

# Single cell omics

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Spring 2021

Daifeng Wang

[daifeng.wang@wisc.edu](mailto:daifeng.wang@wisc.edu)

*Thanks to Ting Jin for slides!*

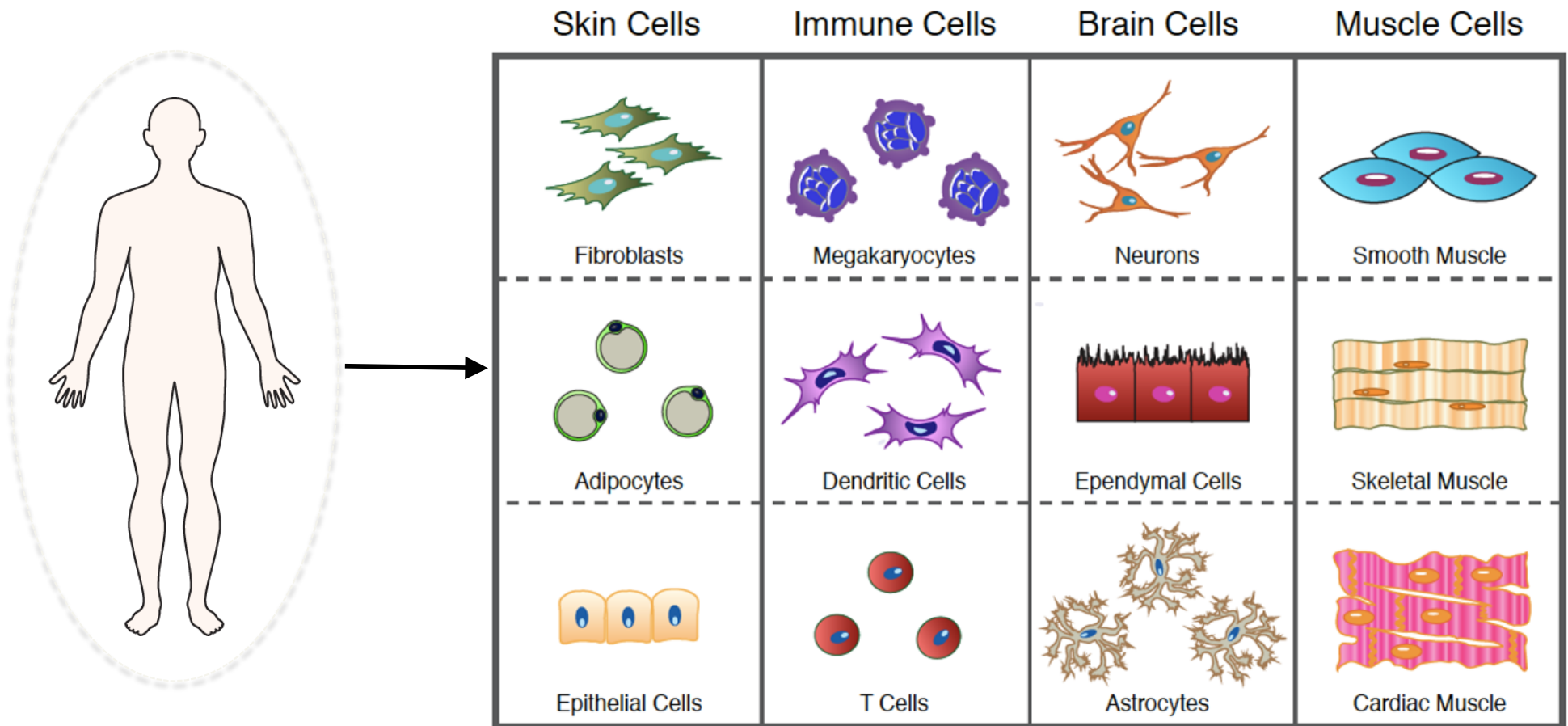
These slides, excluding third-party material, are licensed under [CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/) by Mark Craven, Colin Dewey, Anthony Gitter and Daifeng Wang

# Outline

- Introduction on single cell sequencing
- Single-cell RNA sequencing data processing & analysis
- Cell-type gene regulatory networks
- scATAC-seq data & analysis plus integration with scRNA-seq
- Single cell deconvolution
- Dropout

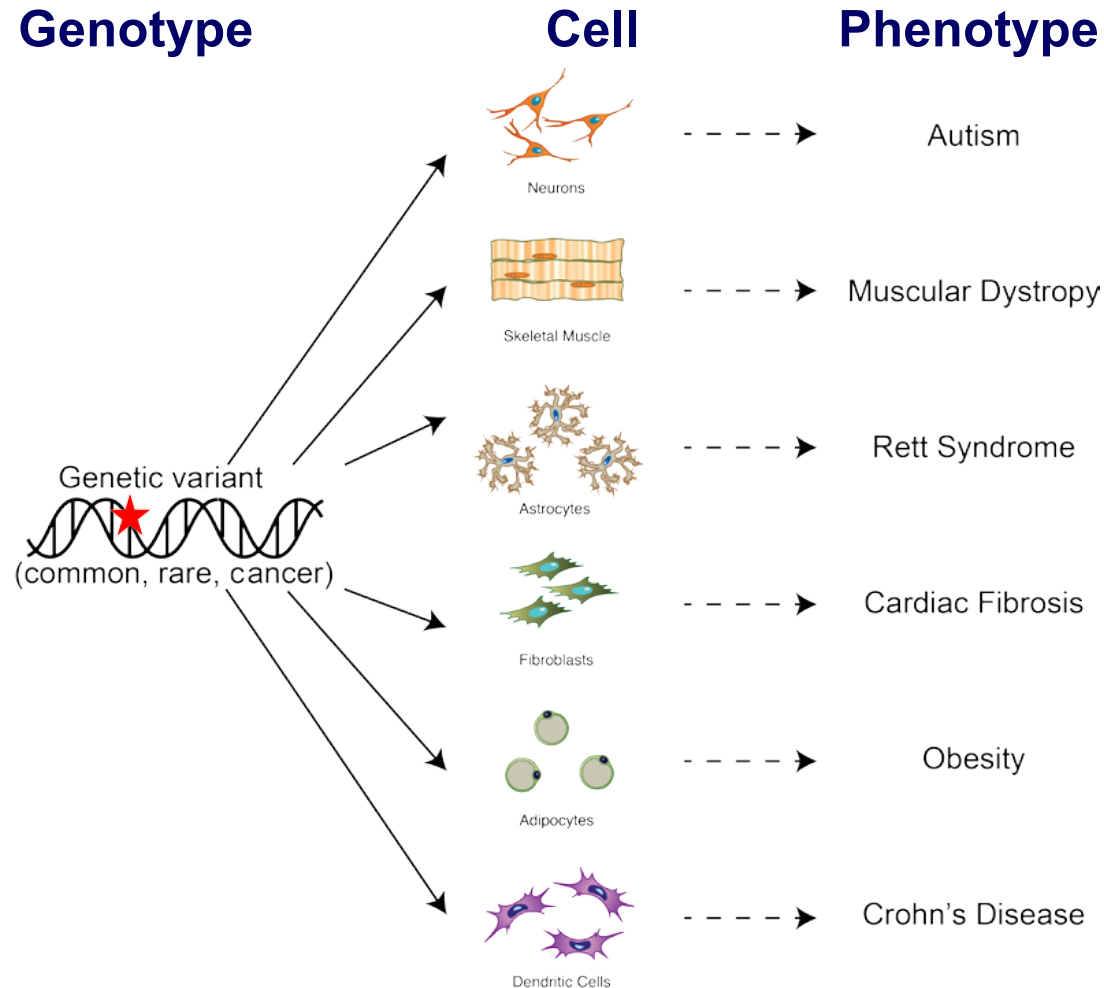


# Why study single cells?



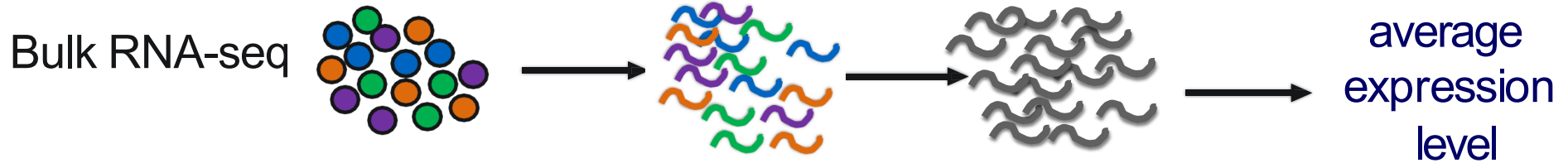
Cells are our core constituents, are classified by characteristic molecules, structures, and functions

# Why study single cells?

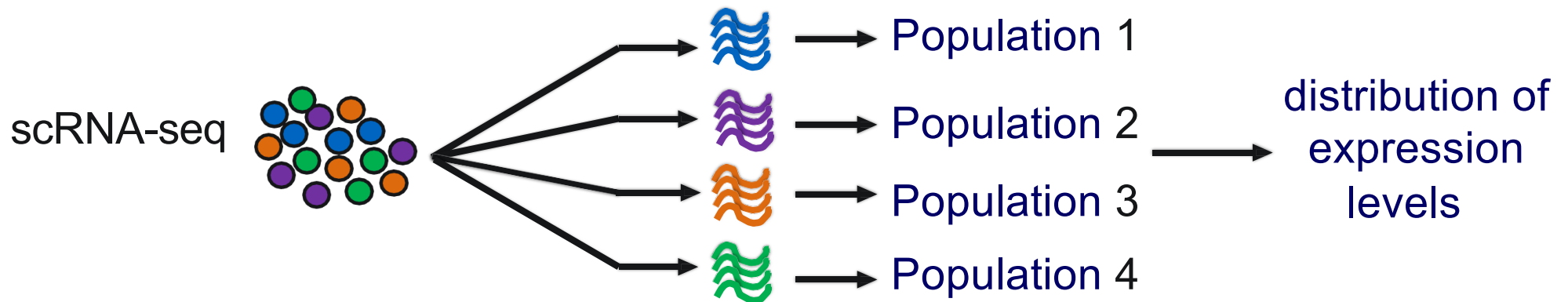


Cells are key intermediate from genotype to phenotype, also are essential for functional dissection of genetic variants

# Bulk vs scRNA-seq

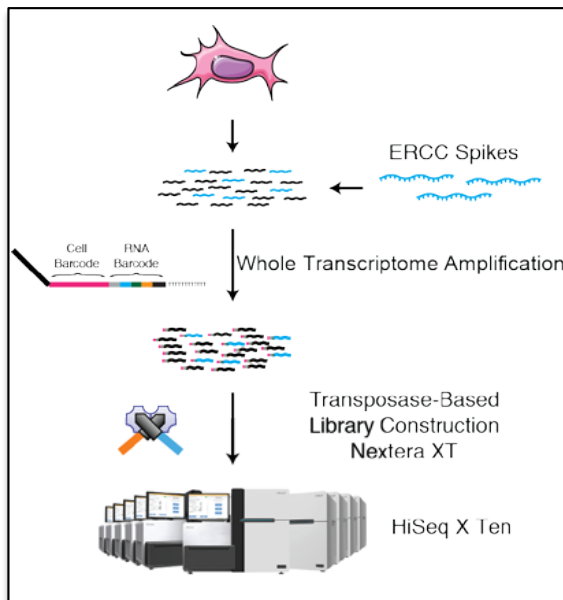


- comparative transcriptomics
- quantifying expression signatures from ensembles
- insufficient for studying heterogeneous system

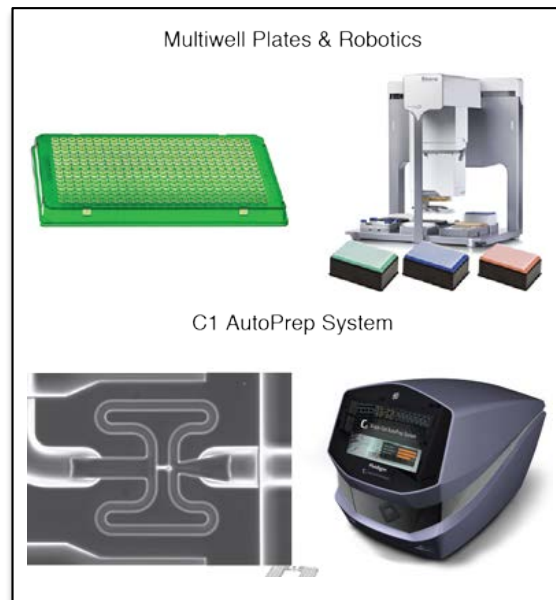


- inference of gene regulatory networks across the cells
- heterogeneity of cell responses
- cell type identification

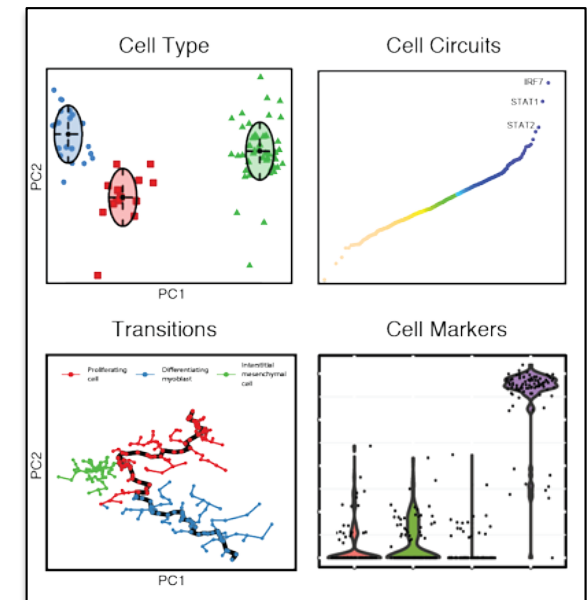
# How can we study single cell?



✓ Core technology



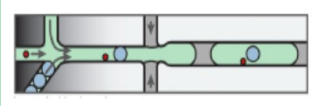




✓ Sample prep



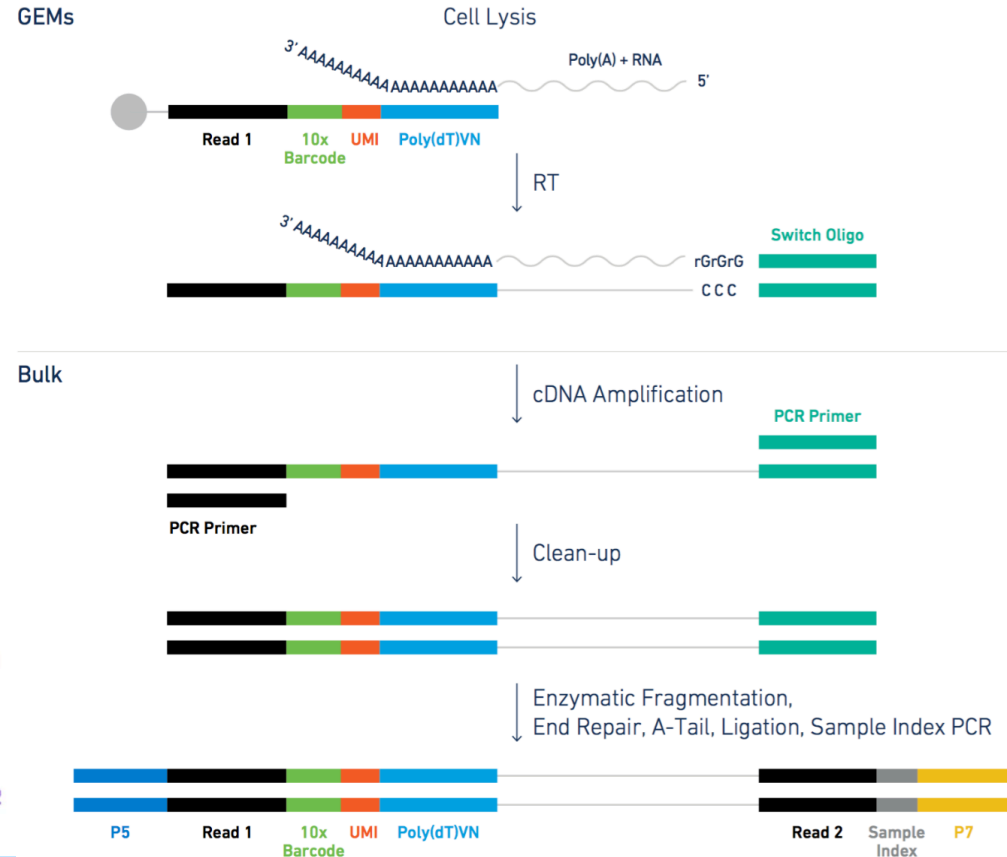
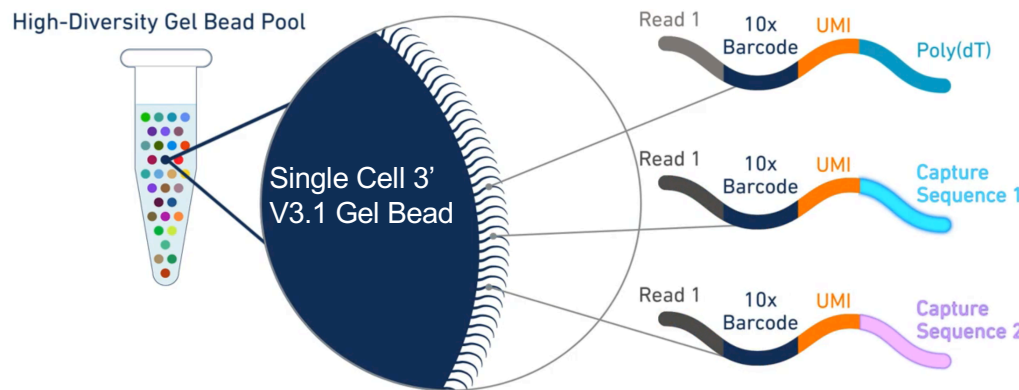
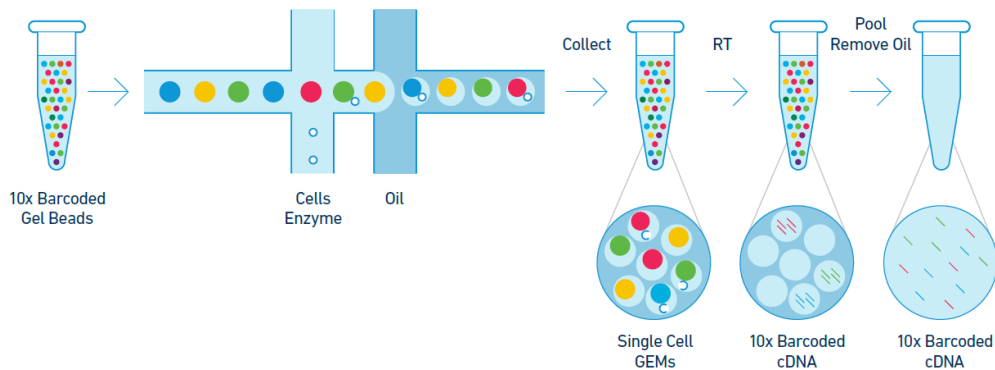
✓ Computation

2012: 18 cells → 2020: ~100,000 cells

# Single cell technology

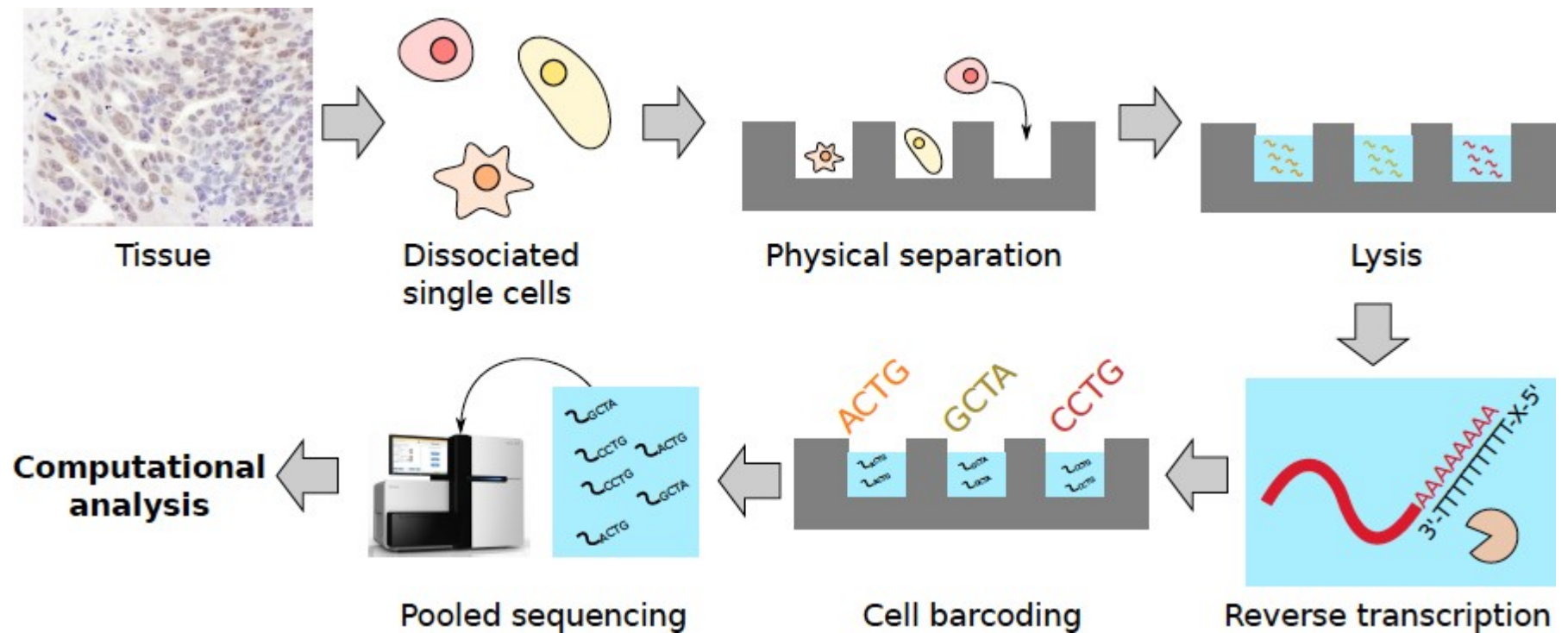
	inDrops	10x Genomics	Drop-seq	Seq-well (Honeycomb)	SMART-seq
Cell capture efficiency	~70-80%	~50-70%	~10%	~80%	~80%
Time to capture 10k cells	~30min	10min	1-2 hours	5-10min	--
Encapsulation type	Droplet 	Droplet 	Droplet 	Nanolitre well 	Plate-based 
Library prep	CEL-seq Linear amplification by IVT	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification
Commercial	Yes	Yes	--	Yes (Summer 2020)	Yes
Cost (~\$ per cell)	~0.06	~0.2	~0.06	~0.15	1
Strengths	<ul style="list-style-type: none"> <li>Good cell capture</li> <li>Cost-effective</li> <li>Real-time monitoring</li> <li>Customizable</li> </ul>	<ul style="list-style-type: none"> <li>Good cell capture</li> <li>Fast and easy to run</li> <li>Parallel sample collection</li> <li>High gene / cell counts</li> </ul>	<ul style="list-style-type: none"> <li>Cost-effective</li> <li>Customizable</li> </ul>	<ul style="list-style-type: none"> <li>Good cell capture</li> <li>Cost-effective</li> <li>Real-time monitoring</li> <li>Customizable</li> </ul>	<ul style="list-style-type: none"> <li>Good cell capture</li> <li>Good mRNA capture</li> <li>Full-length transcript</li> <li>No UMI</li> </ul>
Weaknesses	Difficult to run	Expensive	Difficult to run & low cell capture efficiency	Available Soon	Expensive

# 10x Genomics



- **Step1** : GEM Generation & Barcoding
- GEM: Gel Beads-in-emulsion partition that encapsulates each tiny micro-reaction within the Chromium system
- **Step2** : Post GEM-RT Cleanup & cDNA Amplification
- **Step3**: 3' Gene Expression Library Construction

# Workflow for typical scRNA-seq experiment

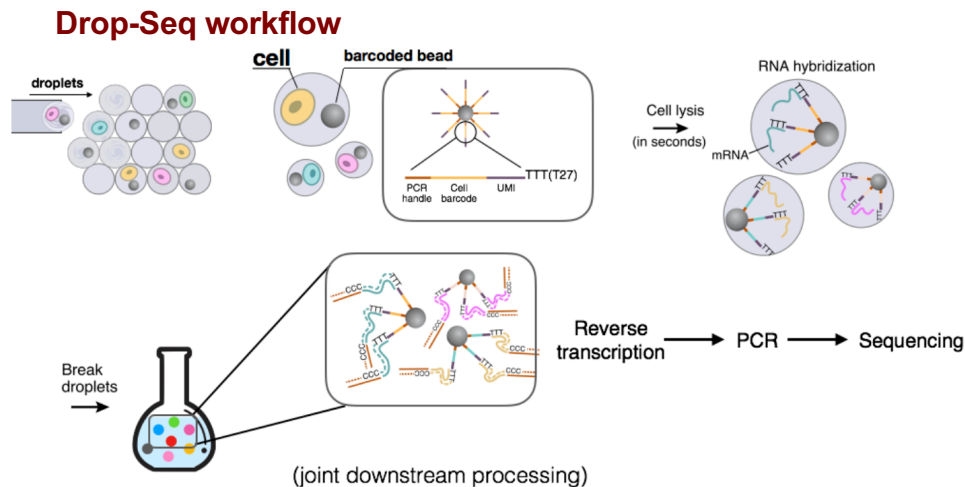




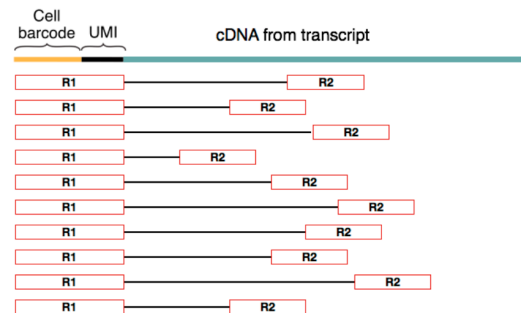
# UMI

## Unique molecular identifier

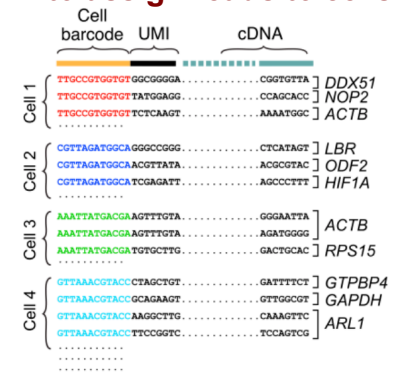
- UMIs are short (4-10bp) random barcodes added to transcripts during reverse-transcription. They enable sequencing reads to be assigned to individual transcript molecules and thus the removal of amplification noise and biases from scRNA-Seq data



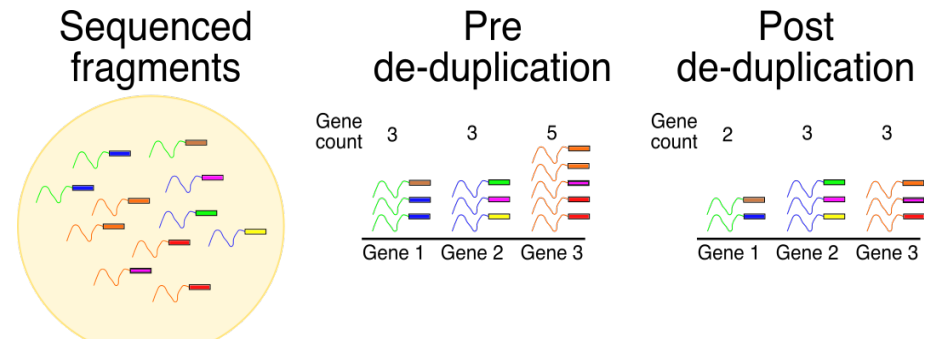
### Biased paired-end reads



### Grouping barcodes to assign reads to cells



- They reduce the amplification noise by allowing (almost) complete de-duplication of sequenced fragments



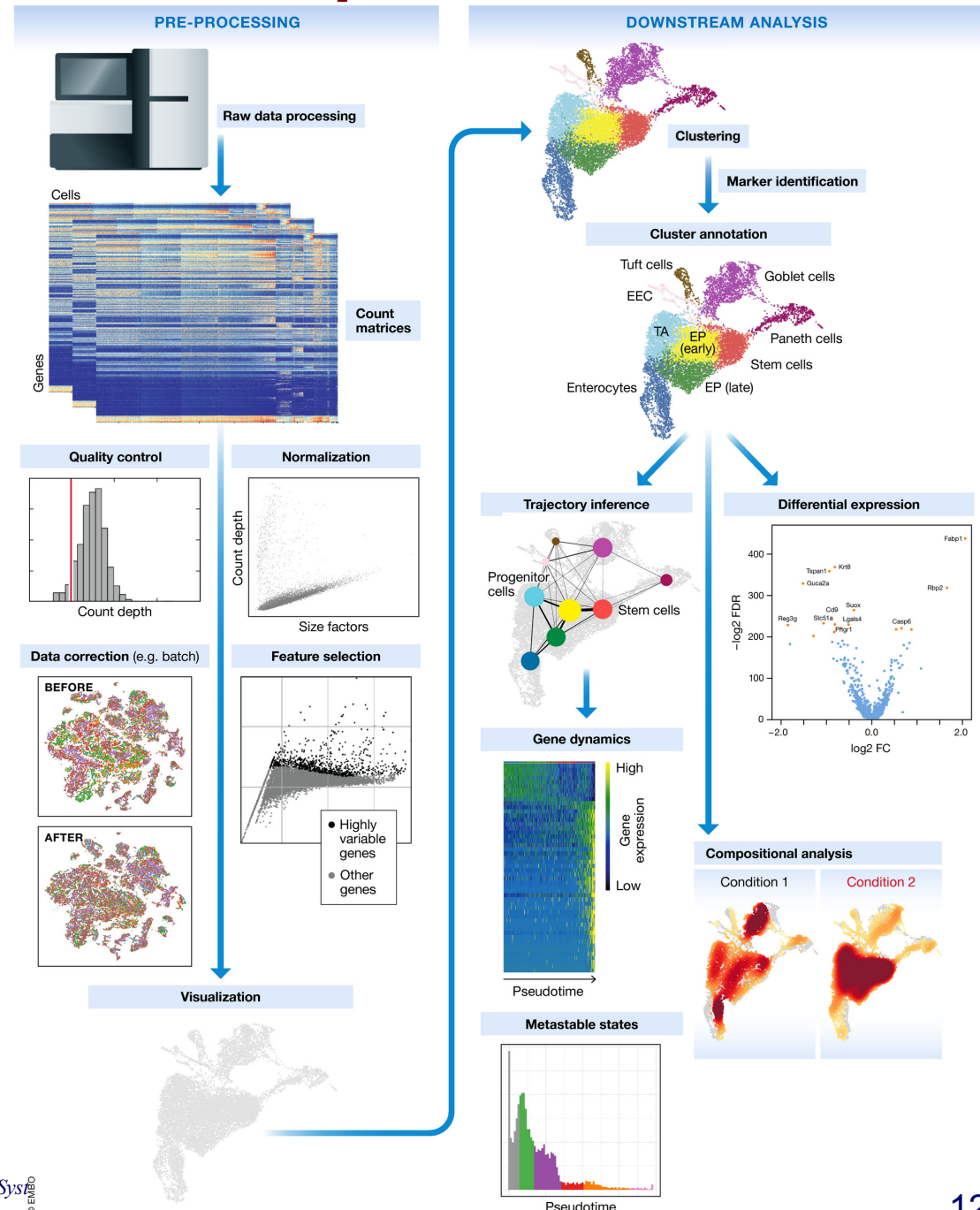


# Single-cell RNA sequencing data processing & analysis

- Dimensional reduction
  - Non-linear: t-SNE, UMAP
  - Linear: PCA
- Cell clustering
  - K-means
  - hierarchical clustering
  - graph-based clustering
- Cell type annotation
  - SingleR
- Cell type markers
  - Differential expression analysis
- Pseudo timing

