

Single cell RNA-seq analysis

Part II: cell types and cell-type gene regulation

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2024

Daifeng Wang

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

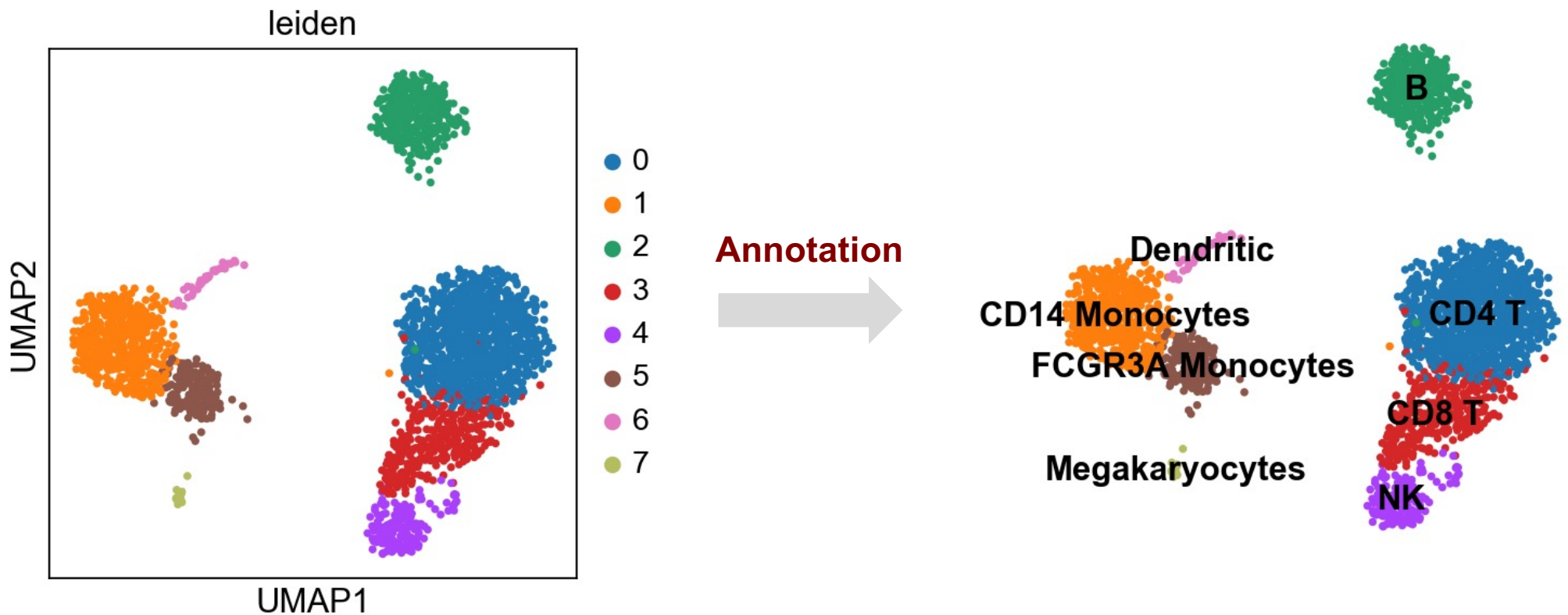
- scRNA-seq data analysis
 - Cell type annotation
 - SingleR
 - Cell type markers identification
 - Pseudo timing
 - Monocle
 - Cell-type gene regulatory networks
 - SCENIC

Outline

- scRNA-seq data analysis
 - **Cell type annotation**
 - SingleR
 - Cell type markers identification
 - Pseudo timing
 - Monocle
 - Cell-type gene regulatory networks
 - SCENIC

Cell type annotation

- Cell types -> cellular functions
- Assign the cell type for each cell



https://btep.ccr.cancer.gov/wp-content/uploads/Celltype_Annotation_final.pdf

https://biocellgen-public.svi.edu.au/mig_2019_sernaseq-workshop/public/clustering-and-cell-annotation.html

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

Cell type annotation tools

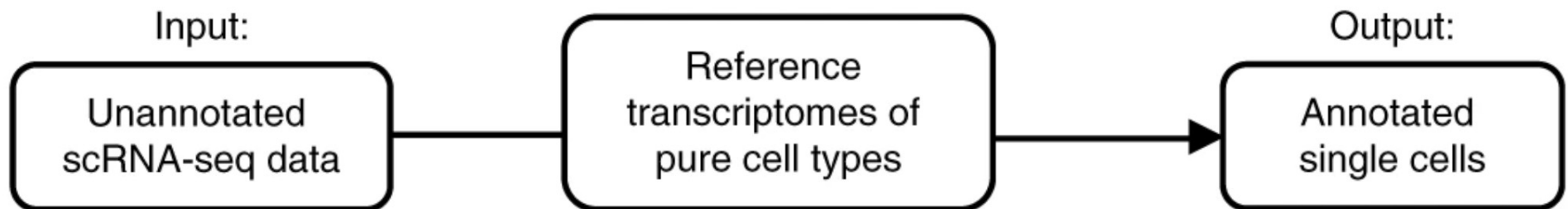
- **Supervised methods:** a training dataset labeled with the corresponding cell population is needed to train the classifier
 - **SingleR**, ACTINN, CaSTle
- **Prior-knowledge based methods:** either a marker gene file is required as an input or a pretrained classifier for specific cell populations is provided
 - DigitalCellSorter, Moana

	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
	Baron Mouse	Baron Human	Muraro	Seegerstolpe	Xin	10X	CEL-Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K
SVM _{rejection}	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 _{CV}												0.94*	0.6*
Garnett _{pretrained}												0.98*	0.54*
Moana												0.93*	0.5*
Garnett _{DE}												0.65	0.37
SCINA _{DE}												0.38	0.47
DigitalCellSorter _{DE}												0	0

https://btep.ccr.cancer.gov/wp-content/uploads/Celltype_Annotation_final.pdf
https://biocellgen-public.svi.edu.au/mig_2019_scmaseq-workshop/public/clustering-and-cell-annotation.html
<https://bioconductor.org/books/release/OSCA/cell-type-annotation.html>

SingleR: Reference-based annotation of scRNA-seq

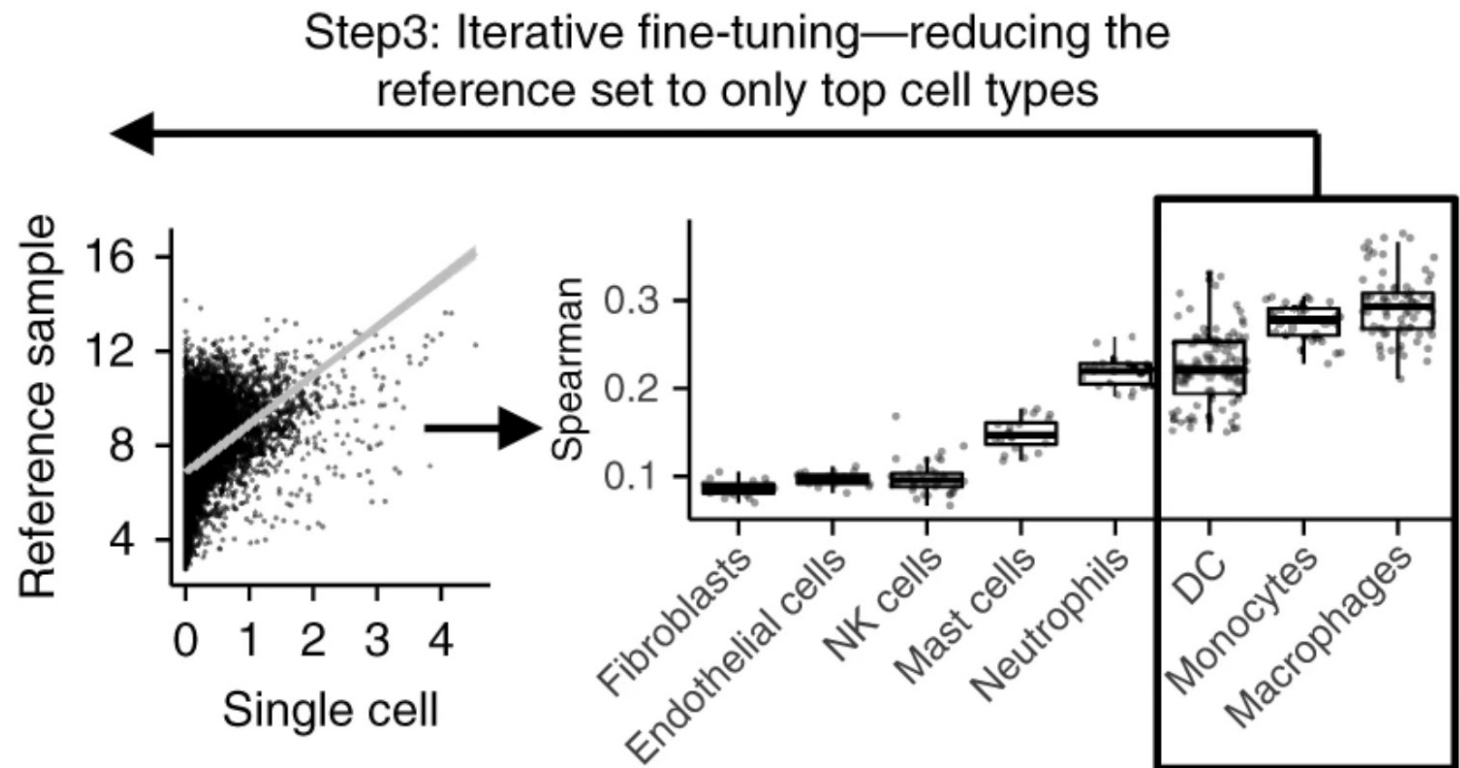
- SingleR pipeline is based on correlating reference bulk transcriptomic data sets of pure cell types with single-cell gene expression.
- Reference set: a comprehensive transcriptomic dataset (microarray or RNA-seq) of **pure** cell types
- Human
 - Human Primary Cell Atlas (HPCA) : 38 main cell types, 169 subtypes, 713 samples
 - Blueprint+Encode: 43 cell types, 259 bulk RNAseq samples
- Mouse
 - Immunological Genome Project (ImmGen) : 20 main cell types, 830 microarray samples
 - mouse RNA-seq samples (brain-specific) : 28 cell types, 358 RNA-seq samples



SingleR: Reference-based annotation of scRNA-seq

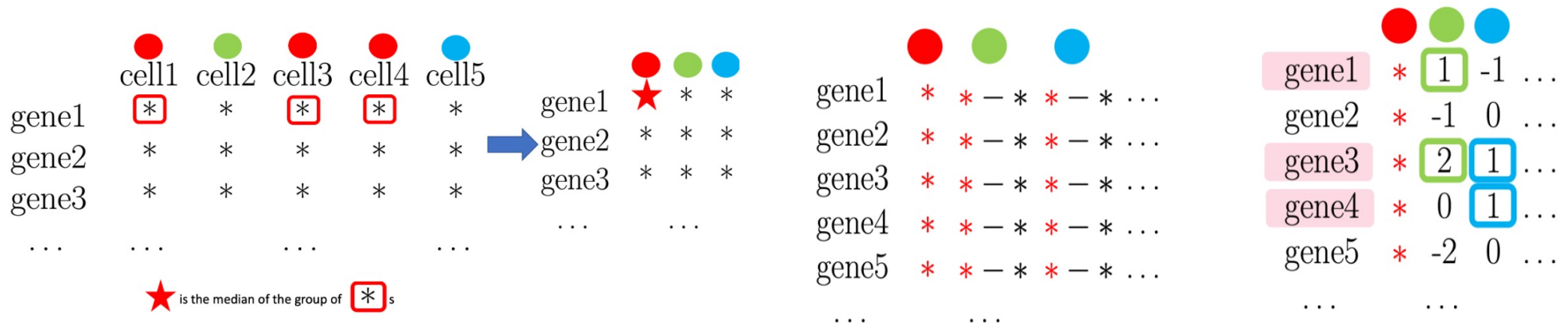
Step 1:
Identifying variable
genes among cell types
in the reference set

