Single cell RNA-seq analysis

Part II: understanding cell types

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2022

Daifeng Wang

daifeng.wang@wisc.edu

Thanks to Ting Jin for slides!

Outline

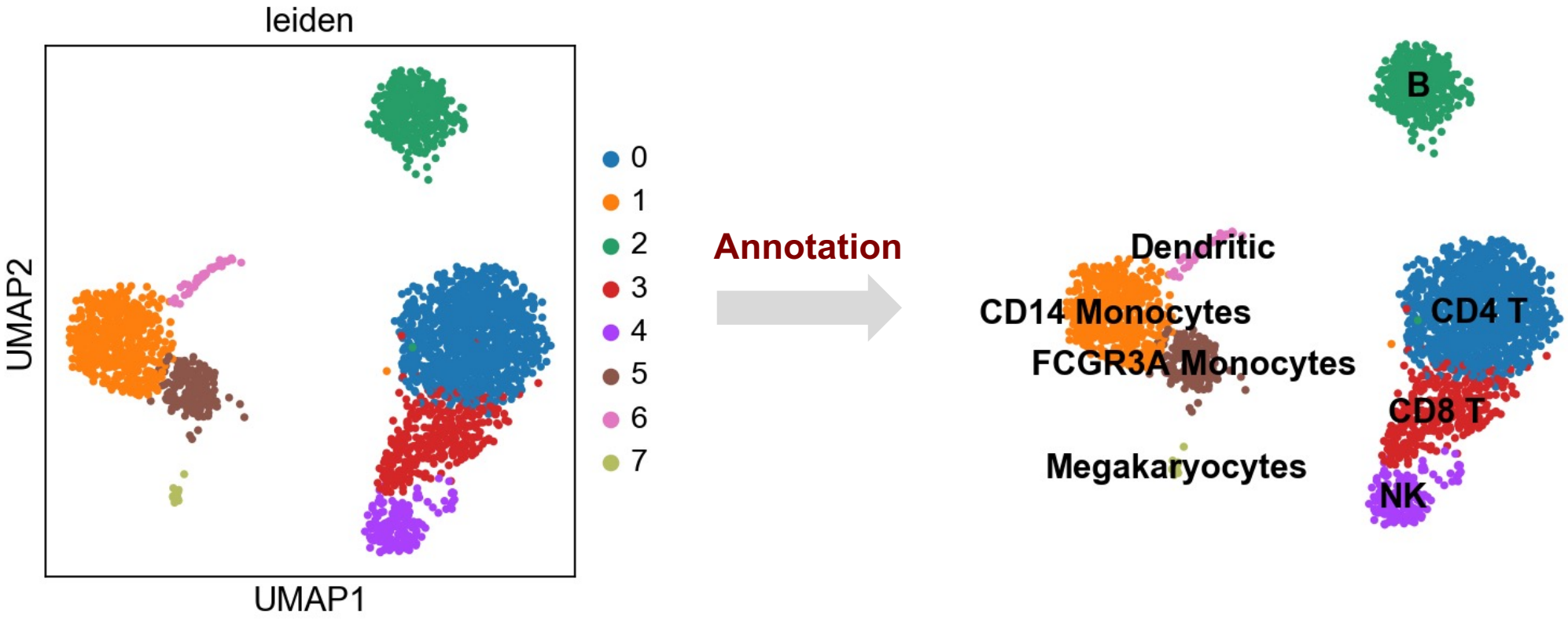
- scRNA-seq data analysis
  - Cell type annotation
    - SingleR
  - Cell type markers identification
  - Pseudo timing
    - Monocle
  - Cell-type gene regulatory networks
    - SCENIC
    - BEELINE
  - Single cell deconvolution
    - CIBERSORTx

Outline

- scRNA-seq data analysis
  - **Cell type annotation**
    - **SingleR**
  - Cell type markers identification
  - Pseudo timing
    - Monocle
  - Cell-type gene regulatory networks
    - SCENIC
    - BEELINE
  - Single cell deconvolution
    - CIBERSORTx

Cell type annotation

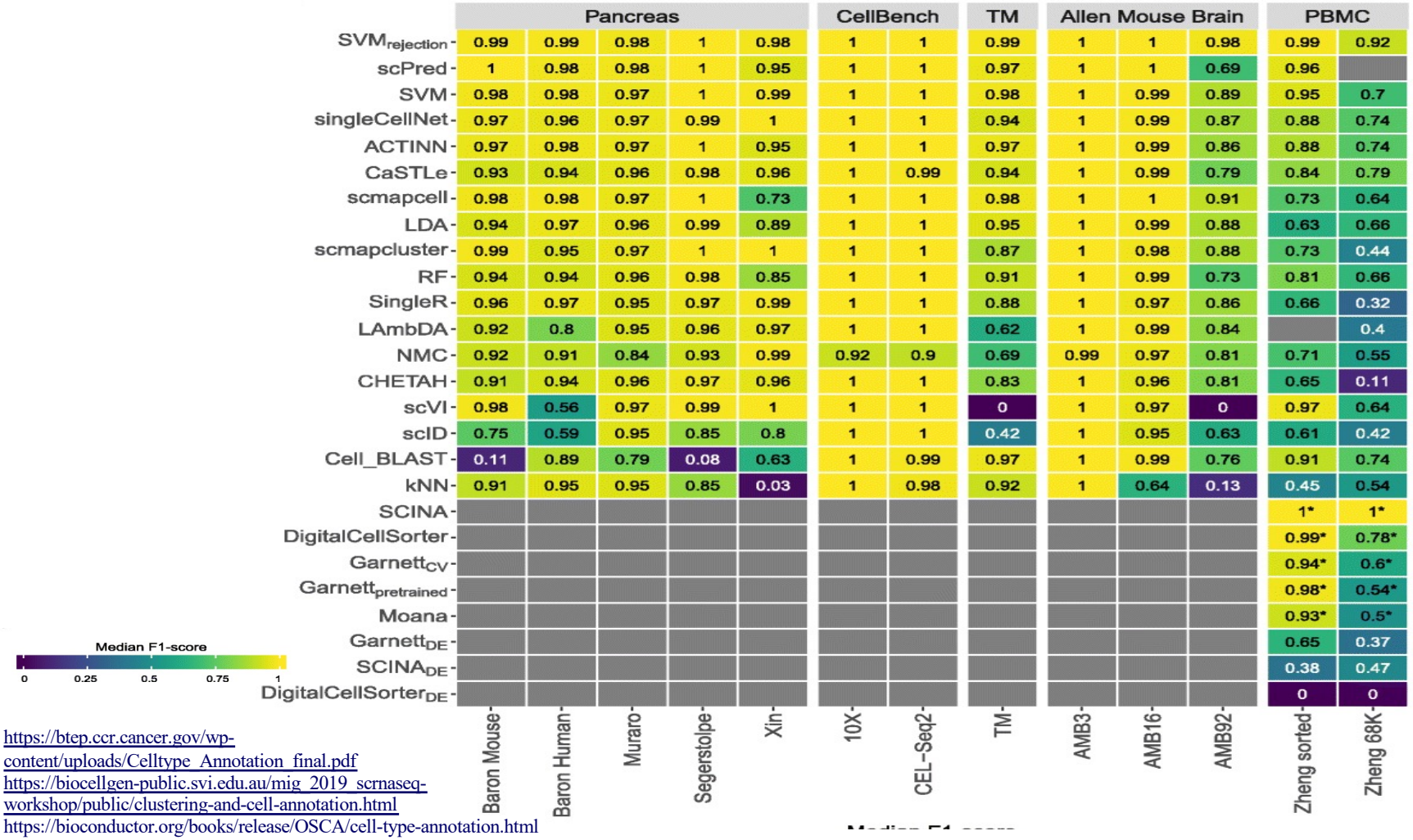
- Cell types -> cellular functions
- Assign the cell type for each cell



https://btep.ccr.cancer.gov/wp-content/uploads/Celltype_Annotation_final.pdf
https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-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



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

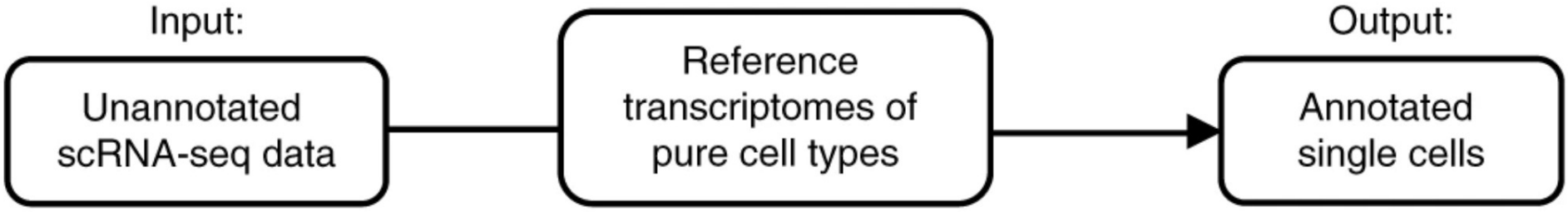
SingleR: Reference-based annotation of scRNA-seq

- 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



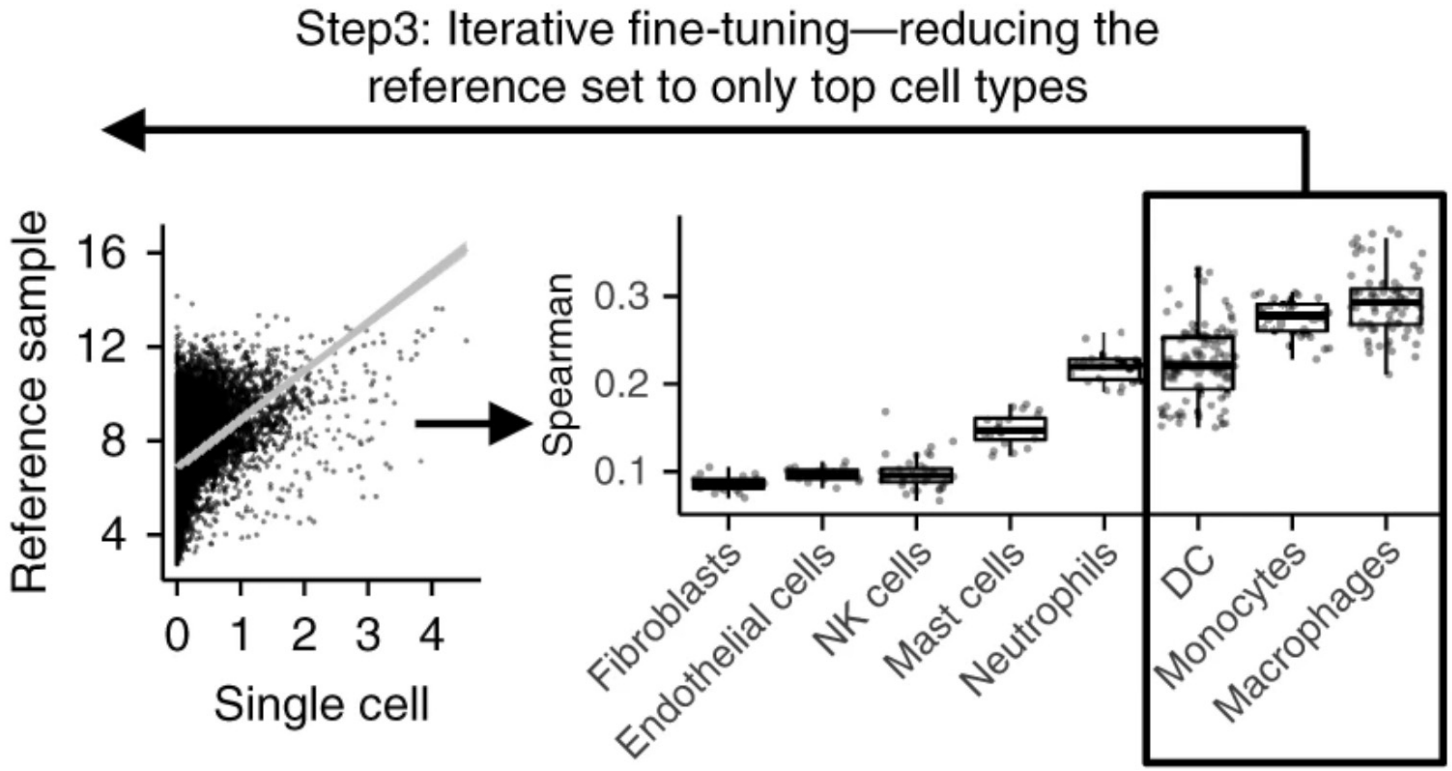
SingleR: Reference-based annotation of scRNA-seq

Step 1:

Identifying variable genes among cell types in the reference set

Step 2:

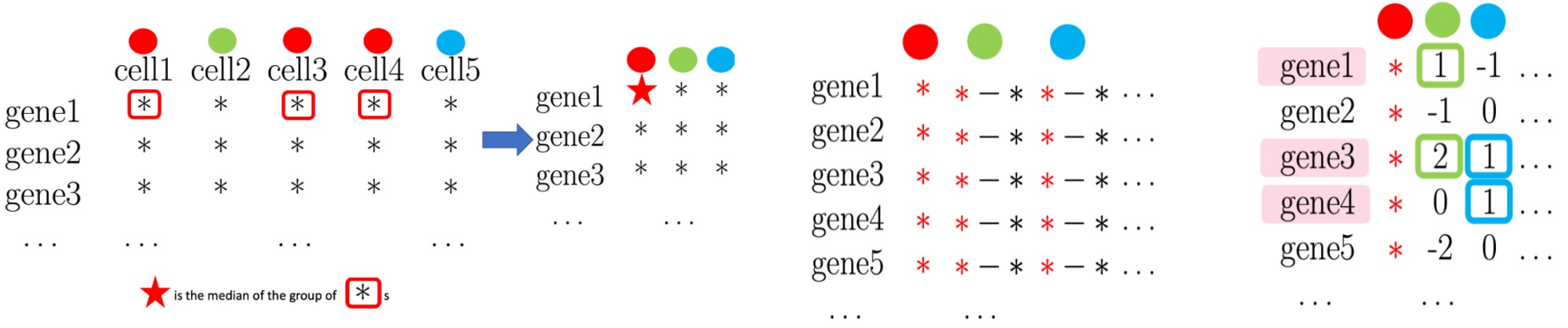
Correlating each single-cell transcriptome with each sample in the reference set



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.