# Single-cell RNA-seq workflow

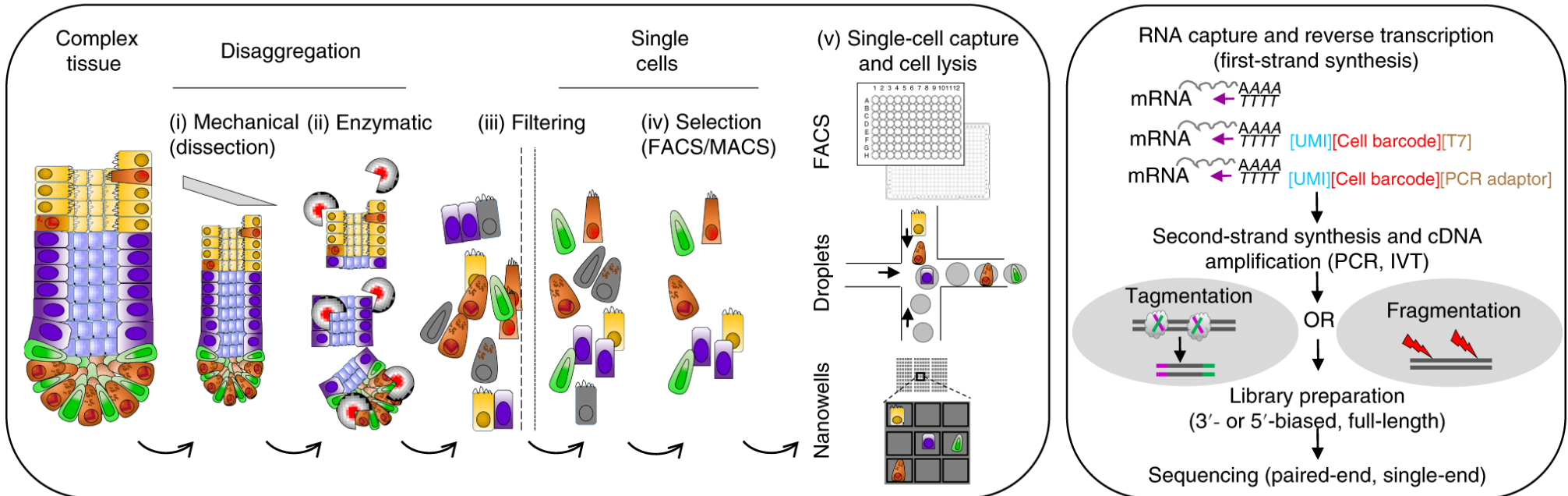
- Generation of the count matrix
- Quality control of the raw counts
- Marker gene identification
  - Single-cell differential expression analysis



# Single-cell RNA-seq process

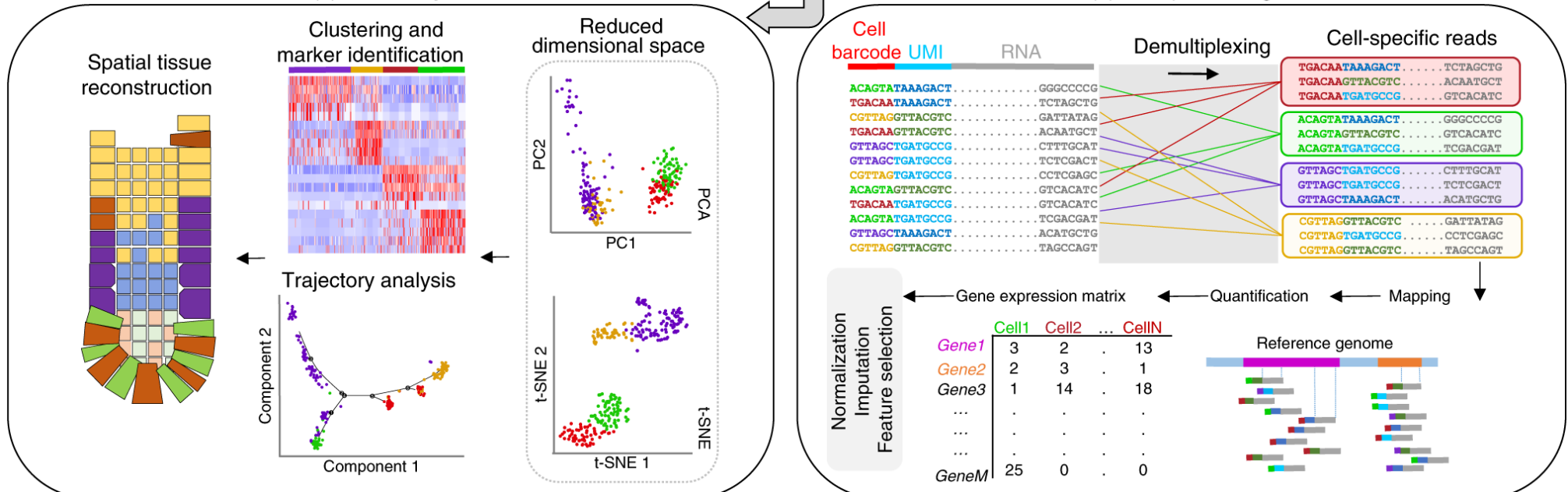
(1) Sample preparation

(2) Single-cell RNA sequencing



(4) Data analysis

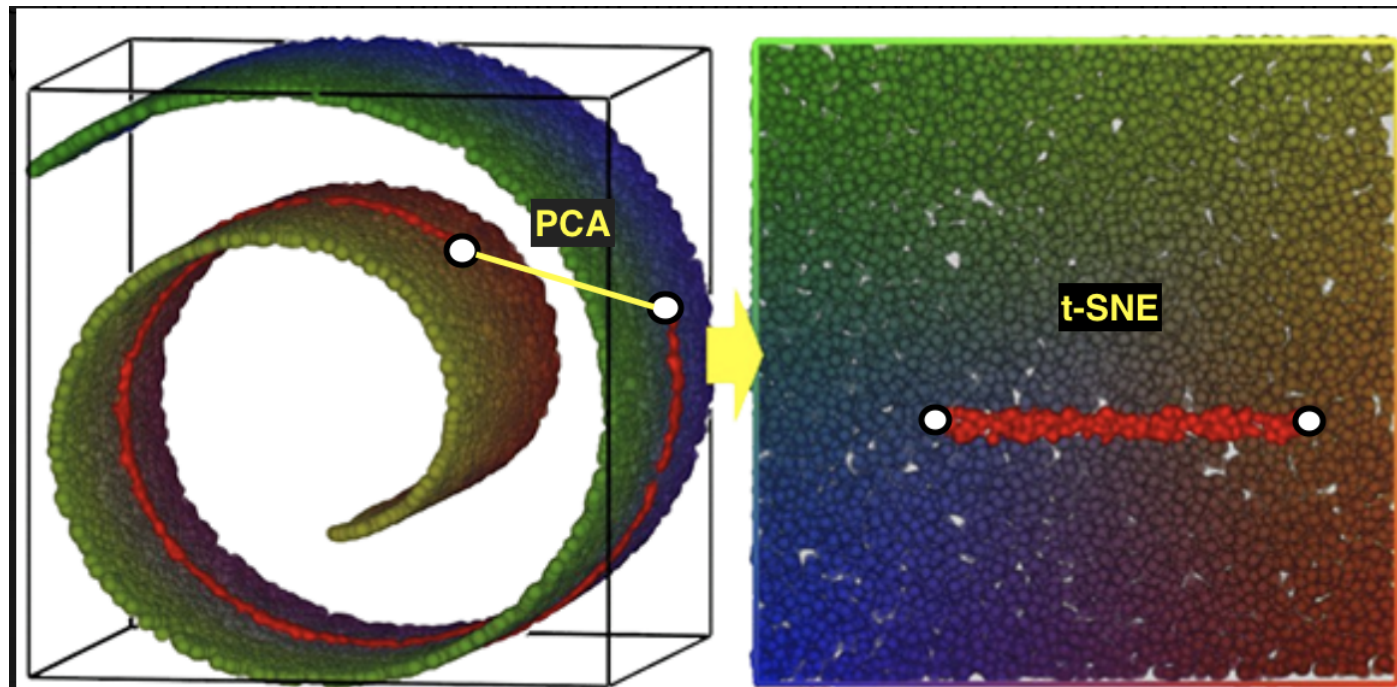
(3) Data processing



# tSNE

## t-Stochastic Neighborhood Embedding

- Compute an  $N \times N$  similarity matrix in the high-dimensional input space
- Define an  $N \times N$  similarity matrix in the low-dimensional embedding space
- Define cost function - sum of  $KL$  divergence between the two probability distributions at each point
- Iteratively learn low-dimensional embedding by minimizing the cost function using gradient descent



Src: <http://web-ext.u-aizu.ac.jp/~shigeo/home.html>

- **Modulization**

Given a collection of points  $X = \{x_1, \dots, x_n\} \subset R^d$ , find a collection of points  $Y = \{y_1, \dots, y_n\} \subset R^{d'}$ , where  $d \ll d'$ .  $p_{ji}$  and  $q_{ji}$  measure the conditional probability that a point  $i$  would pick point  $j$  as it's nearest neighbor, in high (p) and low (q) dimensional space respectively.

- **Similarity matrix at high dimension:**

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\tau_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\tau_i^2\right)} \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

where  $\tau_i^2$  is the variance for the Gaussian distribution centered around  $x_i$

- **Similarity matrix at low dimension:**

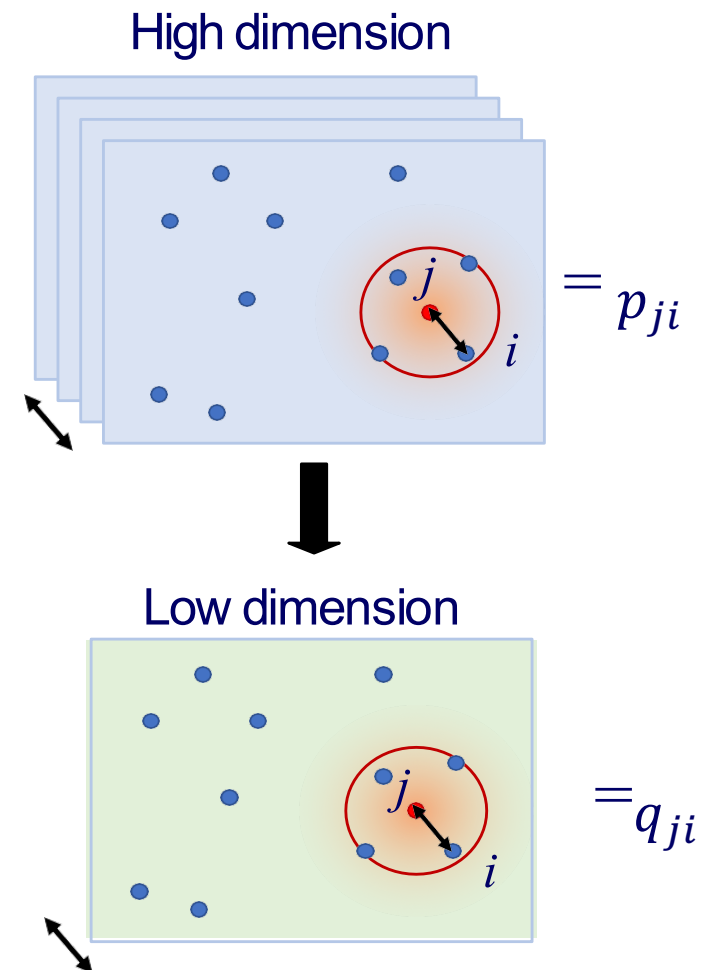
$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}$$

- **Cost Function (Kullback-Leibler divergence):**

$$C = \sum_i KL(P_i || Q_i) = \sum_{i=1}^n \sum_{j=1}^n p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- **Derivatives:**

$$\frac{dC}{dy_i} = 4 \sum_{j=1, j \neq i}^n (p_{ij} - q_{ij})(1 + \|y_i - y_j\|^2)^{-1}(y_i - y_j)$$

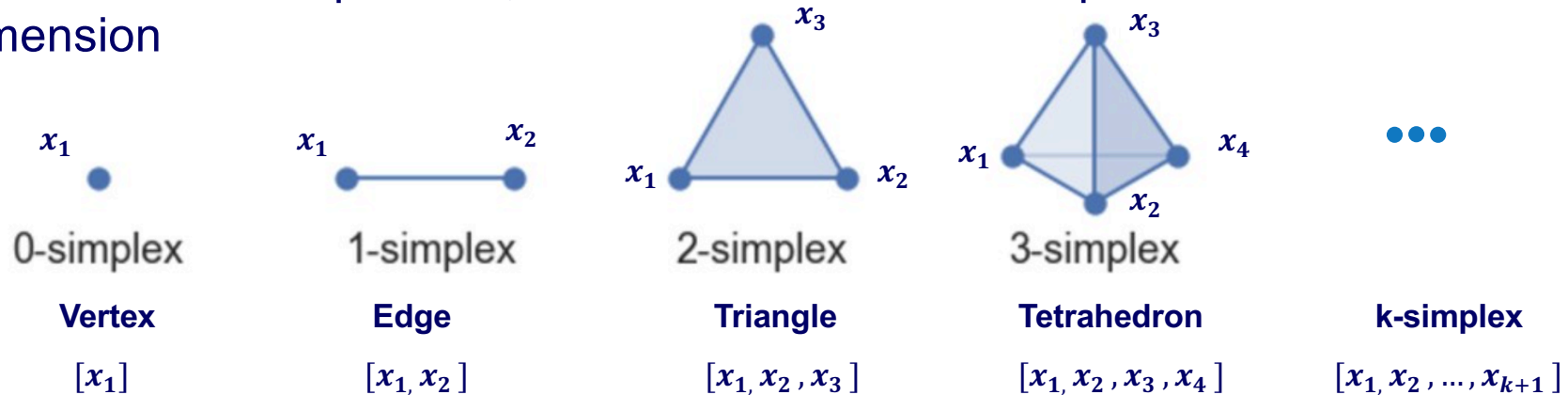




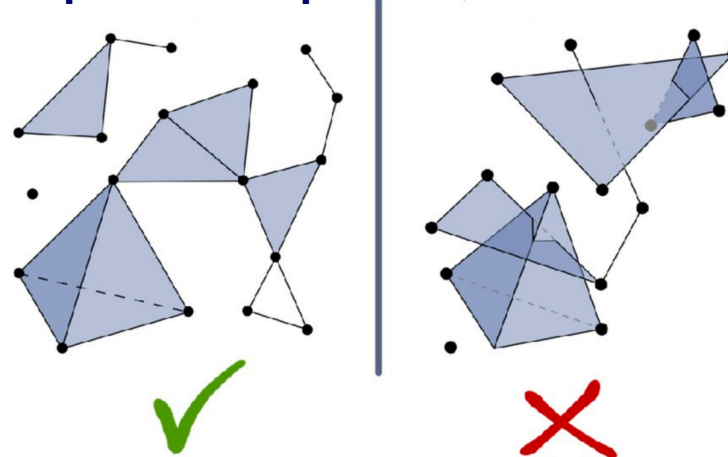
# UMAP

## Uniform **M**anifold **A**pproximation and **P**rojection

- UMAP is based on topological structures in the multidimensional space (simplices)
- Simplicial complexes are a means to construct topological spaces out of simple combinatorial components, and calculate the relative point distances in the low dimension



### Simplicial complex      Collection of simplices



- <https://www.youtube.com/watch?v=nq6iPZVUxZU>
- [https://nbisweden.github.io/excelerate-scRNAseq/session-dim-reduction/lecture\\_dimensionality\\_reduction.pdf](https://nbisweden.github.io/excelerate-scRNAseq/session-dim-reduction/lecture_dimensionality_reduction.pdf)
- [https://umap-learn.readthedocs.io/en/latest/how\\_umap\\_works.html](https://umap-learn.readthedocs.io/en/latest/how_umap_works.html)

- **Modulization**

Let  $X = \{x_1, \dots, x_n\}$  be the input dataset, with a metric (or dissimilarity measure)  $d : X \times X \rightarrow \mathbb{R}_{\gg 0}$ . Given an input hyperparameter  $k$ , for each  $x_i$  we compute the set  $\{x_{i1}, \dots, x_{ik}\}$  of the  $k$  nearest neighbors of  $x_i$  under the metric  $d$ . Given a collection of points  $X$ , find a collection of points  $Y = \{y_1, \dots, y_n\} \subset R^{d'}$ , where  $d \ll d'$ .  $p_{ji}$  and  $q_{ji}$  measure the conditional probability that a point  $i$  would pick point  $j$  as its nearest neighbor, in high ( $p$ ) and low ( $q$ ) dimensional space respectively.

- **Similarity matrix at high dimension:**

$$\rho_i = \min\{d(x_i, x_{i_j}) \mid 1 \leq j \leq k, d(x_i, x_{i_j}) > 0\}$$

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

$$k = 2 \sum_i p_{ij}$$

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j}p_{j|i}$$

- **Similarity matrix at low dimension:**

$$q_{ij} = (1 + a(y_i - y_j)^{2b})^{-1}$$

UMAP uses the family of curves  $1/a \times y^{2b}$  for modelling distance probabilities in low dimensions,  $a$  and  $b$  are hyperparameters

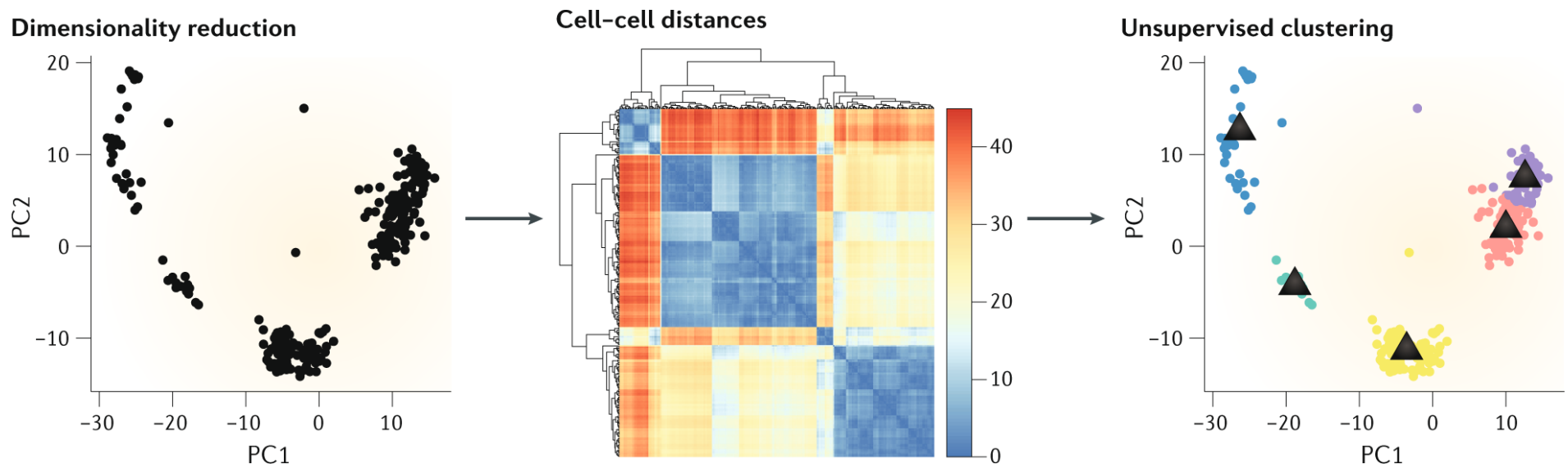
- **Cost Function ( binary cross-entropy (CE)):**

$$CE(X, Y) = \sum_i \sum_j \left[ p_{ij}(X) \log \left( \frac{p_{ij}(X)}{q_{ij}(Y)} \right) + (1 - p_{ij}(X)) \log \left( \frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)} \right) \right]$$

- **Derivatives:**

$$\frac{\delta CE}{\delta y_i} = \sum_j \left[ \frac{2abd_{ij}^{2(b-1)} P(X)}{1 + ad_{ij}^{2b}} - \frac{2b(1 - P(X))}{d_{ij}^2(1 + ad_{ij}^{2b})} \right] (y_i - y_j)$$

# Unsupervised clustering of scRNA-seq



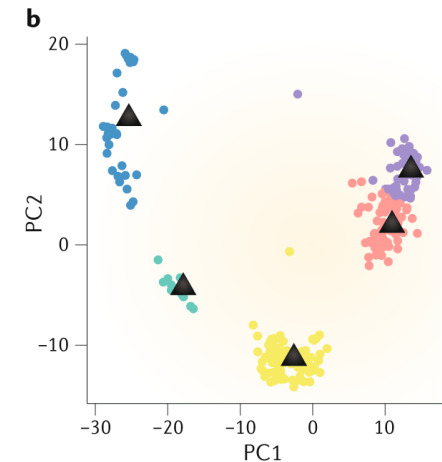
- Feature selection and dimensionality reduction extracts the most informative genes and strongest signals from background noise, respectively
- Cell-cell distances are then calculated in the lower dimensional space and used to either construct a cell-cell distance graph or used directly by clustering algorithms to assign cells to clusters



# Clustering methods for scRNA-seq

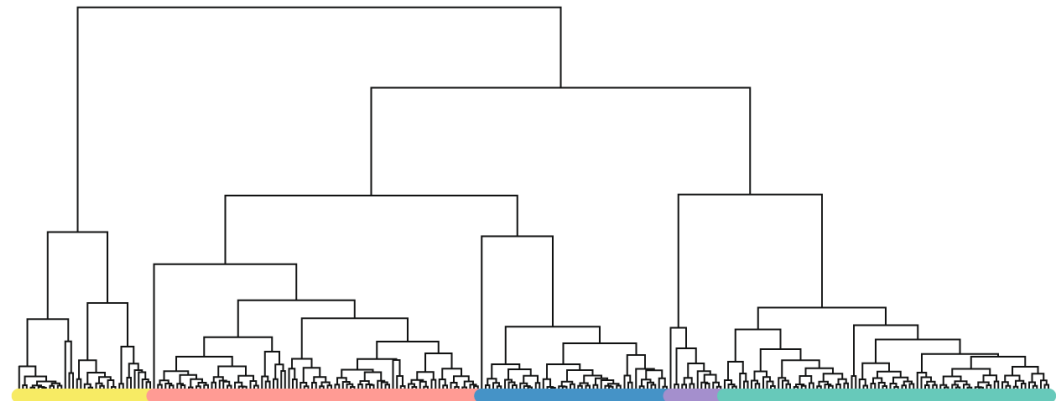
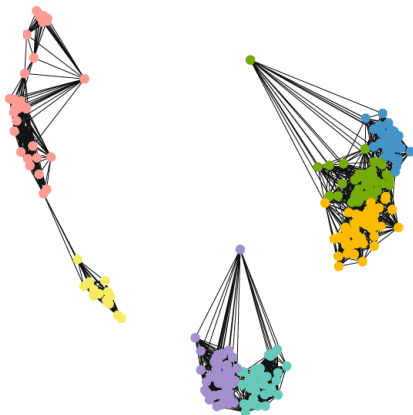
## Unsupervised clustering methods:

- K-means
- hierarchical clustering
- graph-based clustering



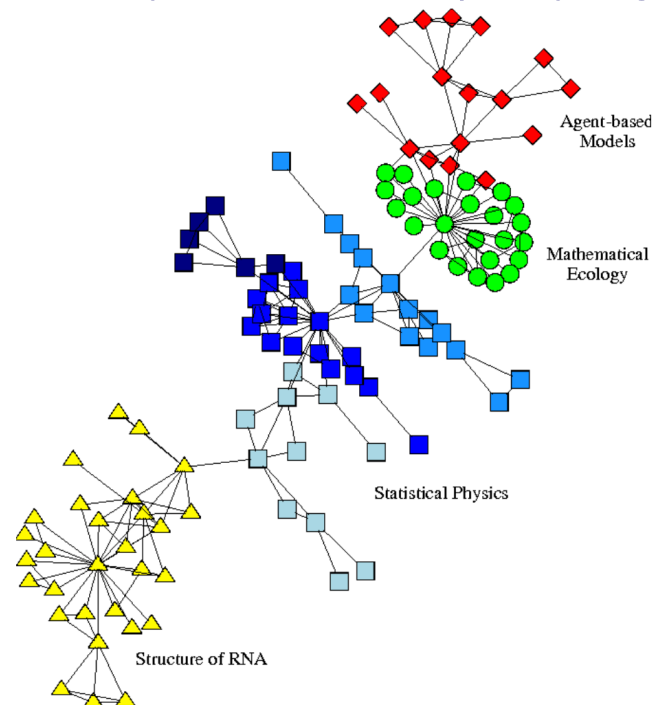
## Tools for graph-based clustering:

- Seurat: Louvain, Leiden, SLM
- igraph: fast greedy, Louvain, optimal, walktrap, spinglass, infomap



# Graph-based clustering

- “A community is subset of actors among whom there are relatively strong, direct, intense, frequent or positive ties”
- Some of **community detection** methods can be applied to scRNA-seq data by building a graph where each vertex represents a cell and (weight of) the edge measures similarity between two cells
- Graph-based clustering is the most popular clustering algorithm in scRNA-seq data analysis, and has been reported to have outperformed other clustering methods in many situations (Freytag et al. 2018)



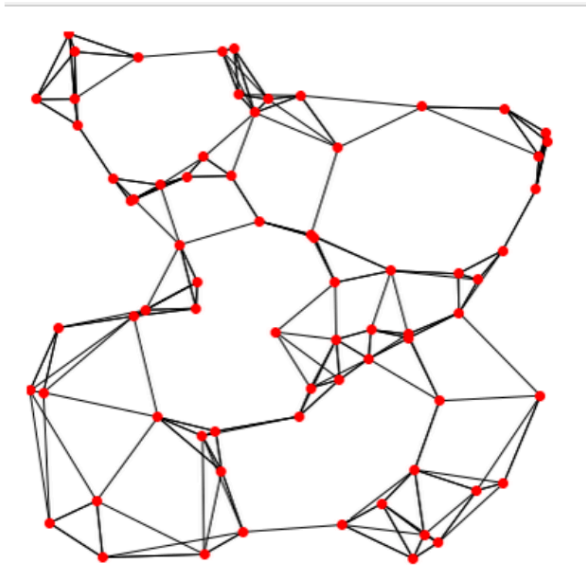
- Freytag, Saskia, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. 2018. “Comparison of Clustering Tools in R for Medium-Sized 10x Genomics Single-Cell Rna-Sequencing Data.” *F1000Research* 7. Faculty of 1000 Ltd.
- Wasserman, S. & Faust, K. (1994) *Social Network Analysis* (Cambridge Univ. Press, Cambridge, U.K.).
- [https://biocellgen-public.svi.edu.au/mig\\_2019\\_scrnaseq-workshop/public/clustering-and-cell-annotation.html#ref-freytag2018comparison](https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/clustering-and-cell-annotation.html#ref-freytag2018comparison)

# Graph-based clustering

## Building a graph

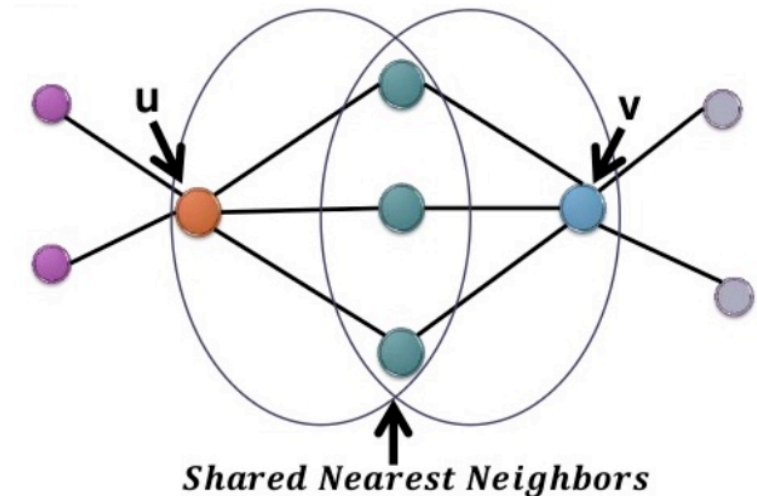
- Step1: Build an unweighted k-Nearest Neighbor (kNN) graph

A graph in which two vertices  $p$  and  $q$  are connected by an edge, if the distance between  $p$  and  $q$  is among the  $k$ -th smallest distances from  $p$  to other objects from  $p$ .



- Step2: Add weights, and obtain a Shared Nearest Neighbor (SNN) graph

A graph in which weights define proximity, or similarity between two nodes in terms of the number of neighbors (i.e., directly connected nodes) they have in common.