Step 2:
Correlating each
single-cell transcriptome
with each sample in the
reference set



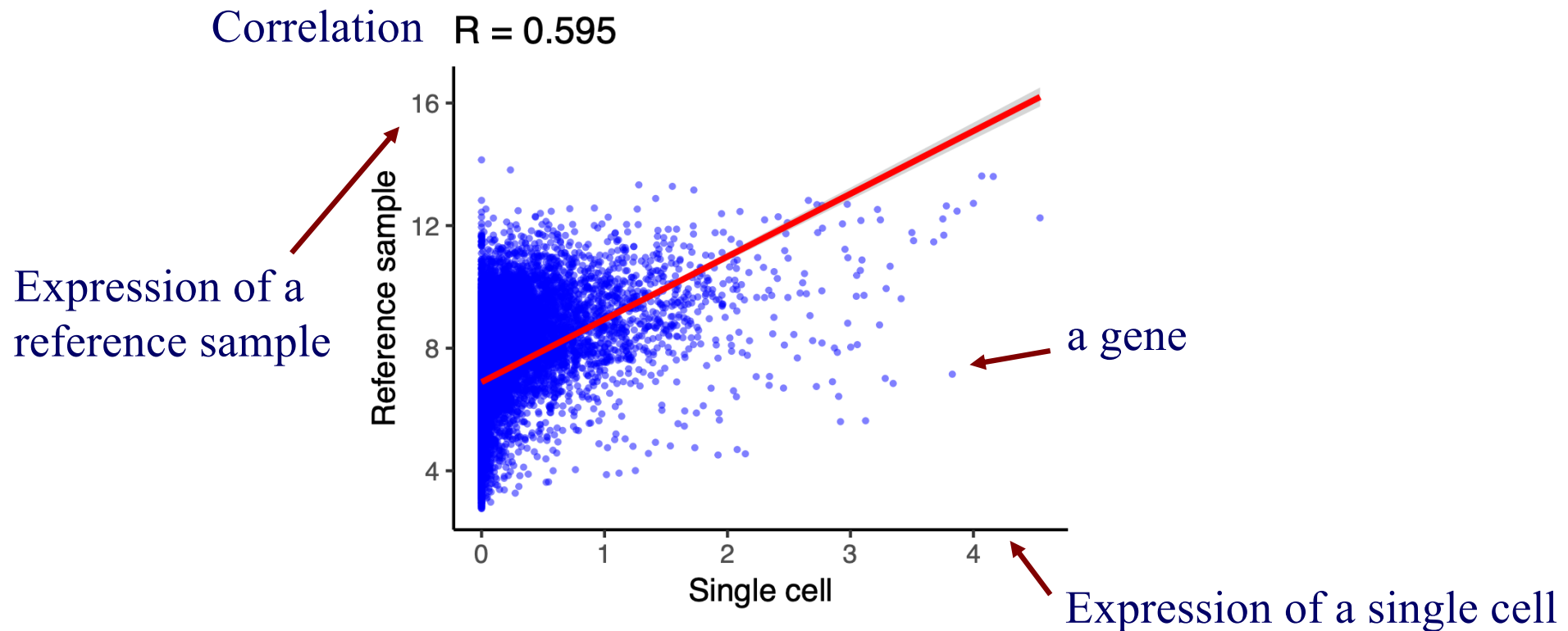
Step 1: Identifying variable genes among cell types in the reference set

- For each cell type, identify the top N variable genes that have a higher median expression in that cell type than in every other cell type
- Take the '**red**' cell type as an example
 - For every gene, **median** expression values grouped by cell type were obtained.
 - Differential expression between each other cell type and the '**red**' cell type was calculated and all genes with positive differential expression values were selected.
 - All selected genes were sorted by differential expression values, and then the top N genes were selected as variable genes for the '**red**' cell type.



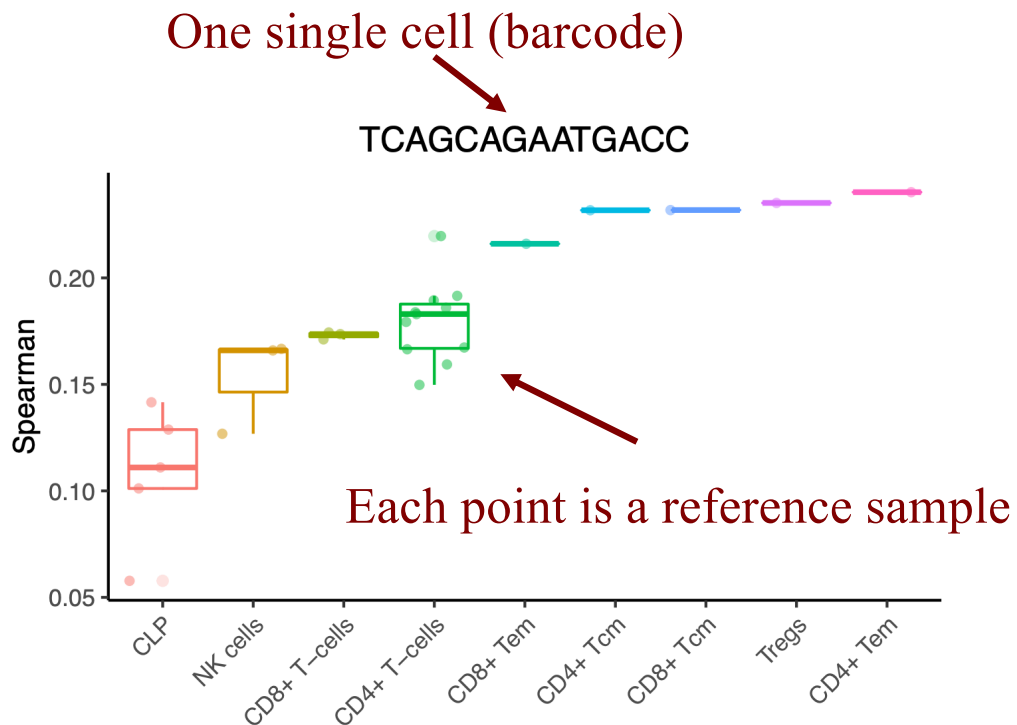
Step 2: Correlating each single-cell transcriptome with each sample in the reference set

- Spearman coefficient is calculated for single cell expression with each of the samples in the reference dataset.
- The correlation analysis is performed only on variable genes in the reference dataset.

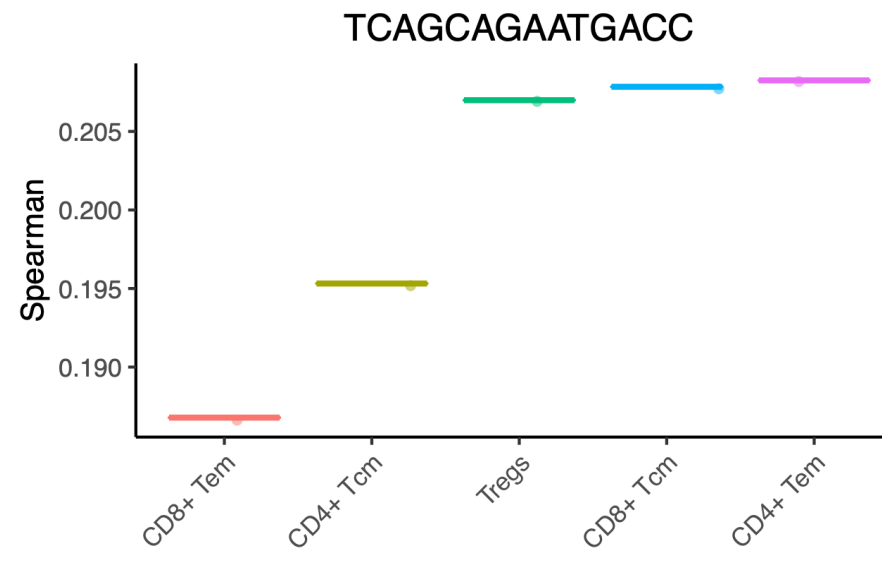


Step 3: Iterative fine-tuning - reducing the reference to only top cell types

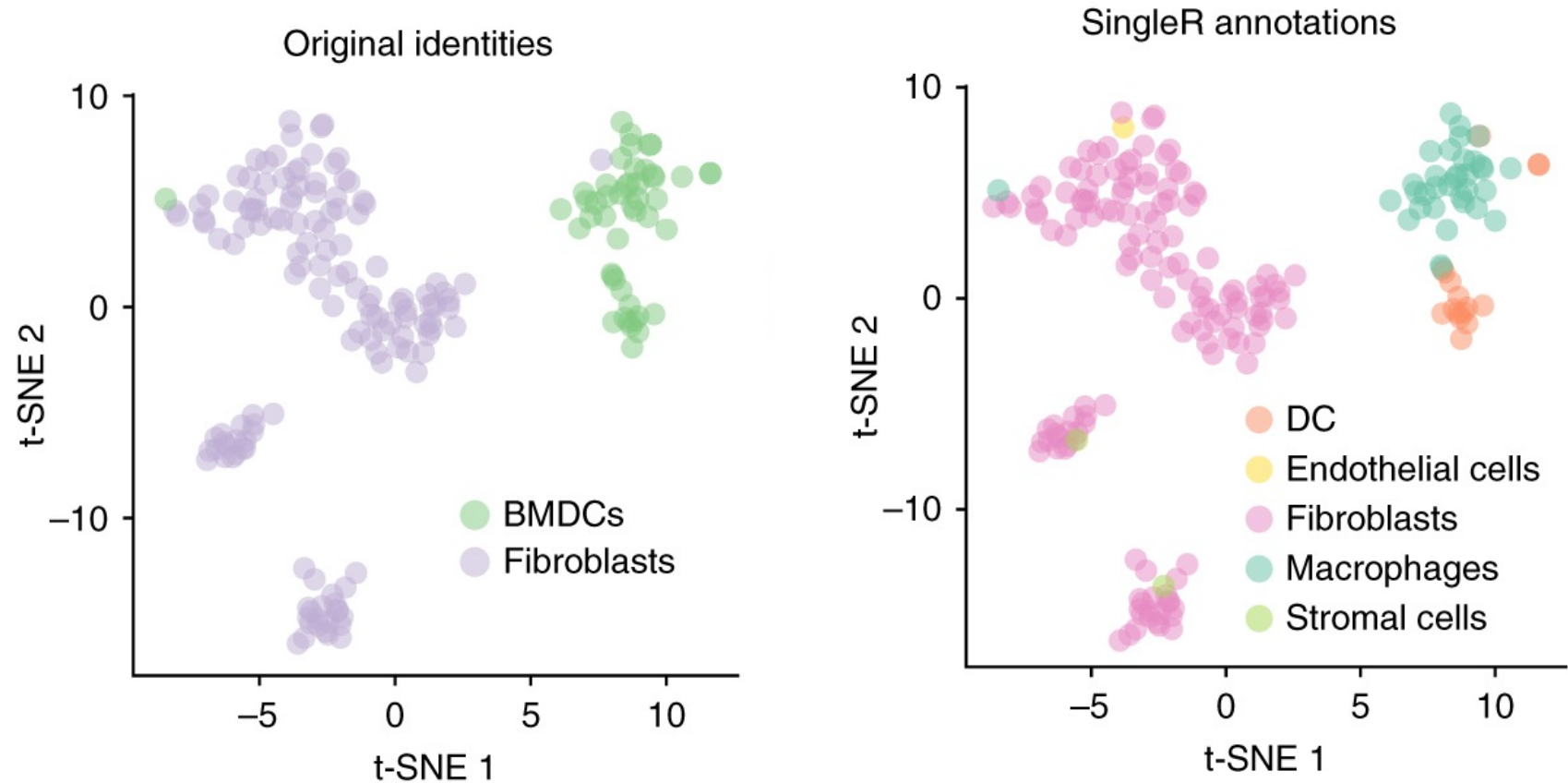
- For a single cell and each cell type, multiple Spearman correlation coefficients are aggregated into a “cell-type score”
 - The SingleR score for each cell type is the 80 percentile in each of the boxplots.
- Cell types with the lowest score or a score below will be removed
- Repeat from step 1 until only one cell type remained