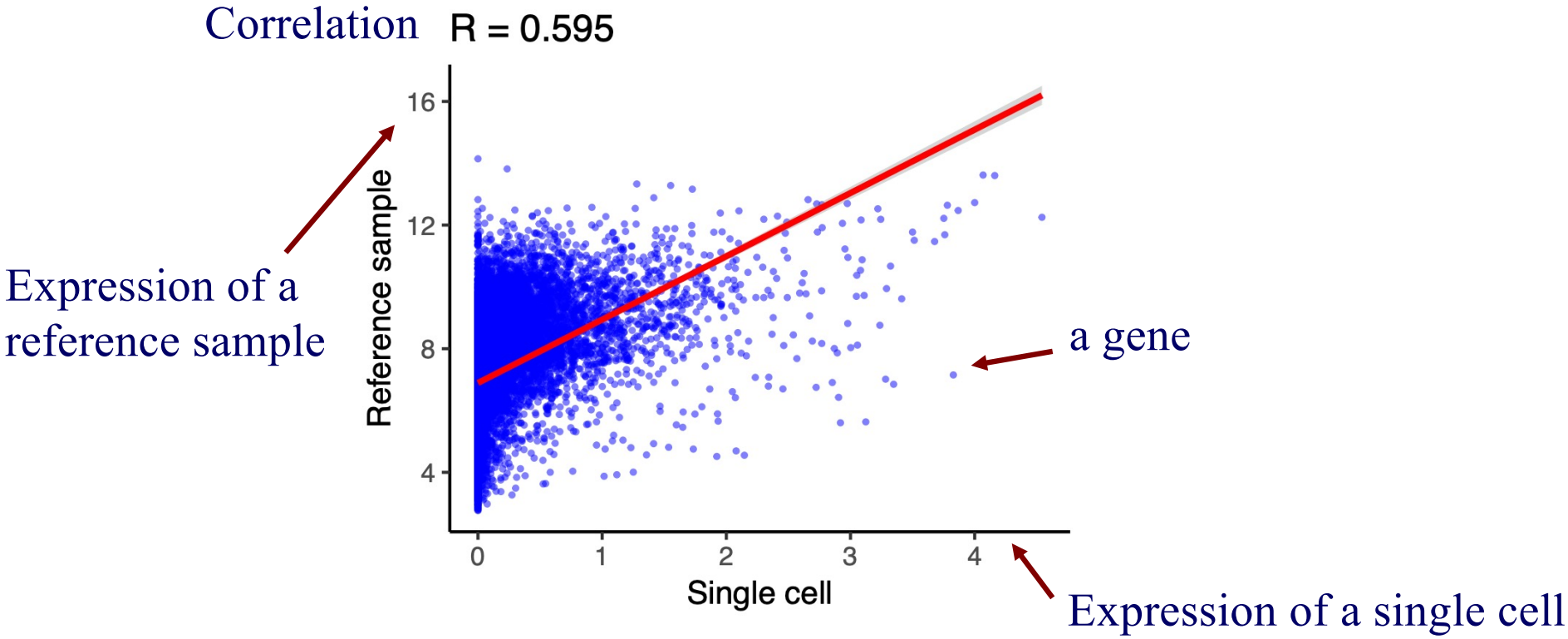


https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/clustering-and-cell-annotation.html

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

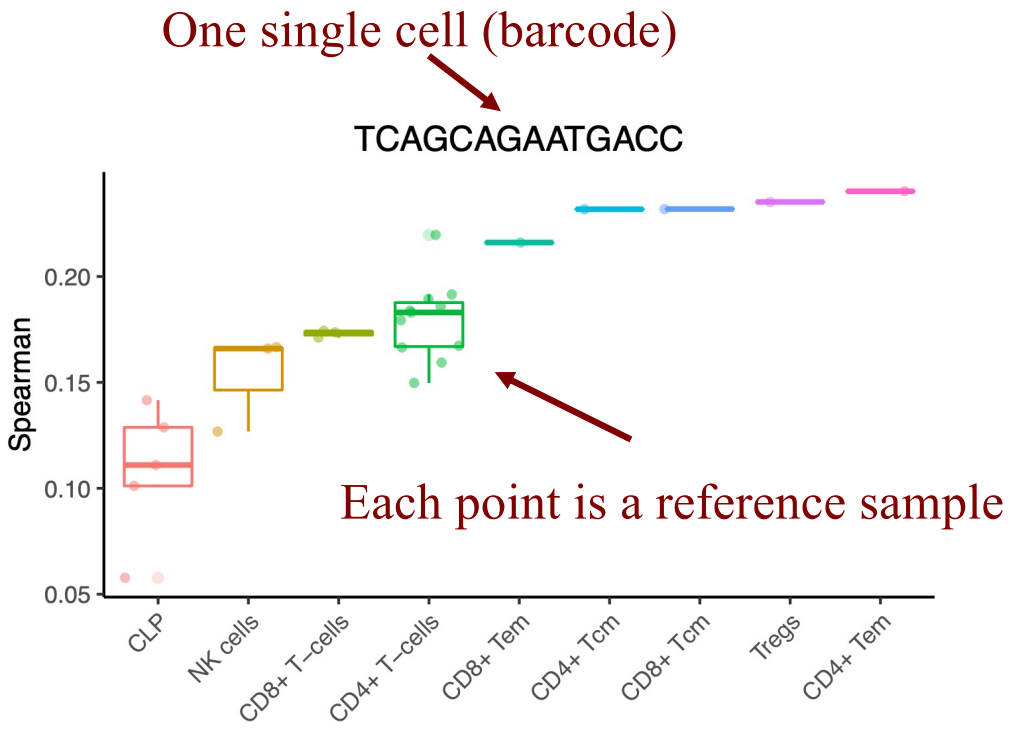
Step 2: Correlating each single-cell transcriptome with each sample in the reference set

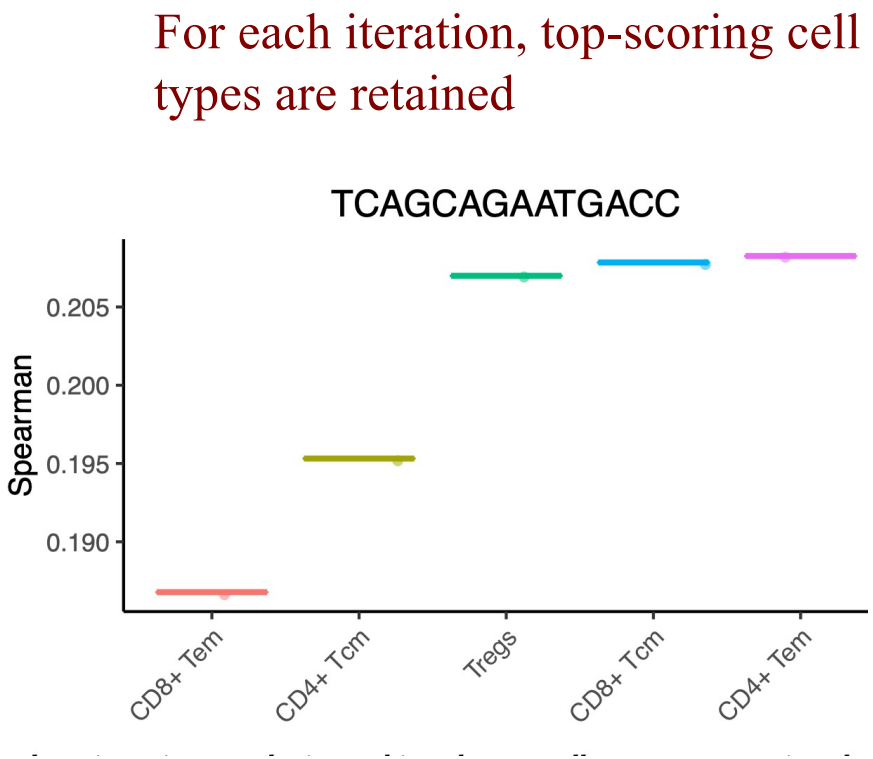
- Spearman coefficient is calculated for single cell expression with each of the samples in the reference dataset.
- The correlation analysis is performed only on variable genes in the reference dataset.



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

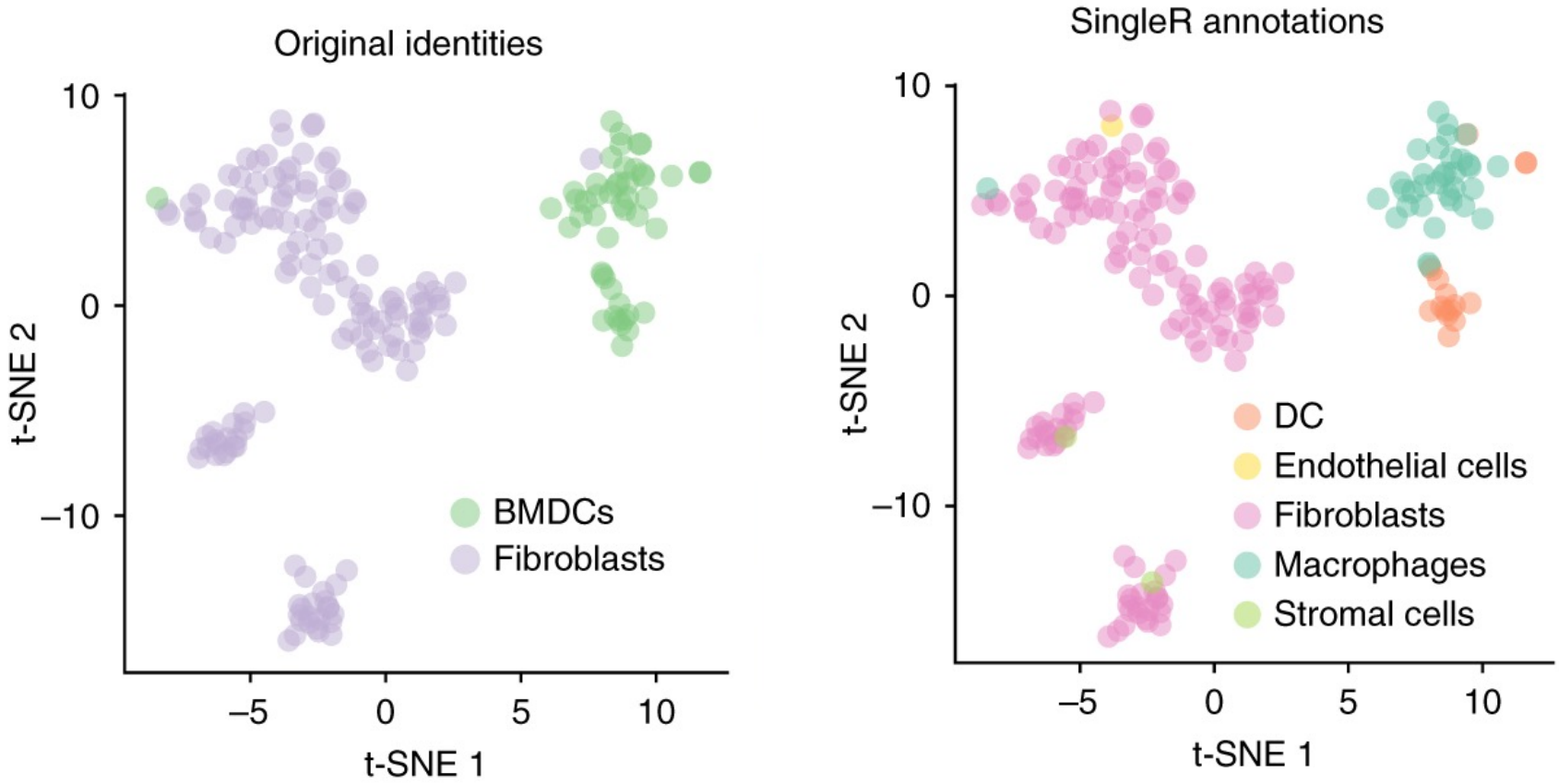
- For a single cell and each cell type, multiple Spearman correlation coefficients are aggregated into a “cell-type score”
  - The SingleR score for each cell type is the 80 percentile in each of the boxplots.
- Cell types with the lowest score or a score below will be removed
- Repeat from step 1 until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





https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/clustering-and-cell-annotation.html

SingleR: Reference-based annotation of scRNA-seq



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

Outline

- scRNA-seq data analysis
  - Cell type annotation
    - SingleR
  - **Cell type markers identification**
  - Pseudo timing
    - Monocle
  - Cell-type gene regulatory networks
    - SCENIC
    - BEELINE
  - Single cell deconvolution
    - CIBERSORTx