# Modularity

- Modularity (Newman and Girvan 2004) is not the only quality function for graph-based clustering, but it is one of the first attempts to embed in a compact form many questions including the definition of quality function and null model etc.

$$Q \propto \sum_{i,j} A_{i,j} \delta(i, j) - \sum_{i,j} \frac{k_i k_j}{2m} \delta(i, j)$$

$A_{i,j}$  : weight between node  $i$  and  $j$

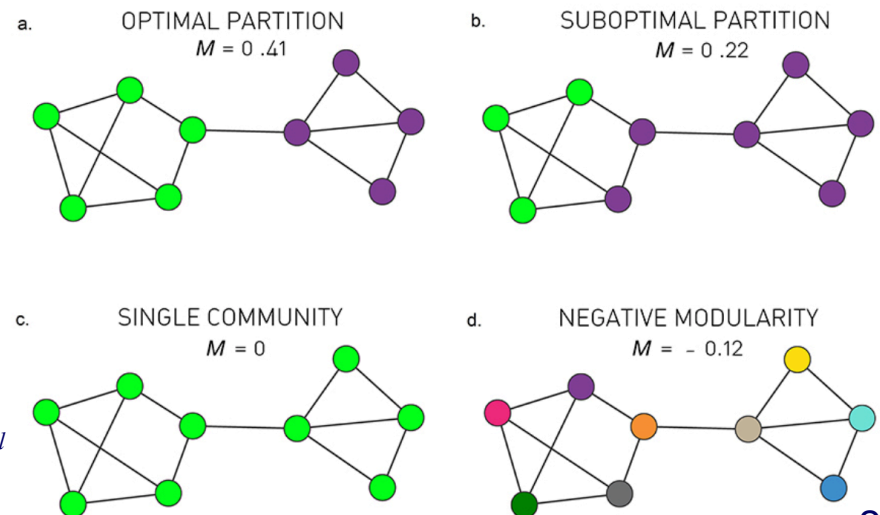
$\delta(i, j)$  : Indicator of whether  $i$  and  $j$  are in the same cluster

$k_i$  : the degree of node  $i$  (the sum of weights of all edges connected to  $i$ )

$m$ : the total weight in the all graph

- Modularity :  $-1 \ll Q \ll 1$
- $Q > 0$  when # edges within groups > #edges within groups in a randomly rewired graph
- $Q > 0.3$  : significantly community structure

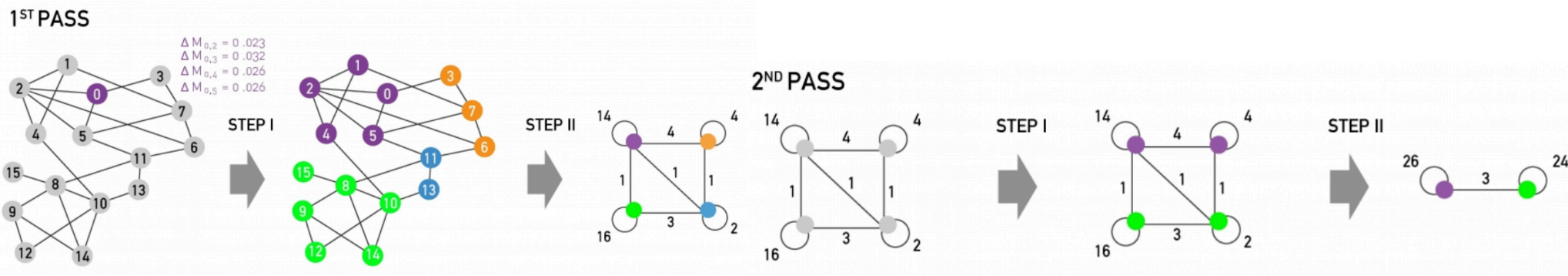
Higher modularity implies better partition



- Newman, Mark EJ, and Michelle Girvan. 2004. "Finding and Evaluating Community Structure in Networks." *Physical Review E* 69 (2). APS: 026113.
- [https://biocellgen-public.svi.edu.au/mig\\_2019\\_scrnaseq-workshop/public/clustering-and-cell-annotation.html#ref-freytag2018comparison](https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/clustering-and-cell-annotation.html#ref-freytag2018comparison)

# Louvain community detection

- Start with every node in its own community
- **Step1:** Modularity optimization
  - Order the nodes and for each node  $i$ , move  $i$  to the community of neighbor  $j$  that leads to maximum  $\Delta Q$
  - If all  $\Delta Q < 0$  the  $i$  remains in its current community
  - Repeatedly cycle through all nodes until  $\Delta Q = 0$
- **Step2:** Community aggregation
  - Create a weighted network of communities from Step1
  - Nodes : communities in Step1
  - Edge weights : sum of weights of edges between communities
  - Edges within a community become two self-loops
- **Repeat:** Apply Step1/ Step2 to resulting network, and so on until  $\Delta Q = 0$





# Cell type annotation

- The most challenging task in scRNA-seq data analysis is arguably the interpretation of the results. Obtaining clusters of cells is fairly straightforward, but it is more difficult to determine what biological state is represented by each of those clusters
- Various computational approaches could exploit prior biological knowledge to assign meaning to an uncharacterized scRNA-seq dataset could be used

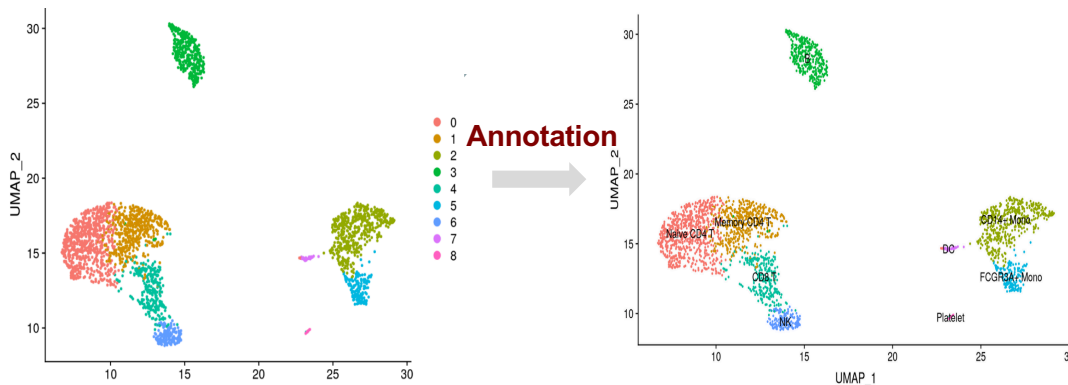
- Types of cell type annotation tools

- Supervised methods:** require a training dataset labeled with the corresponding cell populations in order to train the classifier

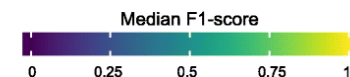
- SingleR, ACTINN, CaSTle

- Prior-knowledge based methods:** either a marker gene file is required as an input or a pretrained classifier for specific cell populations is provided

- DigitalCellSorter, Moana



	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
SVM <sub>rejection</sub>	0.99	0.99	0.98	1	0.98	1	1	0.99	1	1	0.98	0.99	0.92
scPred	1	0.98	0.98	1	0.95	1	1	0.97	1	1	0.69	0.96	
SVM	0.98	0.98	0.97	1	0.99	1	1	0.98	1	0.99	0.89	0.95	0.7
singleCellNet	0.97	0.96	0.97	0.99	1	1	1	0.94	1	0.99	0.87	0.88	0.74
ACTINN	0.97	0.98	0.97	1	0.95	1	1	0.97	1	0.99	0.86	0.88	0.74
CaSTle	0.93	0.94	0.96	0.98	0.96	1	0.99	0.94	1	0.99	0.79	0.84	0.79
scmapcell	0.98	0.98	0.97	1	0.73	1	1	0.98	1	1	0.91	0.73	0.64
LDA	0.94	0.97	0.96	0.99	0.89	1	1	0.95	1	0.99	0.88	0.63	0.66
scmapcluster	0.99	0.95	0.97	1	1	1	1	0.87	1	0.98	0.88	0.73	0.44
RF	0.94	0.94	0.96	0.98	0.85	1	1	0.91	1	0.99	0.73	0.81	0.66
SingleR	0.96	0.97	0.95	0.97	0.99	1	1	0.88	1	0.97	0.86	0.66	0.32
LAmbDA	0.92	0.8	0.95	0.96	0.97	1	1	0.62	1	0.99	0.84		0.4
NMC	0.92	0.91	0.84	0.93	0.99	0.92	0.9	0.69	0.99	0.97	0.81	0.71	0.55
CHETAH	0.91	0.94	0.96	0.97	0.96	1	1	0.83	1	0.96	0.81	0.65	0.11
scVI	0.98	0.56	0.97	0.99	1	1	1	0	1	0.97	0	0.97	0.64
scID	0.75	0.59	0.95	0.85	0.8	1	1	0.42	1	0.95	0.63	0.61	0.42
Cell_BLAST	0.11	0.89	0.79	0.08	0.63	1	0.99	0.97	1	0.99	0.76	0.91	0.74
KNN	0.91	0.95	0.95	0.85	0.03	1	0.98	0.92	1	0.64	0.13	0.45	0.54
SCINA												1*	1*
DigitalCellSorter												0.99*	0.78*
Garnett <sub>CV</sub>												0.94*	0.6*
Garnett <sub>pretrained</sub>												0.98*	0.54*
Moana												0.93*	0.5*
Garnett <sub>DE</sub>												0.65	0.37
SCINA <sub>DE</sub>												0.38	0.47
DigitalCellSorter <sub>DE</sub>												0	0
	Baron Mouse	Baron Human	Muraro	Seegerstolpe	Xin	10X	CEL-Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K



[https://btep.ccr.cancer.gov/wp-content/uploads/Celltype\\_Annotation\\_final.pdf](https://btep.ccr.cancer.gov/wp-content/uploads/Celltype_Annotation_final.pdf)

[https://biocellgen-public.svi.edu.au/mig\\_2019\\_scrnaseq-workshop/public/clustering-and-cell-annotation.html](https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/clustering-and-cell-annotation.html)

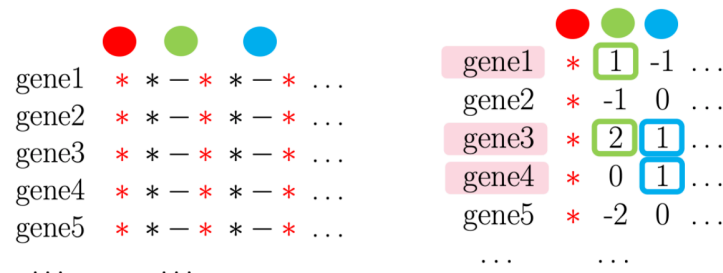
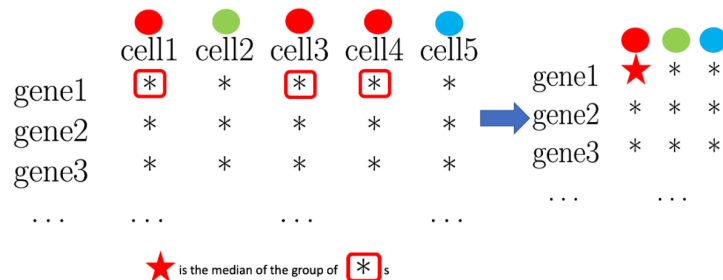
<https://bioconductor.org/books/release/OSCA/cell-type-annotation.html>

# Cell type annotation

## SingleR : Reference-based annotation of scRNA-seq

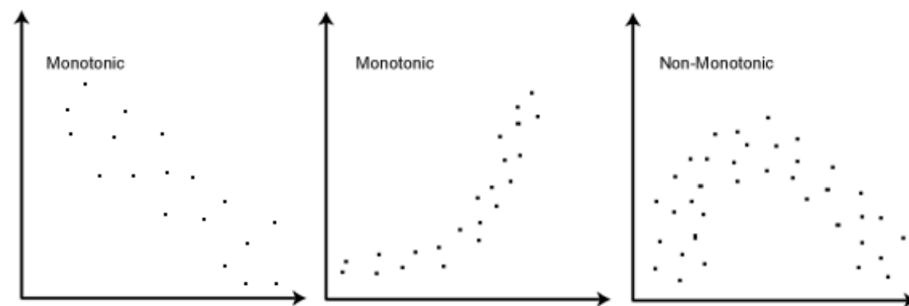
### Step1: Find variable gene

- For every gene, obtain **median** grouped by label
- Select genes that makes at least one label different: If we are looking for the genes that makes label “green” different from label “red”, we subtract the second column by the first, and pick the top  $N$  highest and positive values



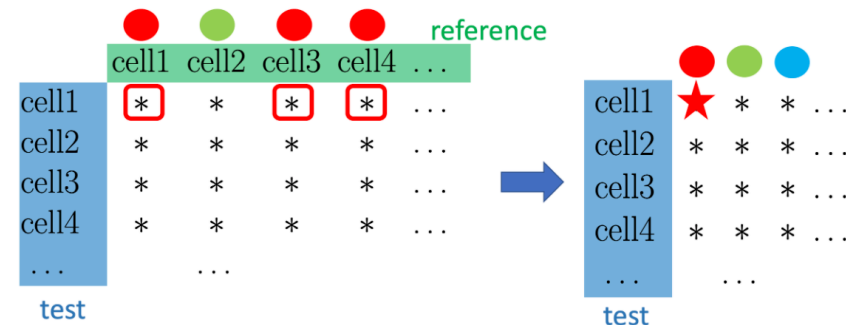
### Step2: Spearman's correlation

- Spearman's correlation  $\in [-1, 1]$  is a measure of the strength of a linear or monotonic relationship between paired data.
- compute the Spearman's correlation for all pairs of cells in the test and reference dataset, and obtain an  $n_{test} \times n_{ref}$  correlation matrix, where  $n$  is the number of cells (see the first matrix in Step3).



### Step3 : Scoring

- We want to know how each cell in the test data is correlated to the labels in the reference data, instead of each reference cell. So we take the correlations of a cell in the test data with all the cells with a certain label in the reference data, and summarize them into one number or a score, in SingleR, the default is to take the 80% quantile.



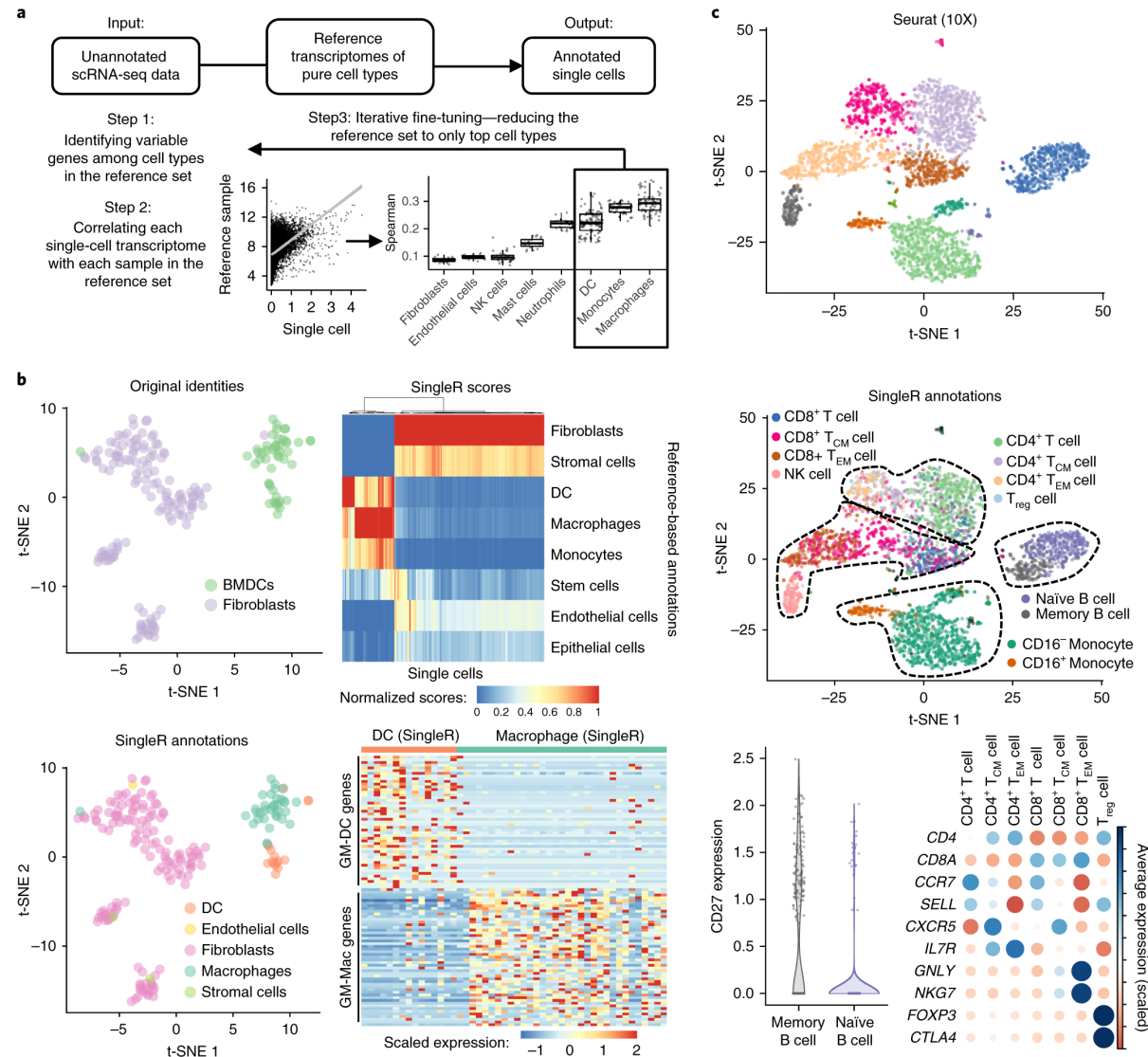
### Step4 : Fine tuning

Each (\*) is a spearman's correlation, ★ is the 80% quantile of (\*)s

# Cell type annotation

## SingleR : Reference-based annotation of scRNA-seq

- correlated single-cell transcriptomes with reference transcriptomic data sets and improved its inferences iteratively



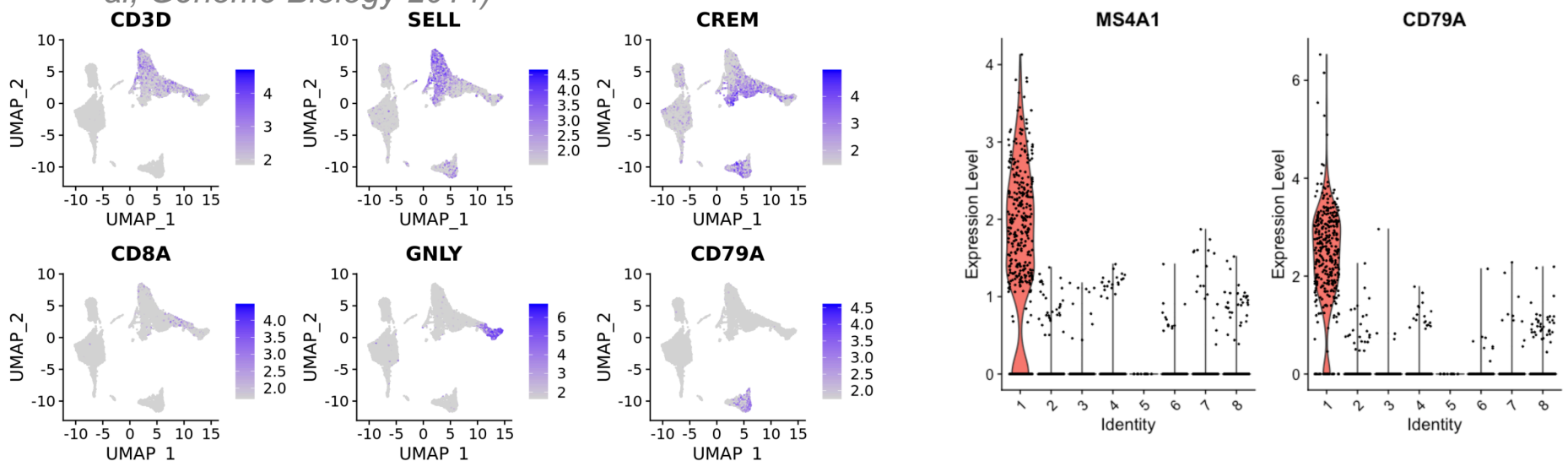
Aran, D., Looney, A.P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).  
<https://doi.org/10.1038/s41590-018-0276-y>



# Cell type markers identification

## Differential expression analysis

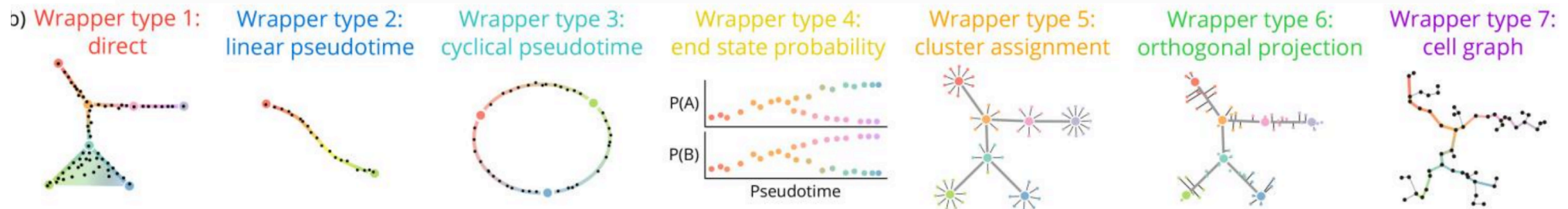
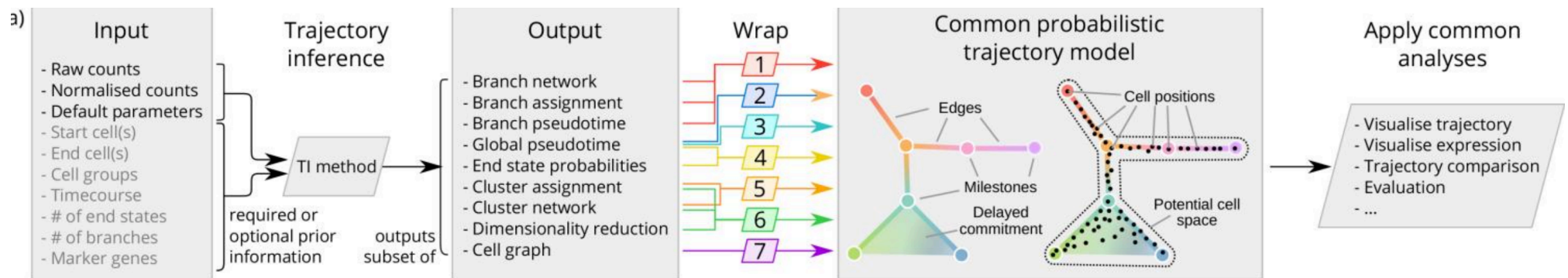
- Non-parametric tests
  - Wilcoxon rank sum test
  - Student's t-test
- Methods specific for scRNA-seq
  - MAST : GLM-framework that treats cellular detection rate as a covariate (*Finak et al, Genome Biology, 2015*)
- Methods for bulk RNA-seq
  - DESeq2 : DE based on a model using the negative binomial distribution (*Love et al, Genome Biology 2014*)



- Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015). <https://doi.org/10.1186/s13059-015-0844-5>
- Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, 550. doi: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
- [https://satijalab.org/seurat/archive/v3.1/immune\\_alignment.html](https://satijalab.org/seurat/archive/v3.1/immune_alignment.html)

# Pseudo timing

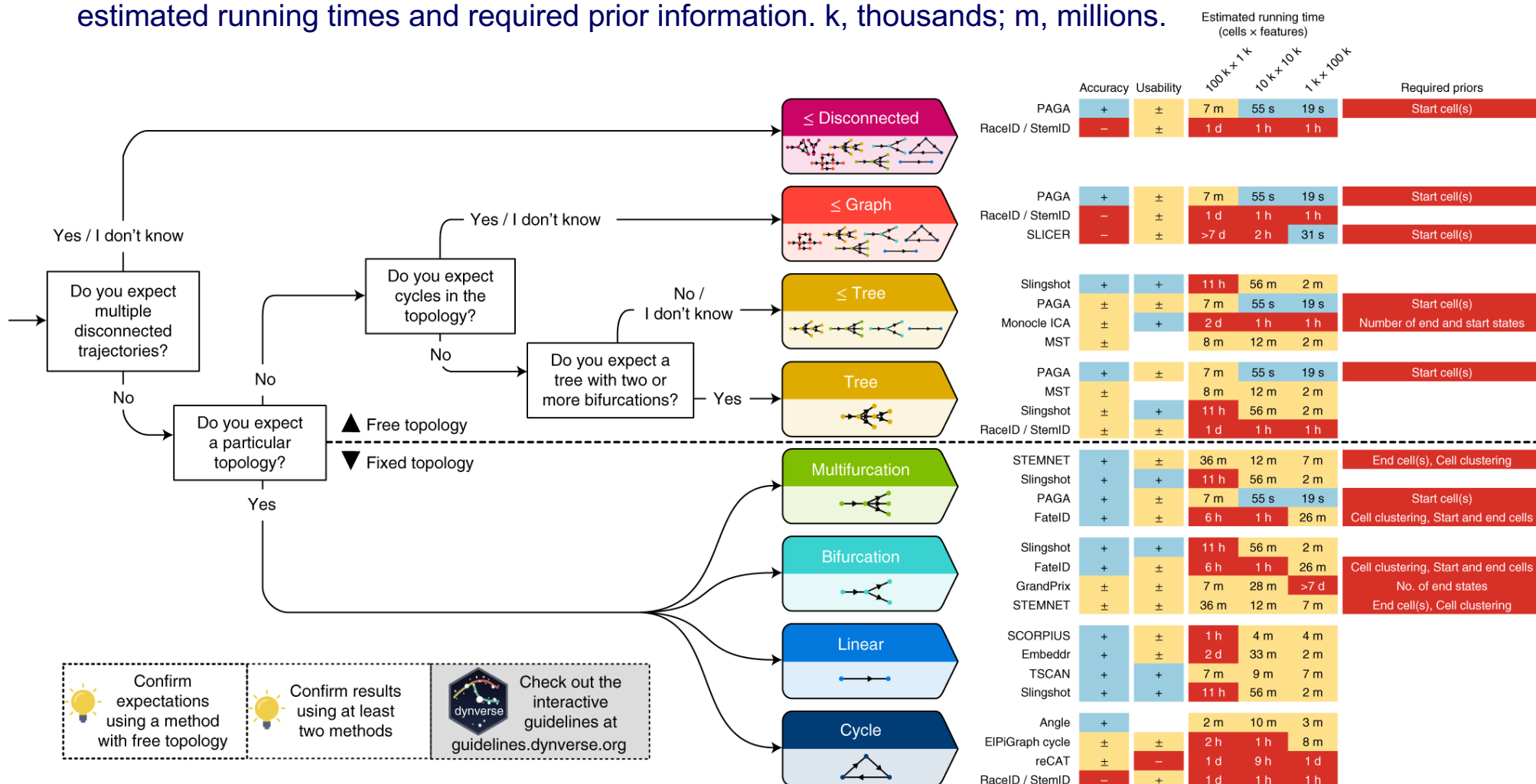
- Many differentiation processes taking place during development: following a stimulus, cells will change from one cell-type to another. We must sample at multiple time-points and obtain snapshots of the gene expression profiles.
- Since some of the cells will proceed faster along the differentiation than others, each snapshot may contain cells at varying points along the developmental progression.
- We use statistical methods to order the cells along one or more trajectories which represent the underlying developmental trajectories, this ordering is referred to as “**pseudotime**”
- Using single-cell -omics data, many **trajectory inference** (TI) methods could computationally order cells along trajectories, allowing the unbiased study of cellular dynamic processes



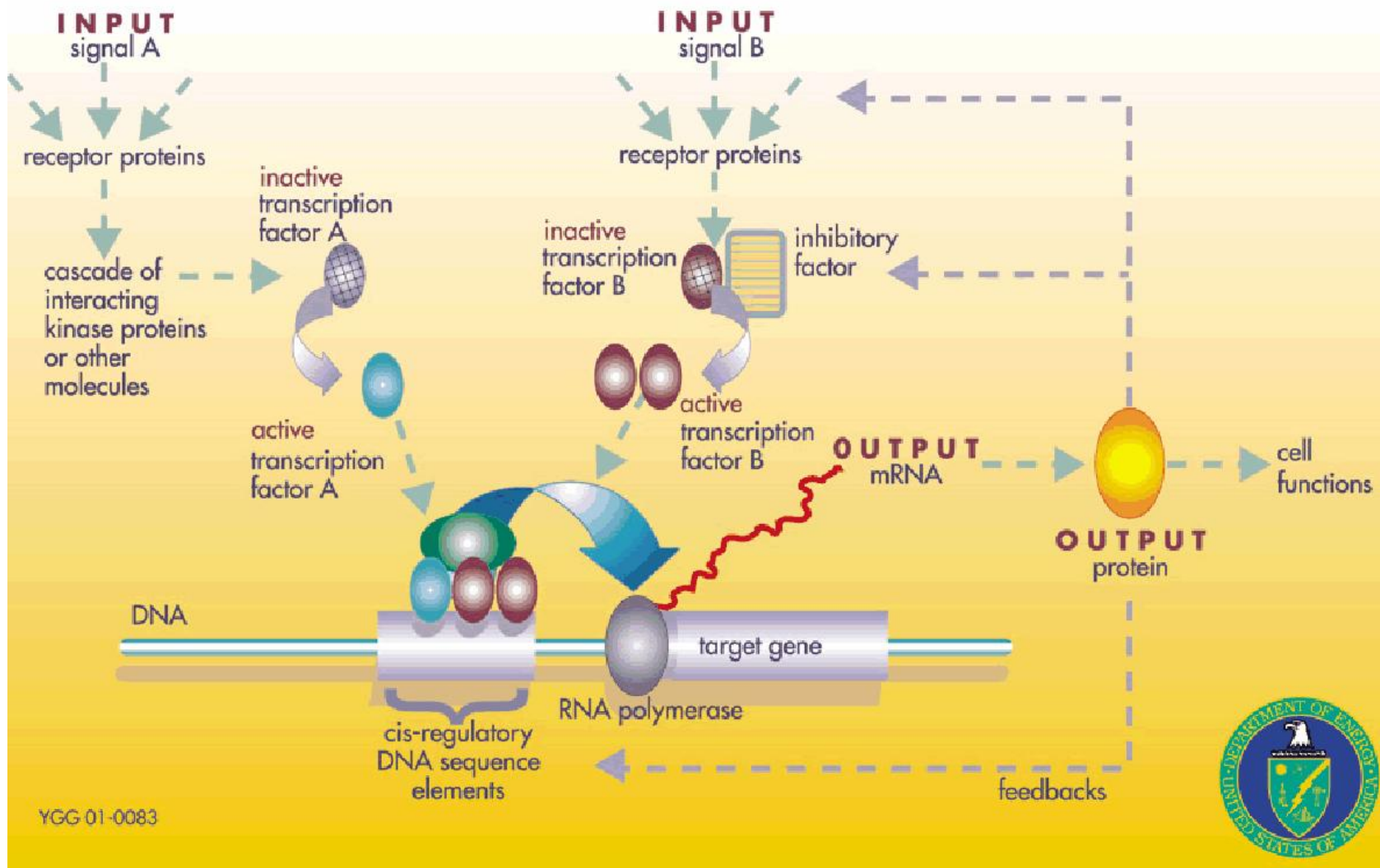
# Pseudo timing

## Practical guideline for method users

- The performance of a method mostly depends on the topology of the trajectory, the choice of TI method will be primarily influenced by the user's existing knowledge about the expected topology in the data. Methods to the right are ranked according to their performance on a particular (set of) trajectory type. Further to the right are shown the accuracy (+: scaled performance  $\geq 0.9$ ,  $\pm$ :  $>0.6$ ), usability scores (+:  $\geq 0.9$ ,  $\pm$ :  $\geq 0.6$ ), estimated running times and required prior information. k, thousands; m, millions.