For each iteration, top-scoring cell types are retained



SingleR: Reference-based annotation of scRNA-seq



Outline

- scRNA-seq data analysis
 - Cell type annotation
 - SingleR
 - **Cell type markers identification**
 - Pseudo timing
 - Monocle
 - Cell-type gene regulatory networks
 - SCENIC

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

Cell type markers identification

Differential testing and visualization in Scanpy

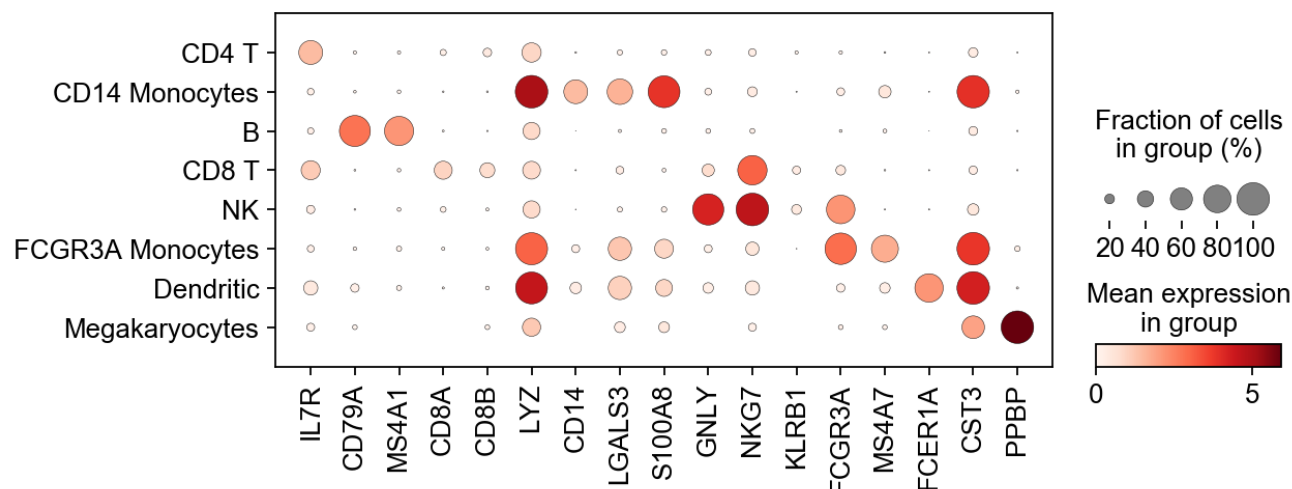


```
sc.tl.rank_genes_groups(adata, 'leiden', method='wilcoxon')
```

```
sc.tl.rank_genes_groups(adata, 'leiden', method='t-test')
```

```
marker_genes = ['IL7R', 'CD79A', 'MS4A1', 'CD8A', 'CD8B', 'LYZ', 'CD14',  
                'LGALS3', 'S100A8', 'GNLY', 'NKG7', 'KLRB1',  
                'FCGR3A', 'MS4A7', 'FCER1A', 'CST3', 'PPBP']
```

```
sc.pl.dotplot(adata, marker_genes, groupby='leiden')
```



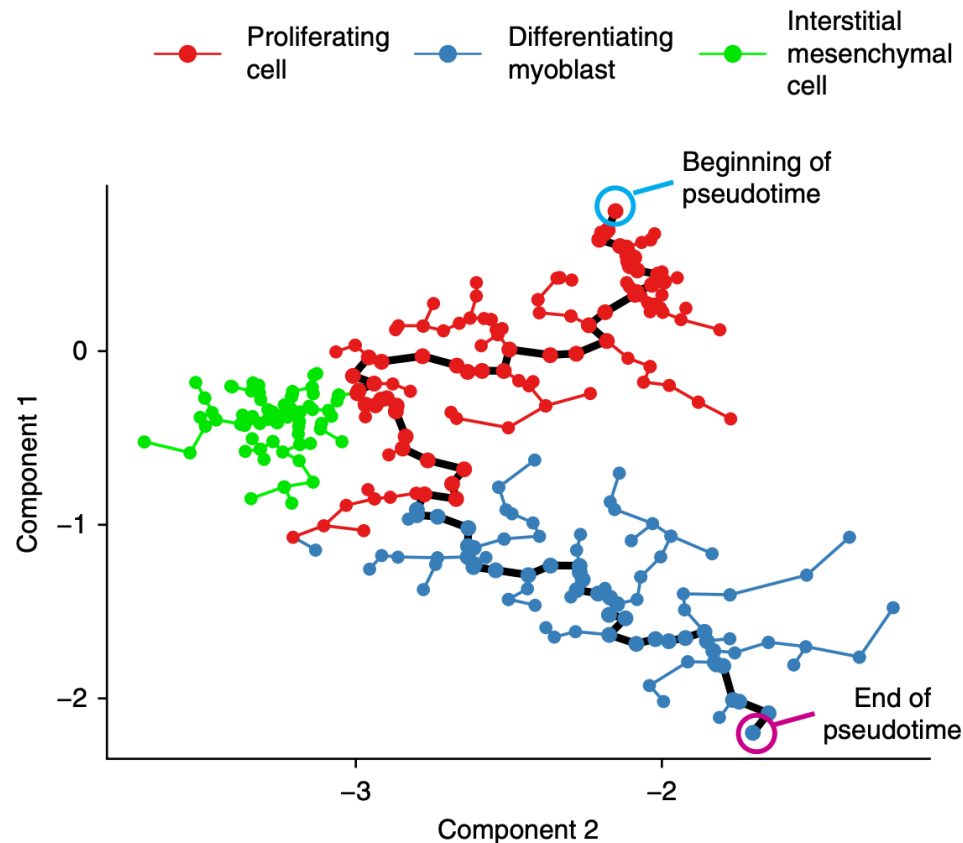
- Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015). <https://doi.org/10.1186/s13059-015-0844-5>
- Love MI, Huber W, Anders S (2014). "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://zenodo.org/record/4317764#.YII7gdPMKCg>

Outline

- scRNA-seq data analysis
 - **Cell type annotation**
 - SingleR
 - Cell type markers identification
 - **Pseudo timing**
 - Monocle
 - Cell-type gene regulatory networks
 - SCENIC

Pseudo timing

- Many cell differentiation processes take place during development
- We order the cells along one or more trajectories representing the underlying developmental processes
- This ordering is called 'pseudotime'
- **Trajectory inference (TI)** aims to reconstruct a cellular dynamic process



<http://cole-trapnell-lab.github.io/monocle-release>

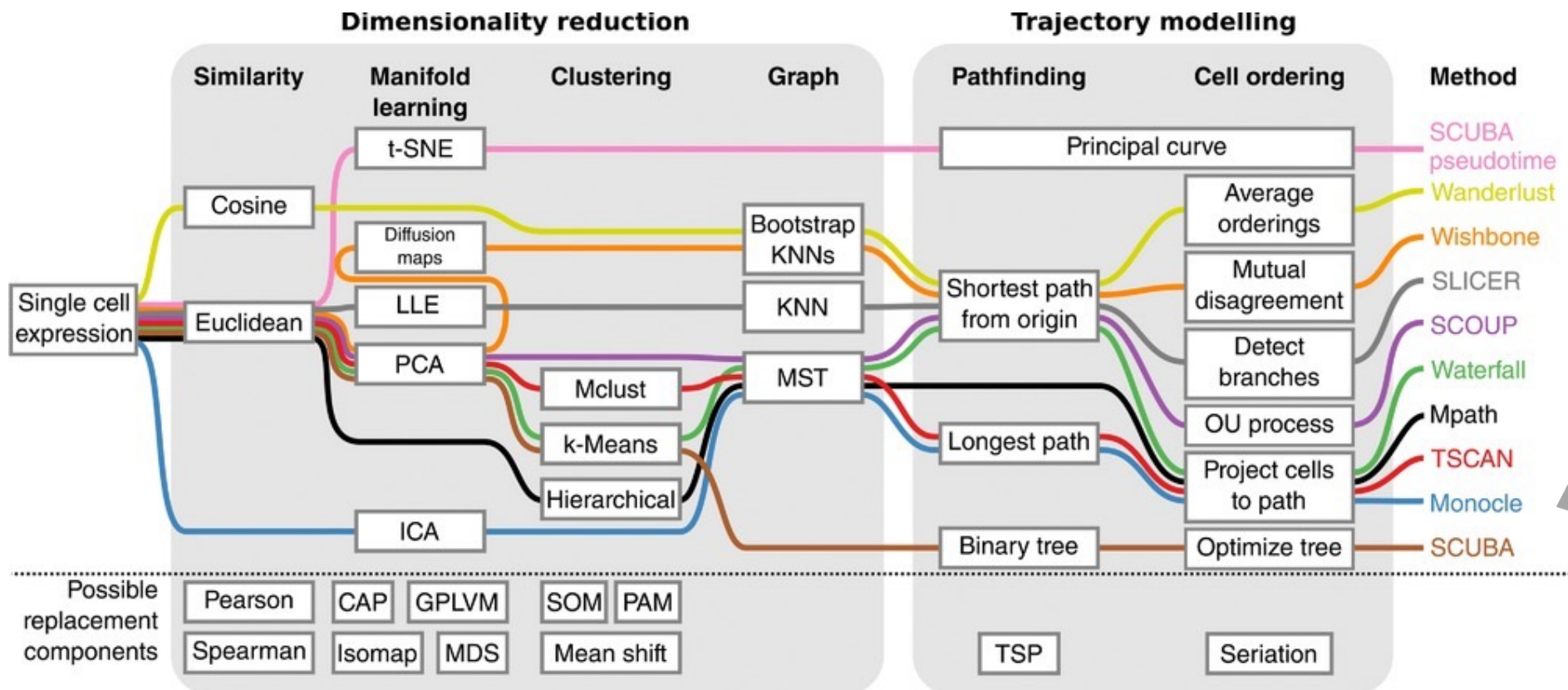
<https://scnaseq-course.cog.sanger.ac.uk/website>

Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019). <https://doi.org/10.1038/s41587-019-0071-9>

Trapnell, C., Cacchiarelli, D., Grimsby, J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014). <https://doi.org/10.1038/nbt.2859>

Pseudo timing

- Using single-cell-omics data, many **trajectory inference (TI)** methods could computationally order cells along trajectories, allowing the unbiased study of cellular dynamic processes



Monocle

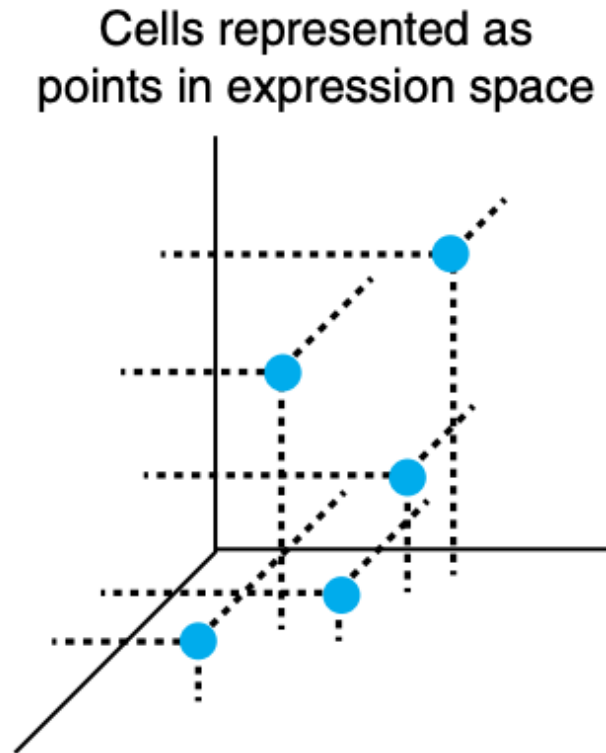
Constructing single cell trajectories

Monocle, an unsupervised algorithm to build single-cell trajectories, and find cell fate decisions and dynamically regulated genes.