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)*

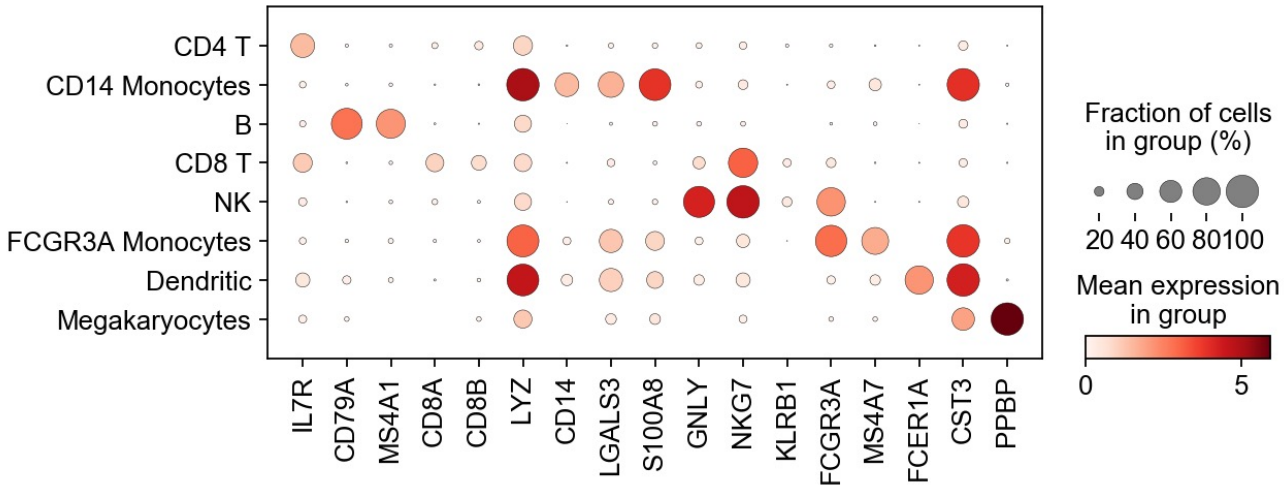
- Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16,** 278 (2015). https://doi.org/10.1186/s13059-015-0844-5
- Love MI, Huber W, Anders S (2014).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



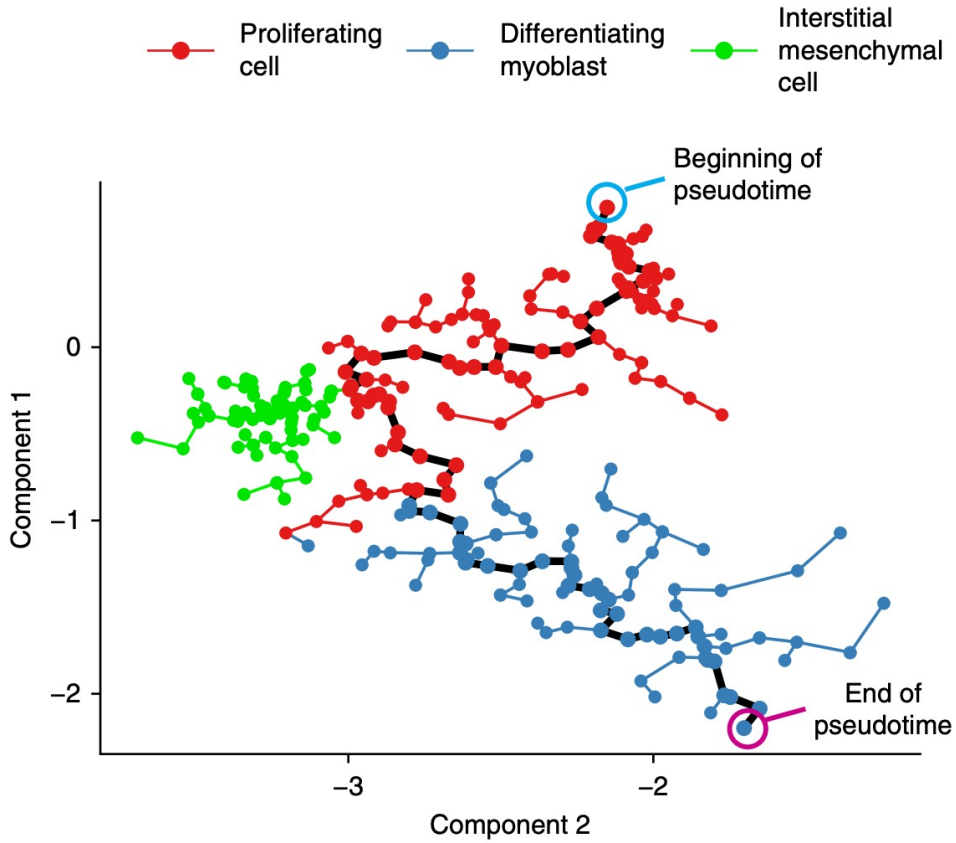
- Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16,** 278 (2015). https://doi.org/10.1186/s13059-015-0844-5
- Love MI, Huber W, Anders S (2014). "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://satijalab.org/seurat/archive/v3.1/immune_alignment.html
- https://zenodo.org/record/4317764#.Yll7gdPMKCg

Outline

- scRNA-seq data analysis
  - **Cell type annotation**
    - **SingleR**
  - Cell type markers identification
  - **Pseudo timing**
    - **Monocle**
  - Cell-type gene regulatory networks
    - SCENIC
    - BEELINE
  - Single cell deconvolution
    - CIBERSORTx

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



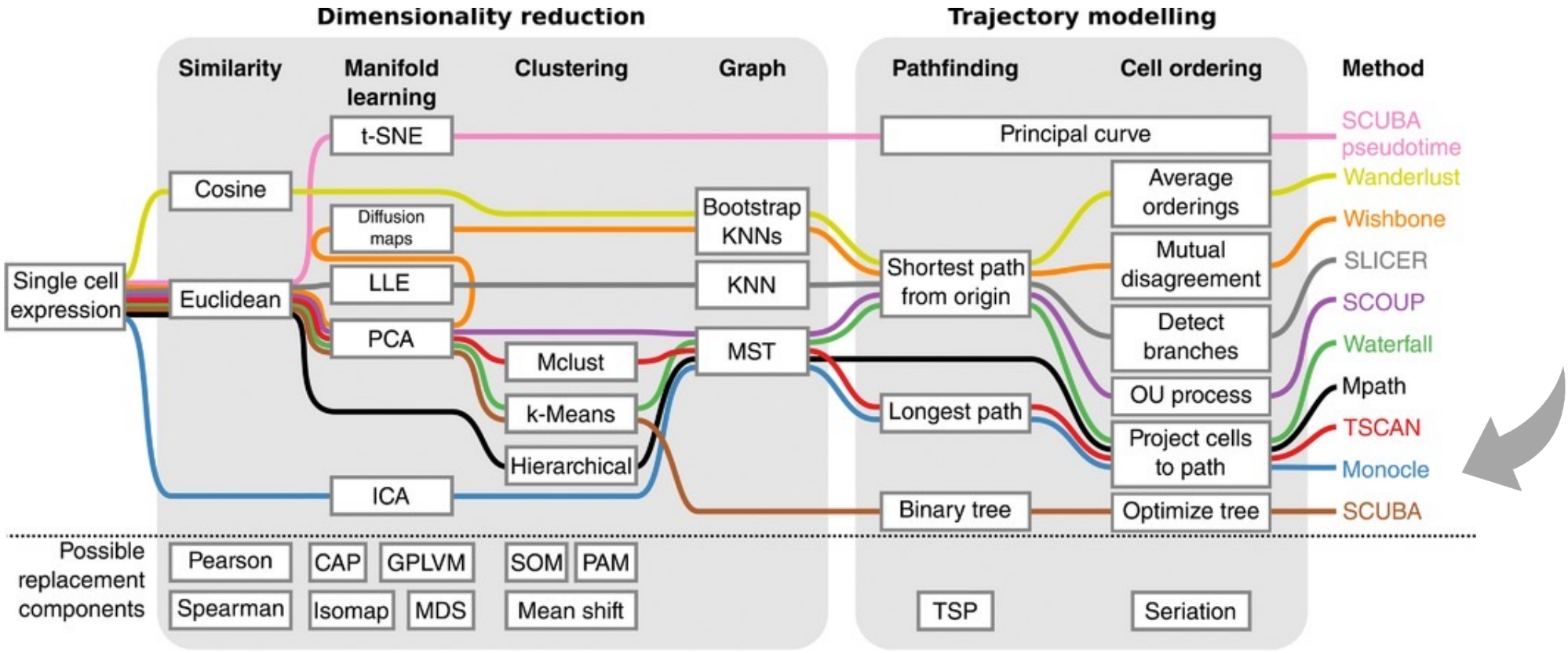
http://cole-trapnell-lab.github.io/monocle-releas
https://scrnaseq-course.cog.sanger.ac.uk/website

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



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

Monocle

Constructing single cell trajectories

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

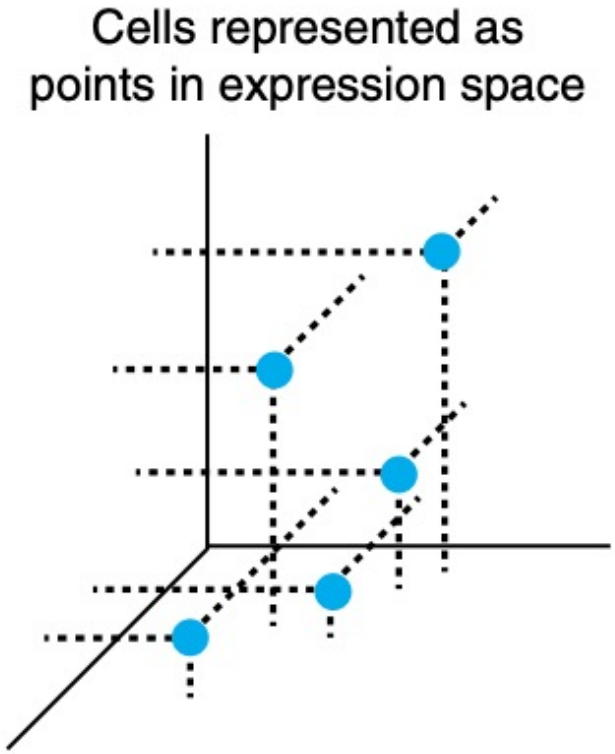
- Step 1: Choose genes that define progress
- Step 2: Reduce data dimensionality
  - independent component analysis (ICA)
- Step 3: Construct minimum spanning tree (MST) on the cells
- Step 4: Find the longest path through the MST
- Step 5: Order cells along the trajectory

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

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

Step 1: Choose genes that define progress

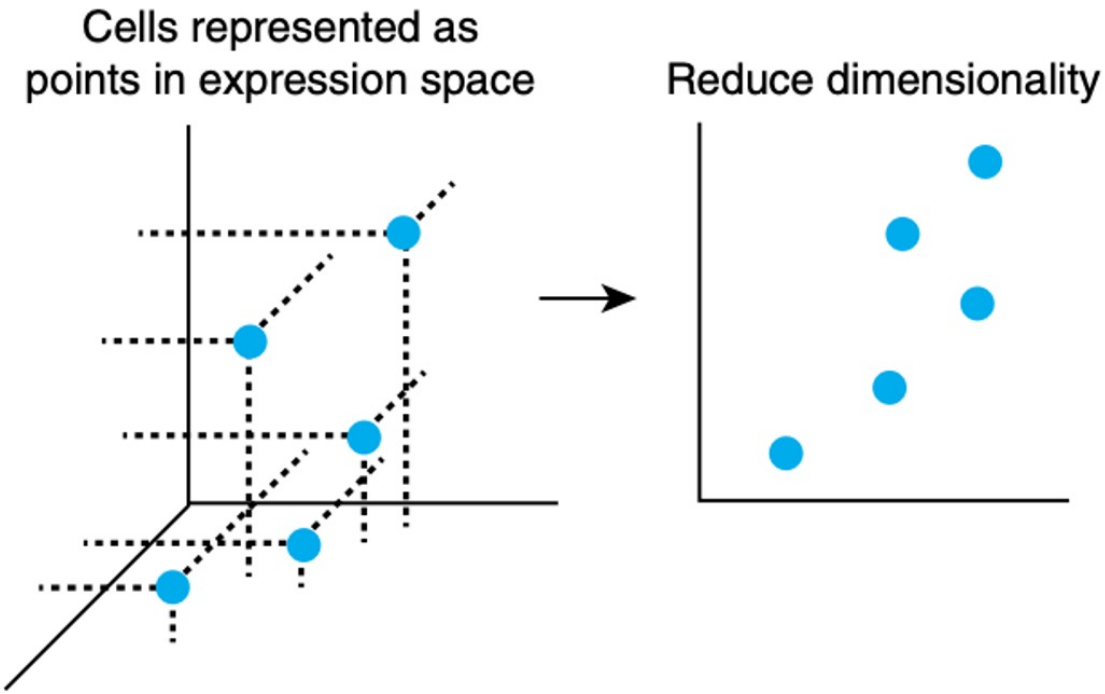
- Represent the expression profile of each cell as a point in a high-dimensional Euclidean space, with one dimension for each gene



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

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 $\boldsymbol{W}$ which maximize independence and unmix sources signal $\boldsymbol{s}$