# Gene regulation

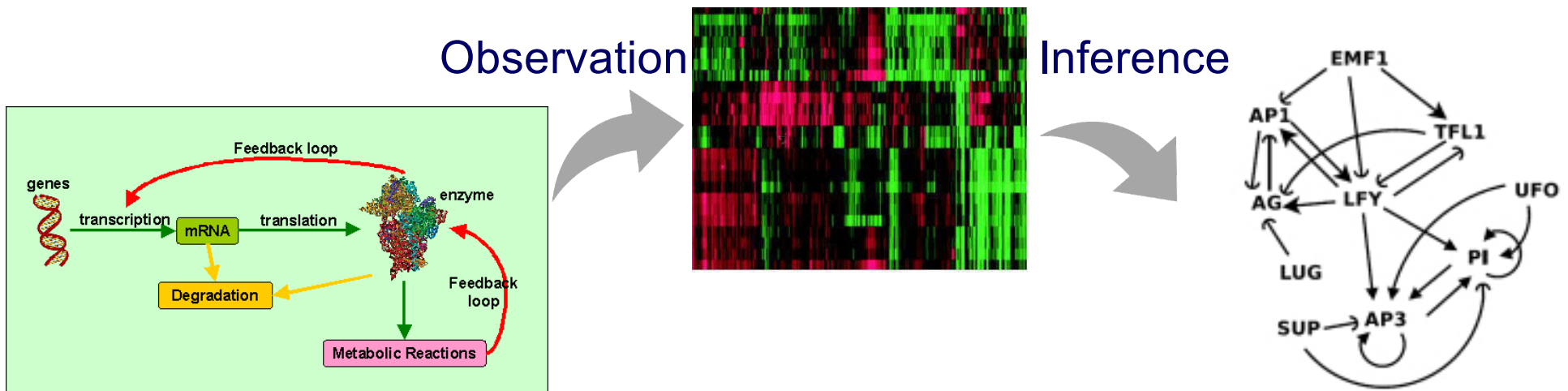
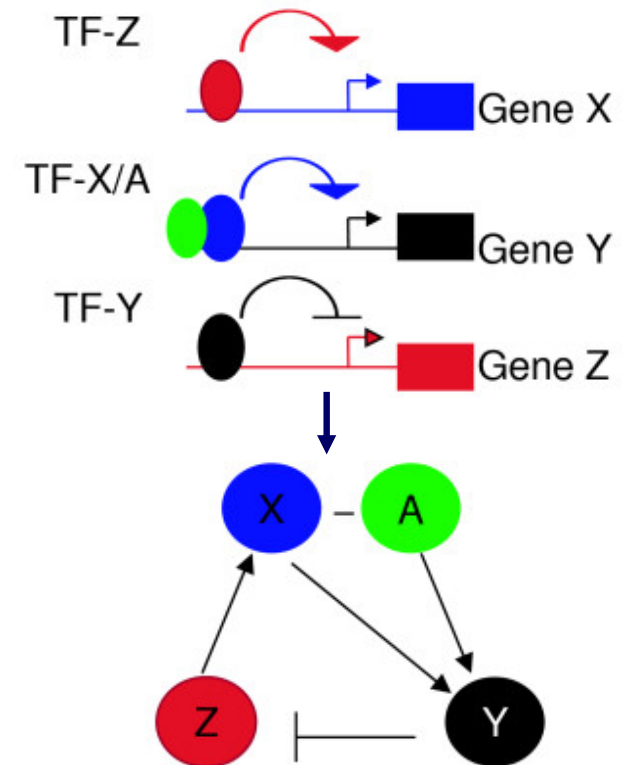


Gene regulation is the process of controlling which genes in a cell's DNA are expressed (used to make a functional product such as a protein).



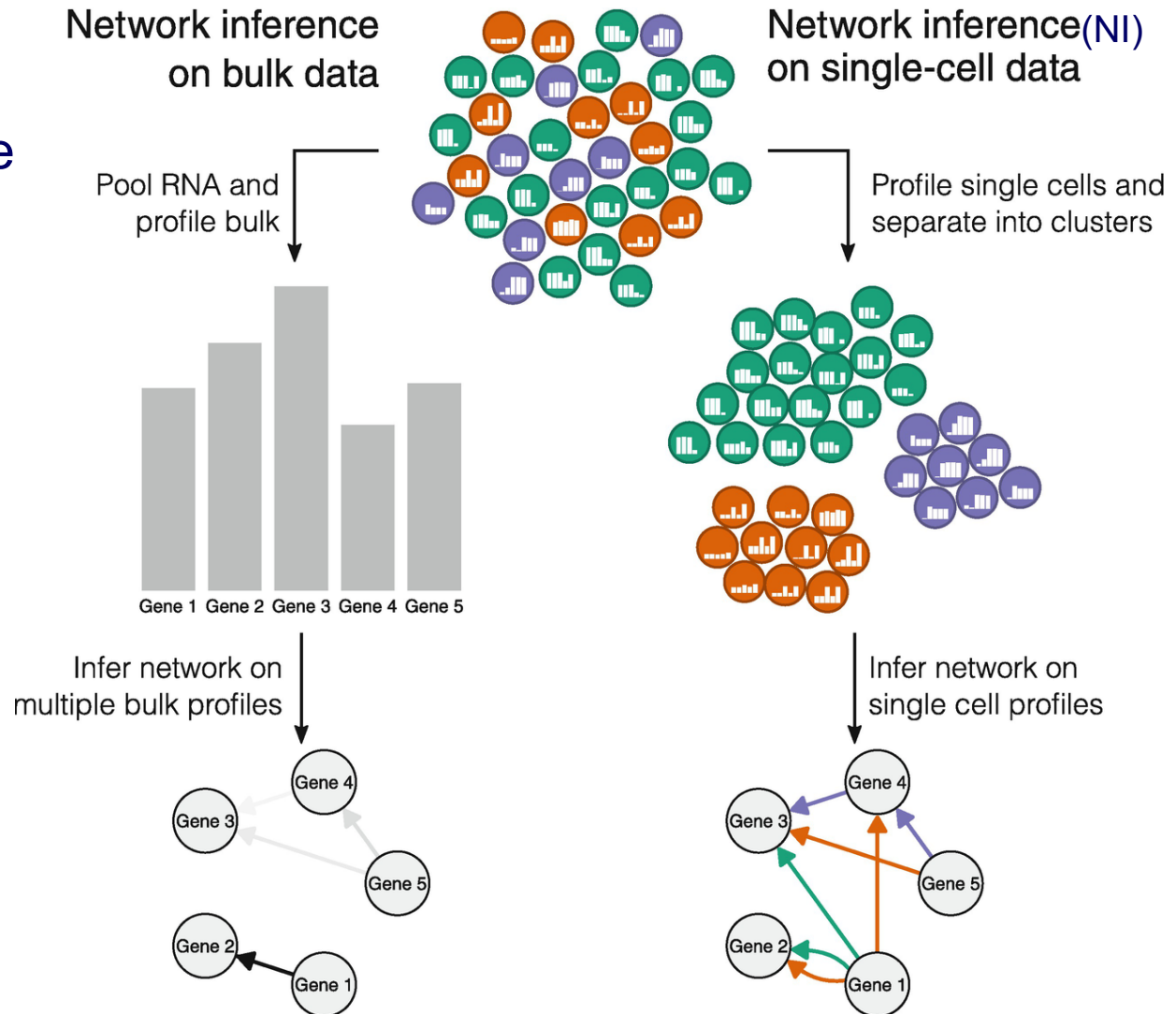
# Gene regulatory network

- Gene regulatory networks (GRNs) are the on-off switches of a cell operating at the gene level
- two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition
- can be inferred from correlations in gene expression data, time-series gene expression data, and/or gene knock-out experiments

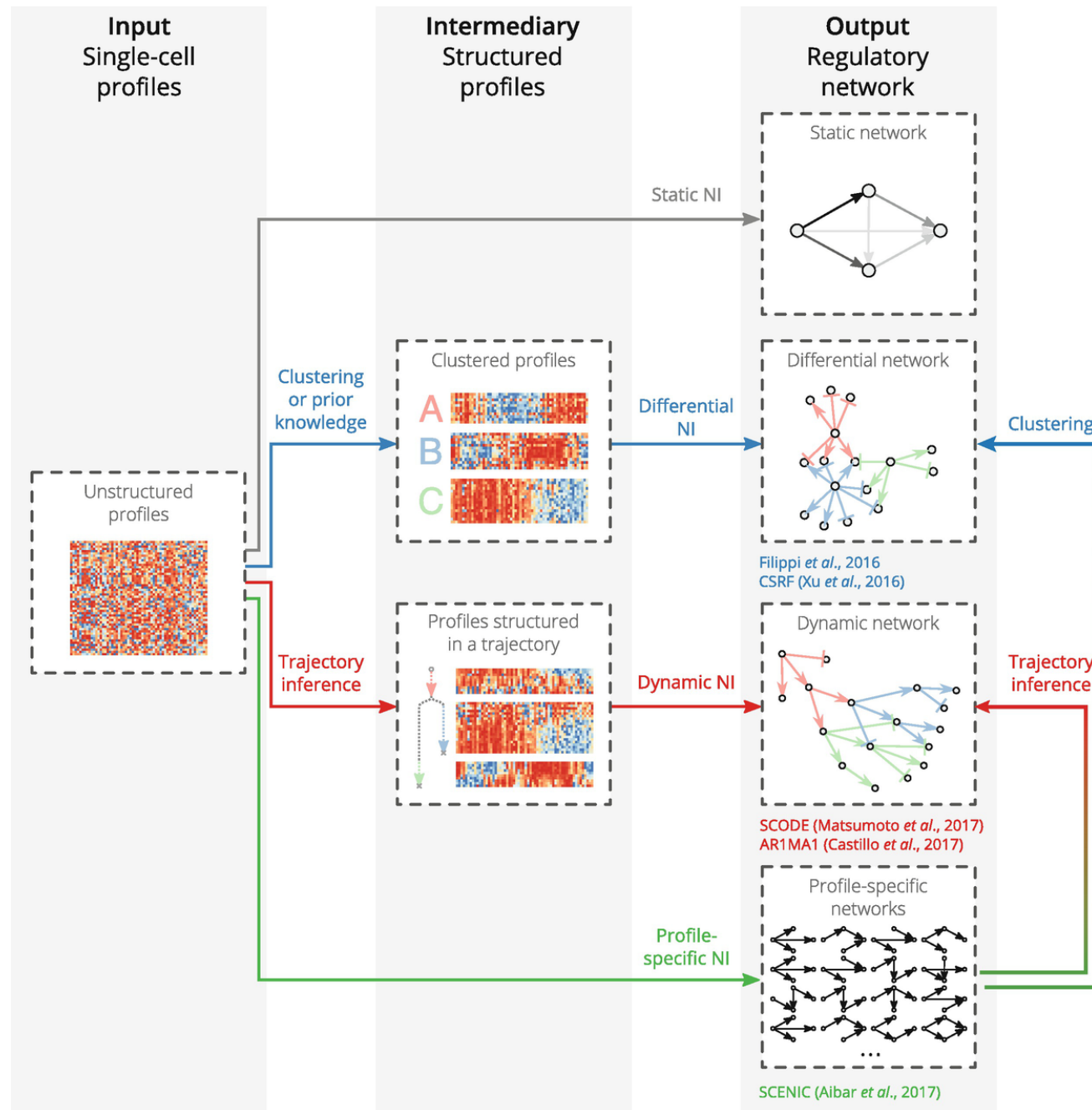


# Cell-type gene regulatory networks

- cell-type-specific GRNs would be key tools for the study of cellular heterogeneity
- cell-type-specific GRNs will reveal key regulatory factors and circuits for specific cell types, facilitating mapping between disease-associated variants and affected cell types



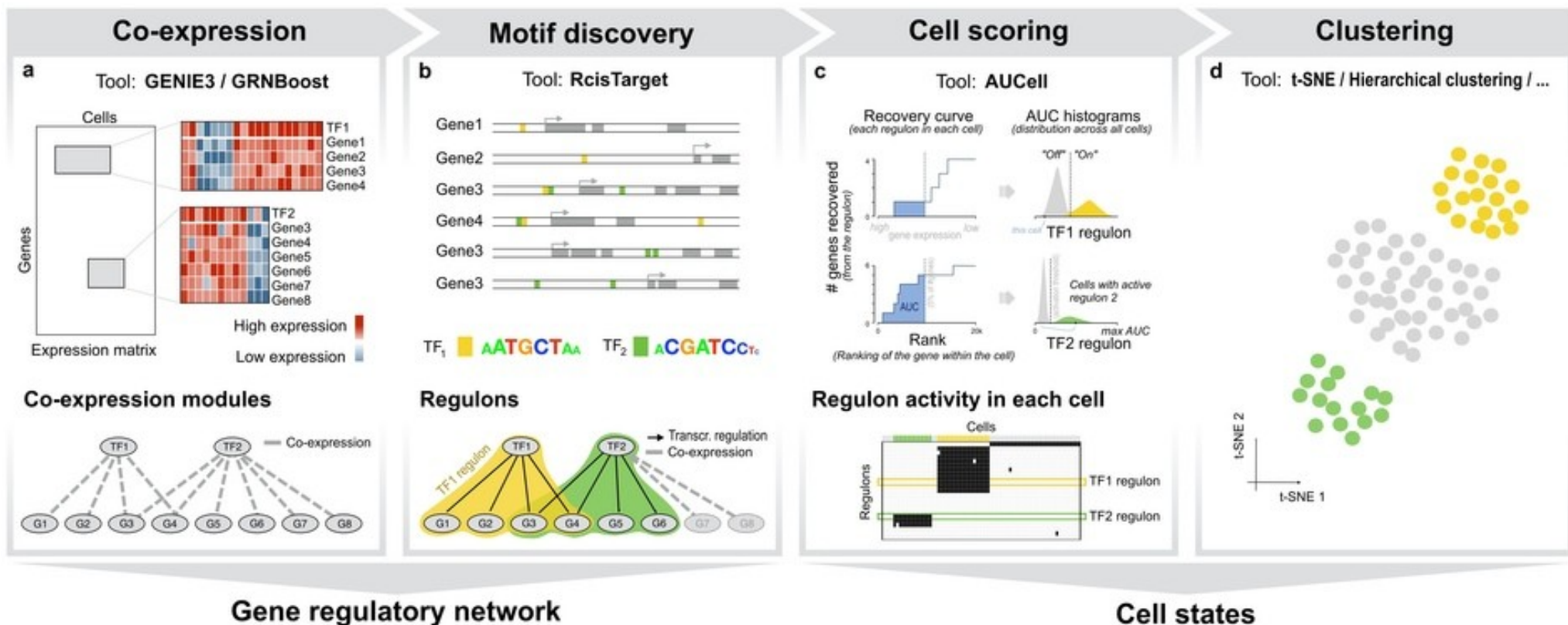
# Network inference algorithms



# SCENIC

## single-cell regulatory network inference and clustering

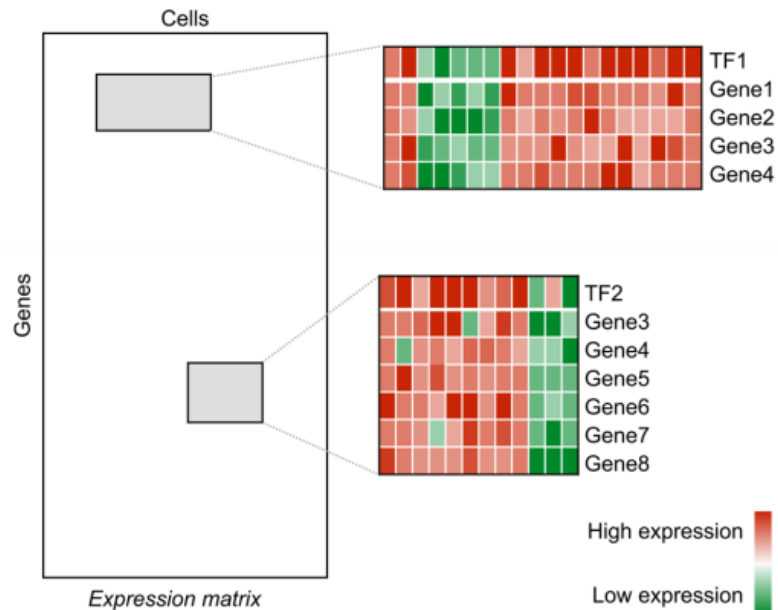
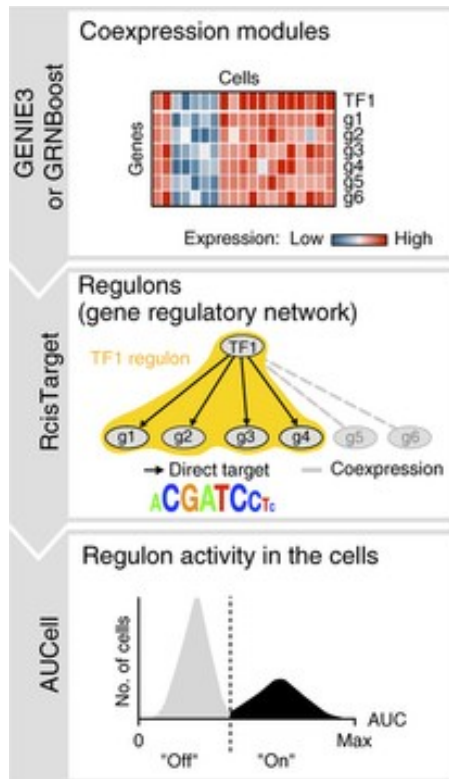
- SCENIC is a tool to simultaneously reconstruct gene regulatory networks and identify stable cell states from single-cell RNA-seq data. The gene regulatory network is inferred based on co-expression and DNA motif analysis, and then the network activity is analyzed in each cell to identify the recurrent cellular states.





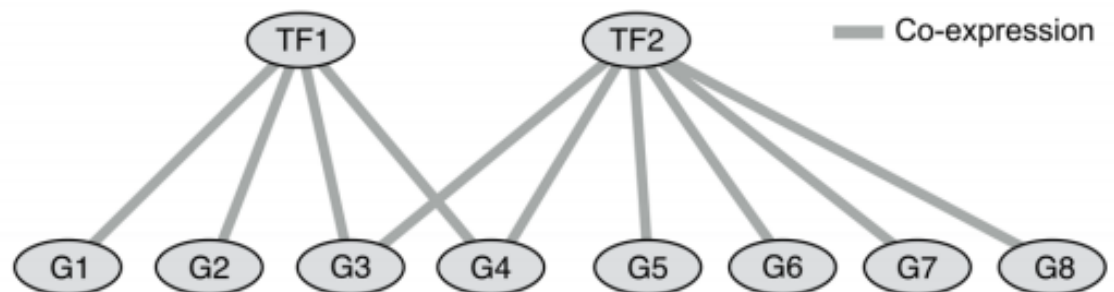
# Step1.TF-based co-expression network

## SCENIC



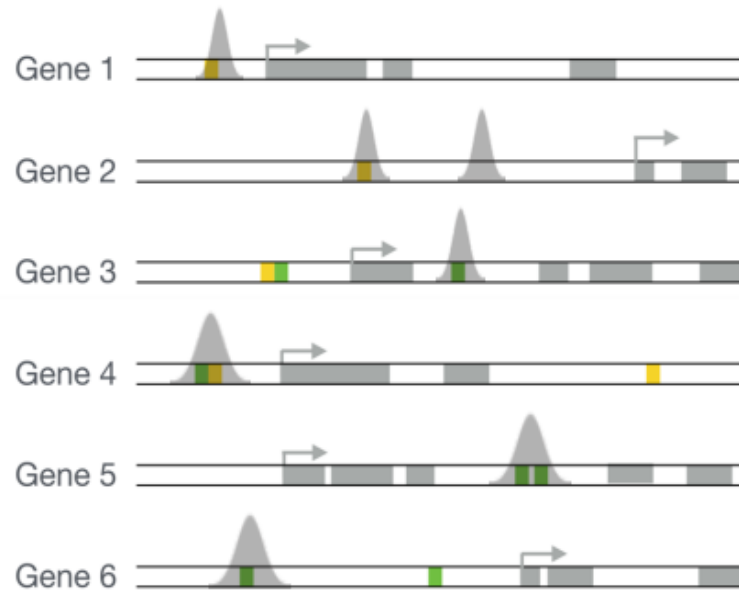
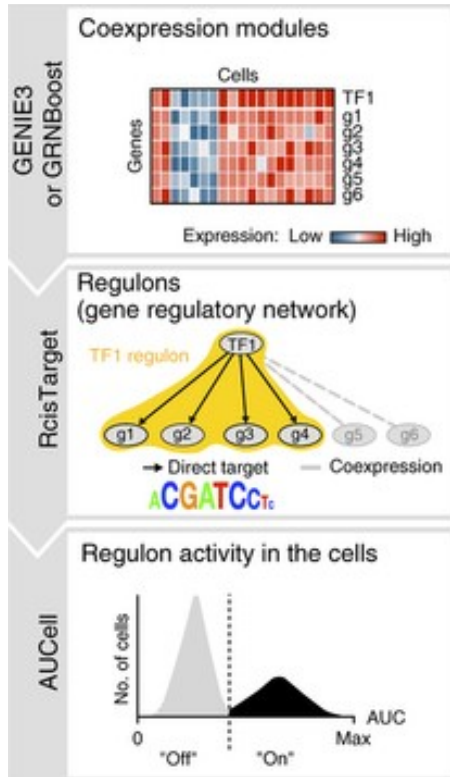
GENIE3  
or  
GRNBoost

## Co-expression modules



# Step2. Gene regulatory network

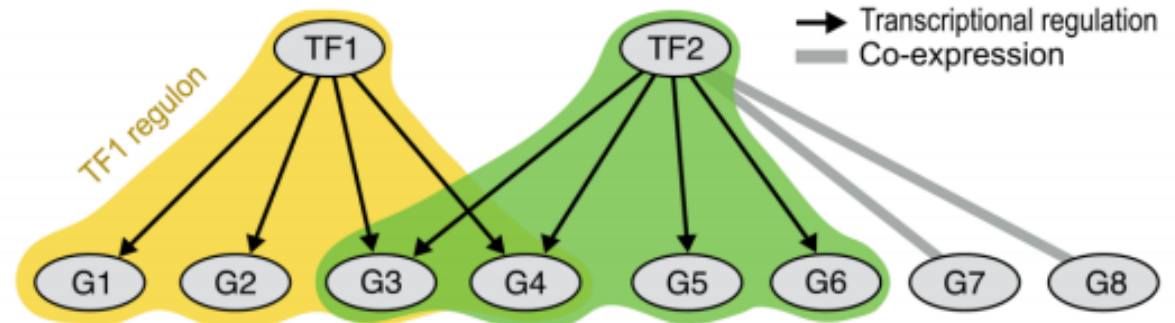
## SCENIC



TF1 **AATGCTAA** TF2 **ACGATCC<sub>TC</sub>**

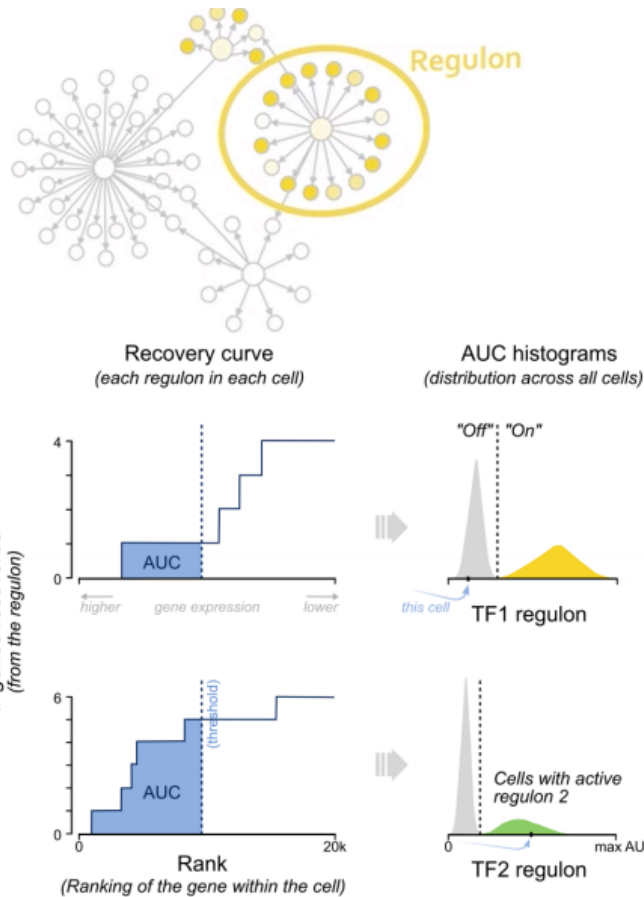
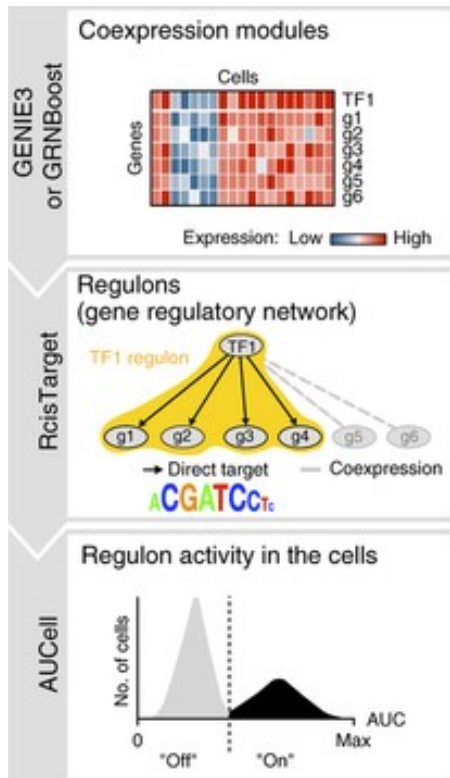
RcisTarget  
*cis-regulatory sequence analysis*