- Step 1: Choose genes that define progress
- Step 2: Reduce data dimensionality
 - independent component analysis (ICA)
- Step 3: Construct minimum spanning tree (MST) on the cells
- Step 4: Find the longest path through the MST
- Step 5: Order cells along the trajectory

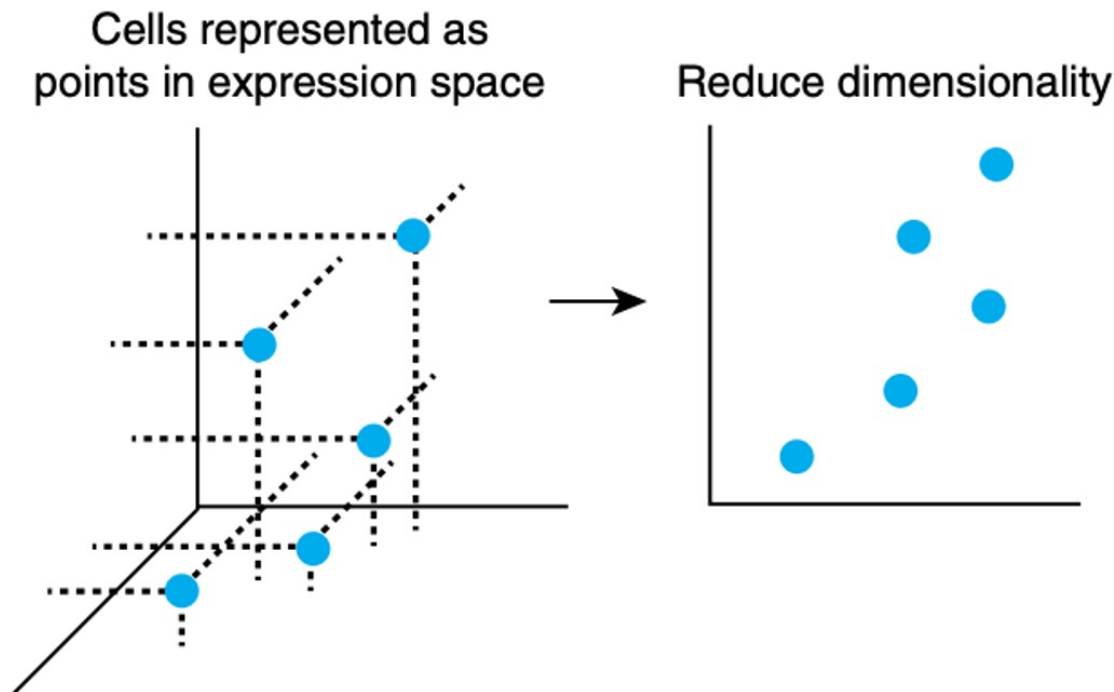
Step 1: Choose genes that define progress

- Represent the expression profile of each cell as a point in a high-dimensional Euclidean space, with one dimension for each gene



Step 2: Reduce data dimensionality

- Reduce dimensionality using independent component analysis (ICA)
- Transform the cell data from a high-dimensional space into a low-dimensional one that preserves essential relationships between cell populations



https://github.com/NBISweden/excelerate-scRNAseq/blob/master/session-trajectories/trajectory_inference_analysis.pdf

https://www.cs.cmu.edu/~tom/10701_sp11/recitations/Recitation_11.pdf

Trapnell, C., Cacchiarelli, D., Grimsby, J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

<https://doi.org/10.1038/nbt.2859>

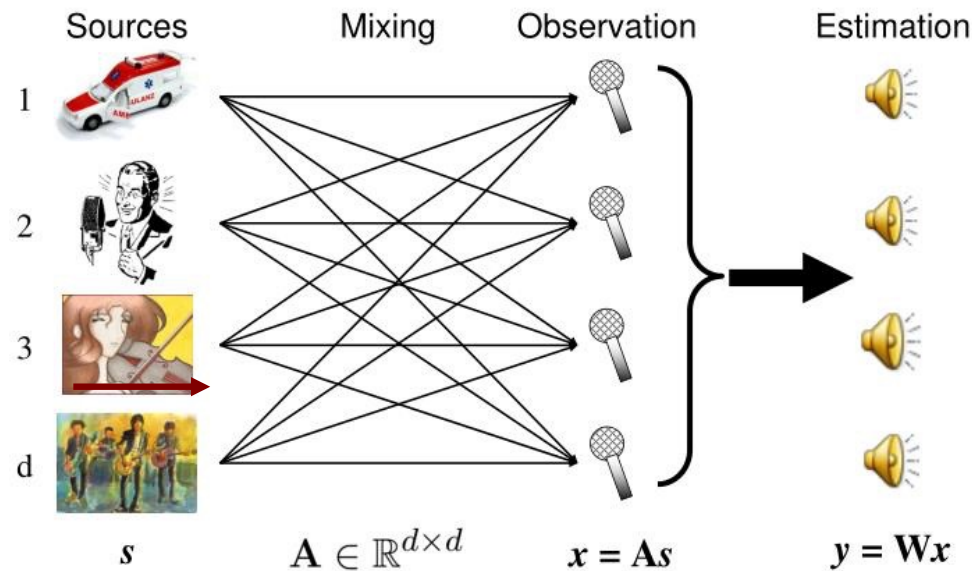
ICA

- Assumption: the mixed sources signals are independent of each other
- Goal: find linear mapping W which maximize independence and unmix sources signal s

$$y = Wx = WA s$$

Mixed variables x Mixing matrix A Source s
 Linear mapping W

Independent Component Analysis
(ICA, The Cocktail Party Problem)



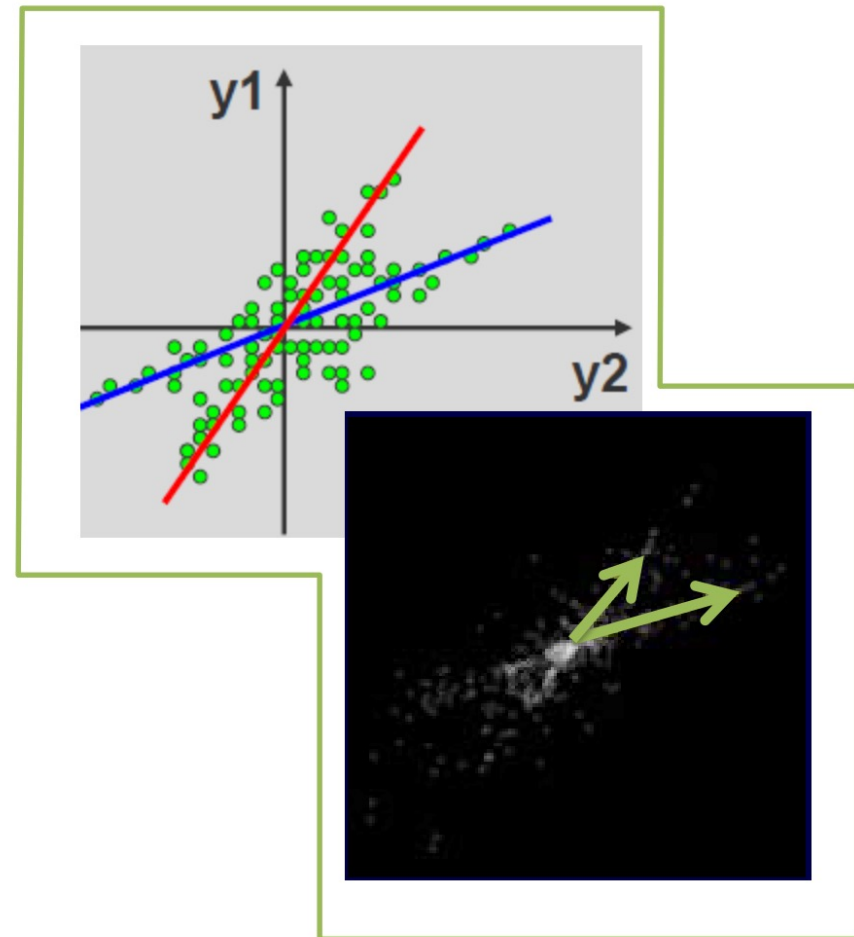
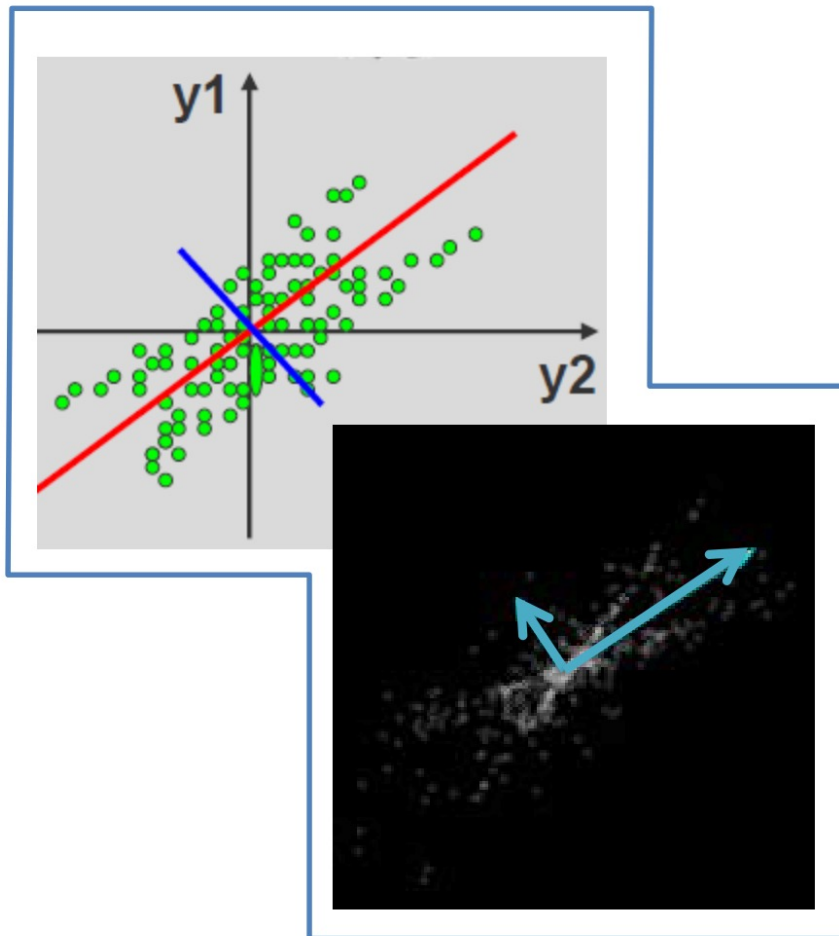
July 8, 2008