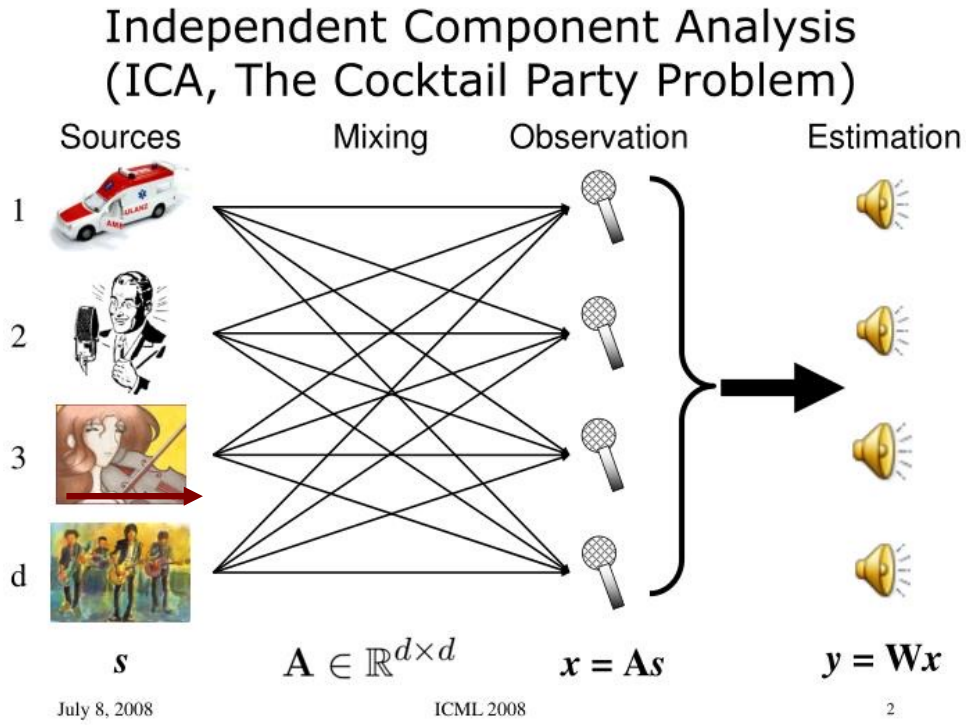
Mixed variables

Mixing matrix

$$y = \boldsymbol{W}\, x = \boldsymbol{W}\, \boldsymbol{A}\, s$$

Source

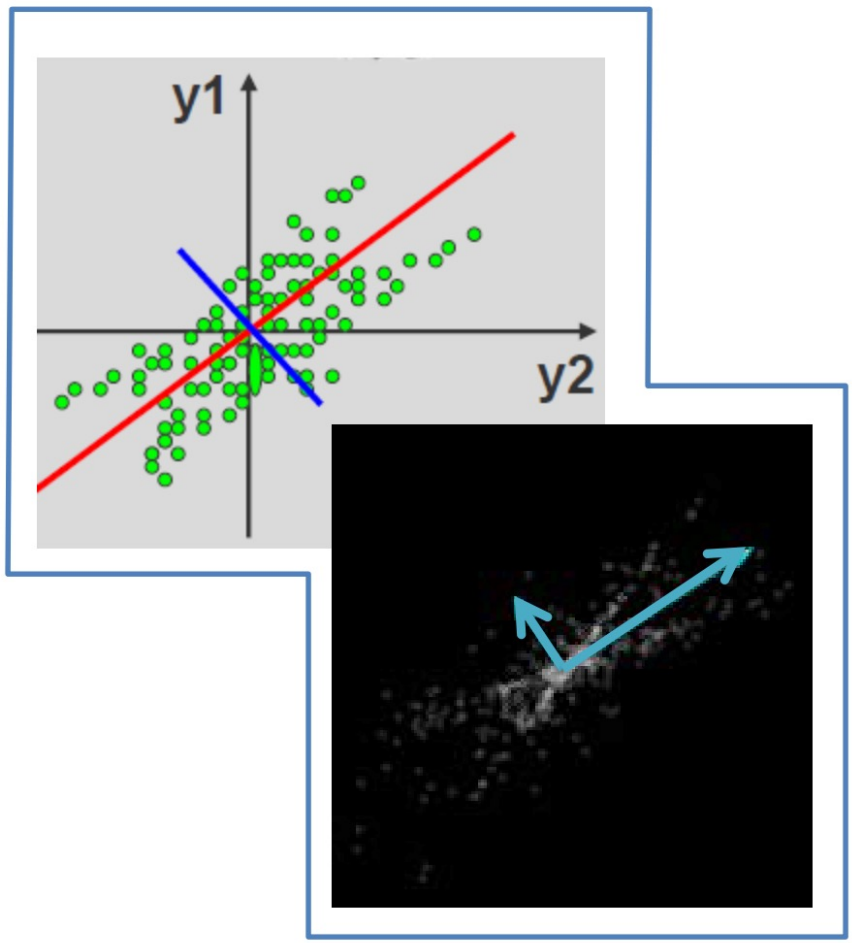
Linear mapping

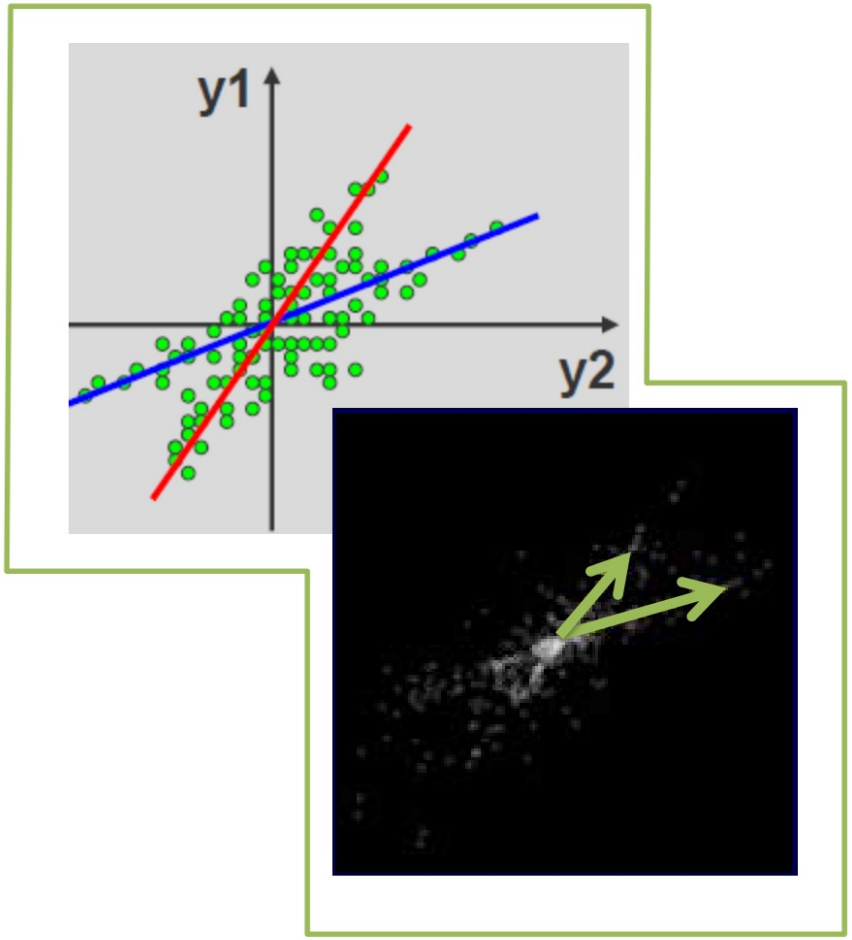


https://github.com/NBISweden/excelerate-scRNAseq/blob/master/session-trajectories/trajectory_inference_analysis.pdf
https://www.slideserve.com/vladimir-kirkland/ica-and-isa-using-schweizer-wolff-measure-of-dependence

ICA vs PCA

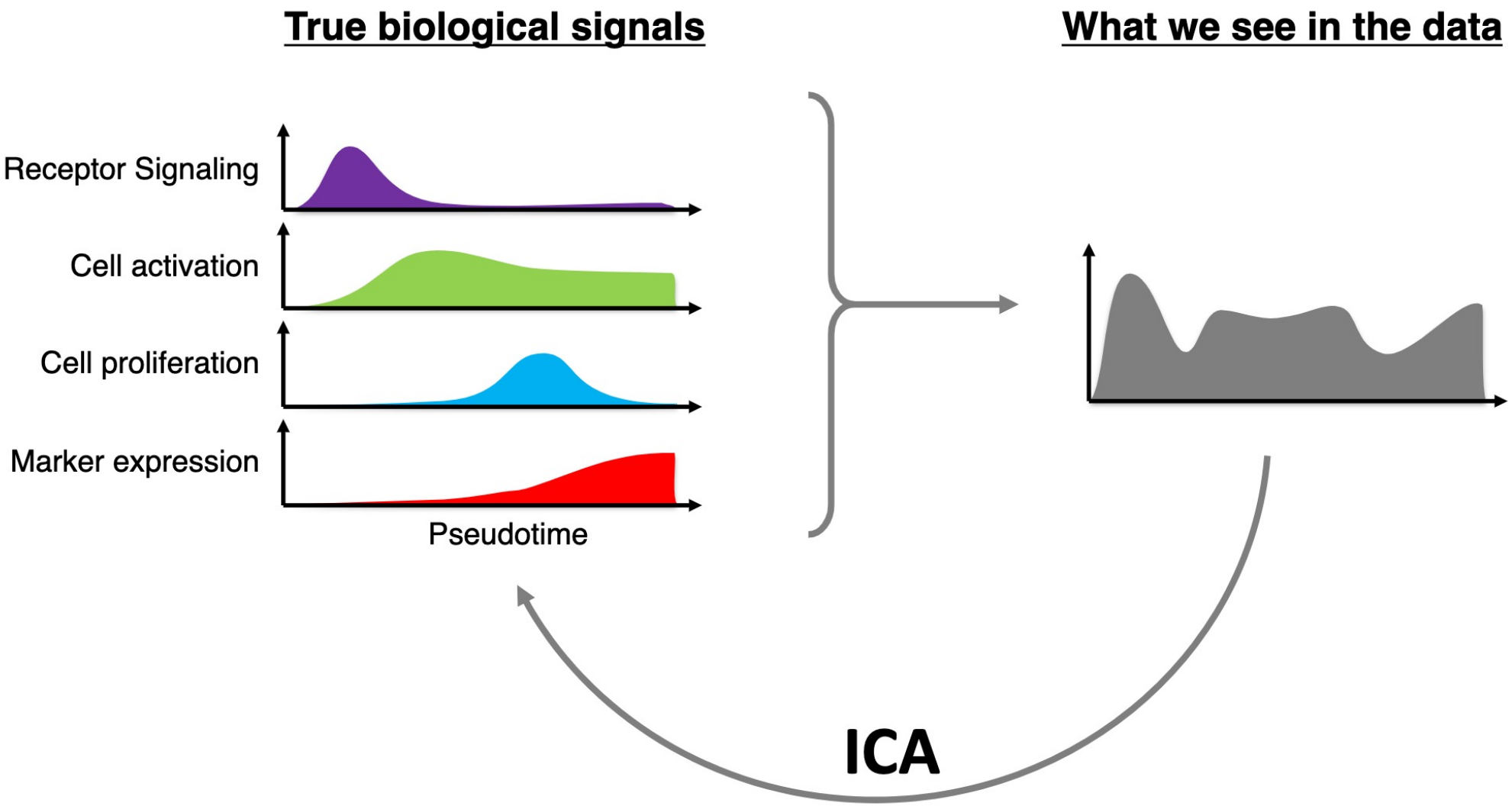
- PCA : Find the directions of maximal variance
- ICA : Find the directions of maximal independence
  - ➢ The values in each source have non-Gaussian distributions





https://github.com/NBISweden/excelerate-scRNAseq/blob/master/session-trajectories/trajectory_inference_analysis.pdf
https://scikit-learn.org/stable/modules/decomposition.html#independent-component-analysis-ica

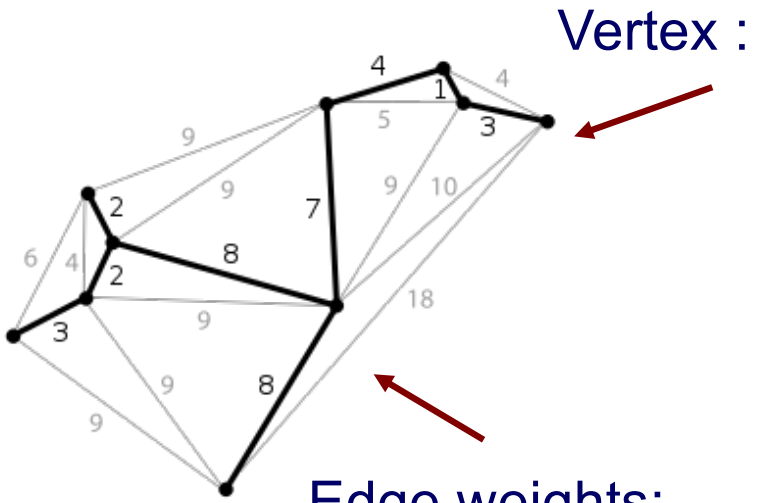
Why ICA

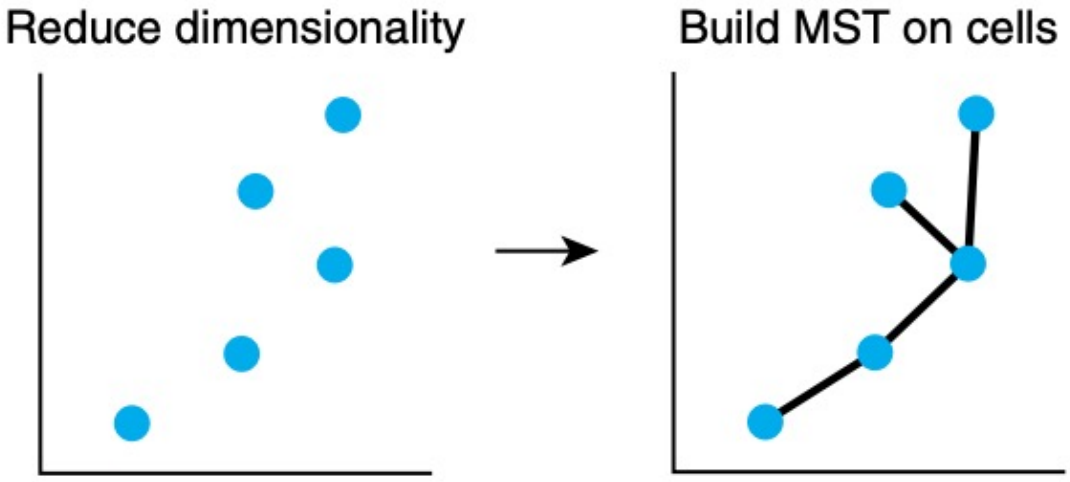


https://github.com/NBISweden/excelerate-scRNAseq/blob/master/session-trajectories/trajectory_inference_analysis.pdf
https://scikit-learn.org/stable/modules/decomposition.html#independent-component-analysis-ica