## Regulons (Gene regulatory network)



# Step3.Activity of the network in each cell

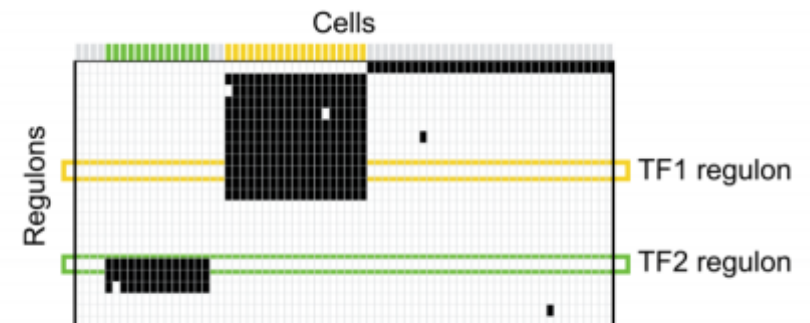
## SCENIC



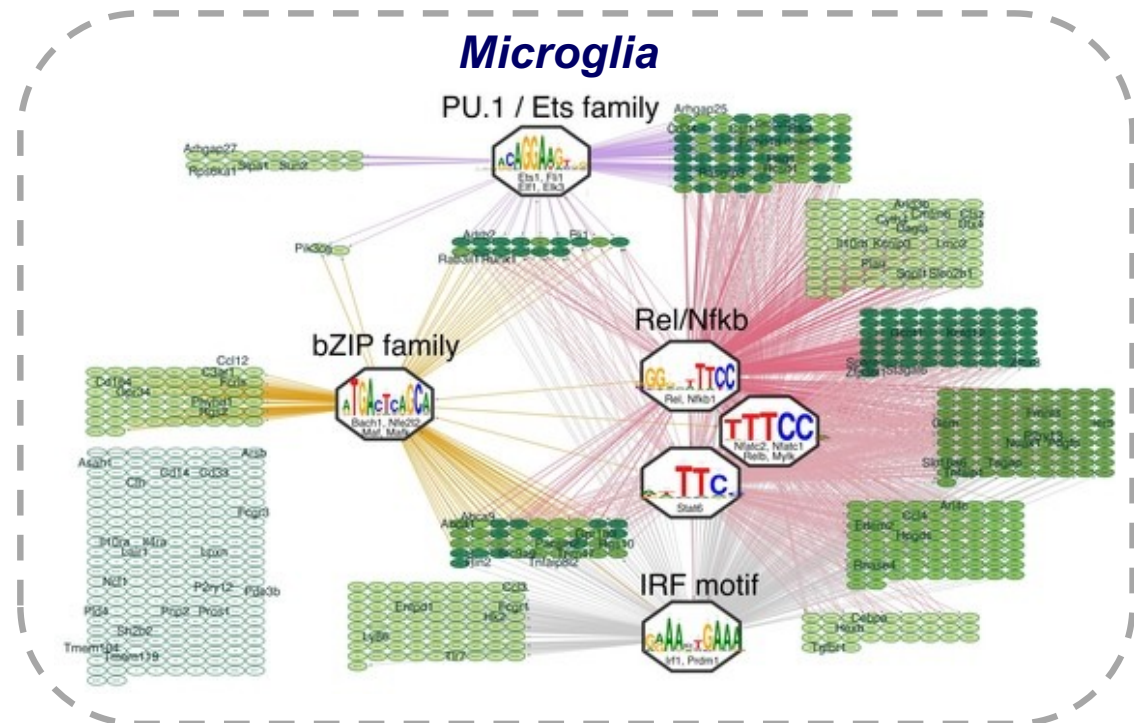
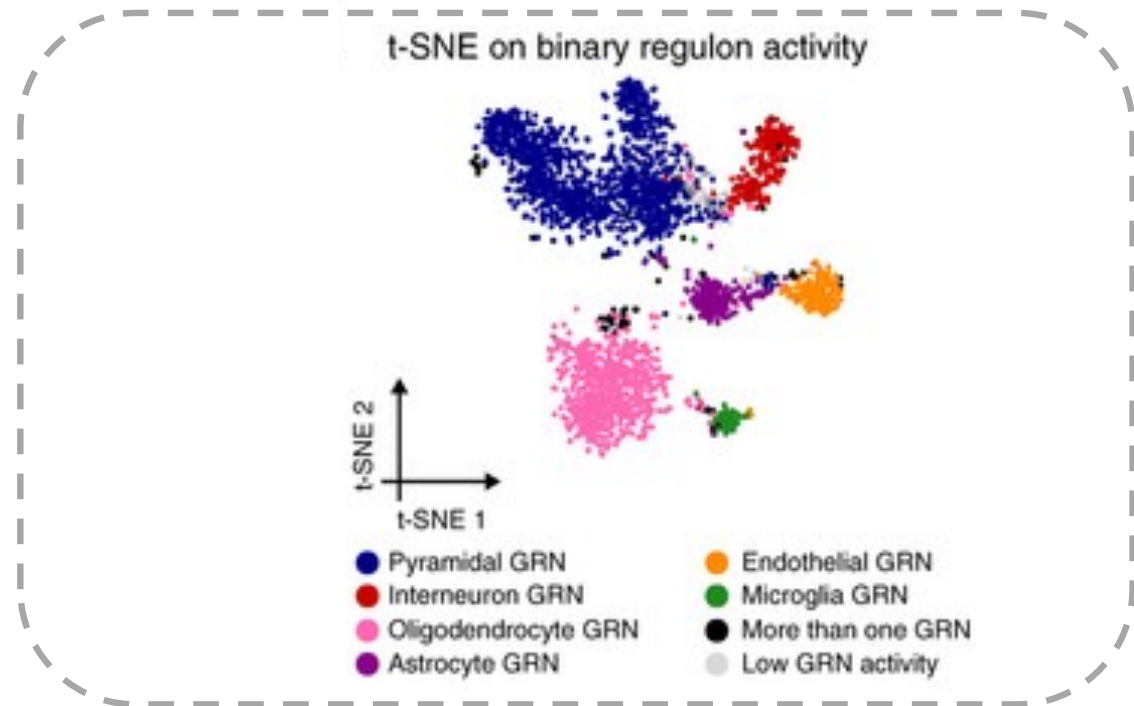
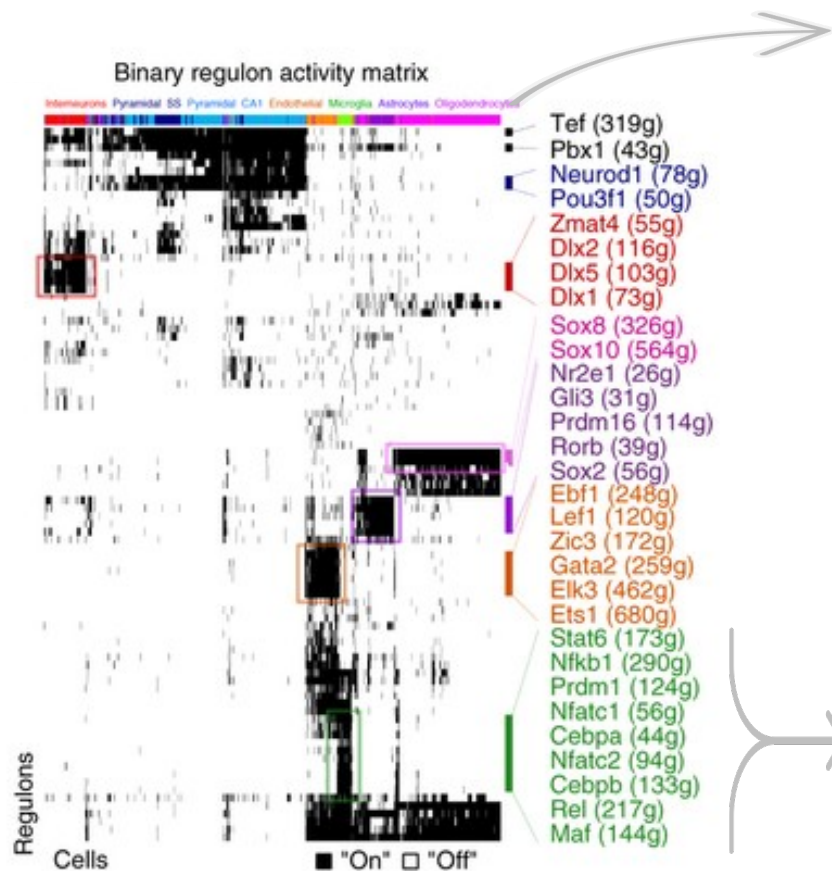
## AUCCell

*Identifying cells with active gene-sets*

**Regulon activity matrix** (Network activity in each cell)



# GRN-based cell states

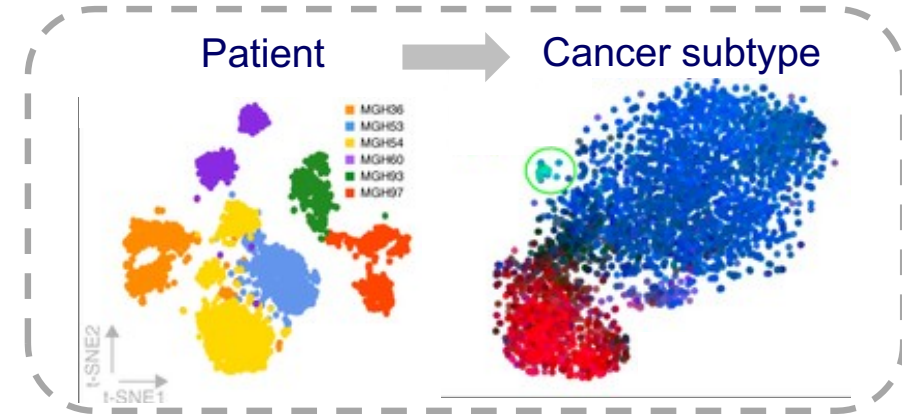
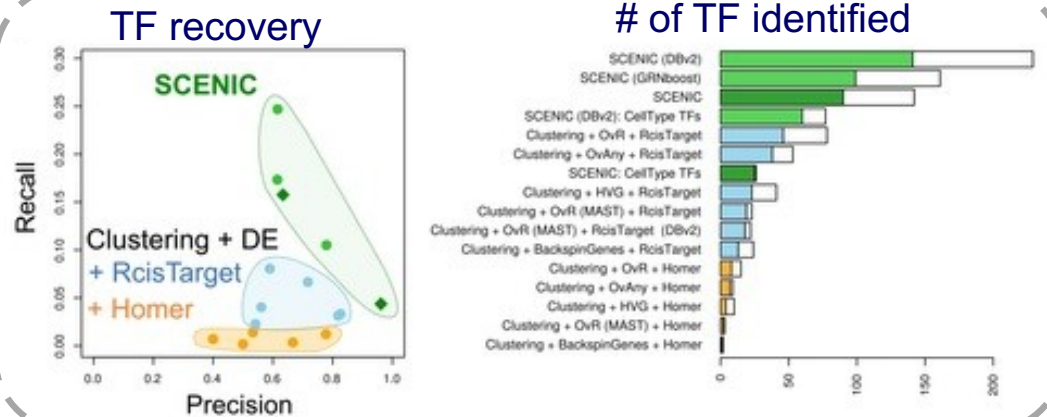




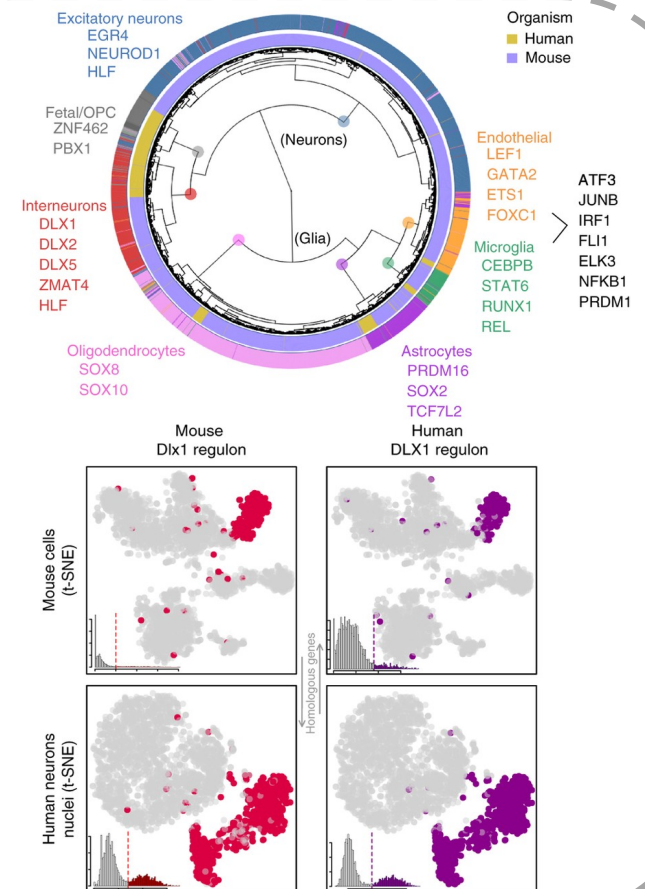
# Further applications

## Biologically-driven dimensionality reduction

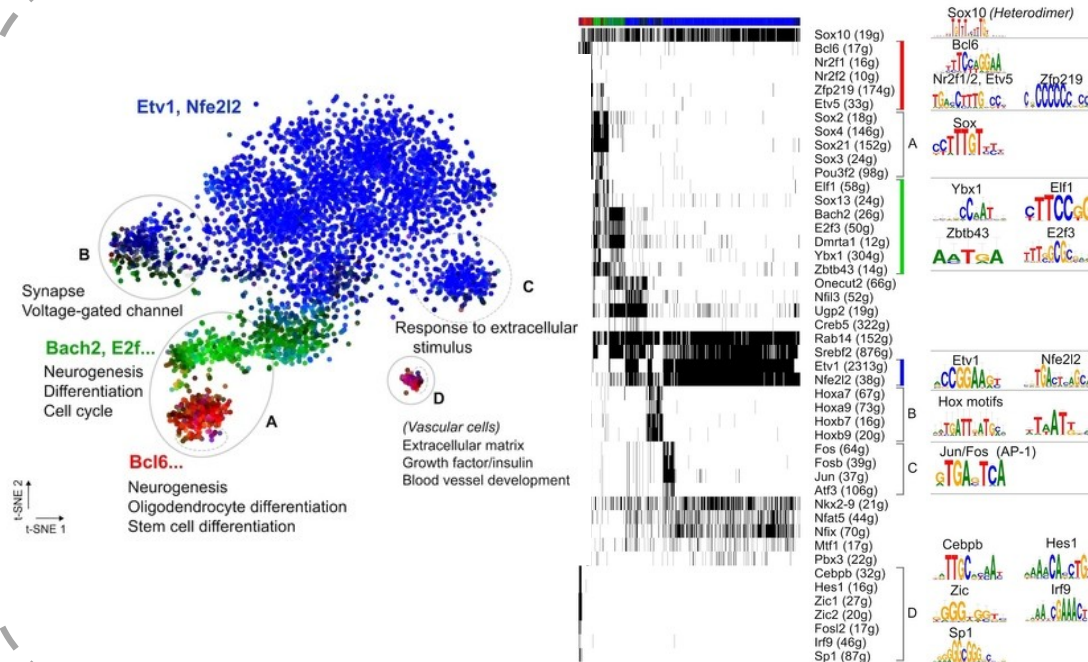
### Cell-type specific master regulators



### Cross-species GRN comparisons



### Dynamic states



# Tools for GRN inference from scRNA-seq

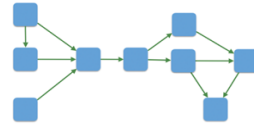
Ground truth of GRNs is usually unknown.

How do we evaluate the performance of existing GRN inference methods from scRNA-seq data?



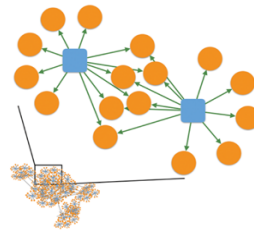
BEELINE

## Predictive/descriptive GRNs



- + Models sequence of events (e.g. TFs that activate other TFs)
- + Some can be used for predictions (e.g. over-expression or knockdown)
- Max ~10-100 TFs/genes
- Requires representation of all/most intermediate states

## Global GRNs



- + Identifies "Master regulators"
- + Putative targets for TF
- High false positive rate
- Mainly based on co-expression

### Tools from bulk RNA-seq (applied to sc-RNA-seq)

- **GENIE3** [33]: Identifies TFs that best predict the expression of each gene
- **WCGNA** [34]: Identifies co-expression modules using clustering on a correlation network

Type of network:

- Directed (TF → any gene)
- Directed (TF → TF)
- Undirected

## Dynamic processes

time-points or "pseudotime"

### Boolean networks

**SCNS** toolkit [9] *F# (run from terminal)*  
Builds a state-transition graph that can be used for predictions. Requires many parameters.

**BoolTraineR** [29] *R*  
Can infer both network structure and Boolean rules without information on trajectories through cell states. Robust to drop-outs.

**SingCellNet** [97] *"Available upon request"*  
Infers regulatory circuits by integrating transcriptional patterns with the cell lineage tree.

### ODE

► **SCODE** [45] *R*  
Models network dynamics and allows simulations. Based on time point data. Recommended to run multiple times to obtain average network. Examples with 100 TFs.

► **InferenceSnapshot** [46] *C++ and Matlab*  
Combines trajectory information and co-expression (Genie3) into an ODE model. Small networks (demo with 6 genes).

### others

#### Trajectory → Network

- **Sinova** [40] *Multiple scripts*  
Analysis pipeline for devel. processes. GRN using co-expression (Pina's method [38]).
- **SCOUP** [43] *C*  
Trajectory inference & identification of regulators and co-expression modules.
- **SCIMITAR** [47] *Python*  
Infers trajectory plus co-regulatory states. Requires TF preselection. Robust to noise.
- **LEAP** [41] *R*  
Starts from ordered cells (no branches). Takes into account delay effects.
- **AR1MA1 - VBEM** [48] *Matlab*  
Bayesian network (activation/inhibition with weights) from ordered cells (no branches).
- **SINCERITIES** [42] *Matlab*  
For time point data (min. 5 time points). Recommends <50TFs.
- **InformationMeasures/PIDC** [44] *Julia*  
Co-expression based on multivariate information measures. Hundreds of genes.

## Cell types/states

"steady-states"

### Co-expression

#### Cell types → Network

- **SINCERA** [52] *R*  
Pipeline for analysis of scRNA-seq. It includes prediction of key regulators for the differentially expressed genes, and possibility to integrate external data.
- **ACTION** [53]\* *Matlab*  
Pipeline for analysis of scRNA-seq. Identifies cell types and cell-type specific transcriptional regulatory networks (TFs regulating marker genes).

#### Network → cell types

- **SCENIC** [54] *R*  
Builds a GRN combining co-expression with TF motif enrichment analysis, and identifies cell types/states based on the activity of the network.

#### Only co-expression modules

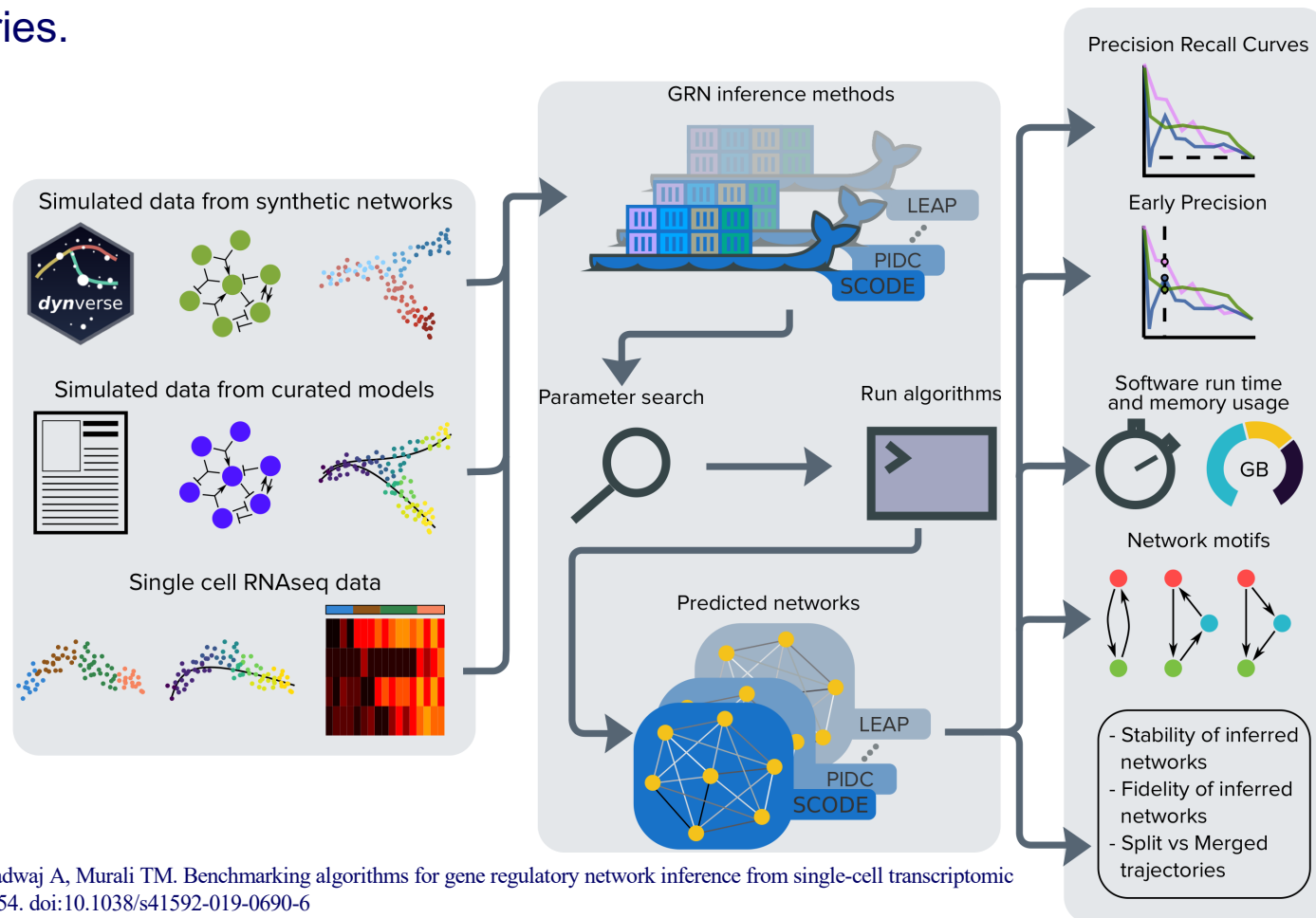
- **nlnet** [98] *R*  
Identifies gene modules based on a distance to measure predictive nonlinear relations (DCOL), which is sensitive and computationally efficient on large matrices.



# BEELINE

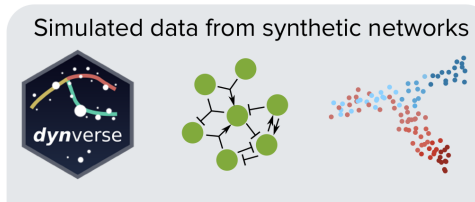
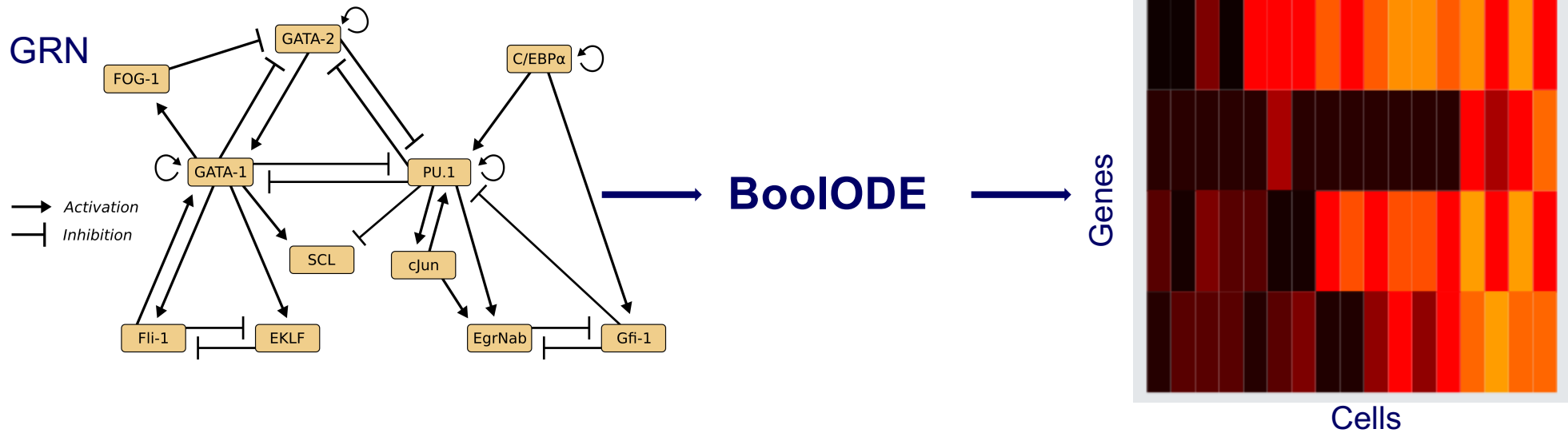
## Benchmarking gene regulatory network inference from single-cell transcriptomic data

- BEELINE is an evaluation framework incorporating 12 diverse GRN inference algorithms to assess the accuracy, robustness, and efficiency of GRN inference techniques for single-cell gene expression data based on well-defined benchmark datasets.
- BoolODE is developed for accurate simulations of Boolean models with predictable trajectories.

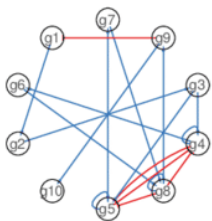


# Input type 1:

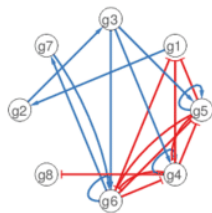
## Simulated datasets from synthetic networks



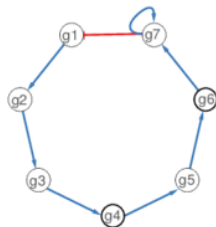
**Bifurcating**  
**Converging**



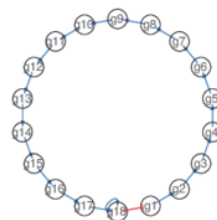
**Trifurcating**



**Linear**



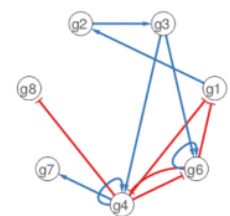
**Linear long**



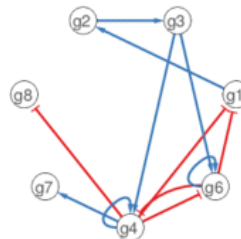
**Cycle**



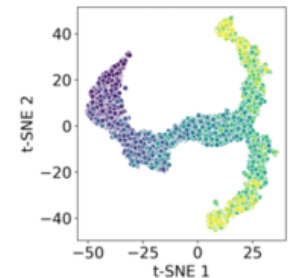
**Bifurcating**



**Bifurcating**

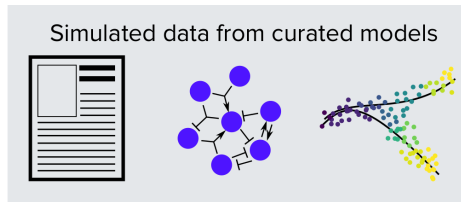


→ **BoolODE** →

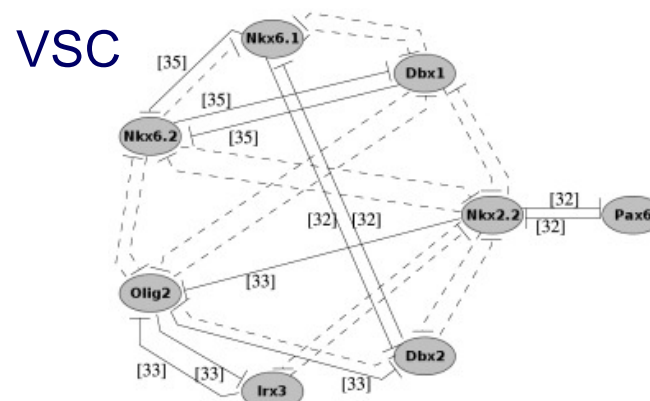
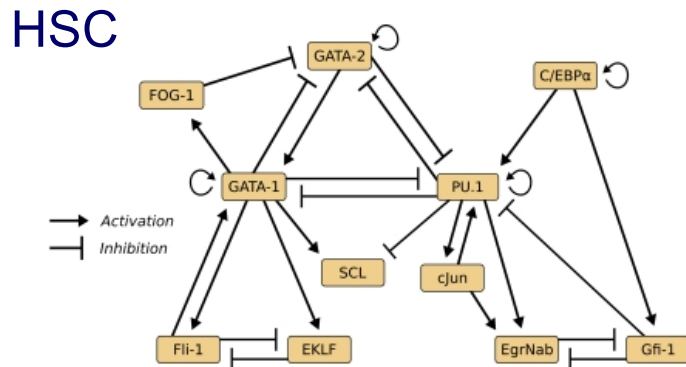
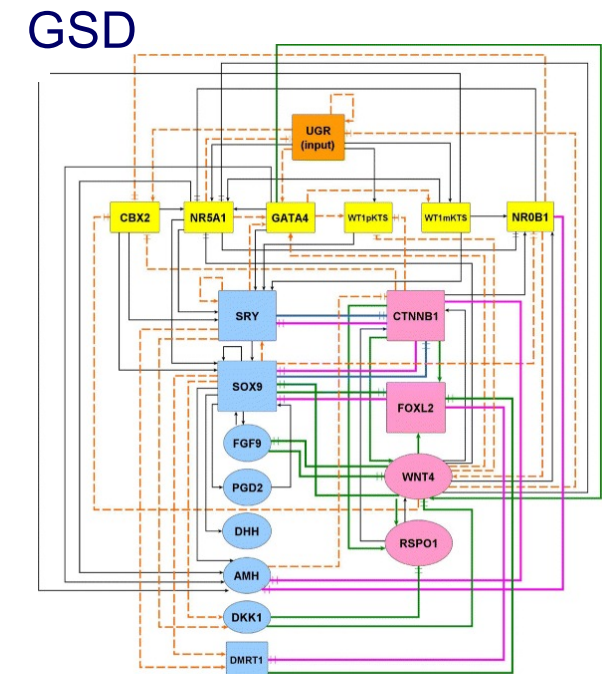
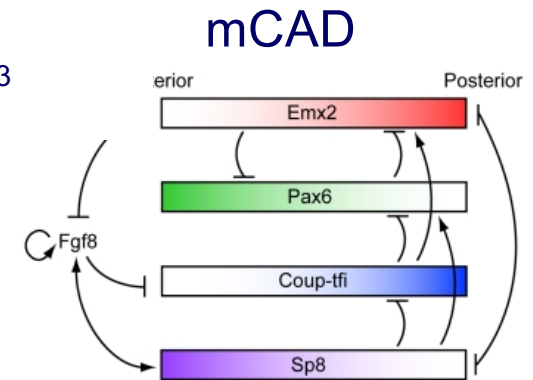
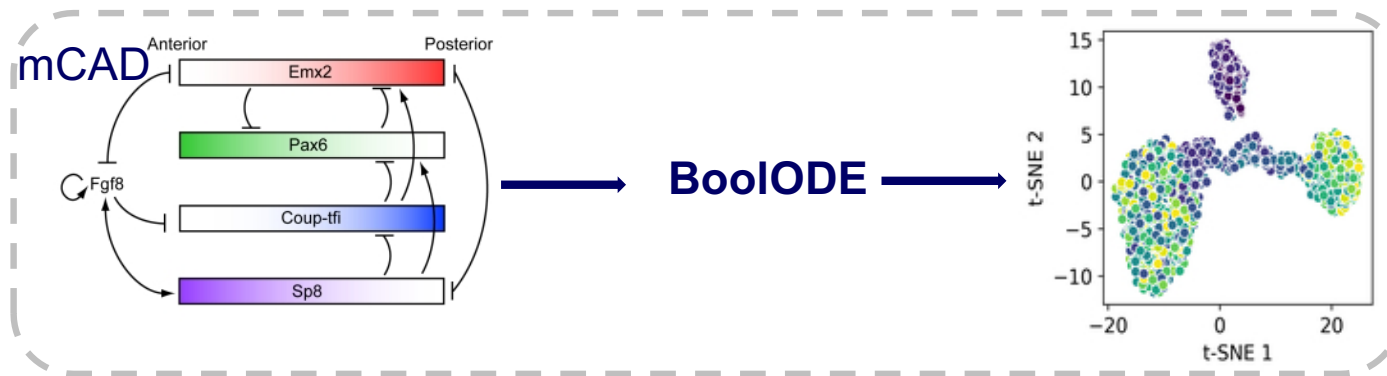


# Input type 2:

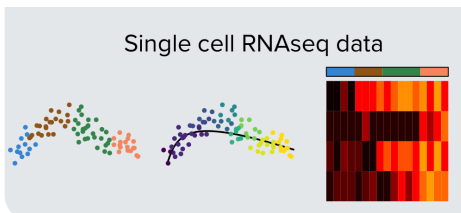
## Simulated datasets from curated models (Boolean)



- Mammalian Cortical Area Development (mCAD)<sup>1</sup>
- Ventral Spinal Cord Development (VSC)<sup>2</sup>
- Hematopoietic Stem Cell Differentiation (HSC)<sup>3</sup>
- Gonadal Sex Determination (GSD)<sup>4</sup>



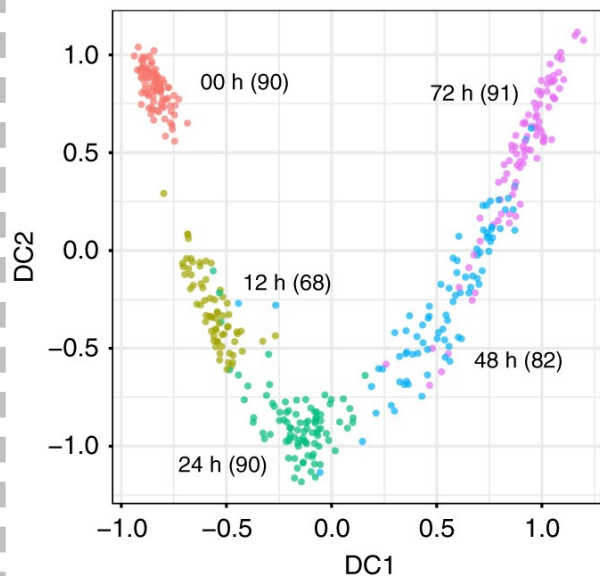
- Giacomantonio CE & Goodhill GJ A boolean model of the gene regulatory network underlying mammalian cortical area development. *PLoS Comput. Biol.* 6, e1000936 (2010).
- Lovrics A et al. Boolean Modelling reveals new regulatory connections between transcription factors orchestrating the development of the ventral spinal cord. *PLoS One* 9, e111430 (2014).
- Krumsiek J, Marr C, Schroeder T & Theis FJ Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. *PLoS One* 6, e22649 (2011).
- Rios O et al. A Boolean network model of human gonadal sex determination. *Theor. Biol. Med. Model.* 12, 26 (2015).