ICML 2008

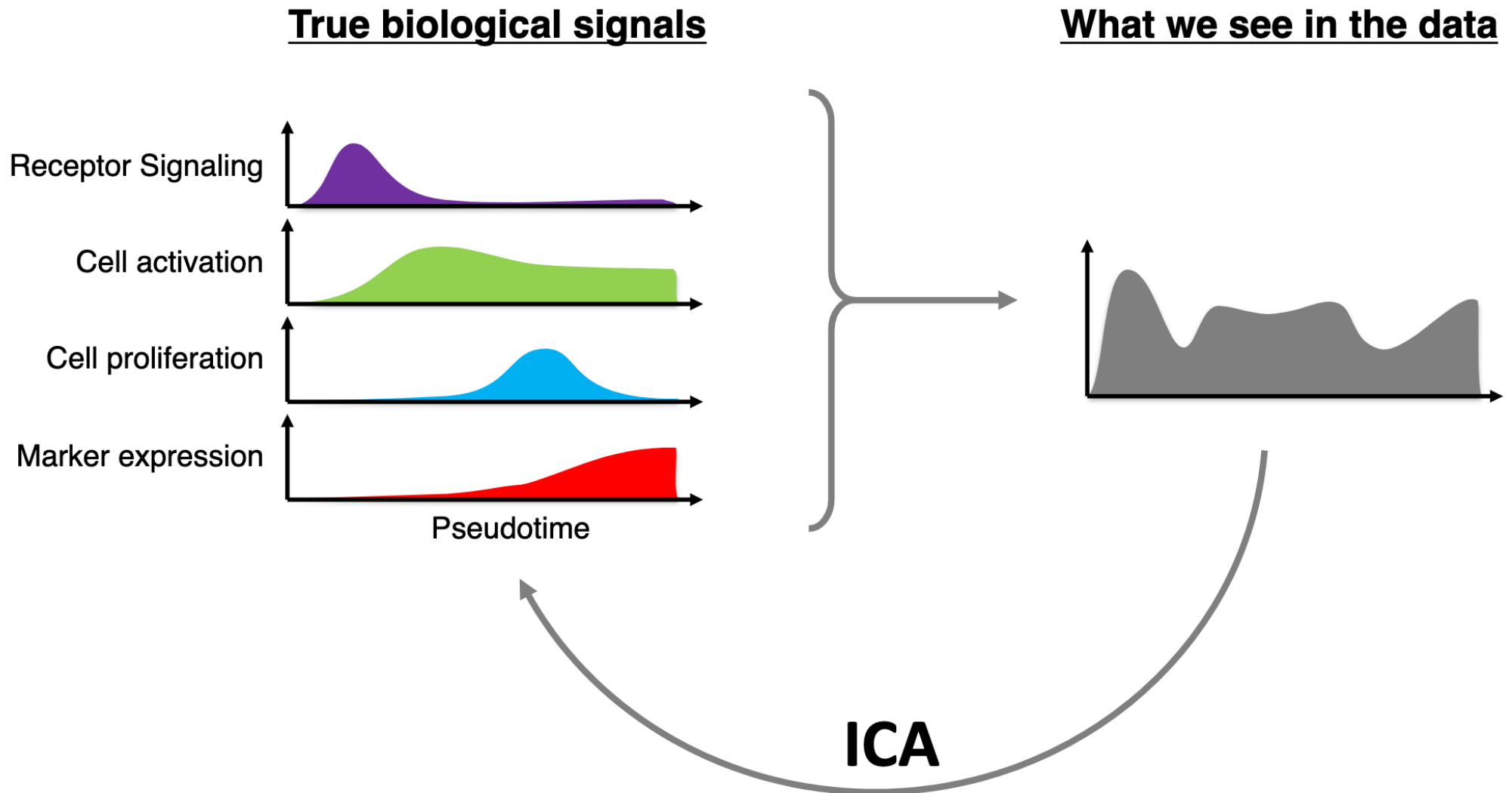
2

ICA vs PCA

- PCA : Find the directions of maximal **variance**
- ICA : Find the directions of maximal **independence**
 - The values in each source have non-Gaussian distributions



Why ICA



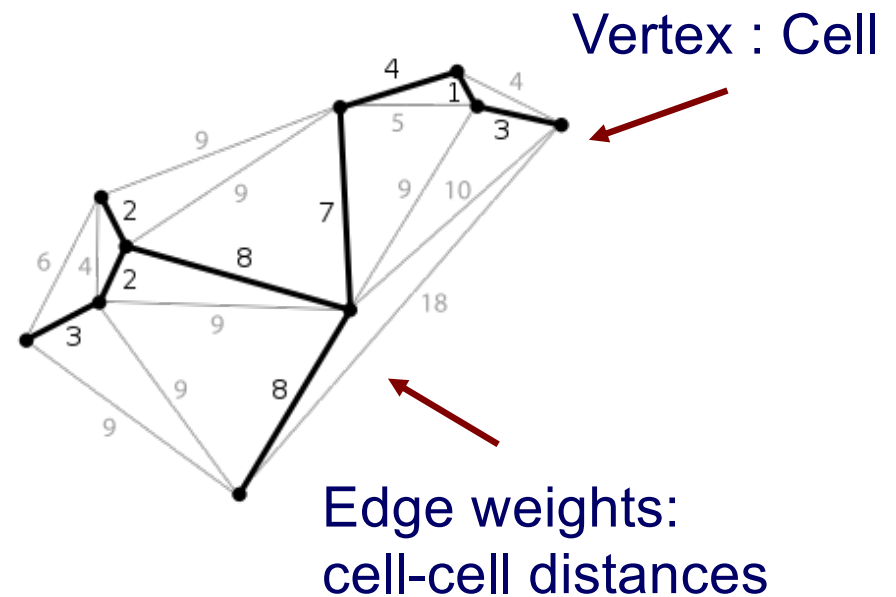
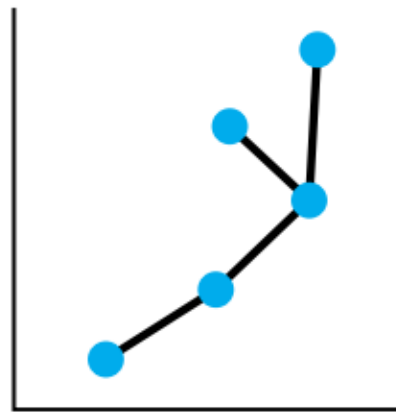
Step 3: Construct minimum spanning tree (MST) on the cells

- Minimum spanning tree (MST)
 - The undirected graph connecting all vertices with the smallest sum of all distances
 - No cycles

Reduce dimensionality



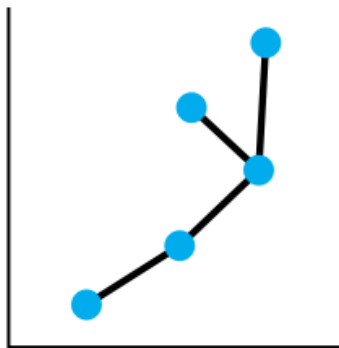
Build MST on cells



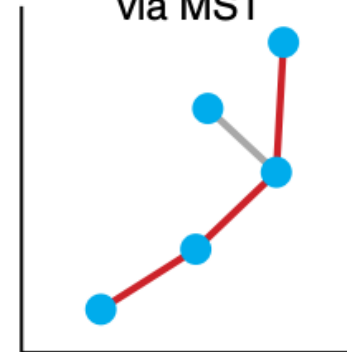
Step 4: Find the longest path through the MST

- Correspond to the longest sequence of similar cells (e.g., gene expression)

Build MST on cells



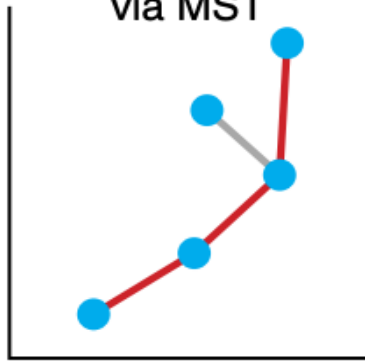
Order cells in pseudotime via MST



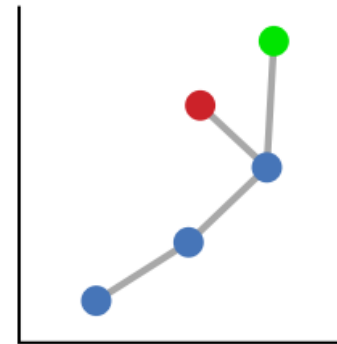
Step 5: Order cells along the trajectory

- Produce a 'trajectory' of an individual cell's progress through differentiation

Order cells in pseudotime
via MST

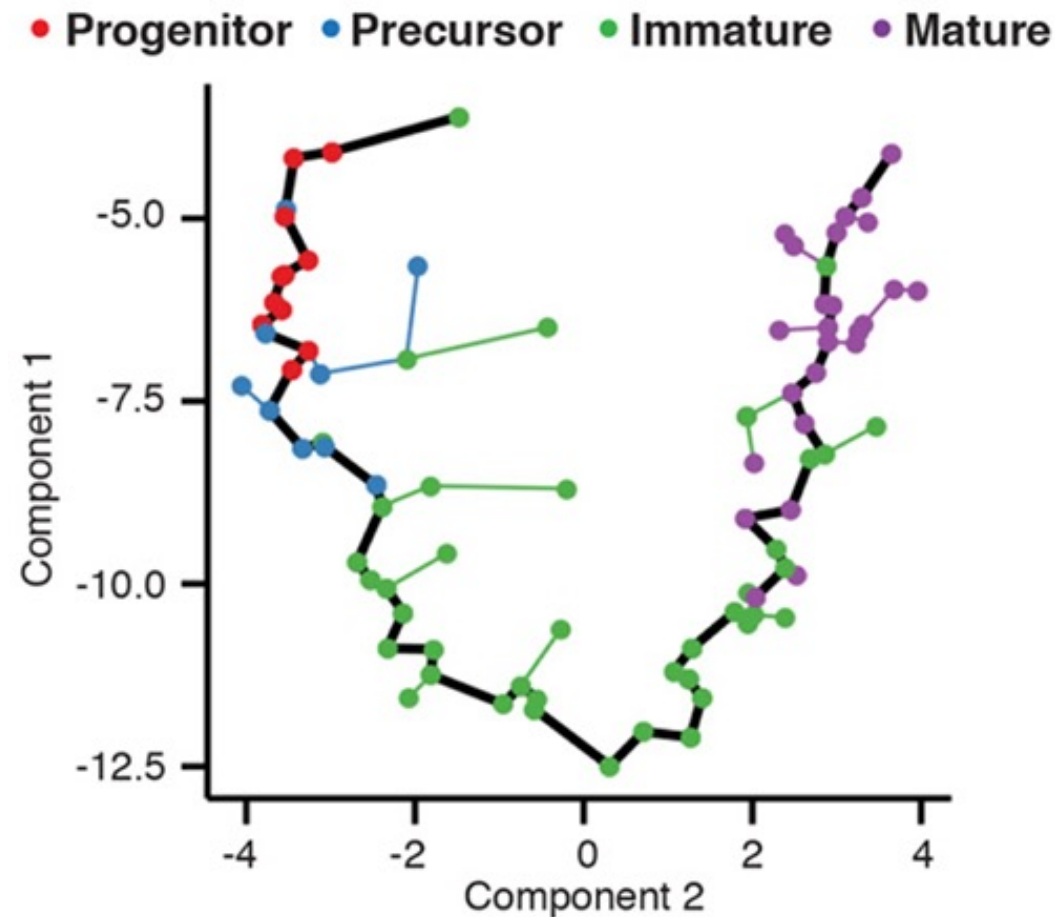


Label cells by type






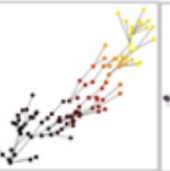
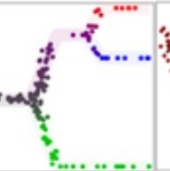
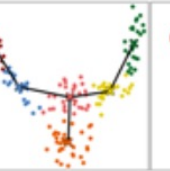
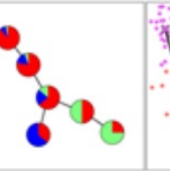
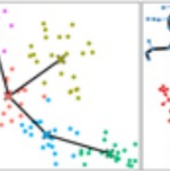
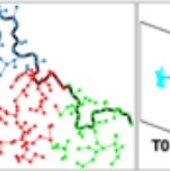
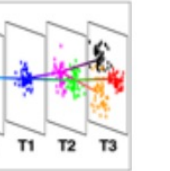
Developmental trajectory of olfactory neurons in mice

- Each point is a cell, which is connected to an MST
- The pseudotime value of each cell is measured as the distance along the trajectory from its position back to the beginning



Pseudo timing

- The performance of TI methods mostly depend on the topology of the trajectory in the single-cell data.