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

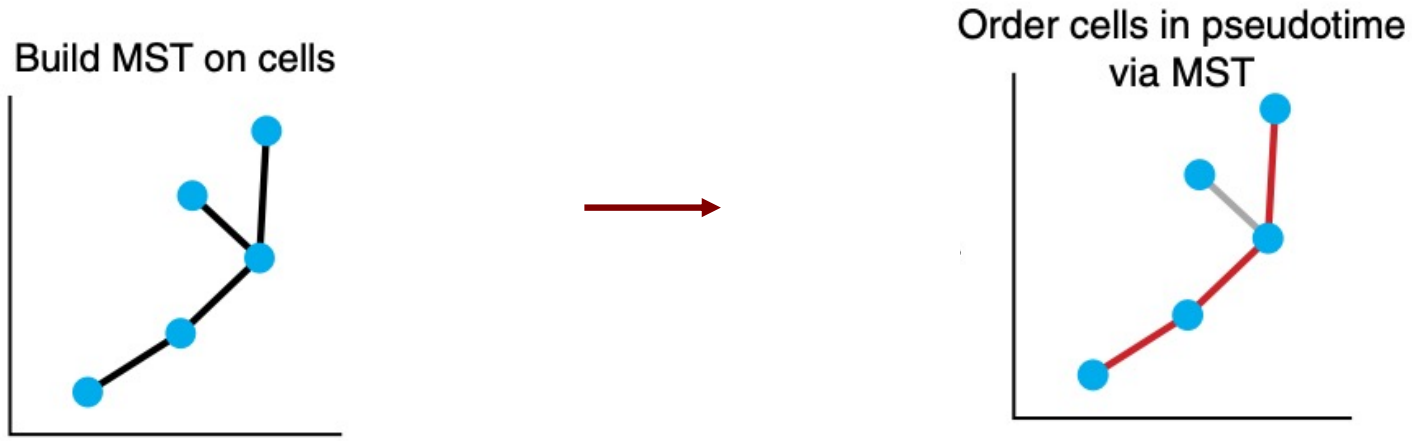




Trapnell, C., Cacchiarelli, D., Grimsby, J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32, 381–386 (2014). https://doi.org/10.1038/nbt.2859
https://en.wikipedia.org/wiki/Minimum_spanning_tree

Step 4: Find the longest path through the MST

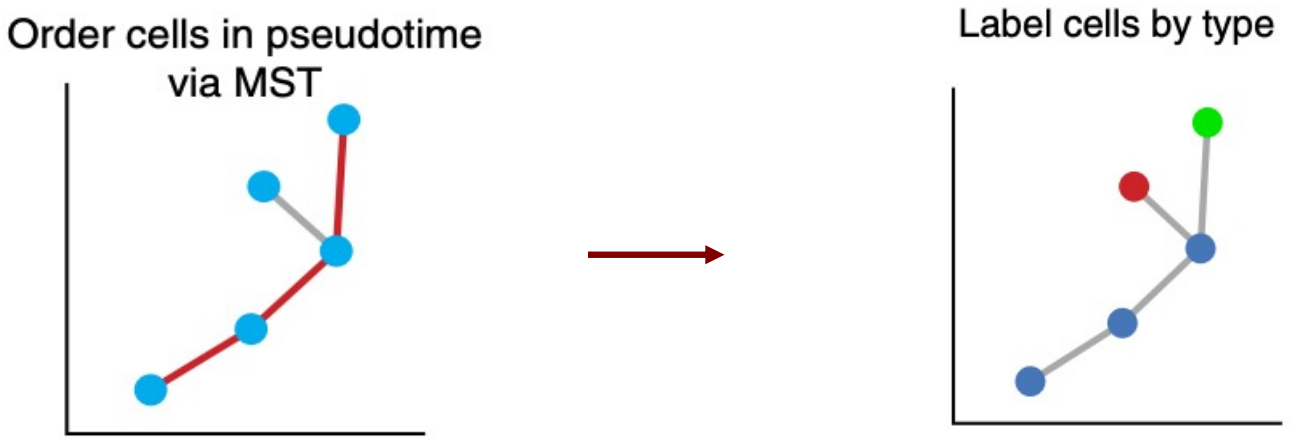
- Correspond to the longest sequence of similar cells (e.g., gene expression)



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

Step 5: Order cells along the trajectory

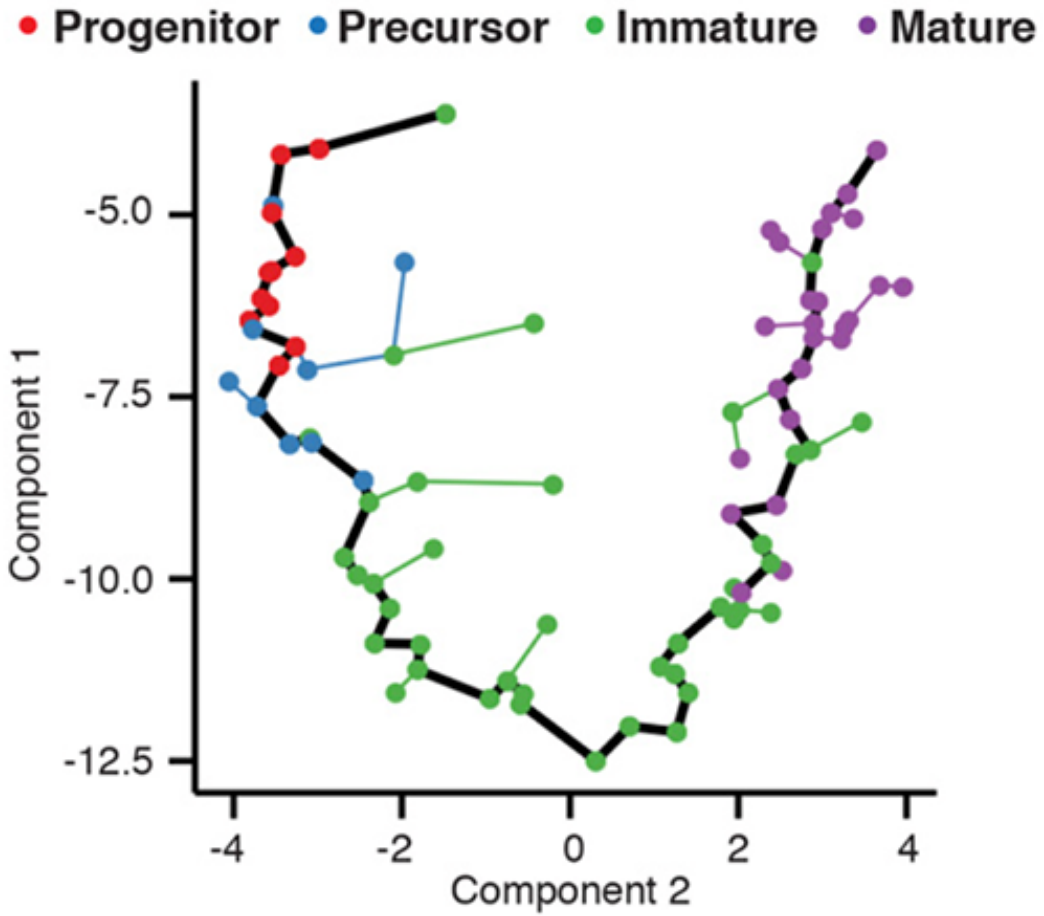
- Produce a ‘trajectory’ of an individual cell’s progress through differentiation



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

Developmental trajectory of olfactory neurons in mice

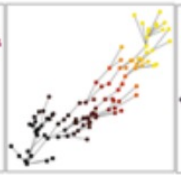
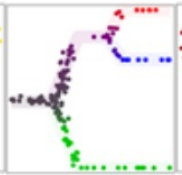
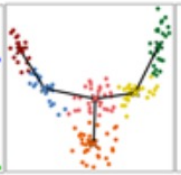
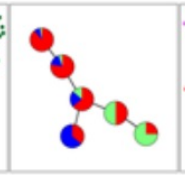
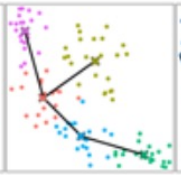
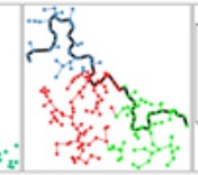
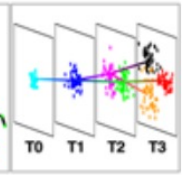
- Each point is a cell, which is connected to an MST
- The pseudotime value of each cell is measured as the distance along the trajectory from its position back to the beginning



https://cole-trapnell-lab.github.io/projects/sc-trajectories/

Pseudo timing

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

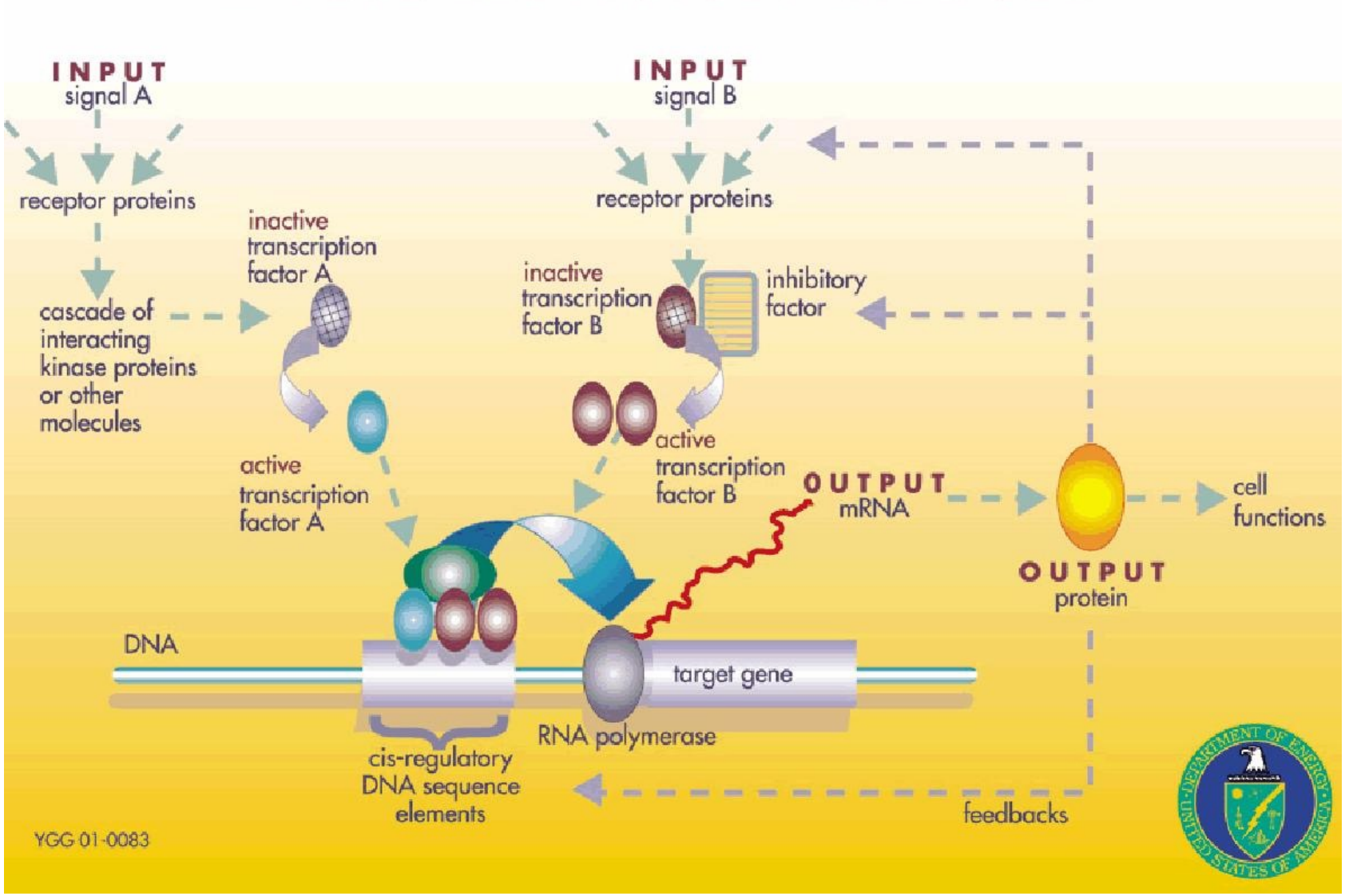
Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting 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**
    - BEELINE
  - Single cell deconvolution
    - CIBERSORTx

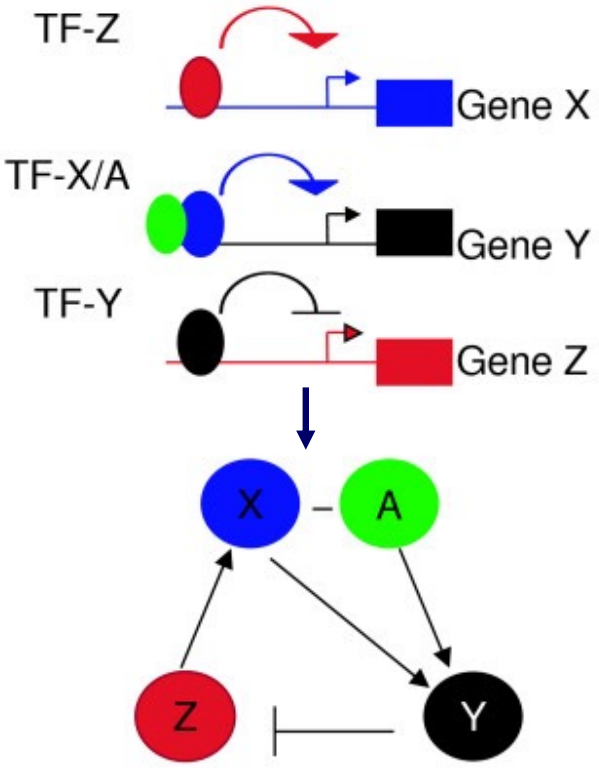
Gene regulation



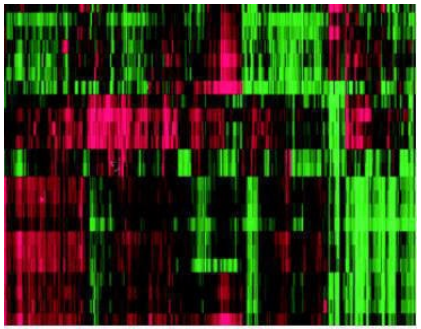
Gene regulation is the process of controlling which genes in a cell's DNA are expressed (used to make a functional product such as a protein).