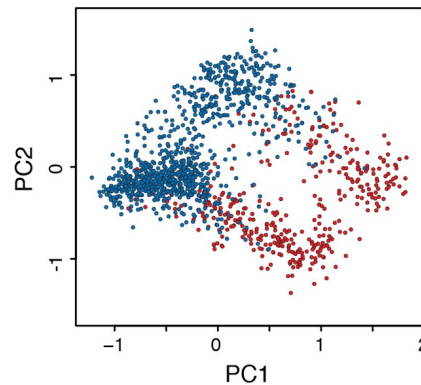
# Input type 3: Experimental scRNA-seq datasets

## Mouse

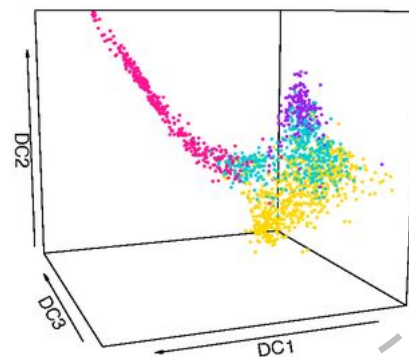
### Embryonic stem cells



### dendritic cells

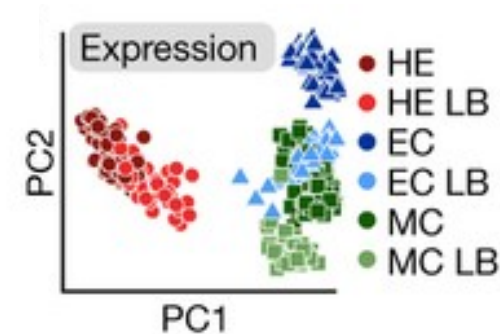


### hematopoietic stem cells

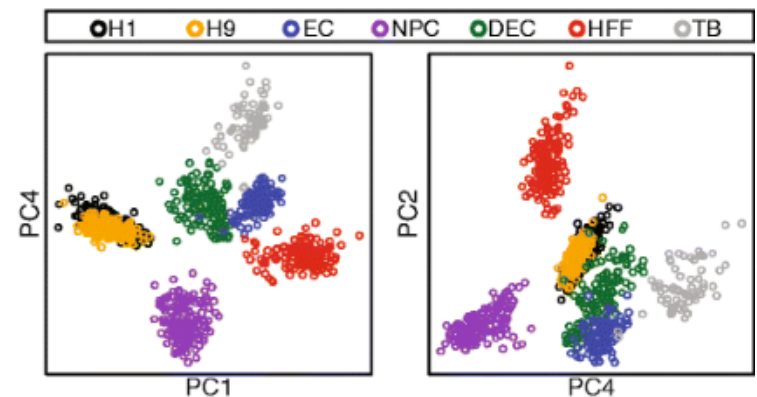


## Human

### hepatocytes

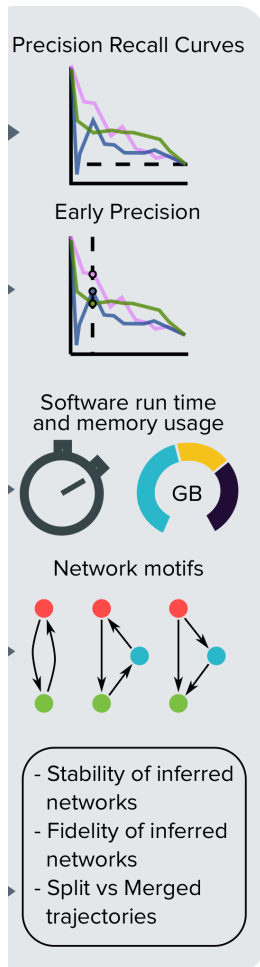


### Embryonic stem cells



- Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 128, 20–31 (2016).
- Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* 9, 619 (2018).
- Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369 (2014).
- Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546, 533–538 (2017).
- Chu, L. F. et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17, 173 (2016).

# BEELINE summary



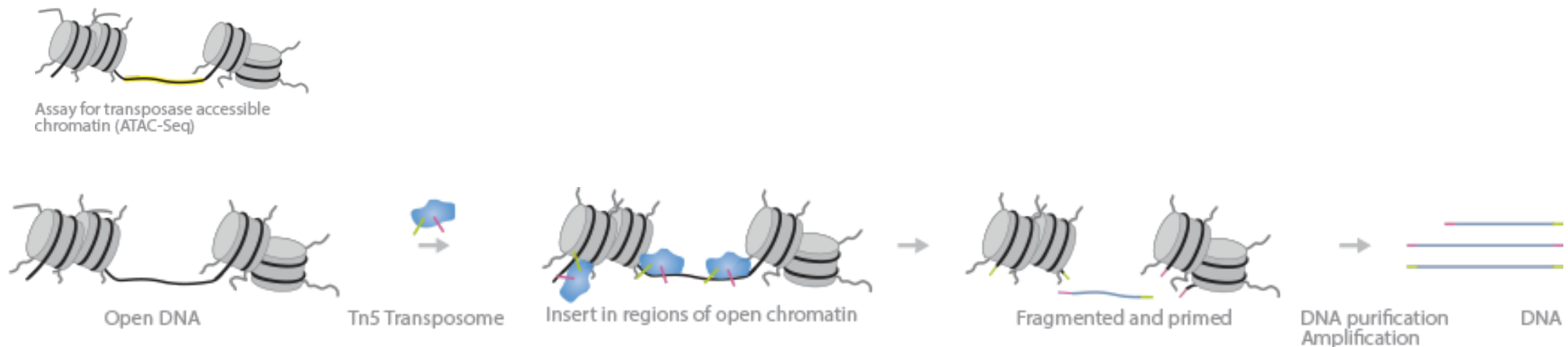
		Properties				Accuracy			Stability			Scalability (genes)								
Category		Addl. Inputs	Time ordered?	Directed?	Signed?	Synthetic	Curated	scRNA-Seq	Datasets	Runs	Dropouts	Pseudotime	Time				Memory			
													100	500	1k	2k	100	500	1k	2k
PIDC	MI	-	✗	✗	✗							-	1s	1m	5m	30m	0.1G	0.1G	0.5G	1G
GENIE3	RF	-	✗	✓	✗							-	5m	1h	3h	12h	1G	2G	2G	2G
GRNBOOST2	RF	-	✗	✓	✗							-	1m	10m	30m	1h	0.1G	0.1G	0.5G	1G
SCODE	ODE+Reg	ODE parameters	✓	✓	✓							-	1m	5m	5m	30m	1M	0.1G	0.1G	0.5G
PPCOR	Corr	-	✗	✗	✓							-	1s	1s	1s	1s	1M	0.1G	0.1G	0.1G
SINCERITIES	Reg	-	✓	✓	✓							-	1s	1m	5m	10m	0.1G	0.1G	0.1G	0.5G
SCRIBE	MI	Type of RDI	✓	✓	✗			-				-	5m	2h	6h	-	0.1G	0.1G	0.1G	-
SINGE	GC	Regression parameters	✓	✓	✗			-				-	3h	>1d	>1d	-	0.5G	0.5G	1G	-
LEAP	Corr	Lag	✓	✓	✗			-				-	1s	1s	1m	5m	1M	0.1G	0.1G	0.5G
GRISLI	ODE+Reg	Regression parameters	✓	✓	✗			-				-	5m	1h	3h	-	0.5G	>4G	>4G	-
GRNVBEM	Reg	-	✓	✓	✓			-				-	1m	>1d	-	-	0.1G	2G	-	-
SCNS	Bool	Boolean model parameters	✓	✓	✓			-				-	-	-	-	-	-	-	-	-

</

# ATAC-seq

Assay for Transposase-Accessible Chromatin using sequencing

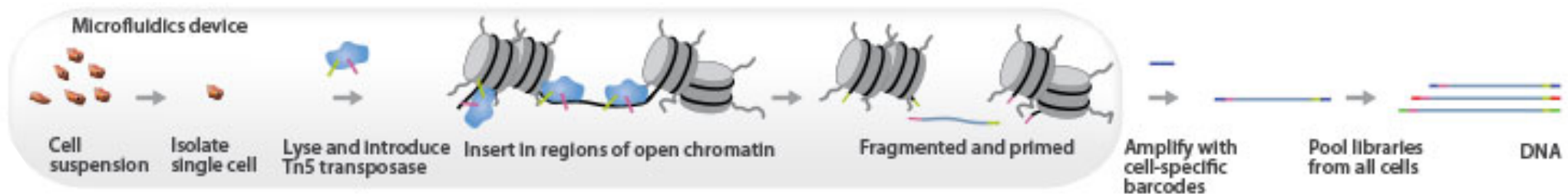
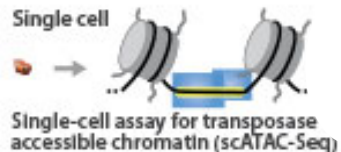
- In ATAC-Seq, genomic DNA is exposed to Tn5, a highly active transposase
- Tn5 simultaneously fragments DNA, preferentially inserts into open chromatin sites, and adds sequencing primers (a process known as tagmentation)
- The sequenced DNA identifies the open chromatin and data analysis can provide insight into gene regulation



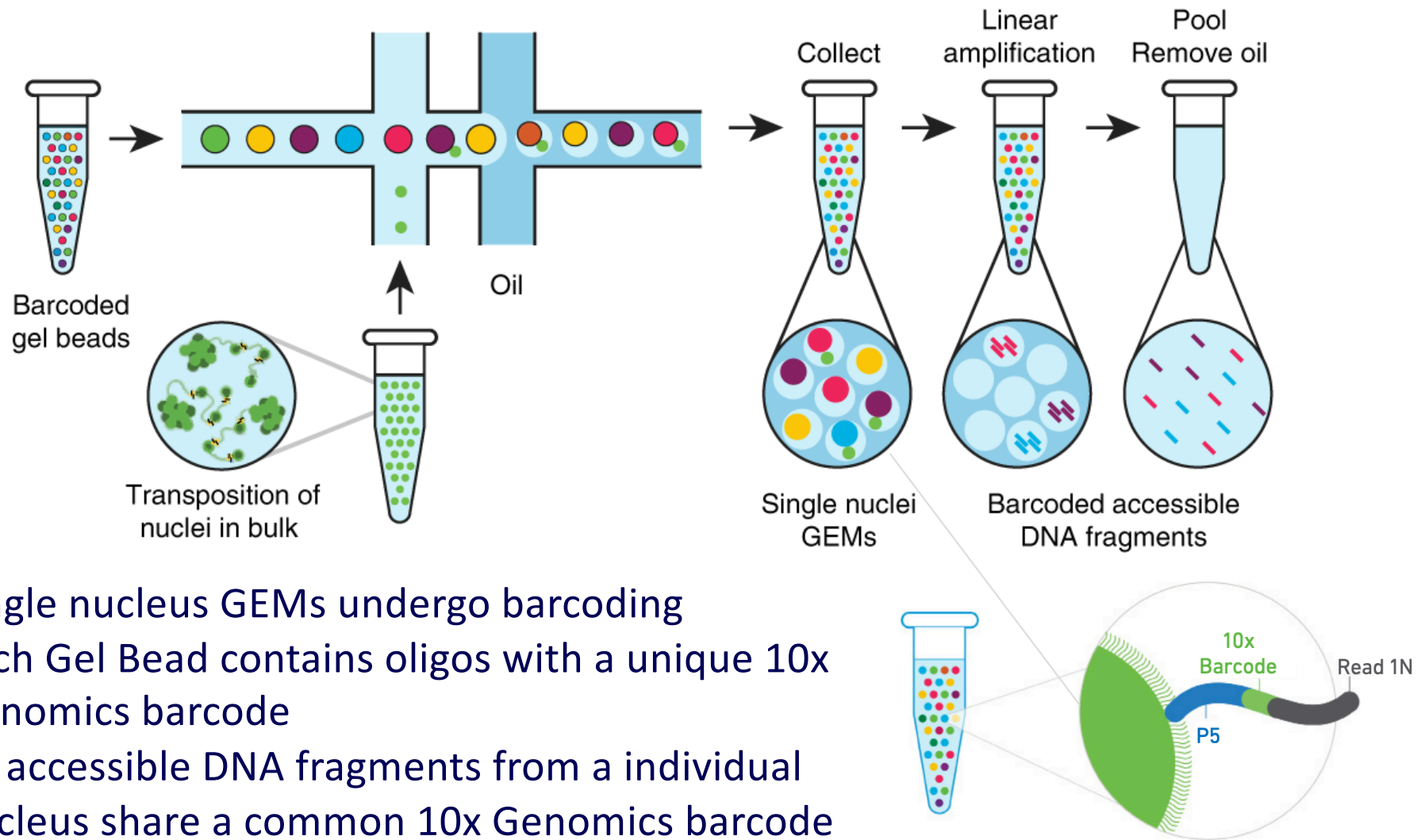


# Single-Cell ATAC-seq

- scATAC-Seq utilizes Tn5 transposase and barcoding of individual cells to profile chromatin accessibility at single cell resolution
- Single cells could be captured by combinatorial cell indexing strategies or with use of a microfluidic device



# scATAC-seq using 10x Genomics technology



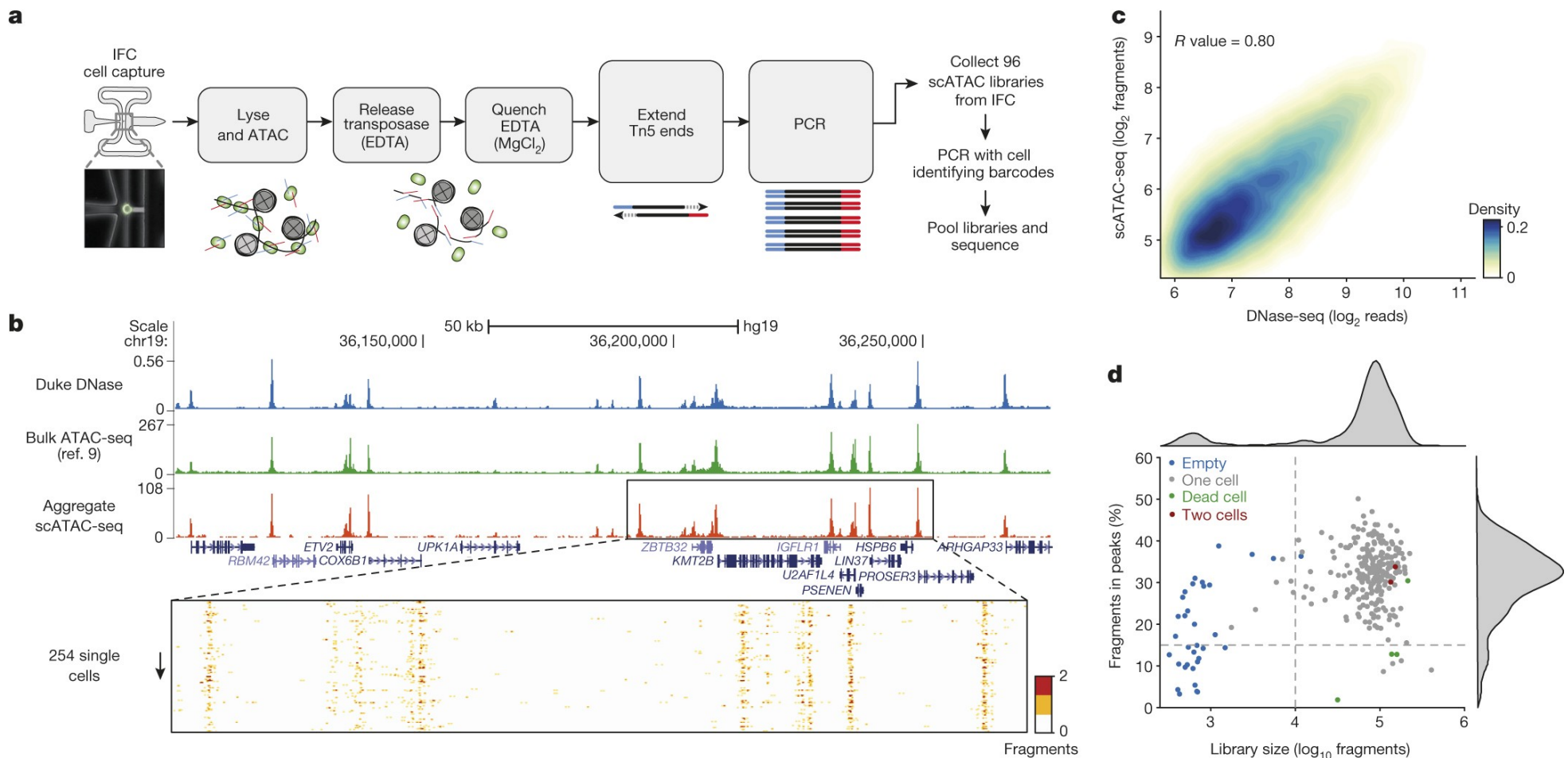
Satpathy, A.T., Granja, J.M., Yost, K.E. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019). <https://doi.org/10.1038/s41587-019-0206-z>  
<https://www.10xgenomics.com/videos/ef93x01cw0?autoplay=true>

**10x Genomics barcode: 16bp long**  
~750,000 different barcodes

# scATAC-seq

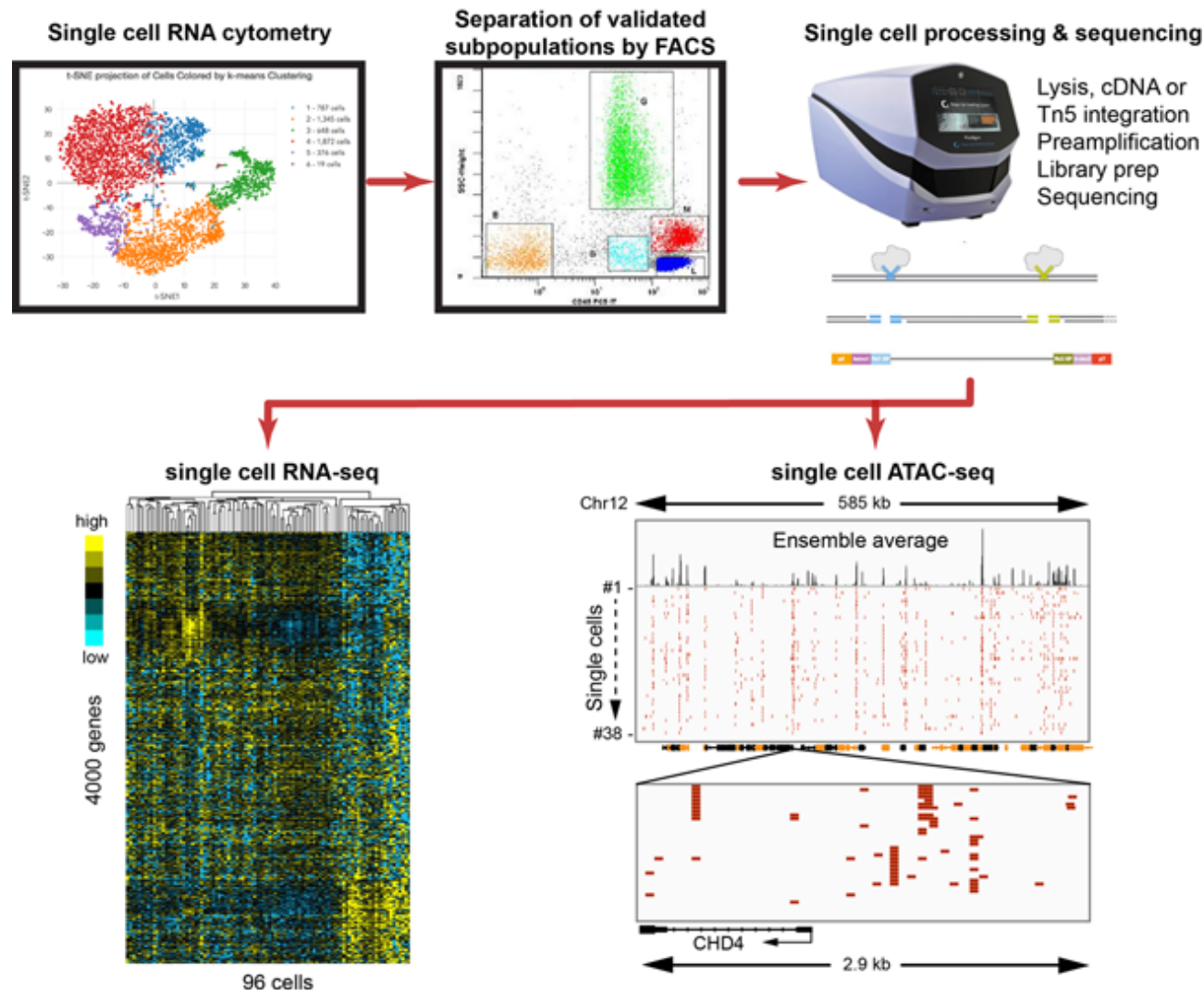
Workflow for measuring single epigenomes using scATAC-seq on a microfluidic device

- Develop technology for single cell epigenomic
- Single-cell ATAC-seq provides an accurate measure of chromatin accessibility genome-wide



# scATAC-seq + scRNA-seq

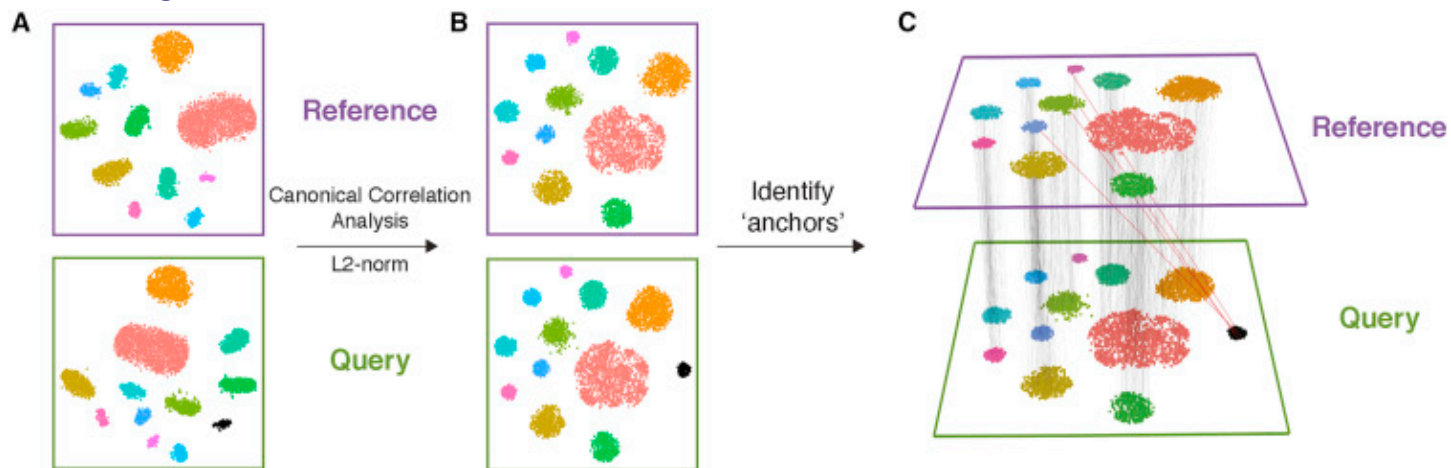
- linking single-cell epigenomics and single-cell transcriptomics



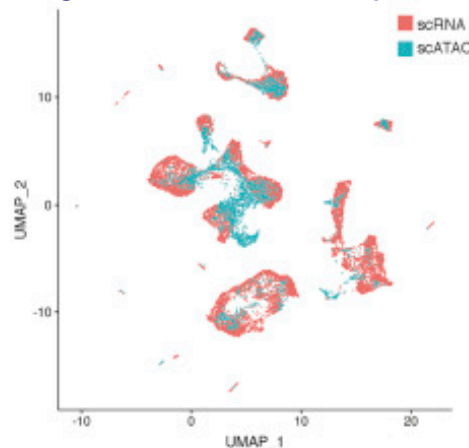
# Seurat

## Integrating scATAC-seq with scRNA-seq to annotate cell types

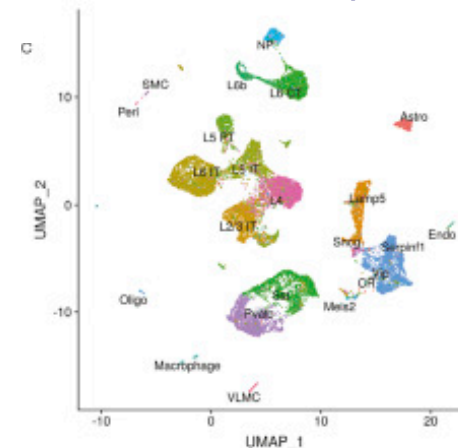
- a strategy to “anchor” diverse datasets together, enabling integrate single-cell measurements not only across scRNA-seq technologies, but also across different modalities



Integrated PBMC scATAC-seq and scRNA-seq



Annotated PBMC scATAC-seq and scRNA-seq

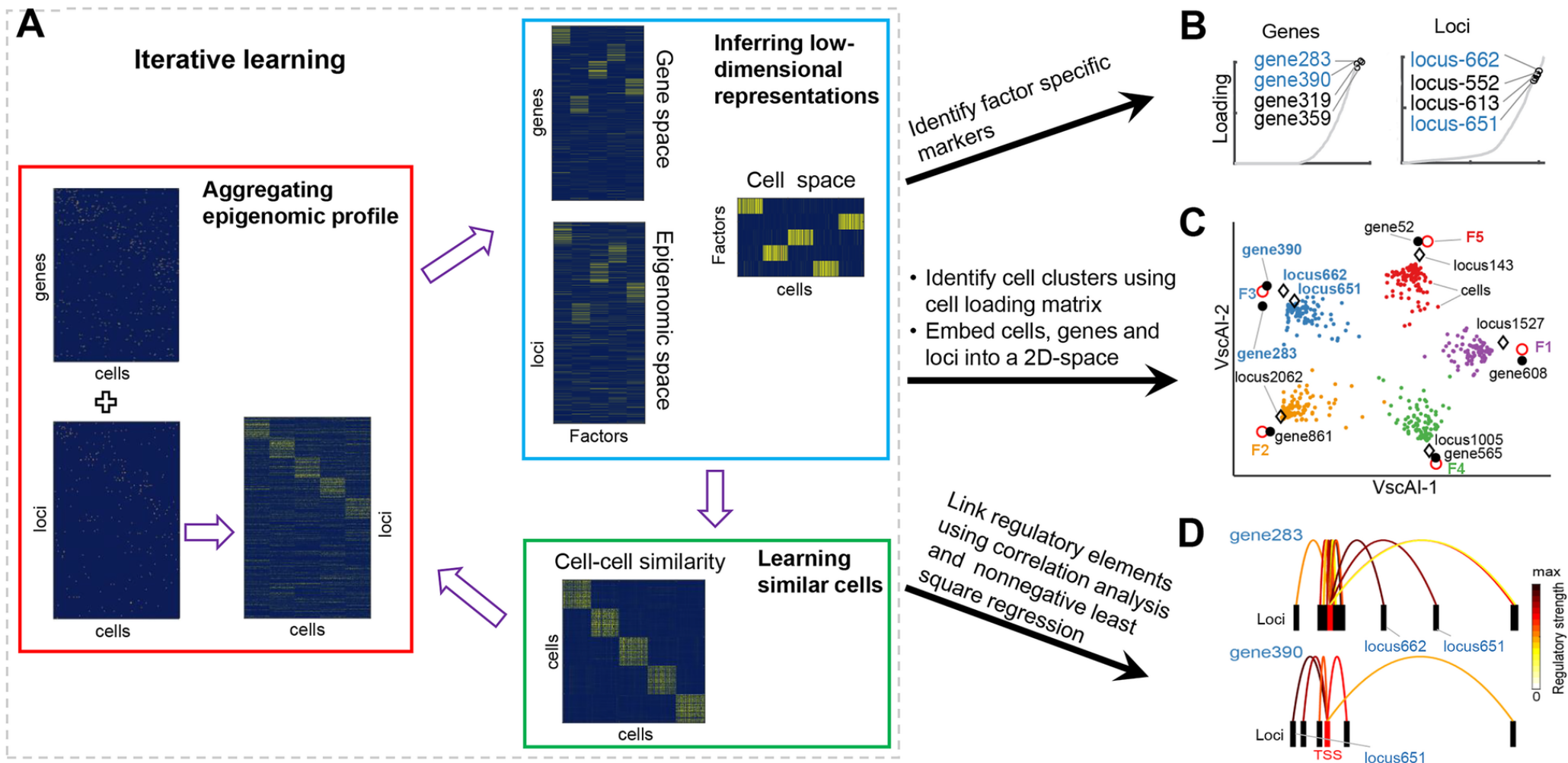




# scAI

## single-cell aggregation and integration

- scAI is an unsupervised approach for integrative analysis of gene expression and chromatin accessibility or DNA methylation profiles measured in the same individual cells