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	-	+	-	-
Scalability w.r.t. cells	-	-	±	±	-	±	+	+	-	±
Scalability w.r.t. genes	+	+	+	+	-	+	±	±	±	+
Code and documentation	-	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	-	±	-	+	+	+

<http://cole-trapnell-lab.github.io/monocle-release/docs/#constructing-single-cell-trajectories>

<https://indico.math.cnrs.fr/event/3780/contributions/3242/attachments/2195/2550/Slides-maugis-181018.pdf>

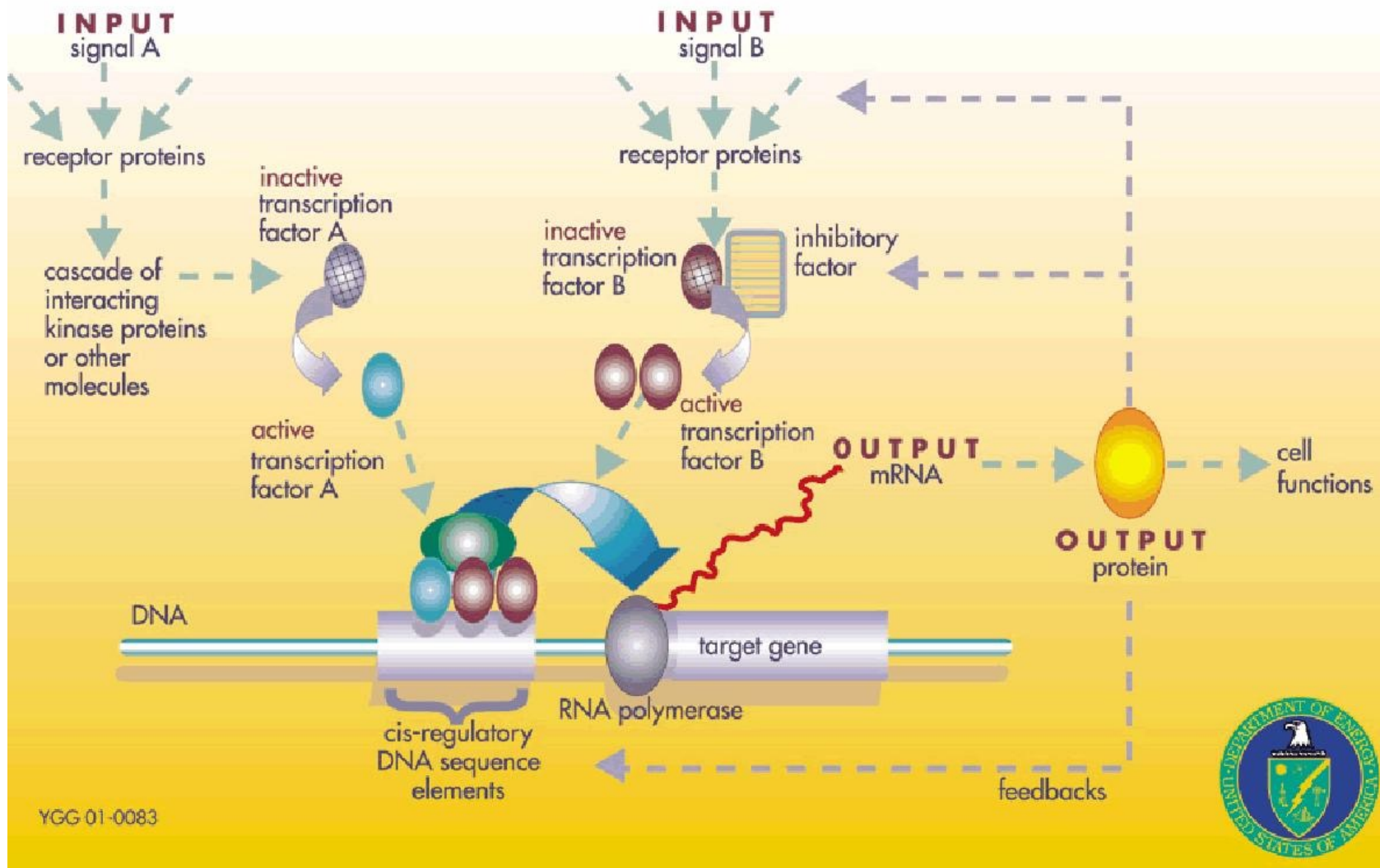
<https://scrnaseq-course.cog.sanger.ac.uk/website/biological-analysis.html#pseudotime-analysis>

Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019). <https://doi.org/10.1038/s41587-019-0071-9>

Outline

- scRNA-seq data analysis
 - Cell type annotation
 - SingleR
 - Cell type markers identification
 - Pseudo timing
 - Monocle
 - **Cell-type gene regulatory networks**
 - **SCENIC**

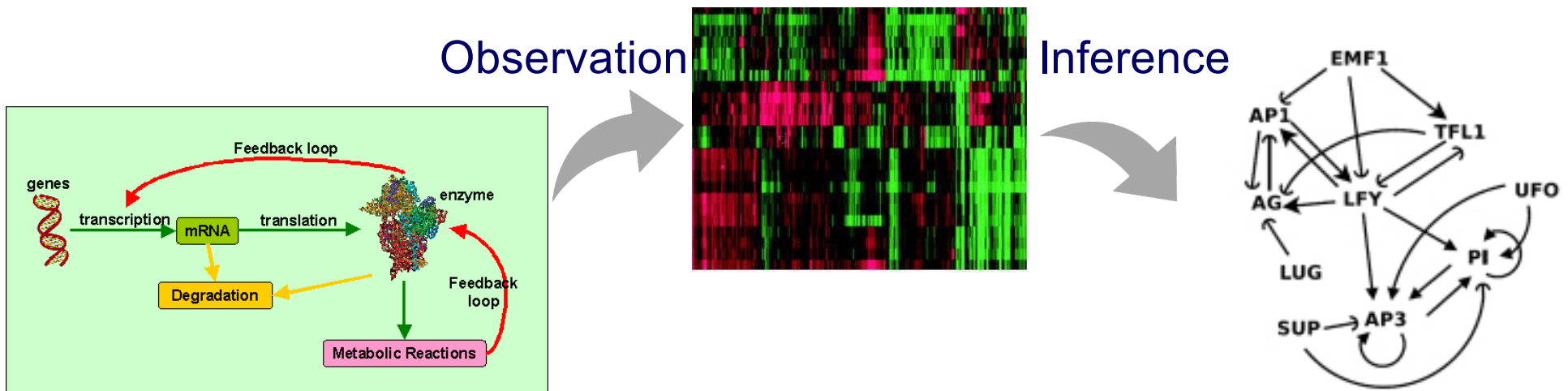
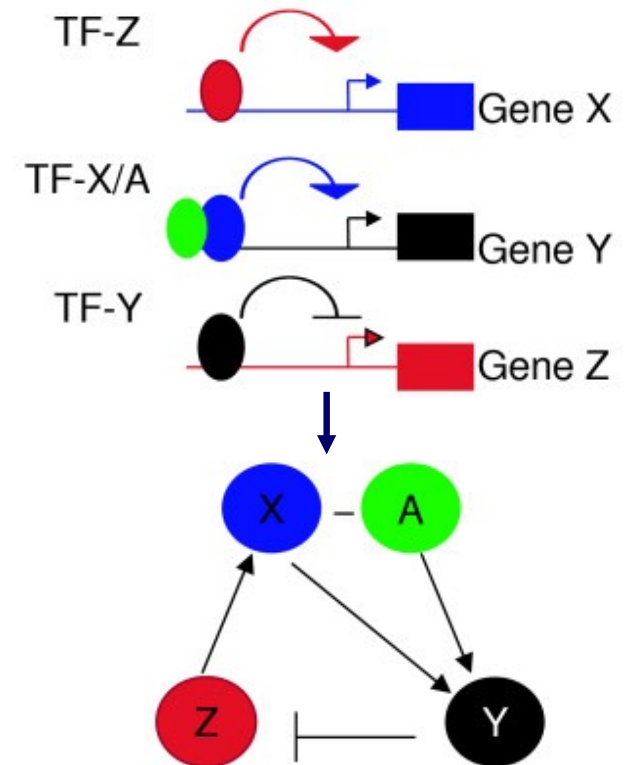
Gene regulation



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).

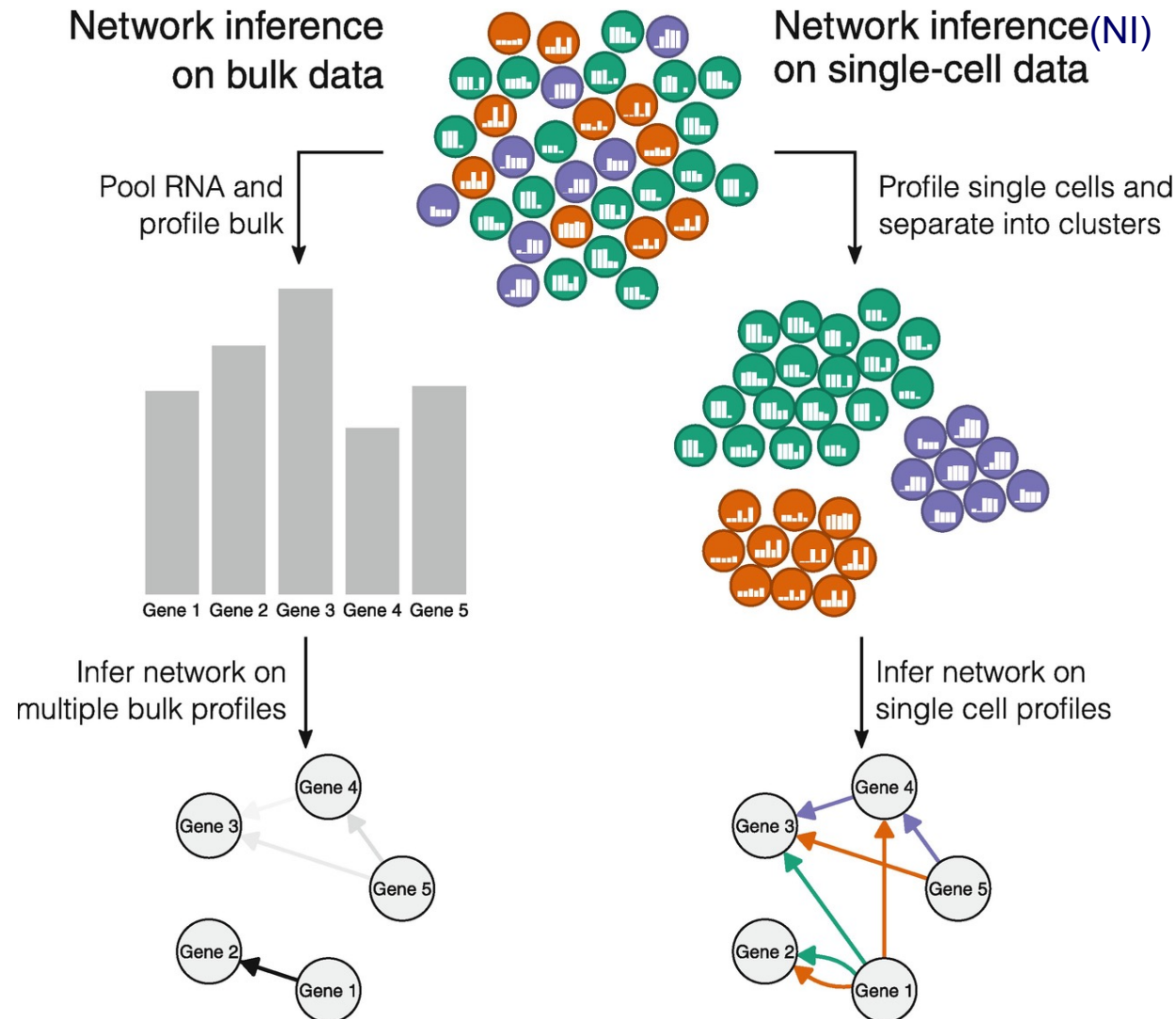
Gene regulatory network

- Gene regulatory networks (GRNs) like 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
- GRN 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

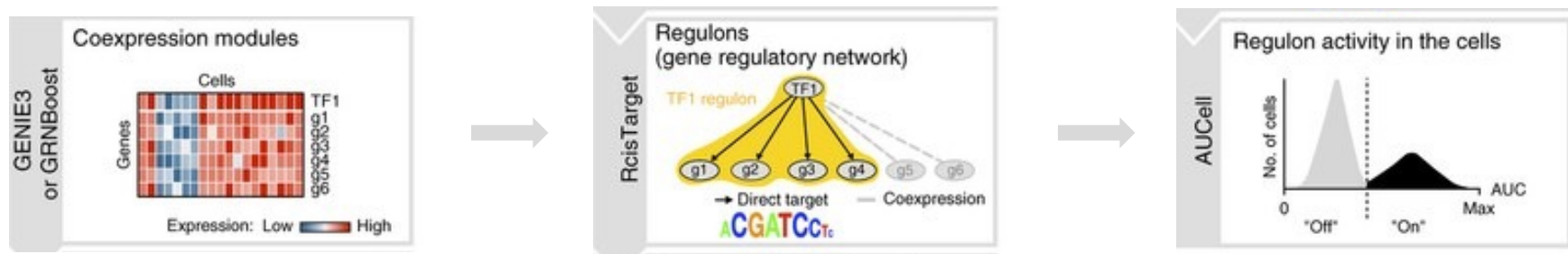
- Cell-type-specific GRNs would be key tools for the study of cellular heterogeneity