Gene regulatory network

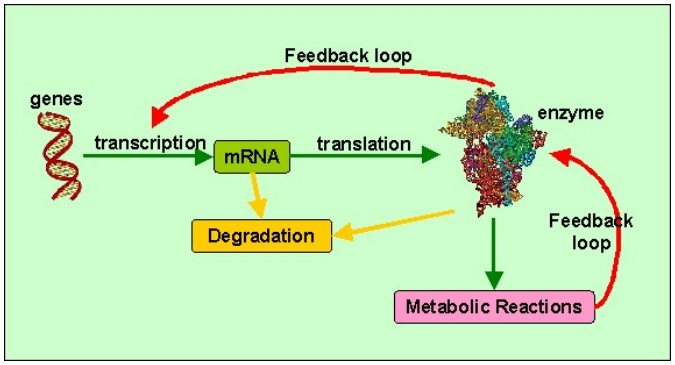
- Gene regulatory networks (GRNs) like on-off switches of a cell operating at the gene level
- Two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition
- GRN can be inferred from correlations in gene expression data, time-series gene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

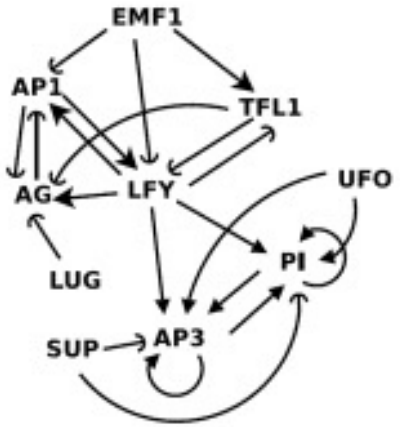


Observation



Inference

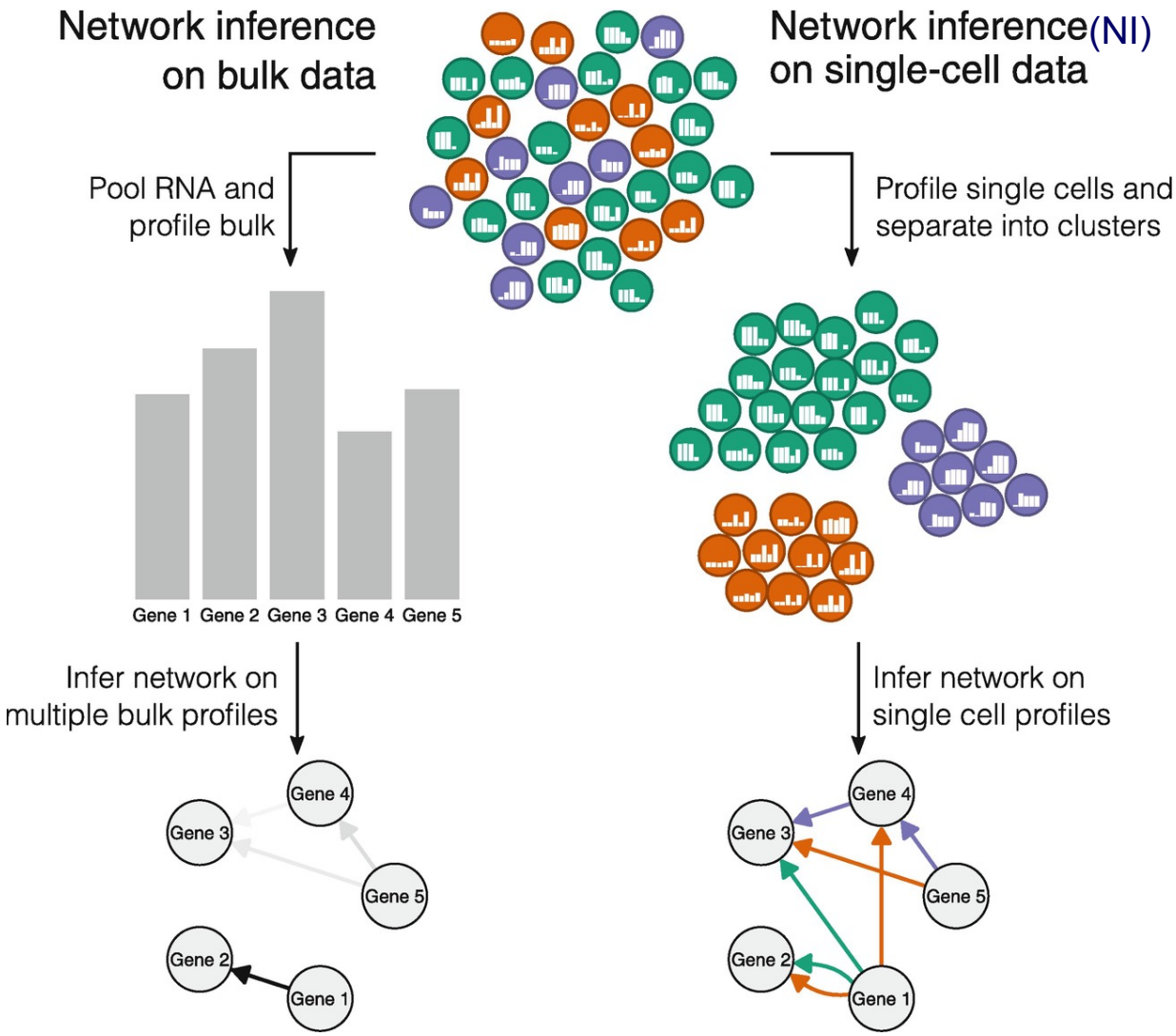




https://www.cs.purdue.edu/homes/ayg/TALKS/STC_CHICAGO10/Introductory_material/regulatory_networks.ppt

Cell-type gene regulatory networks

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



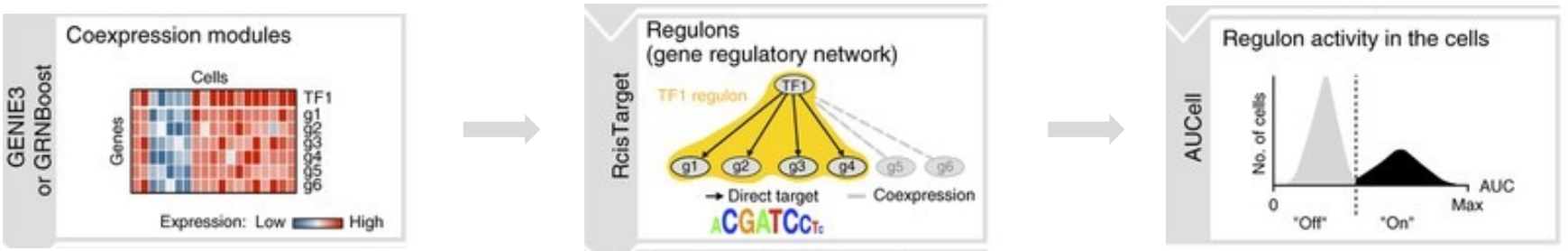
Todorov H., Cannoodt R., Saelens W., Saeys Y. (2019) Network Inference from Single-Cell Transcriptomic Data. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-8882-2_10

SCENIC

single-**c**ell r**e**gulatory **n**etwork **i**nference and **c**lustering

- 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

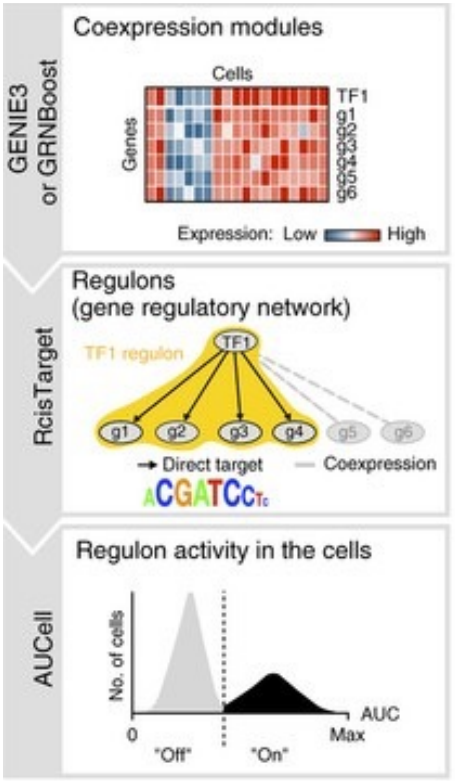
- The gene regulatory network is inferred based on co-expression and DNA motif analysis, and then the network activity is analyzed in each cell to identify the recurrent cellular states.