# scAI method

- ‘To deconvolute heterogeneous single cells from both transcriptomic and epigenomic profiles, scAI aggregates the sparse/binary epigenomic profile in an unsupervised manner to allow coherent fusion with transcriptomic profile while projecting cells into the same representation space using both the transcriptomic and epigenomic data’

low-dimensional representations via the matrix factorization model:

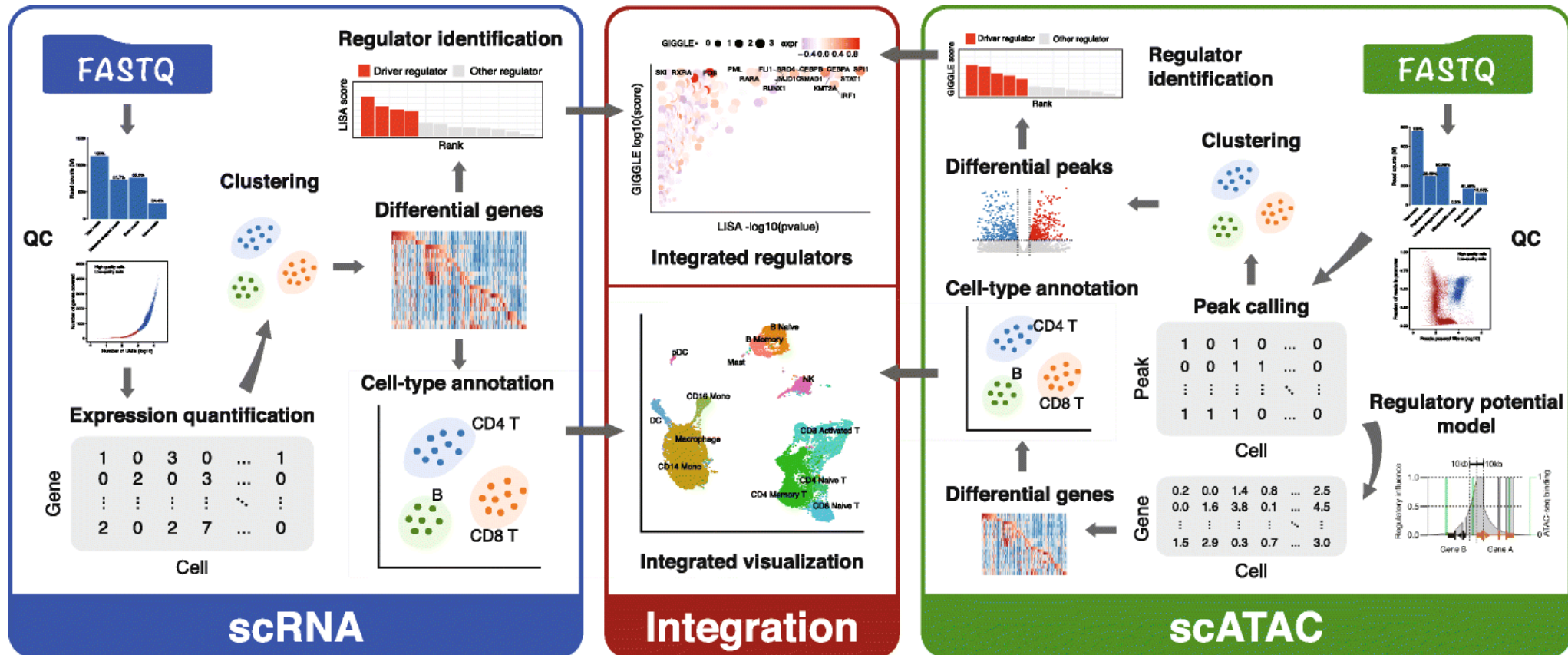
$$\min_{W_1, W_2, H, Z \geq 0} \alpha \|X_1 - W_1 H\|_F^2 + \|X_2 (Z \circ R) - W_2 H\|_F^2 + \lambda \|Z - H^T H\|_F^2 + \gamma \sum_j \|H_{.j}\|_1^2$$

- $X_1$  ( $p$  genes in  $n$  cells): normalized scRNA-seq data matrix
- $X_2$  ( $q$  loci in  $n$  cells) : single-cell chromatin accessibility or DNA methylation data matrix
- $W_1, W_2$  : the gene loading and locus loading matrices with sizes  $p \times K$  and  $q \times K$  ( $K$  is the rank)
- $H$  : the cell loading matrix with size  $K \times n$
- $Z$  : cell-cell similarity matrix
- $R$  : a binary matrix generated by a binomial distribution with a probability
- s.  $\alpha, \lambda, \gamma$  : regularization parameters
- $\circ$  : dot multiplication

# MAESTRO

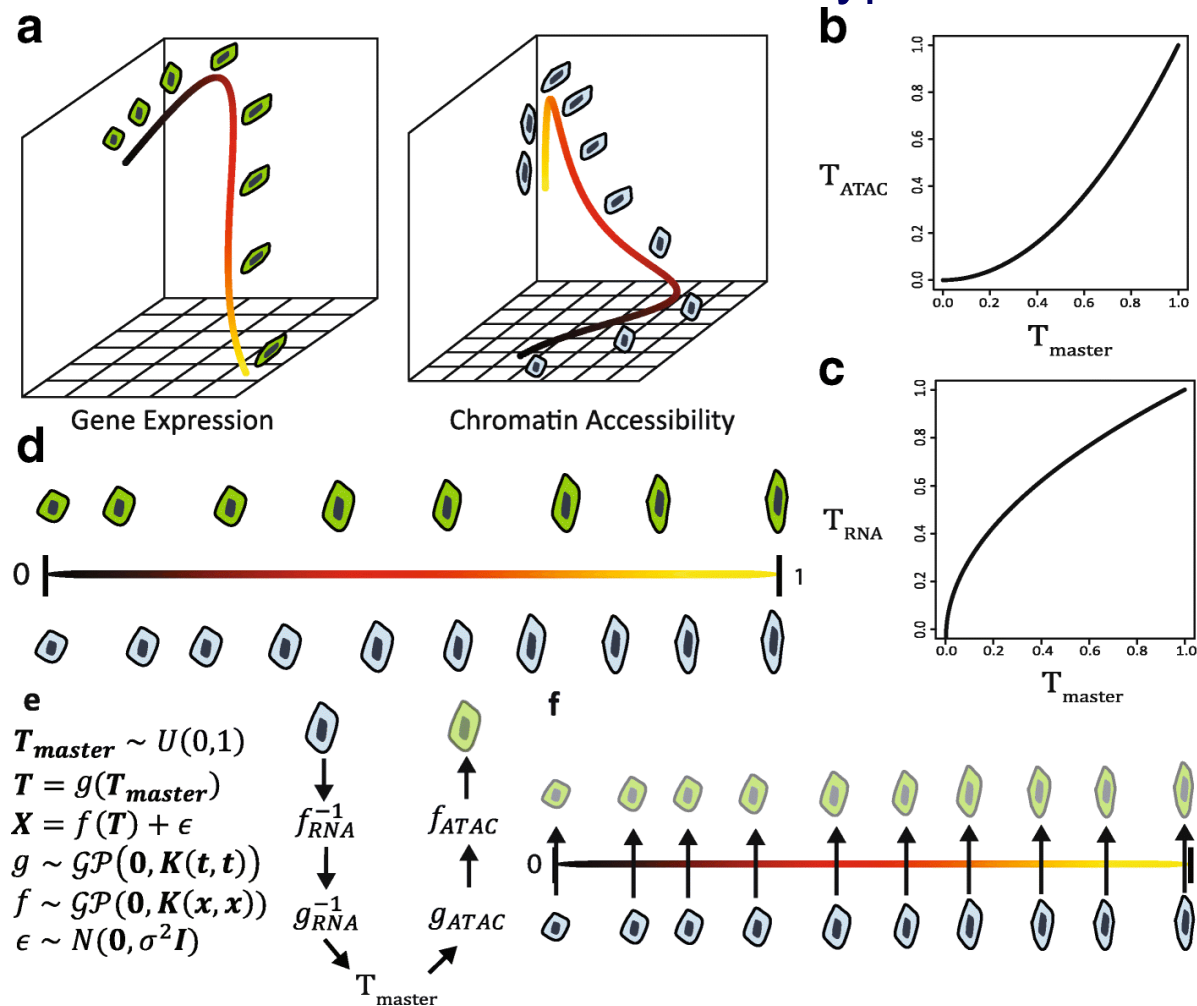
## Model-based Analyses of single-cell transcriptome and regulome

- a comprehensive computational workflow for integrative analysis of scRNA-seq and scATAC-seq data from multiple platforms.



# MATCHER

- an approach for integrating multiple types of single cell measurements
- MATCHER uses manifold alignment to infer single cell multi-omic profiles from transcriptomic and epigenetic measurements performed on different cells of the same type



# Single cell deconvolution

## Pros for Bulk-seq

- Can assay entire sample at once
- Can help identify transcription changes in individual cell types
- Huge amount of data out there already
- Cheap

## Cons

- Lose single cell information

**Bulk**  
\$200/sample (Novogene)



Bulk transcriptomic analyses lose single cell information

How to computationally figure out what went into the mixture?



**Single cell**  
\$4000 ~ 10000/sample



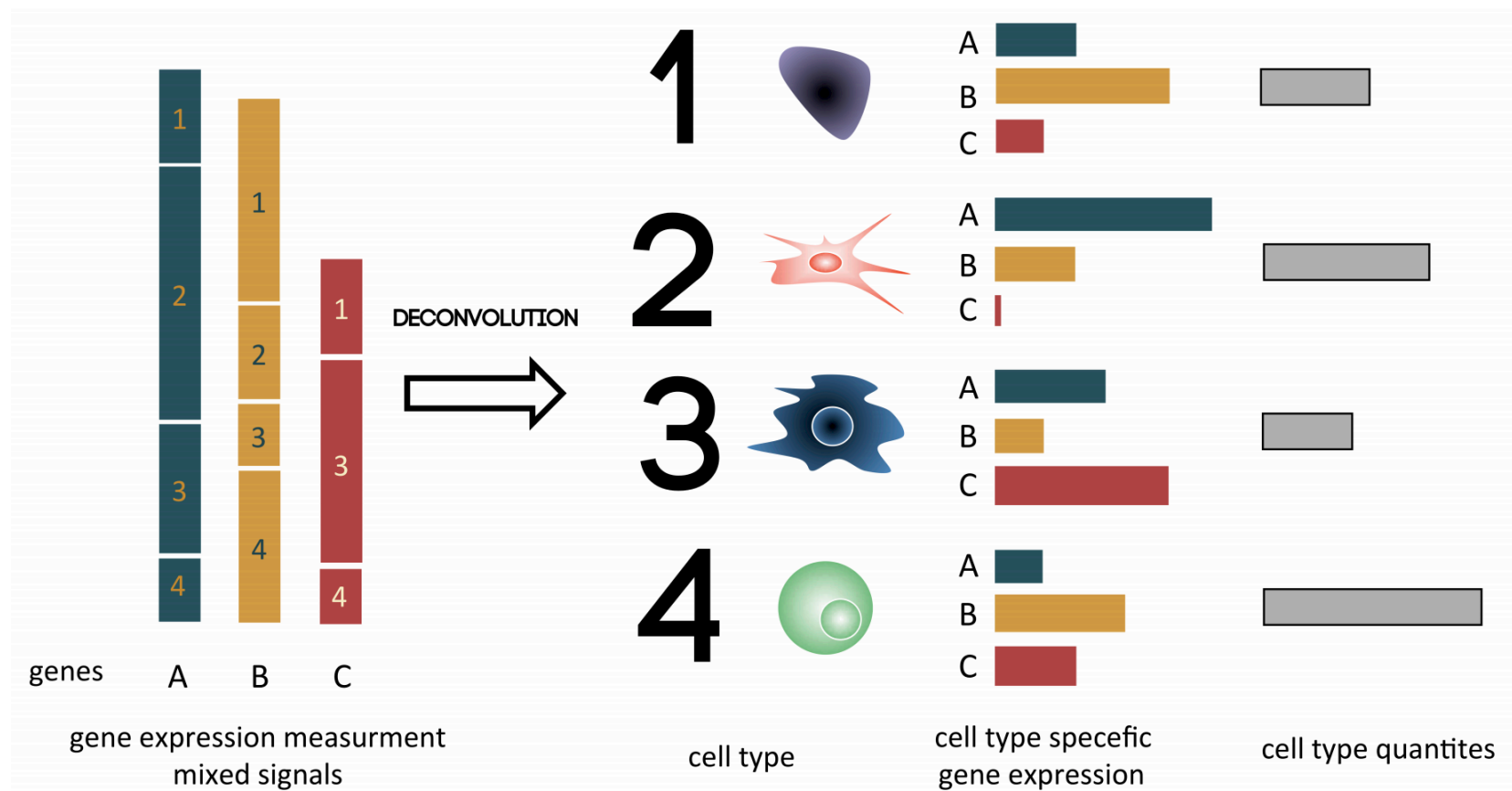
Single cell transcriptomic analyses retain single cell

[https://projects.iq.harvard.edu/files/chanbioinformatics/files/cell\\_type\\_deconvolution.pdf](https://projects.iq.harvard.edu/files/chanbioinformatics/files/cell_type_deconvolution.pdf)

ASHG 2019 scRNAseq  
HiPlex oral presentation  
[https://www.youtube.com/watch?v=YIRemO\\_TE3Y](https://www.youtube.com/watch?v=YIRemO_TE3Y)

# Single cell deconvolution

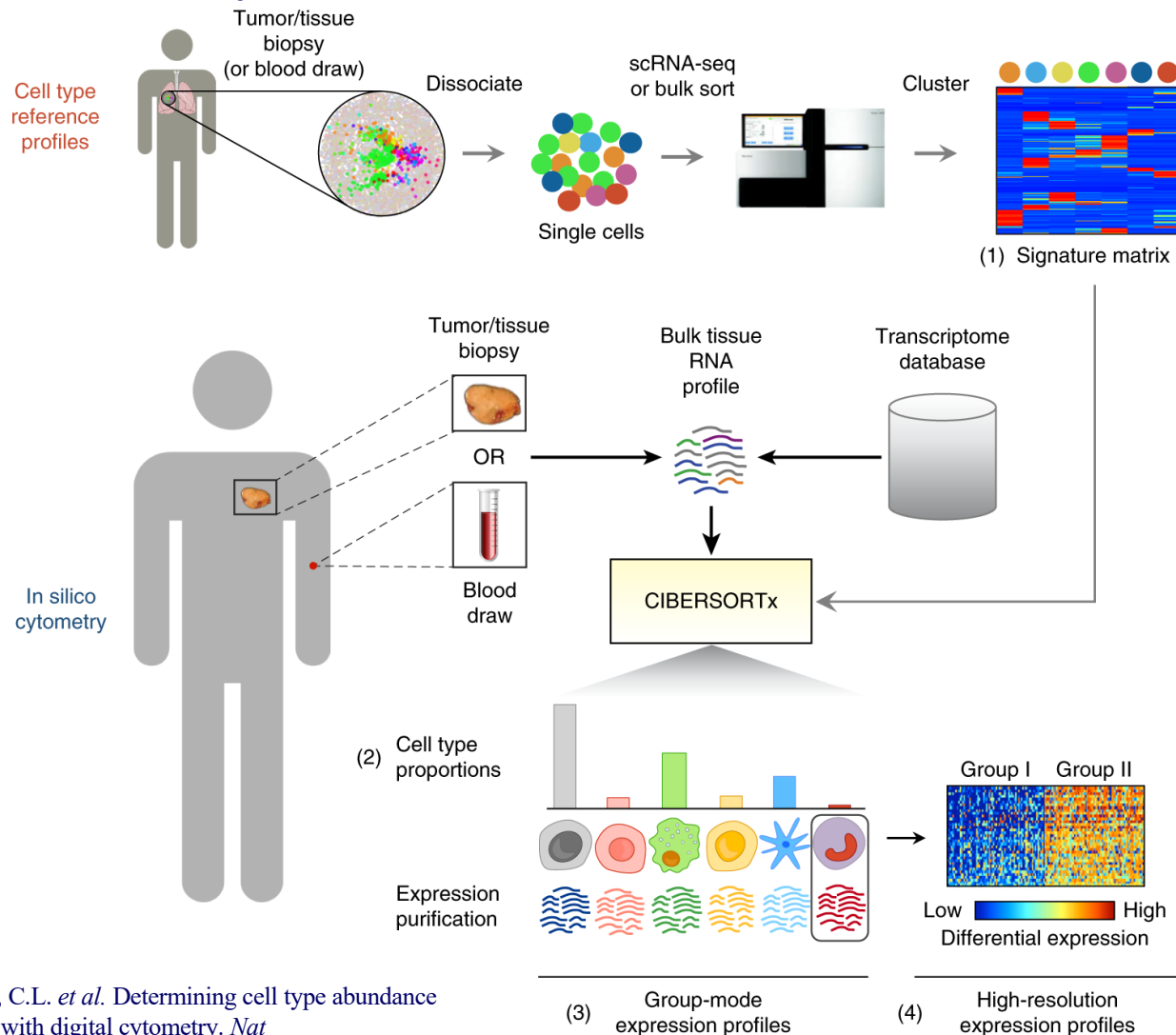
- ‘a system of equations that describe the expression of each gene in a heterogeneous sample as a linear combination of the expression levels of that gene across the different cell subsets present in the sample, weighted by their relative cell fractions’



# CIBERSORTx

cell-type identification by estimating relative subsets of RNA transcripts

- a machine learning method to infer cell-type-specific gene expression profiles without physical cell isolation

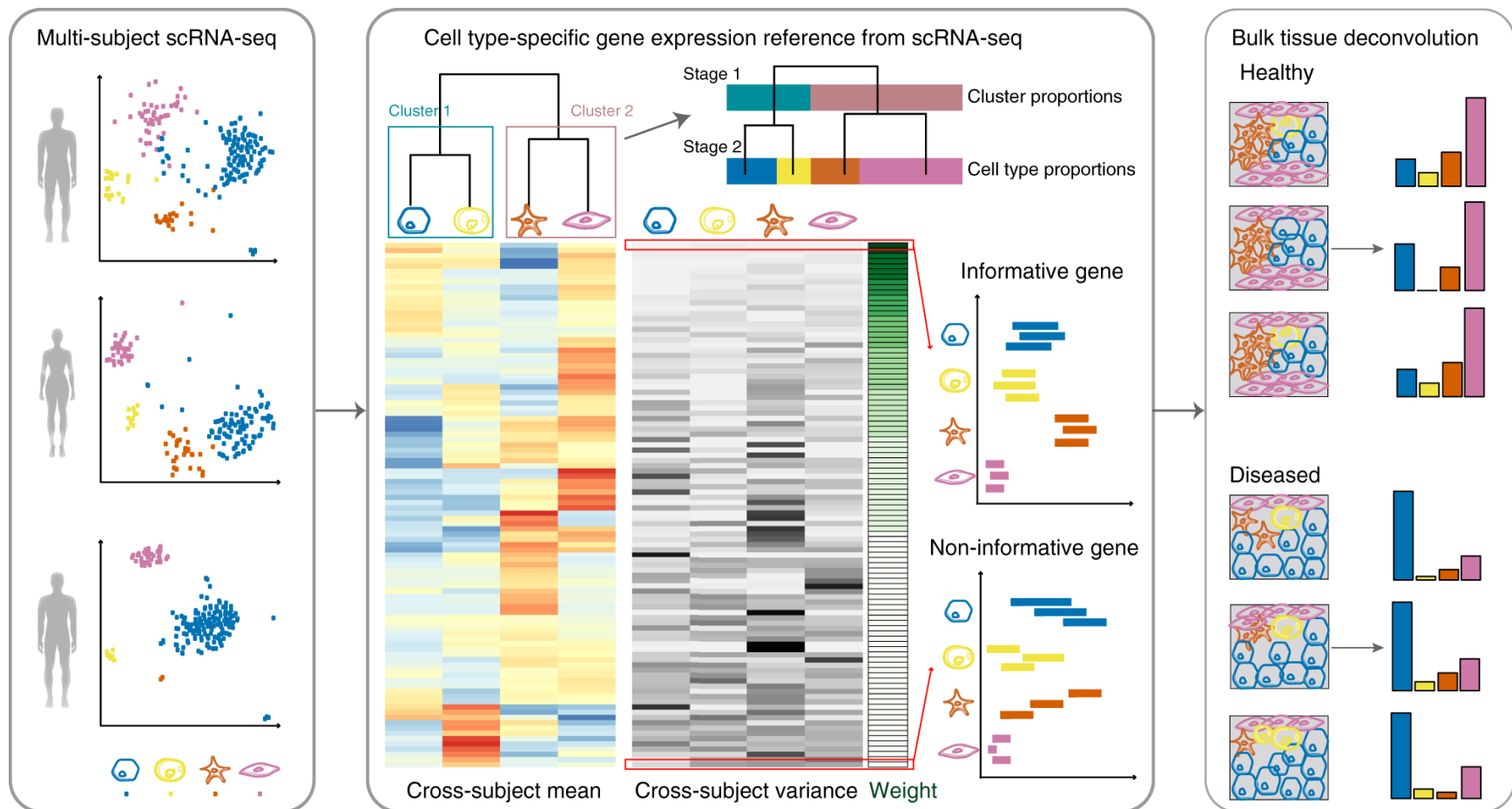




# MuSiC

## Multi-Subject Single Cell deconvolution

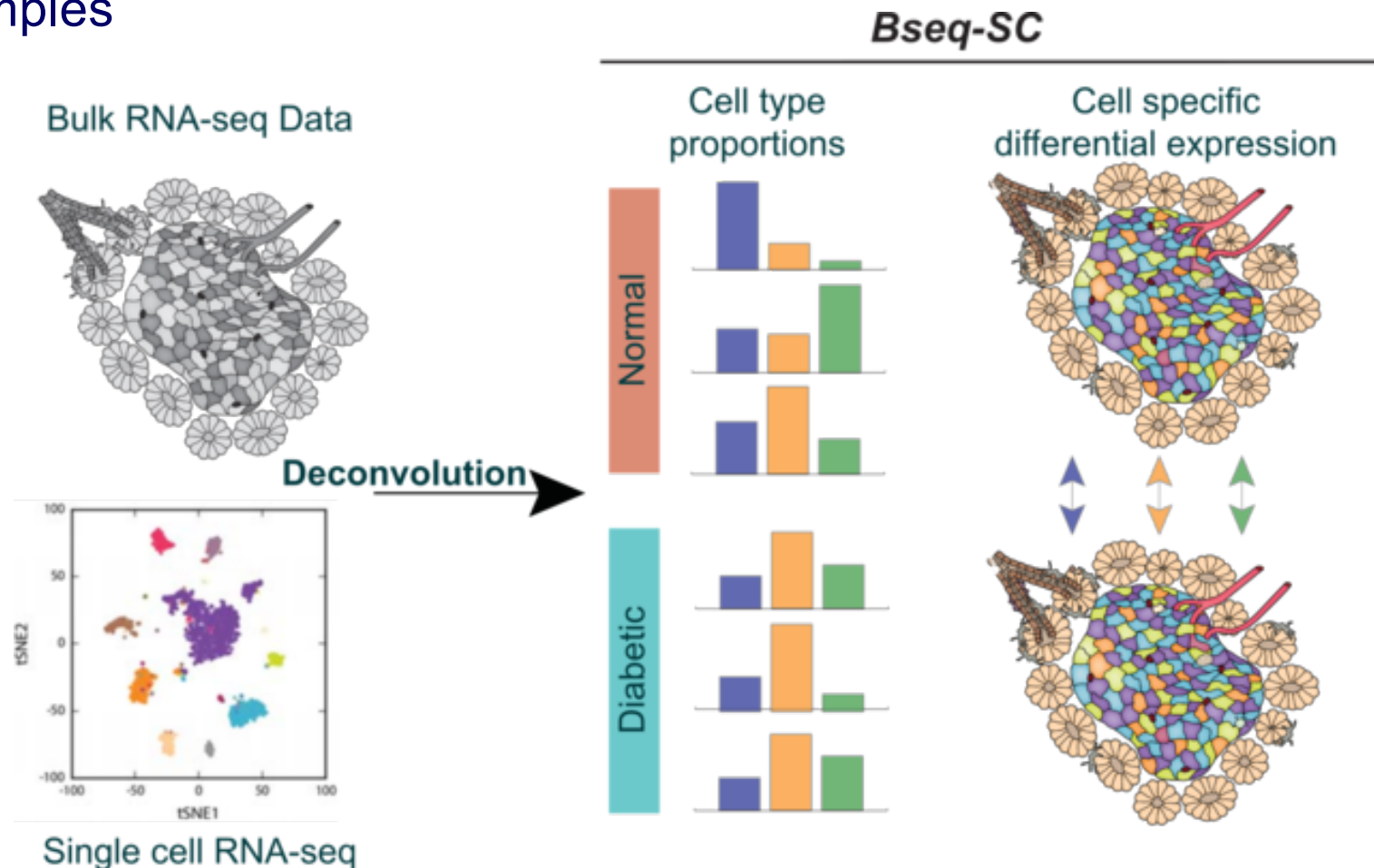
- a method that utilizes cell-type specific gene expression from single-cell RNA sequencing (RNA-seq) data to characterize cell type compositions from bulk RNA-seq data in complex tissues



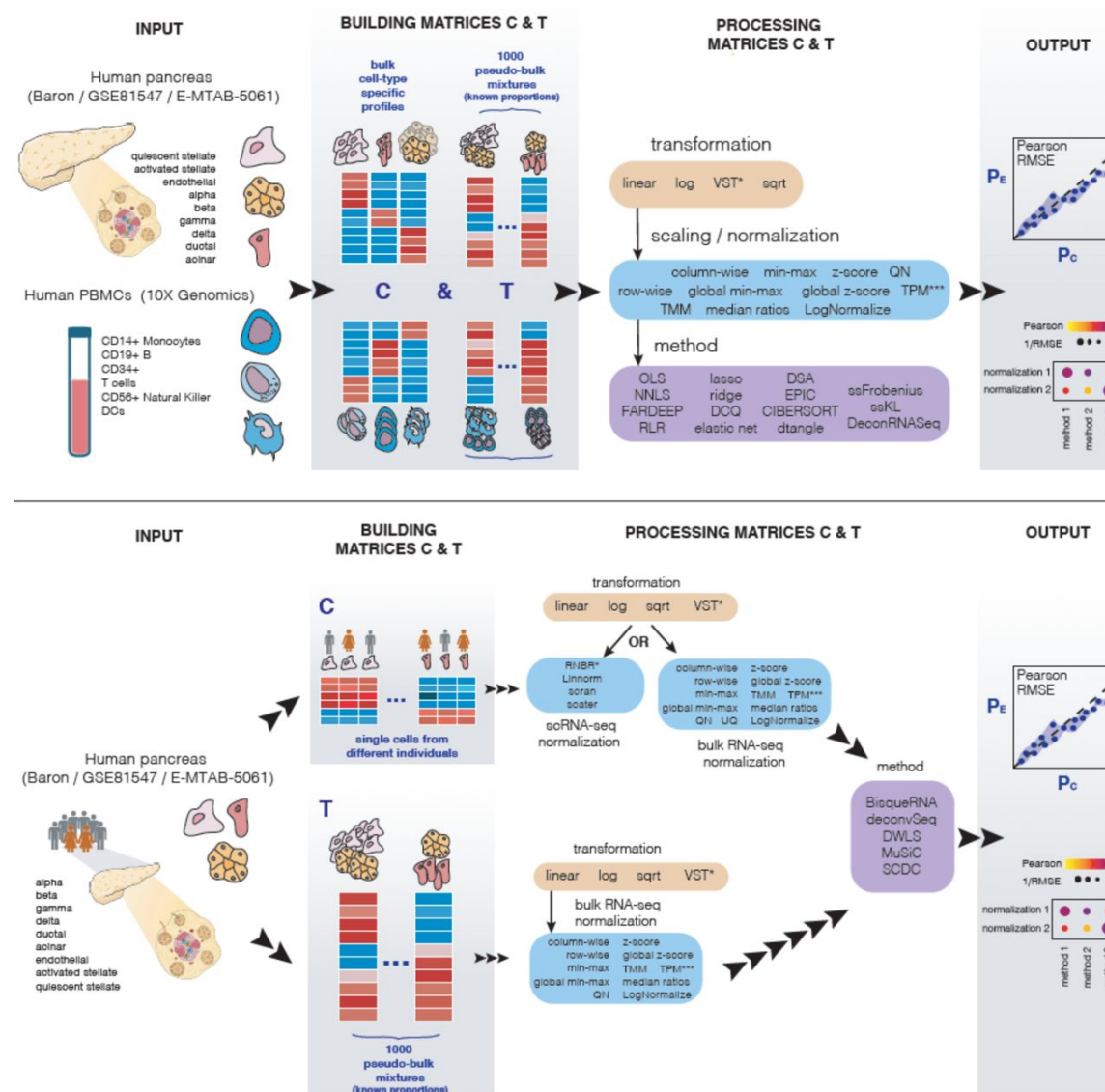
# BSEQ-sc

## Deconvolution of Bulk Sequencing Experiments using Single Cell Data

- a bioinformatics analysis pipeline that leverages single-cell sequencing data to estimate cell type proportion and cell type-specific gene expression differences from RNA-seq data from bulk tissue samples



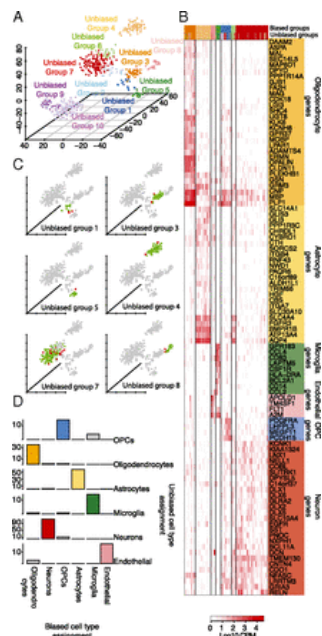
# Comprehensive benchmarking of computational deconvolution of transcriptomics data