SCENIC

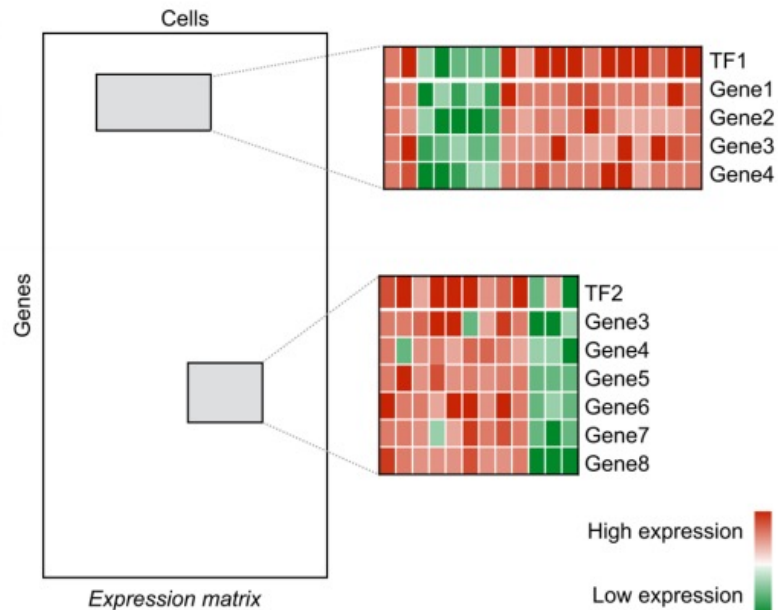
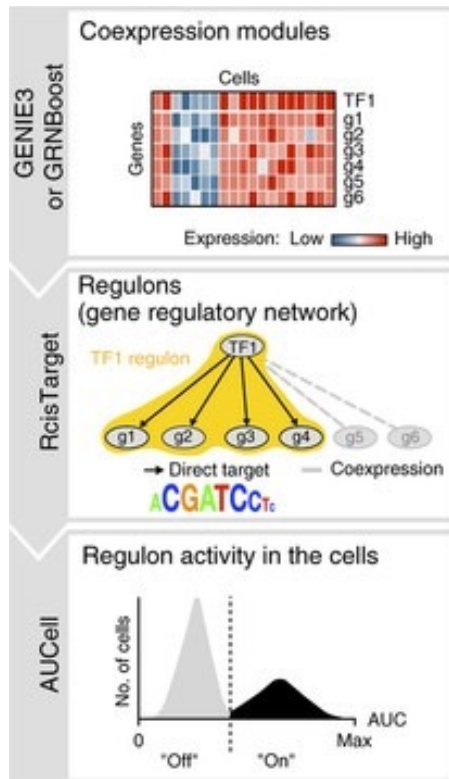
single-cell regulatory network inference and clustering

- Simultaneously reconstruct gene regulatory networks and identify stable cell states from single-cell RNA-seq data, based on three tools
 - GENIE3 or GRNboost
 - RcisTarget
 - AUCell
- 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.



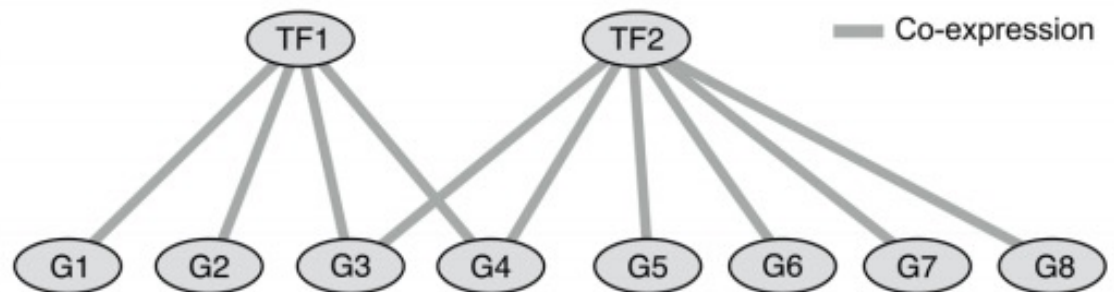
Step 1: TF-based co-expression network

SCENIC



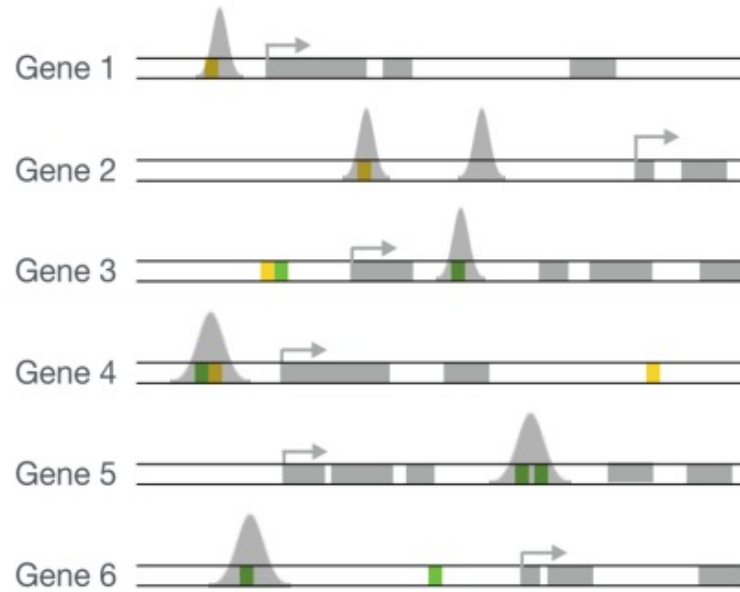
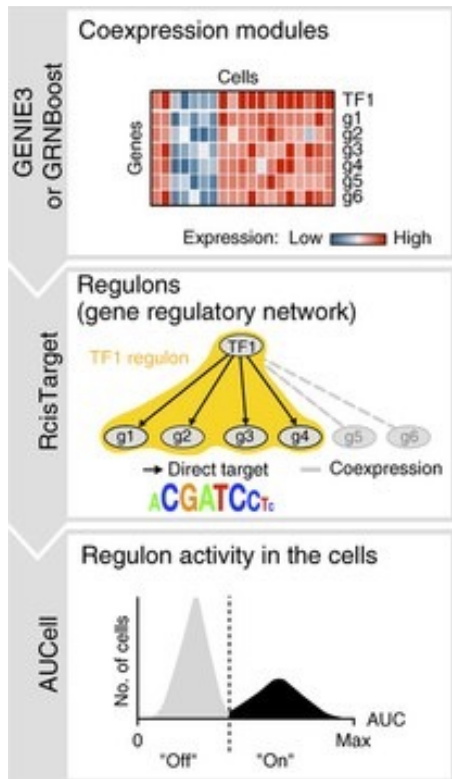
GENIE3
or
GRNBoost

Co-expression modules



Step 2: Identification of transcription factor binding motifs

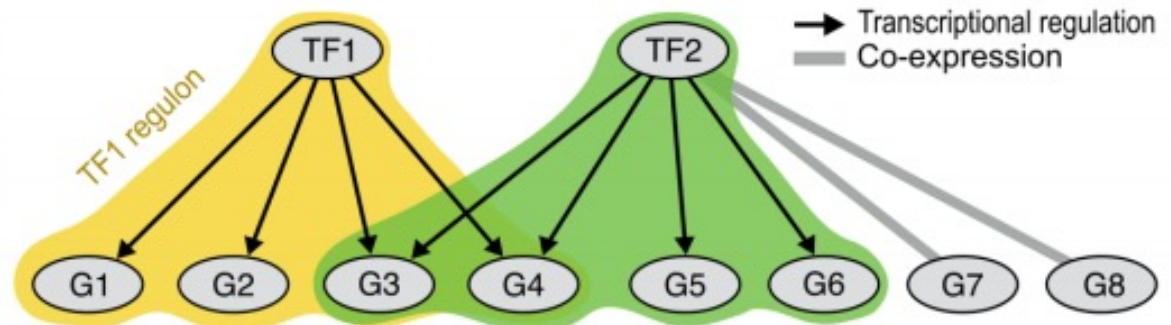
SCENIC



RcisTarget
cis-regulatory sequence analysis

TF₁ AATGCTAA TF₂ ACGATCC_{Tc}

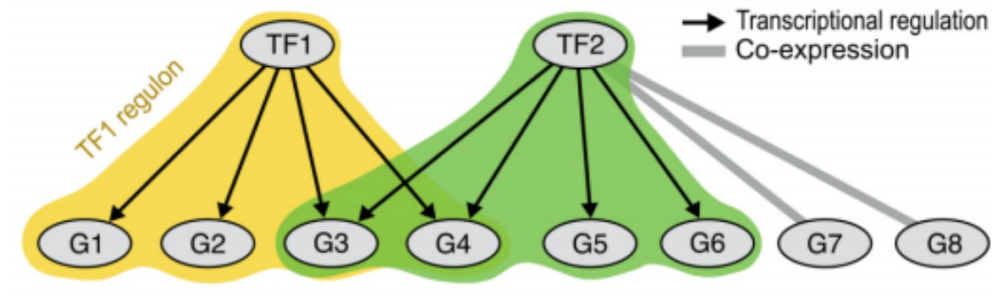
Regulons (Gene regulatory network)



Regulon

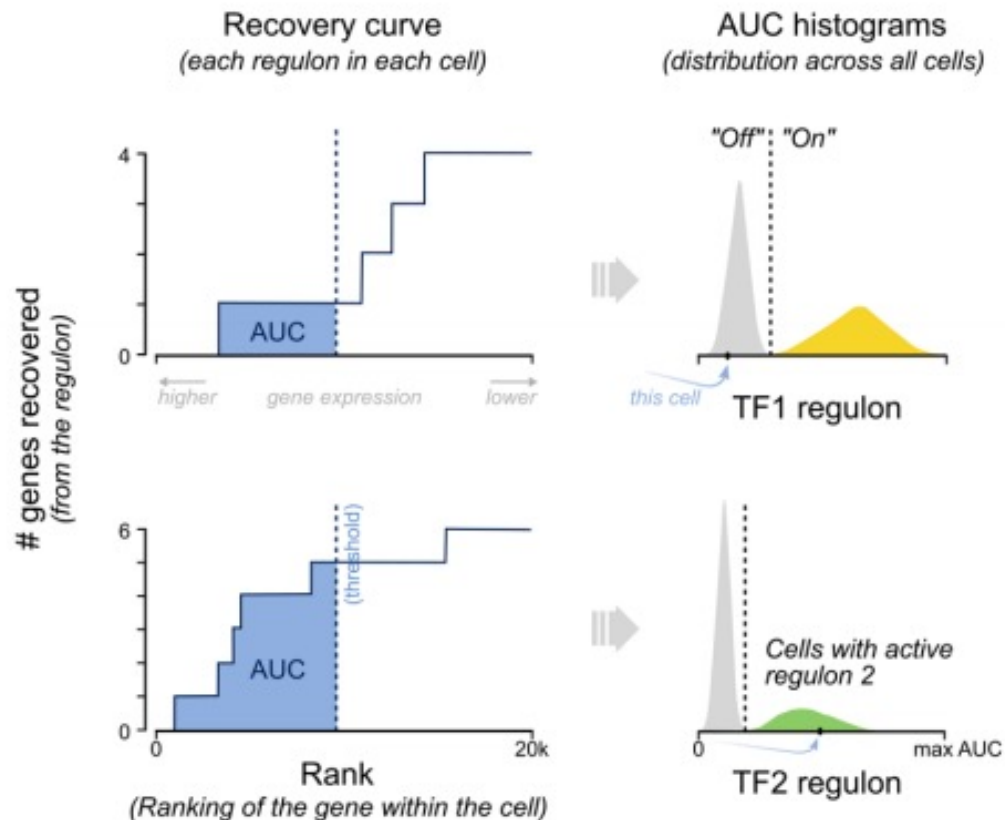
- Regulon: a group of genes that are regulated as a unit