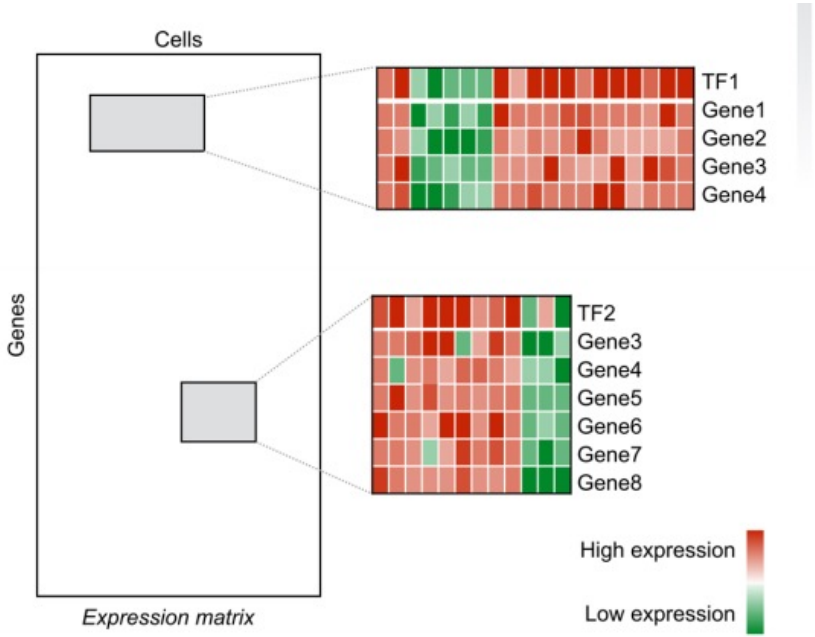


Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

Step 1: TF-based co-expression network

SCENIC

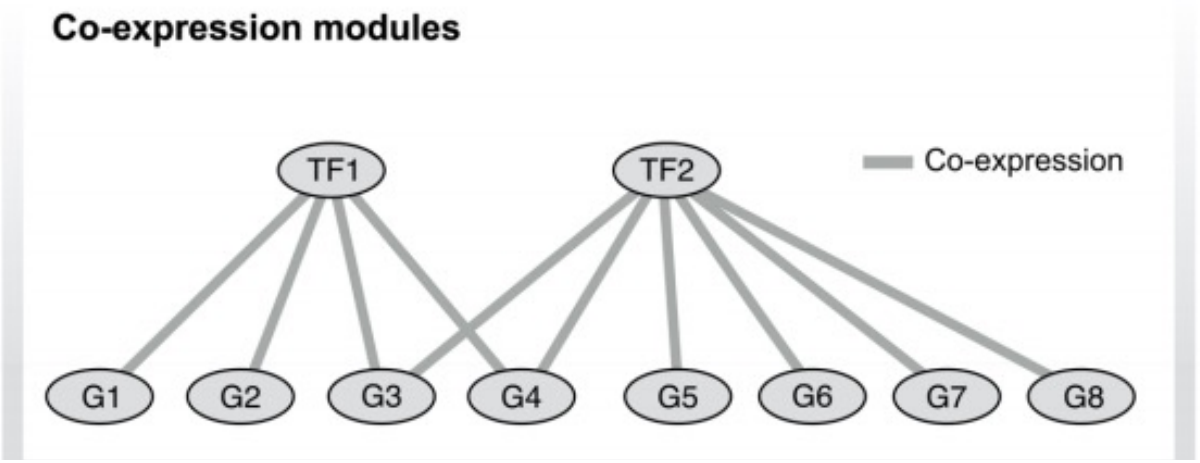




GENIE3

or

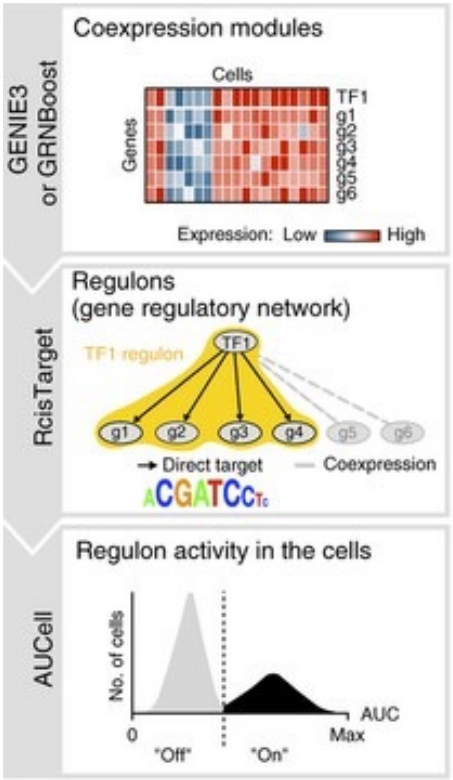
GRNBoost

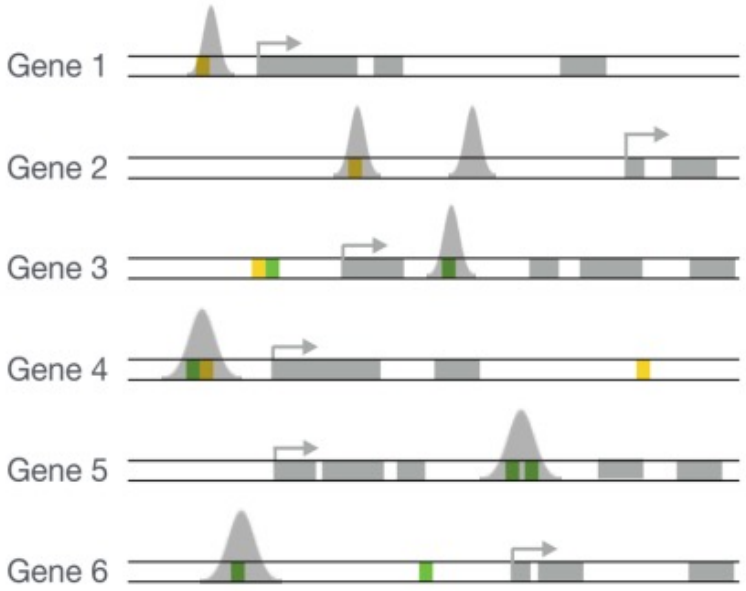


Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

Step 2: Identification of transcription factor binding motifs

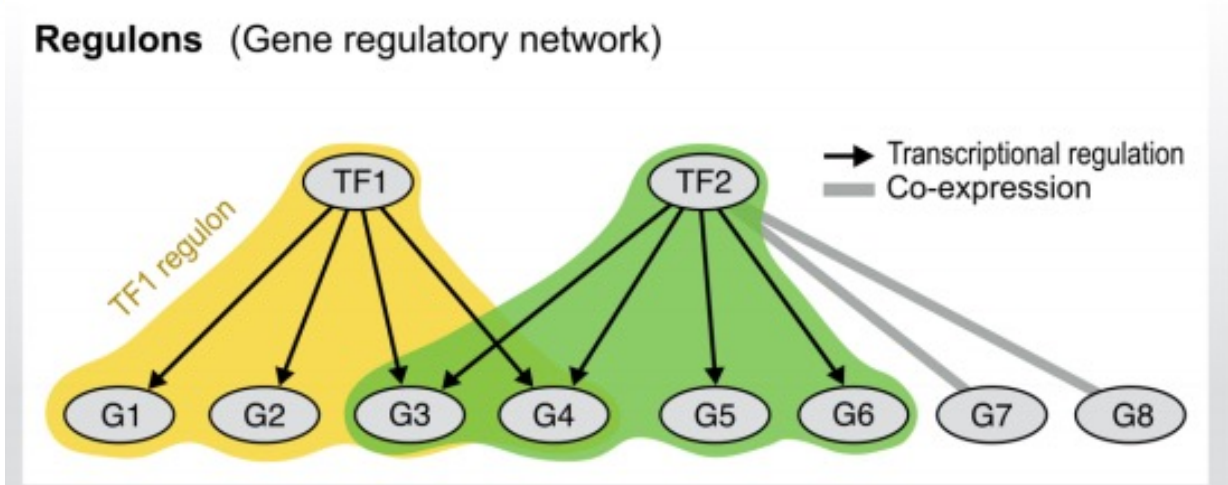
SCENIC





RcisTarget

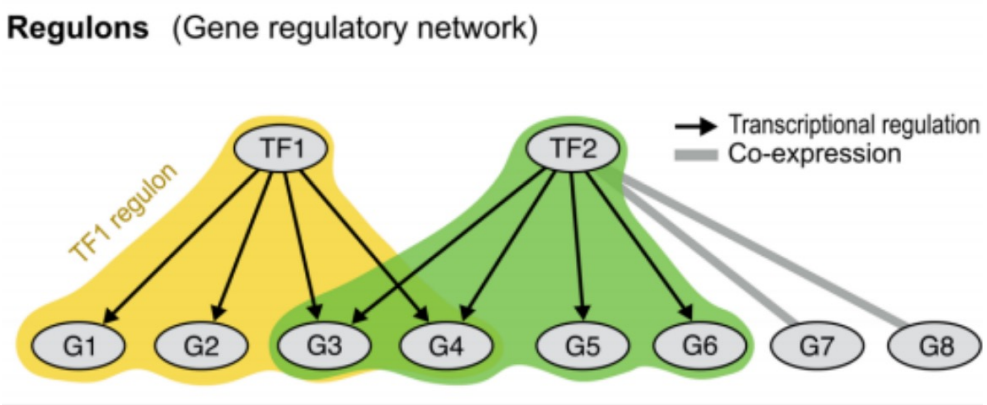
cis-regulatory sequence analysis



Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

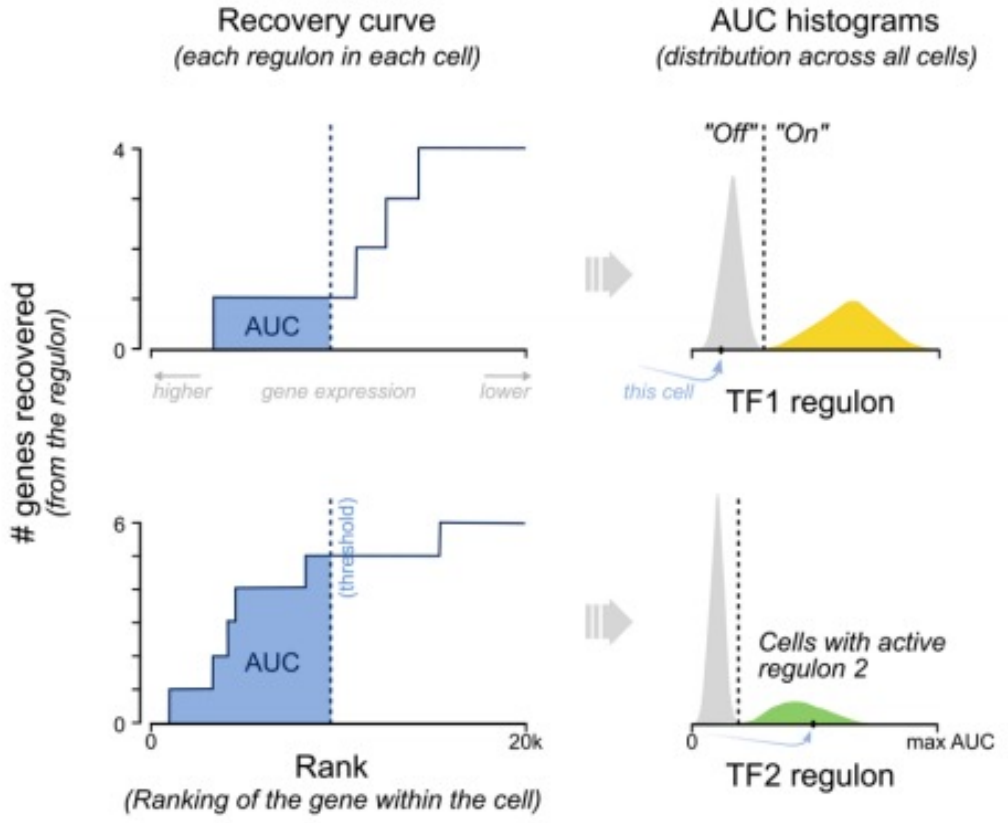
Regulon

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



AUCell score

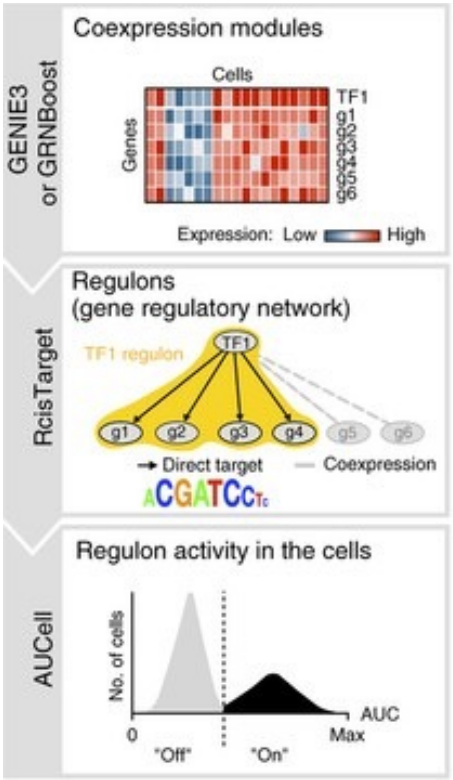
- AUCell uses the “Area Under the Curve” (AUC) to calculate whether a critical subset of the input gene set is enriched within the expressed genes for each cell.
- AUCell score: measure how active a regulon is in a cell
  - Step 1: For each cell, build gene-expression ranking
  - Step 2: Calculate enrichment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