# Single cell transcriptome data resource for human brain

- 8 excitatory and 8 inhibitory adult neuronal subtypes (i.e., cell expression clusters)
- Major adult non-neuronal types: astrocytes, endothelial, microglia, oligodendrocytes, and oligodendrocyte progenitor (OPC), pericyte
- Developmental neuronal and non-neuronal types

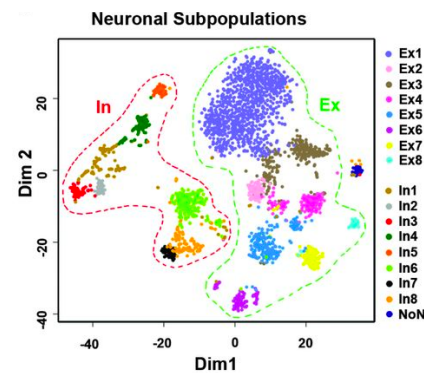
Read-count based; e.g., Transcripts Per Kilobase Million (TPM)



~400 cells  
(Darmanis et al.,  
PNAS, 2015)

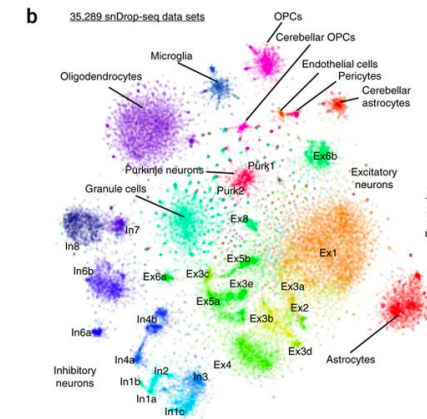


~ 900 cells (PsychENCODE)



~3000 cells (Lake et al.,  
Science, 2016)

Molecular-count based; e.g., Unique molecular identifiers (UMI)



~10319 cells (Lake et al., Nature Biotech, 2018)

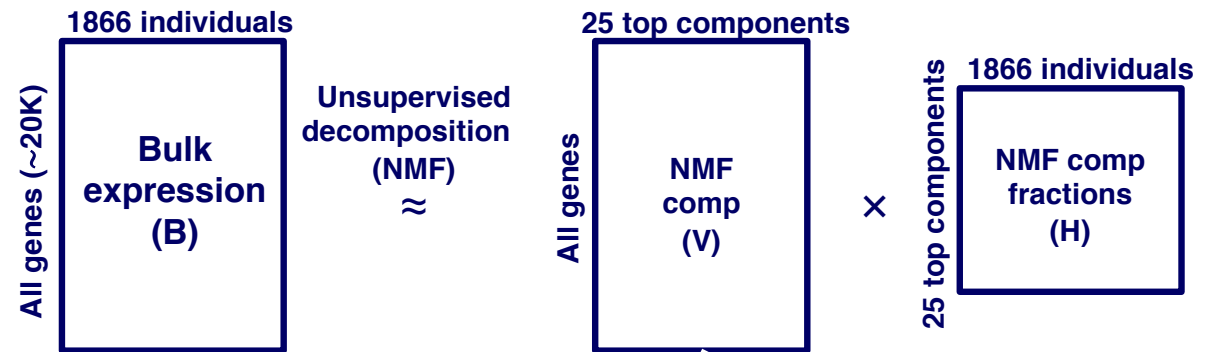


~ 17,093 cells (PsychENCODE)

# Single cell deconvolution

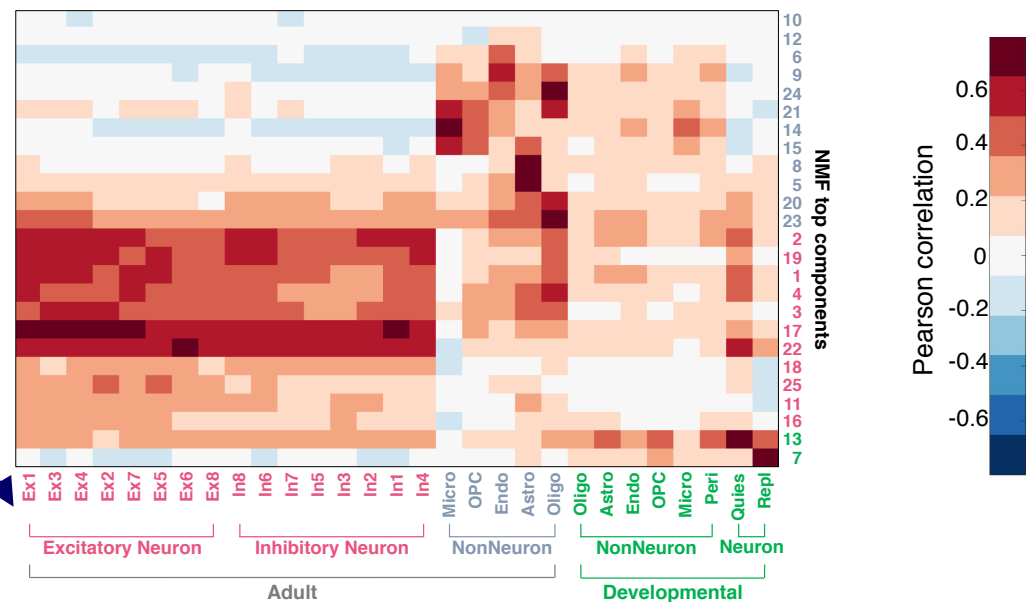
## Step 1: unsupervised learning to see brain cell types

Non-negative matrix factorization (NMF)



Single cell signatures

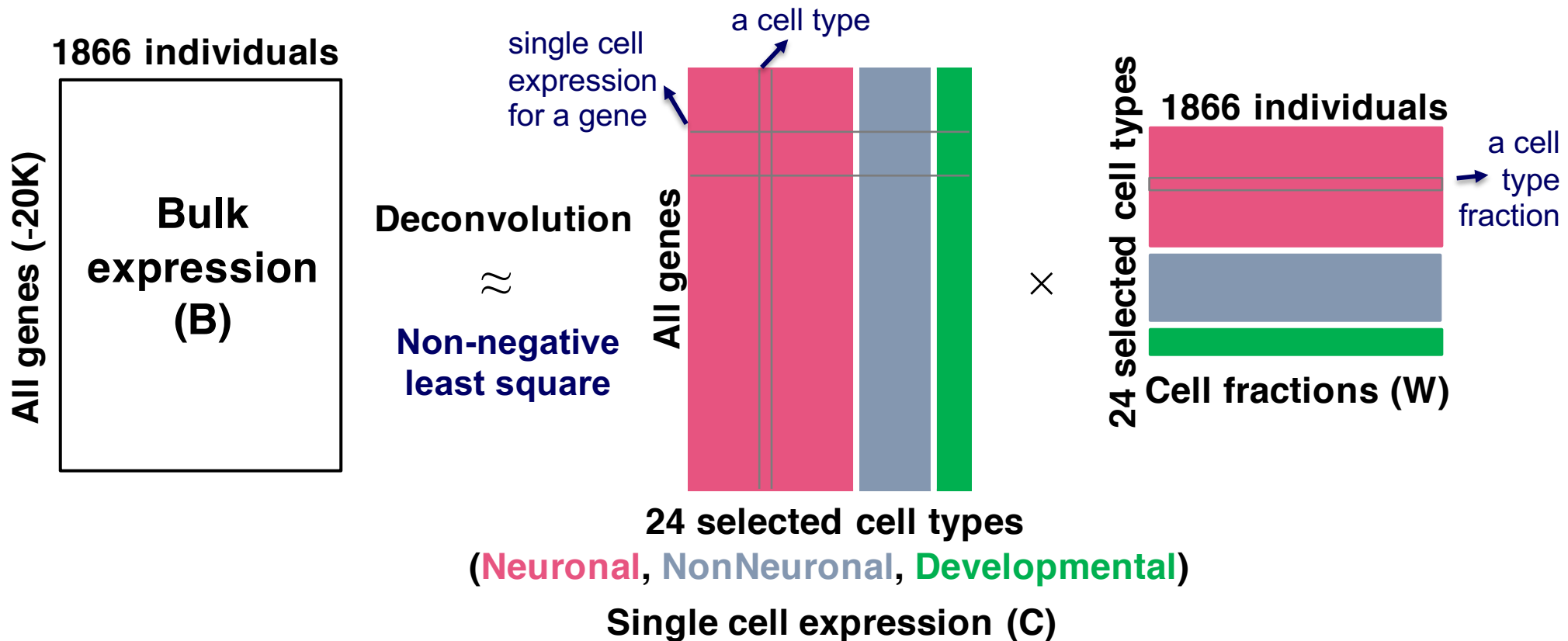
- ~14,000 cells (Lake et al., Science, 2016&2018)
- ~400 cells (Darmanis et al., PNAS, 2015)
- ~18,000 cells (PsychENCODE)





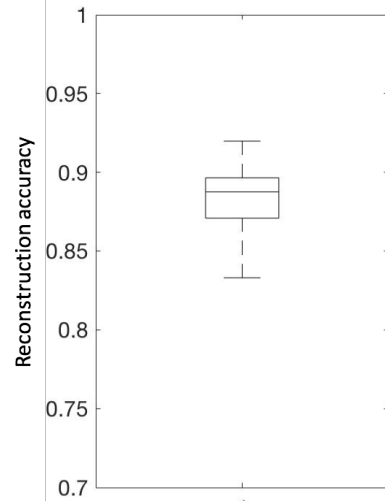
# Single cell deconvolution

## Step 2: supervised learning to estimate cell fractions

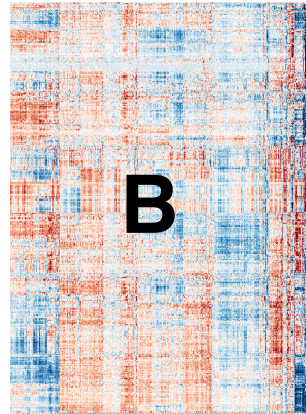


# Cell fractions explain cross-population variation in human brain

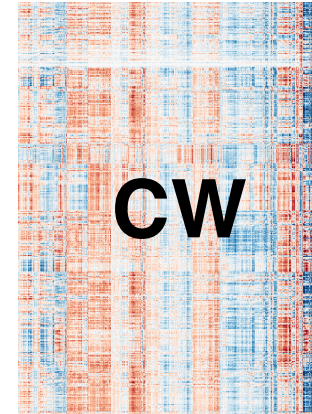
Individual and cross-population reconstruction accuracy via deconvolution



1 -



-

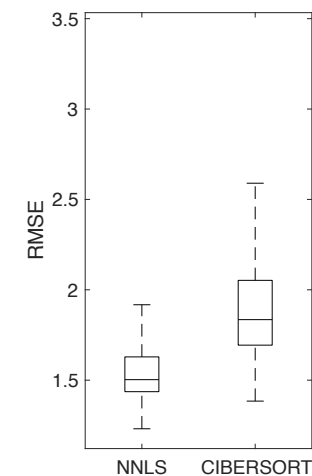
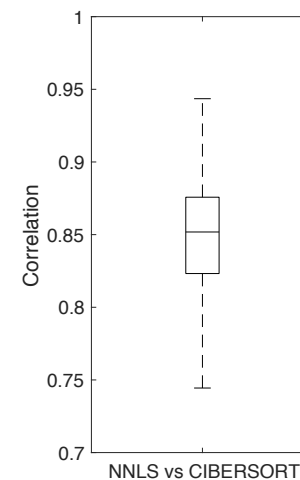
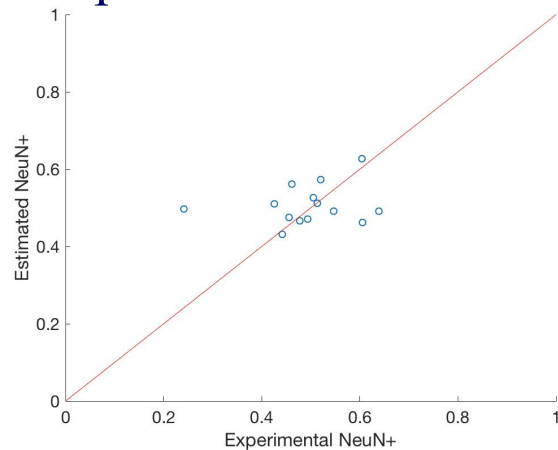


2

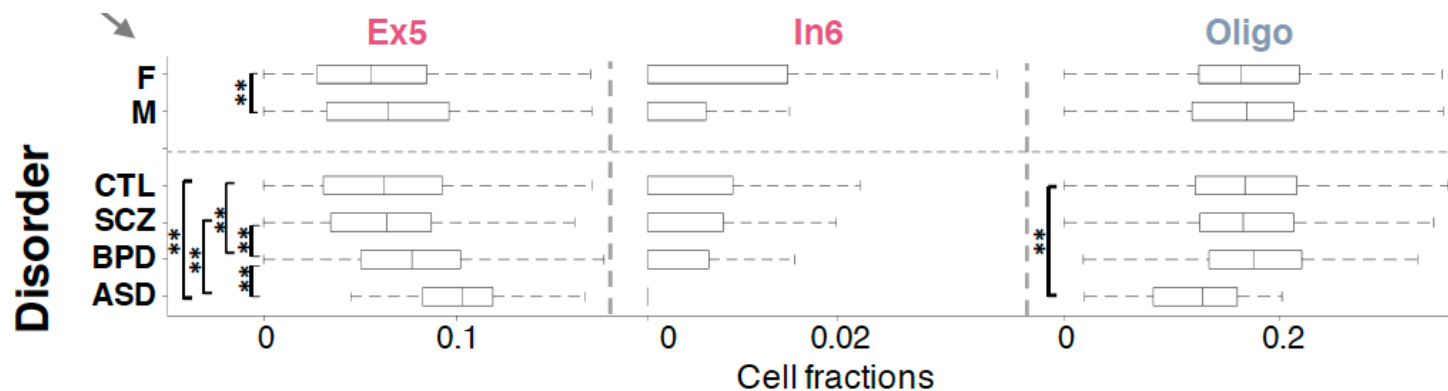
$$/ \| \mathbf{B} \|^2 > 85\%$$

Comparison with existing methods

Experimental validation

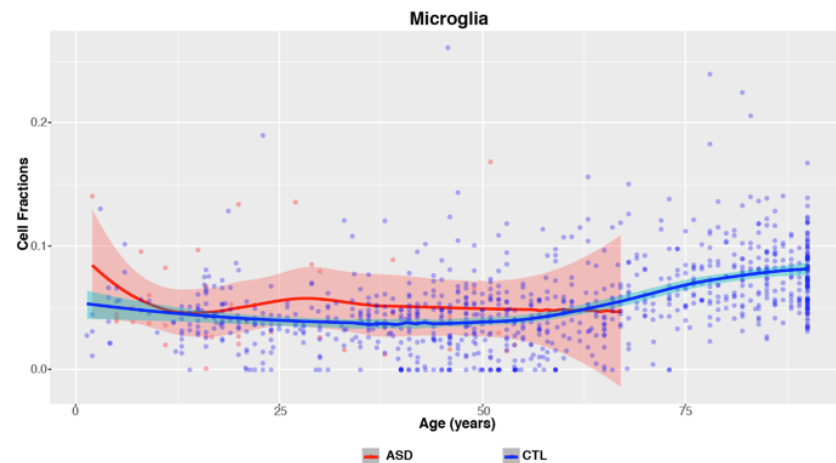
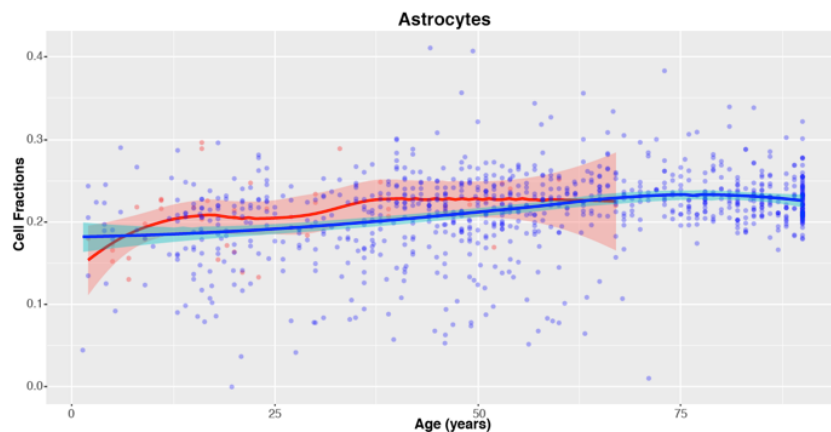


# Neuronal and glial cell fraction changes in gender and disorders



Excitatory to Inhibitory imbalance at neuronal subtype level for ASD\*

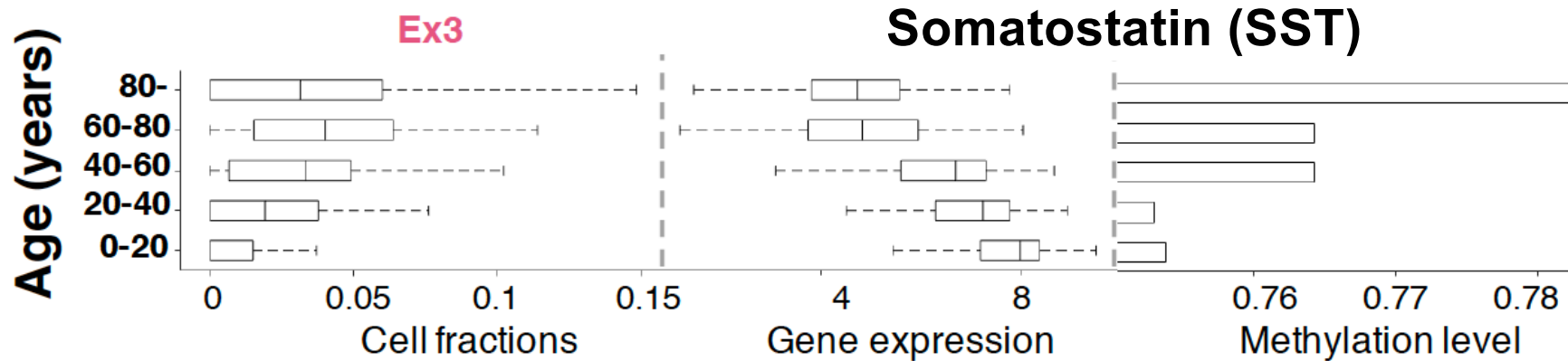
## Astrocyte and Microglia increase in ASD\*\*



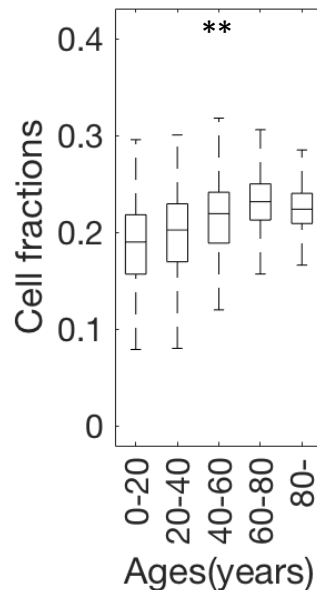
\* Rubenstein et al., Model of autism: increased ratio of excitation/inhibition in key neural systems, Genes Brain Behav. 2003

\*\* Gandal et al., Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap, Science 2018

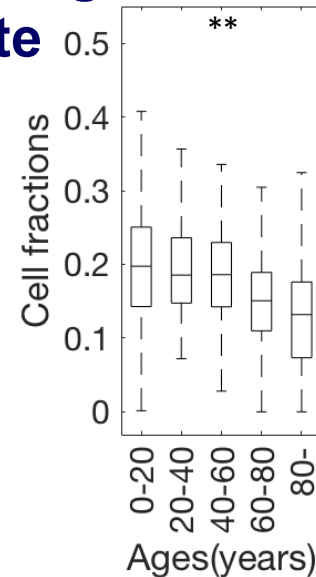
# Cell fraction changes in Age



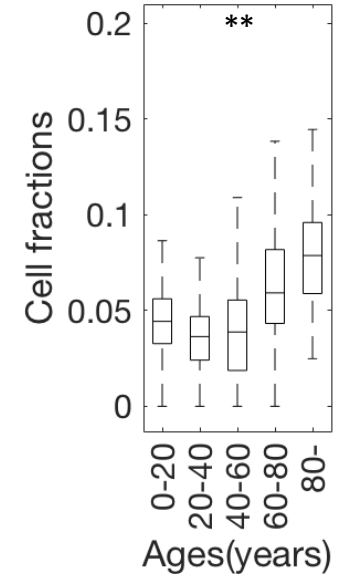
## Astrocyte



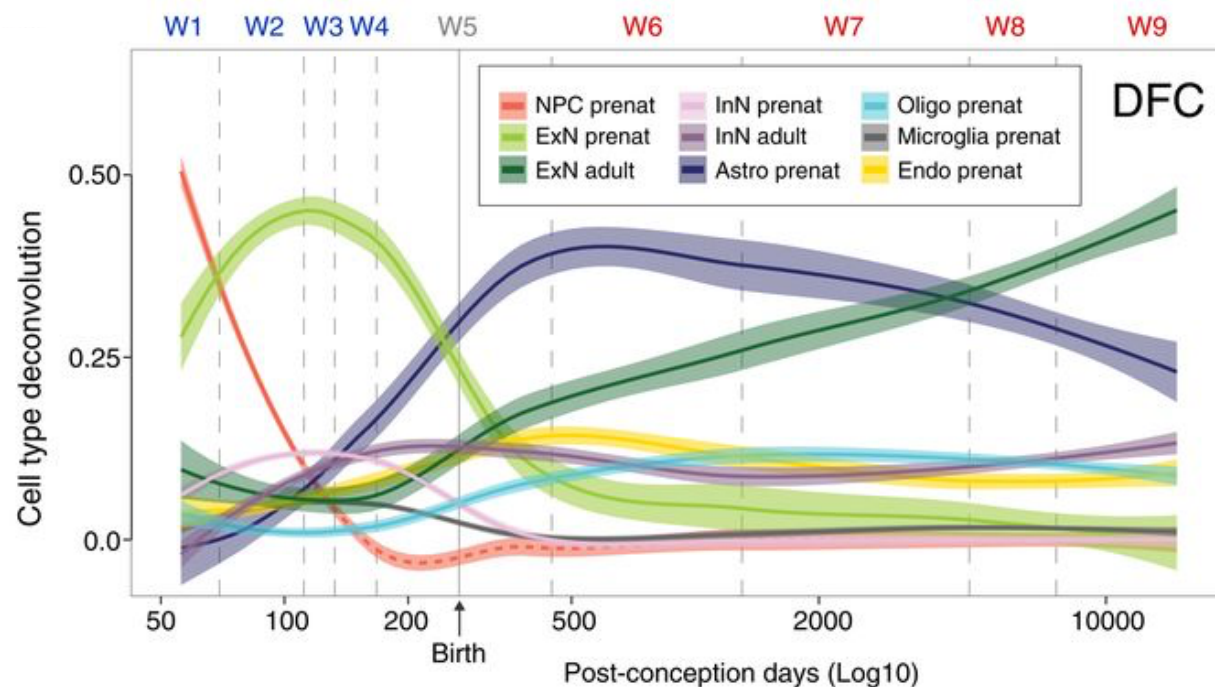
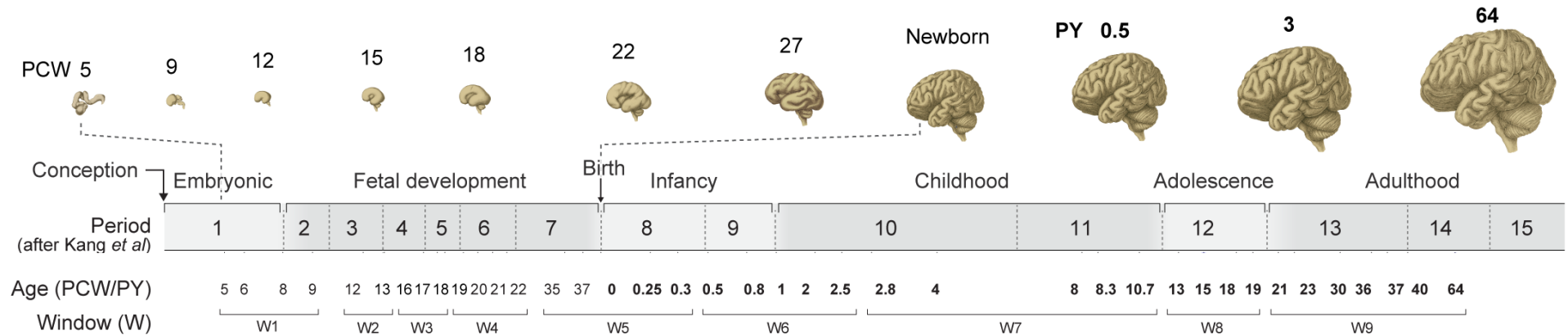
## Oligodendrocyte



## Microglia



# Cell fraction changes in human brain development





# Drop-outs in single cell UMI

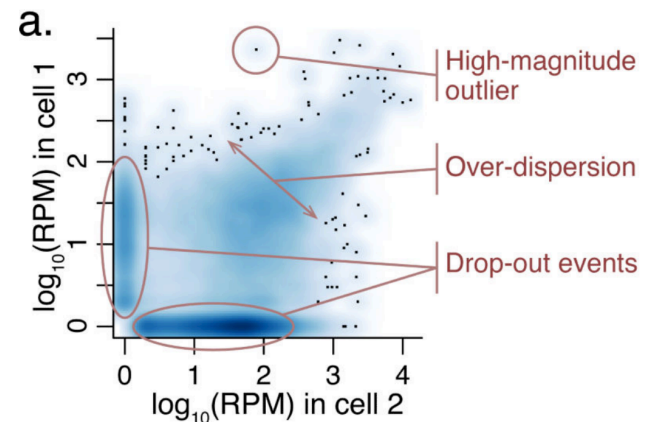
- a gene is observed at a moderate or high expression level in one cell but is not detected in another cell

## Why do dropouts occur in single cell?

- technical artifacts
- cell type differences
- statistical sampling
- biological factors

## What should we do about dropouts?

- Impute before learning
- ignore zero inflation
- preprocess/reduce dimensions



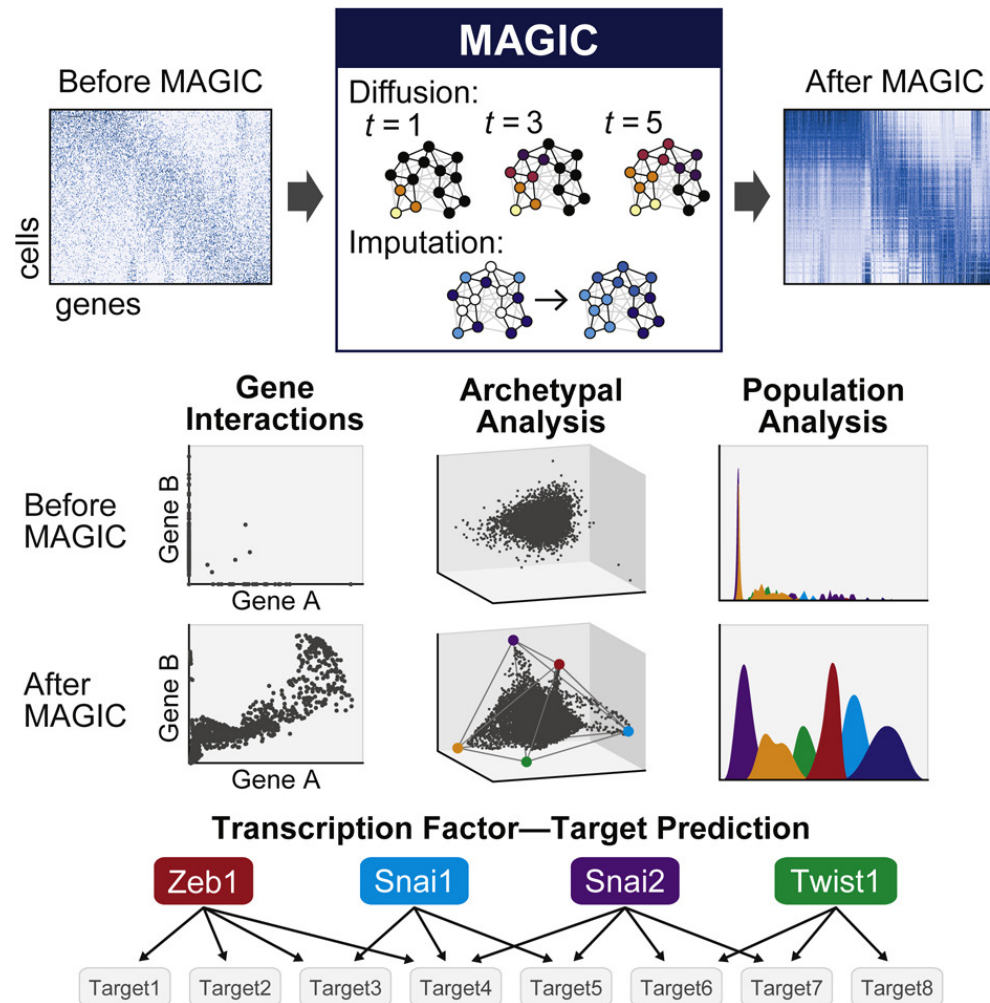
## Methods

- MAGIC
- Droplet
- DrImpute
- scDoc

# MAGIC

## Markov affinity-based graph imputation of cells

- an algorithm for denoising high-dimensional data most commonly applied to single-cell RNA sequencing data. MAGIC learns the manifold data, using the resultant graph to smooth the features and restore the structure of the data.



# Resources

## Tutorial

- <https://github.com/hbctraining/scRNA-seq>
- <https://bioconductor.org/books/release/OSCA/>
- <http://data-science-sequencing.github.io/>
- [https://broadinstitute.github.io/2019\\_scWorkshop/](https://broadinstitute.github.io/2019_scWorkshop/)
- [https://biocellgen-public.svi.edu.au/mig\\_2019\\_scrnaseq-workshop/public/index.html](https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/index.html)

## Tools

- <https://github.com/seandavi/awesome-single-cell>