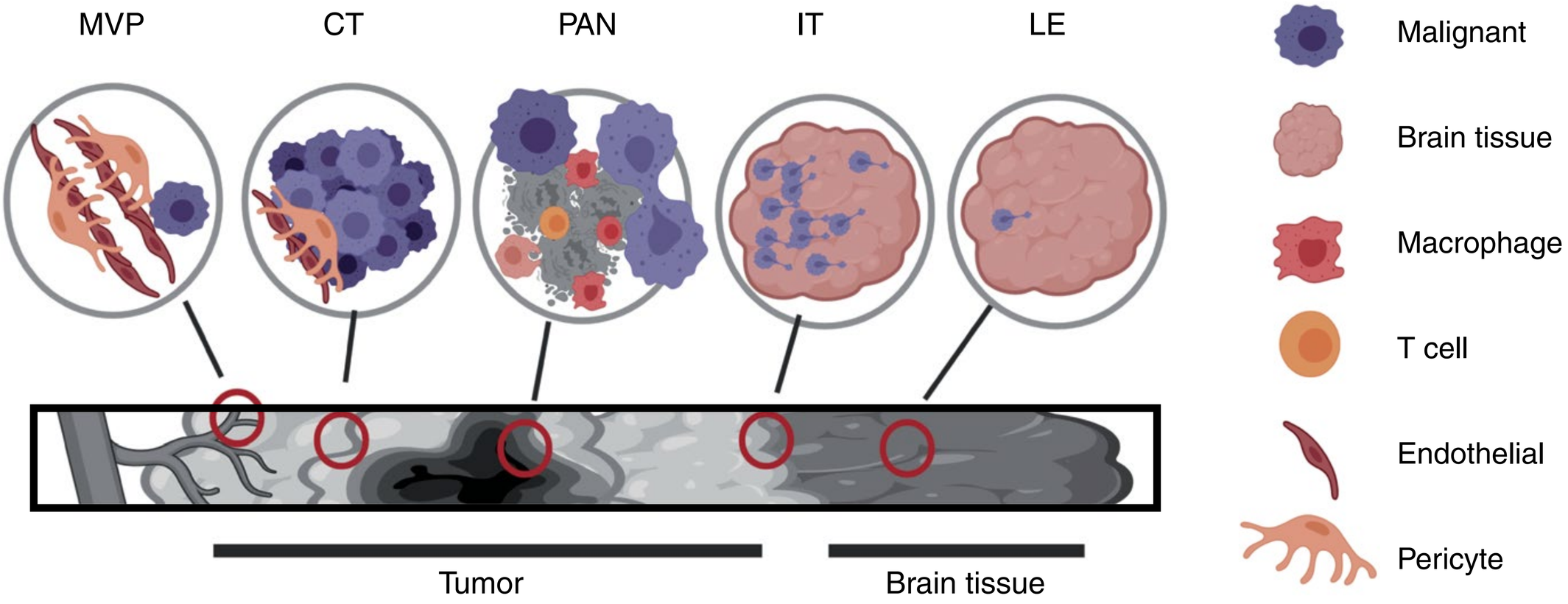
# Rare cell types in tumour samples play an important role in cancer progression



# Macrophages can make a significant difference



# Deconvolution of tissue locations within spatial transcriptomics data



# Today’s lecture: Single-cell Part 2

- **Advanced topics in single-cell RNA-seq analysis**
  - Probabilistic Topic model as a principled framework for modelling
  - Optimal transport and trajectory inference
  - RNA velocity analysis
- **Multimodal data integration methods**
  - scDNA and scRNA joint analysis
  - Single-cell epigenomics and transcriptomics
  - Spatial transcriptomics
- **Don’t abandon bulk sequencing data**