Regulons (Gene regulatory network)



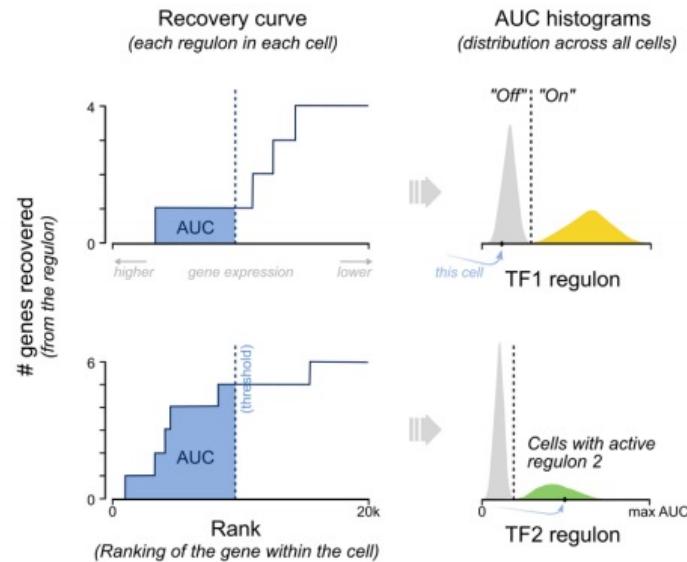
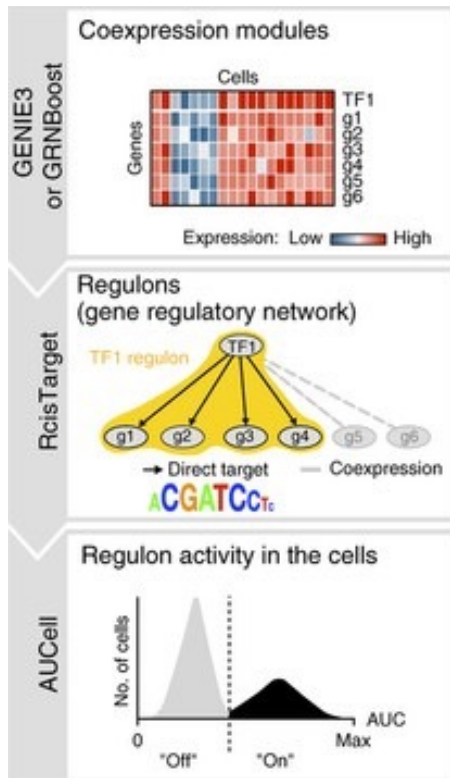
AUCCell score

- AUCCell uses the “Area Under the Curve” (AUC) to calculate whether a critical subset of the input gene set is enriched within the expressed genes for each cell.
- AUCCell score: measure how active a regulon is in a cell
 - Step 1: For each cell, build gene-expression ranking
 - Step 2: Calculate enrichment for the gene signatures (AUC)
 - Step 3: Determine the cells with given regulon



Step 3: Regulon activities in each cell

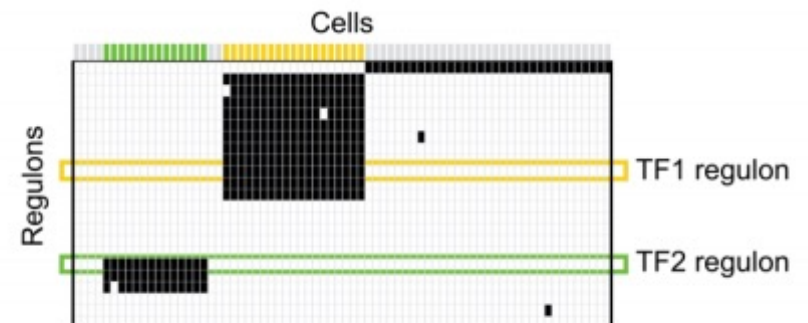
SCENIC



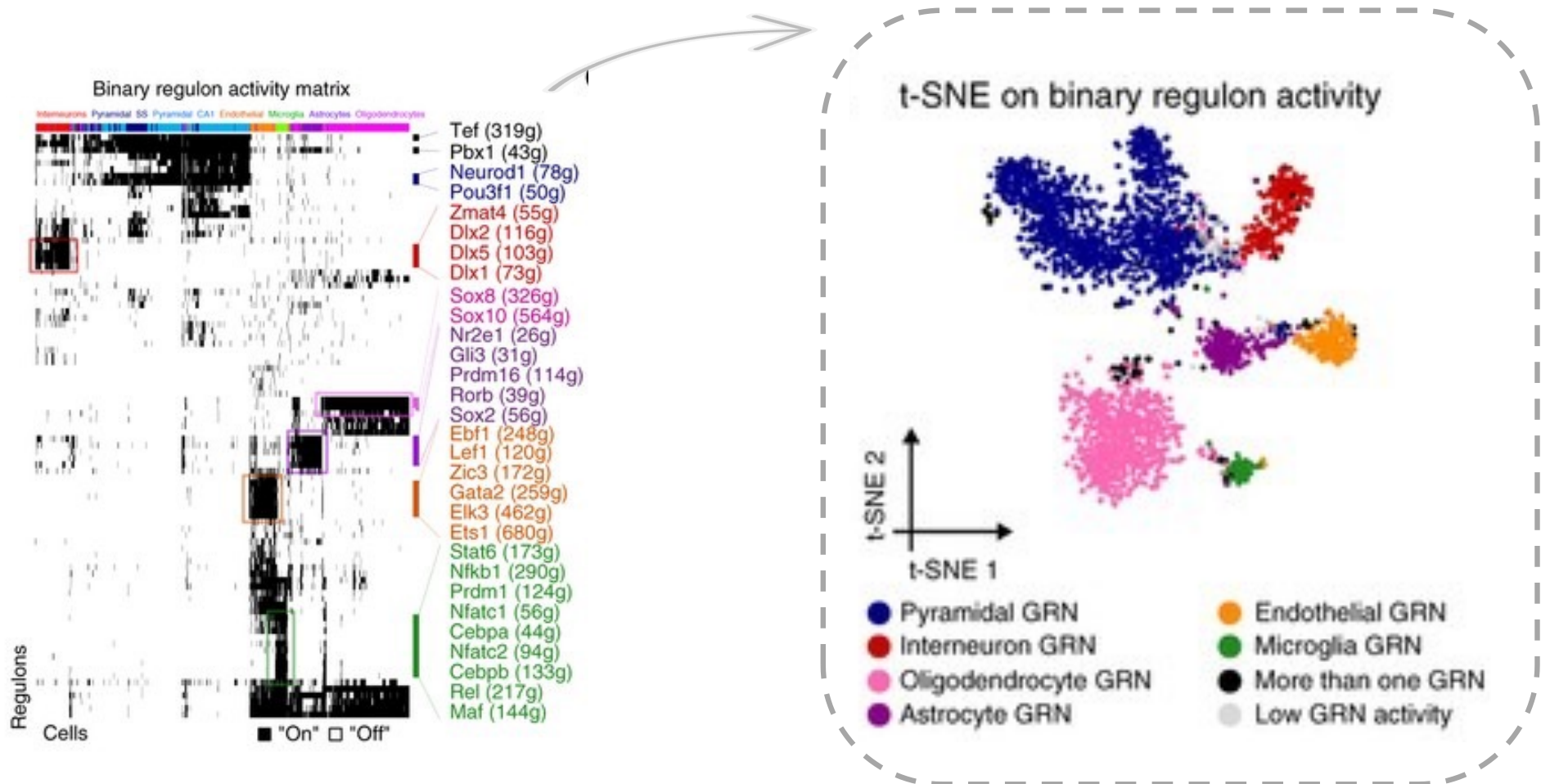
AUCCell

Identifying cells with active gene-sets

Regulon activity matrix (Network activity in each cell)

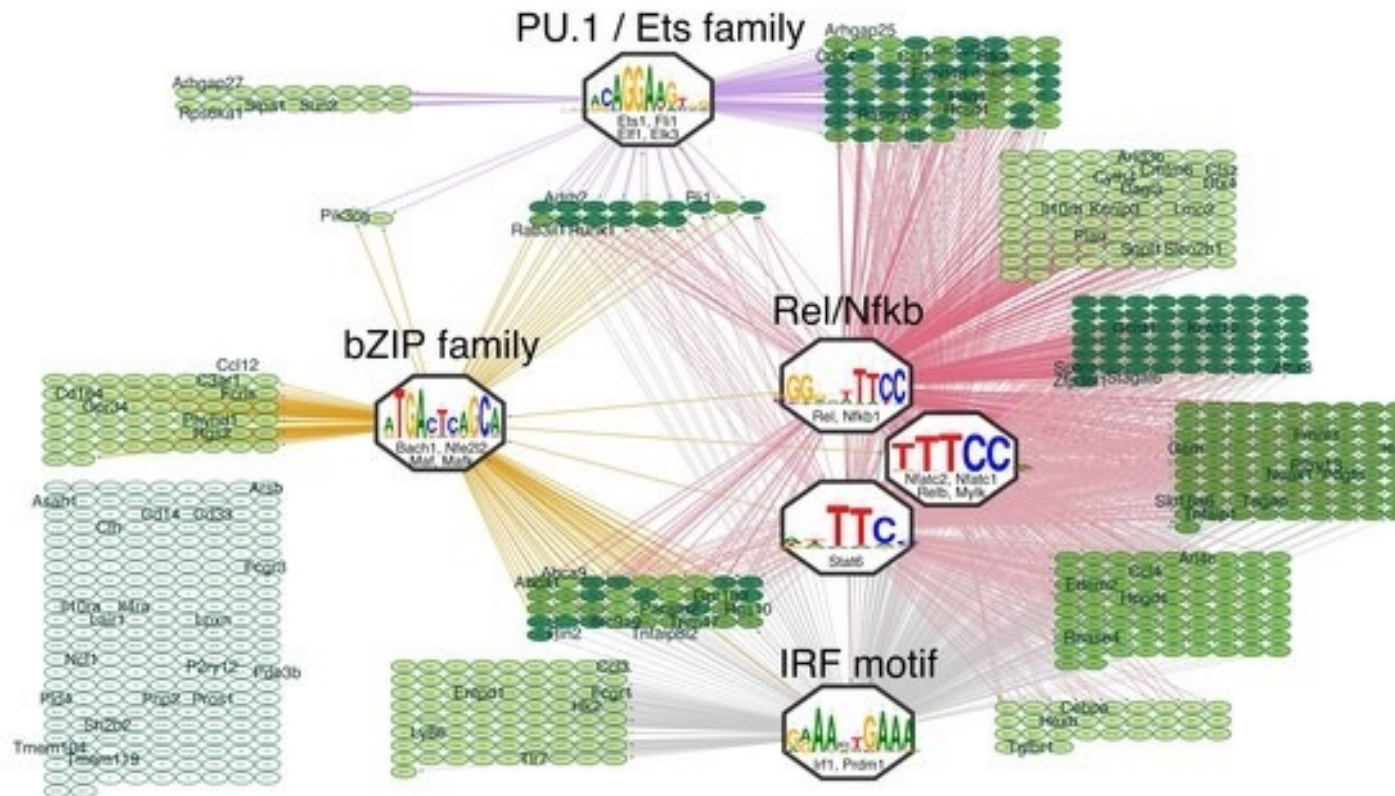


Top regulons on the Mouse brain



Microglia GRN on the Mouse brain

- The regulons associated to microglia can be summarized based on the binding motif of the associated TF .
- The predicted network for microglia contains many well-known regulators of microglial fate and/or microglial activation, including PU.1, Nfkb, Irf, and AP-1/Maf.



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://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/index.html

Tools

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