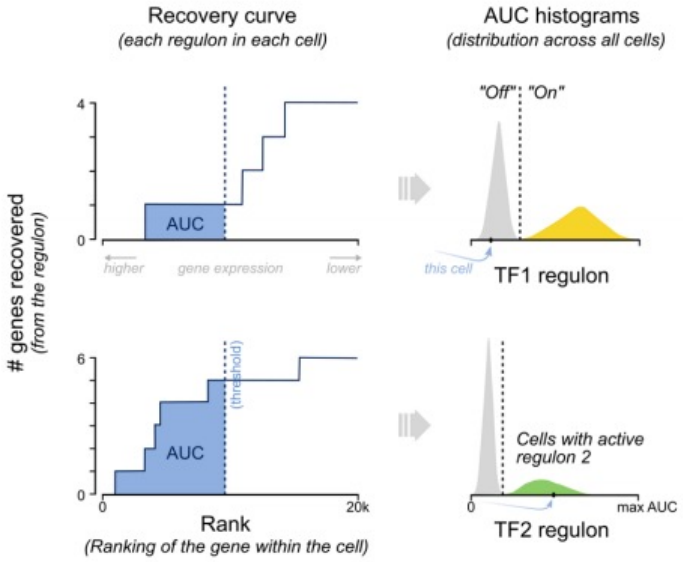


Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

Step 3: Regulon activities in each cell

SCENIC

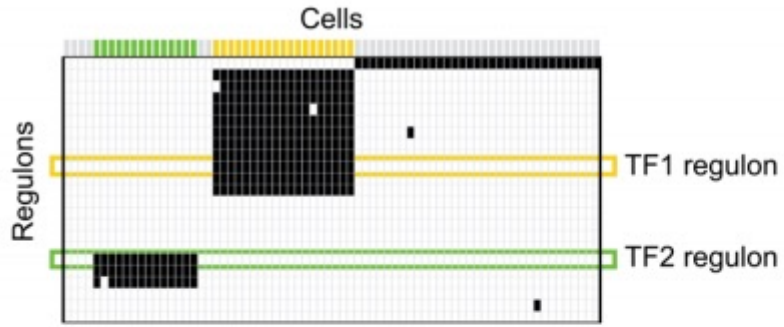




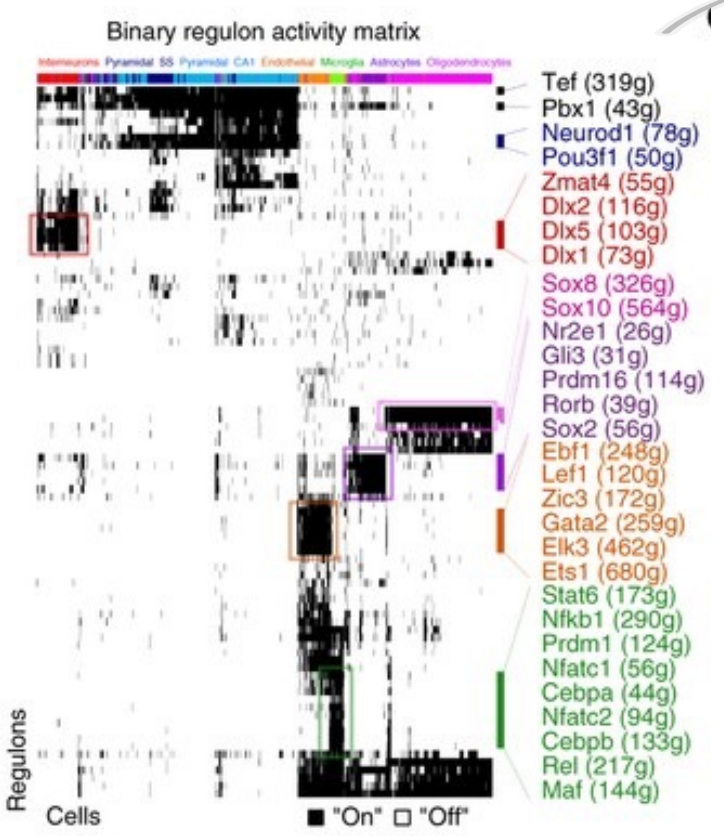
AUCell

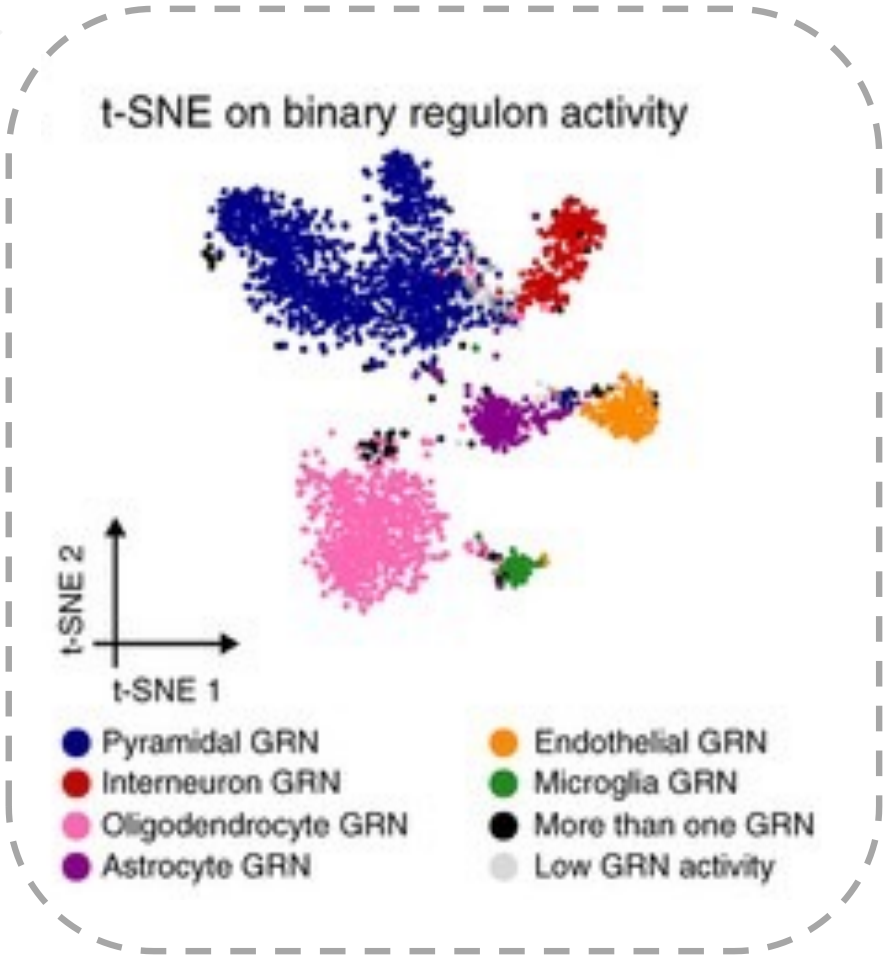
Identifying cells with active gene-sets

Regulon activity matrix (Network activity in each cell)



Top regulons on the Mouse brain

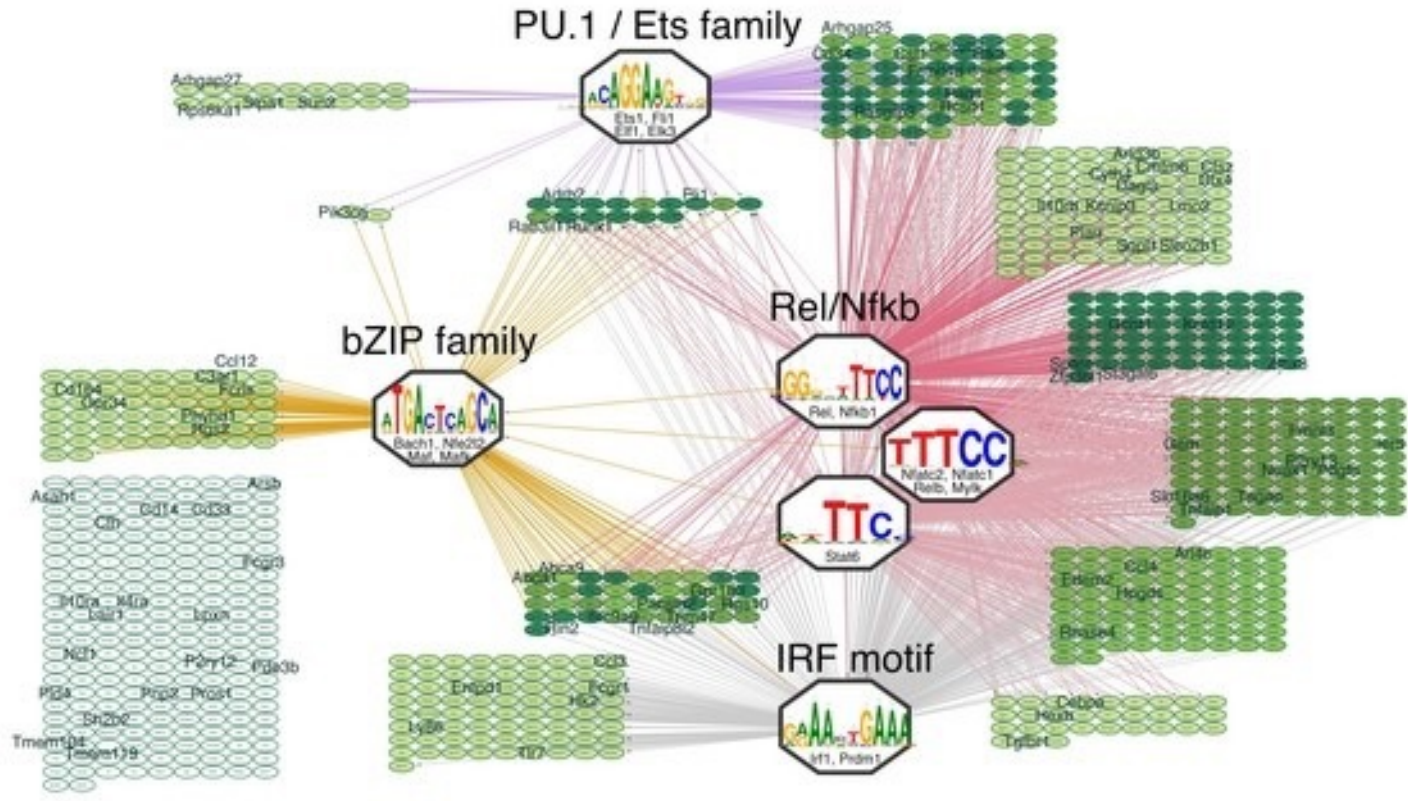




Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

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.



Aibar et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nature Methods. doi: 10.1038/nmeth.4463.

Evaluate GRN inference methods

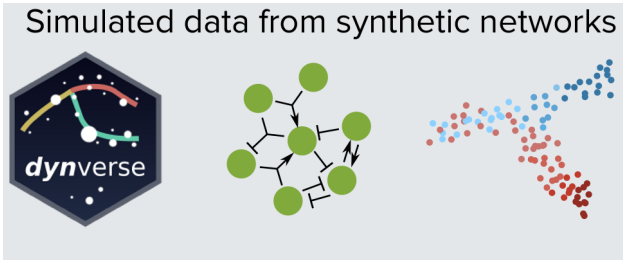
- Ground truth of GRNs is usually unknown.
- How do we evaluate the performance of existing GRN inference methods from scRNA-seq data?

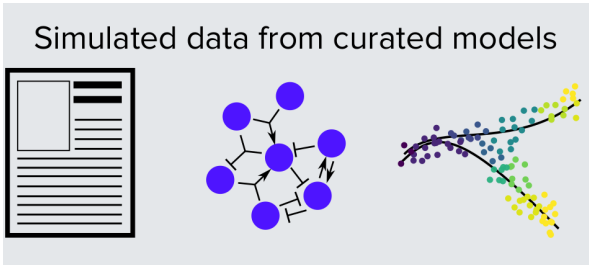
BEELINE

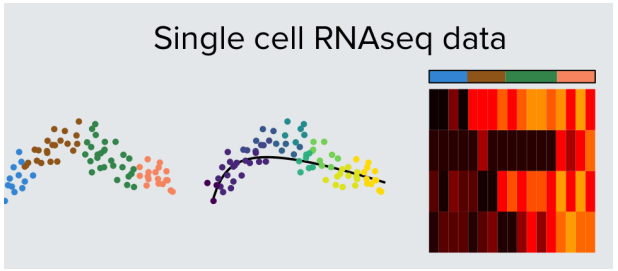
BEELINE

Benchmarking g**e**n**e** regu**l**atory network **i**nference from si**n**gle-c**e**ll 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.
- Step 1: preprocessing
  - Input type 1: Simulated data from synthetic networks
  - Input type 2: Simulated data from curated models
  - Input type 3: Experimental single-cell RNA-seq datasets
- Step 2: Run GRN inference algorithms
- Step 3: postprocessing and evaluation





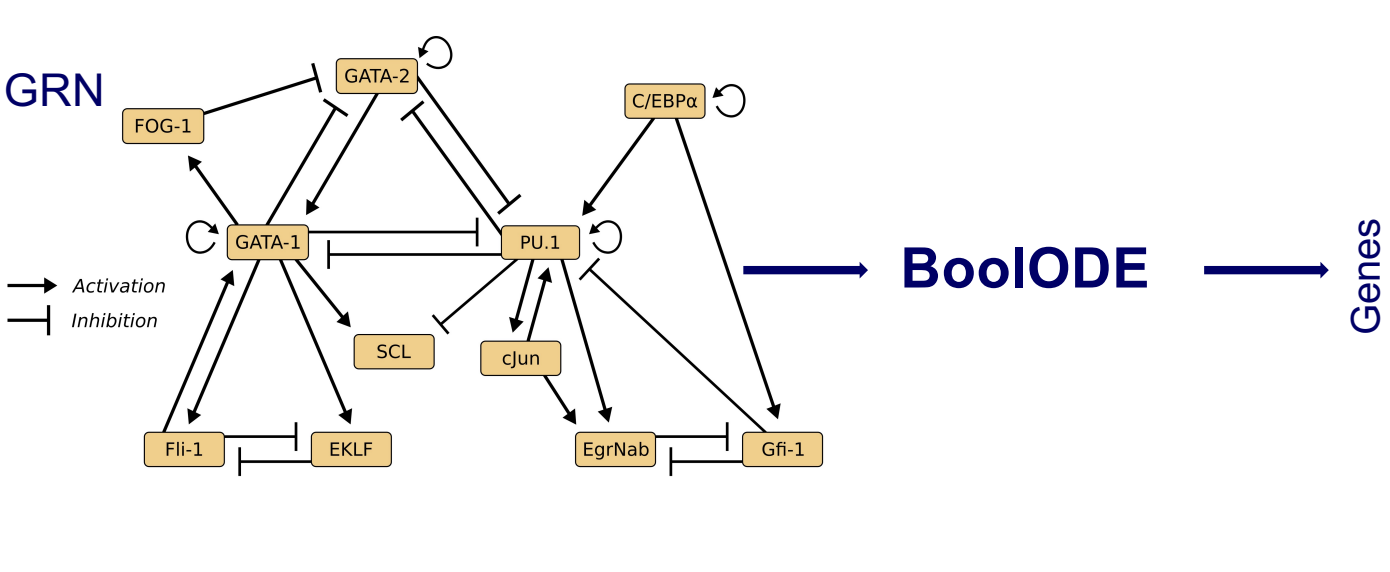


Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6

BoolODE for simulation

Generate inputs for BEELINE

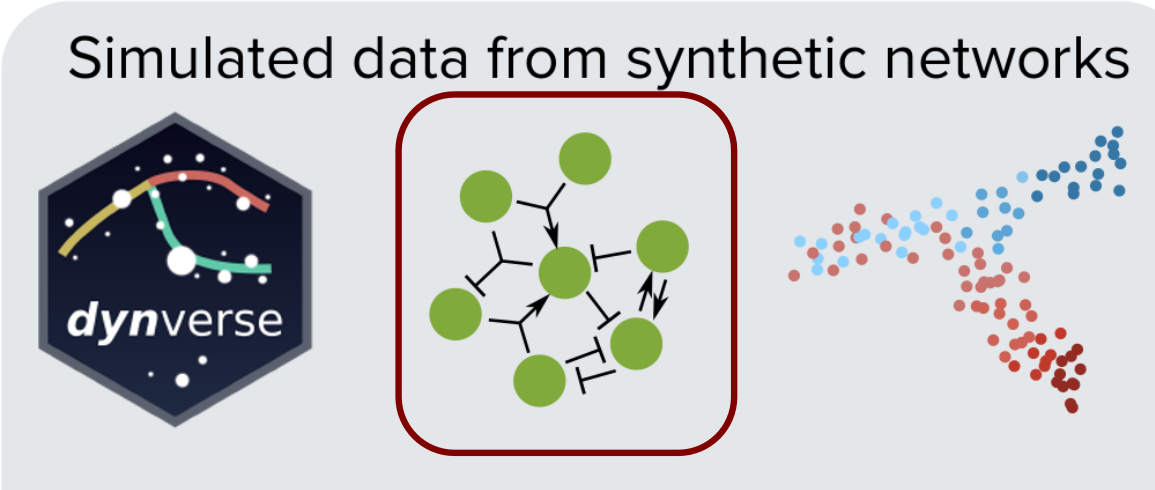
- Convert a Boolean model into a system of stochastic differential equations
- Read a model definition file, and outputs the simulated expression data for a given model.



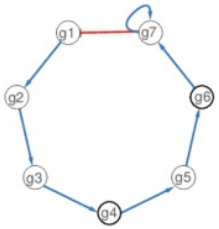
Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6

Input type 1:
Simulated datasets from synthetic networks

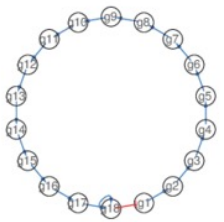
Synthetic network types:
- Linear
- Linear long
- Cycle
- Bifurcating
- Bifurcating converging
- Trifurcating



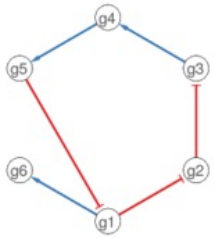
Linear



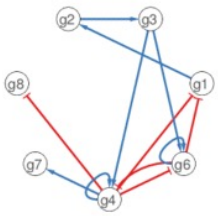
Linear long



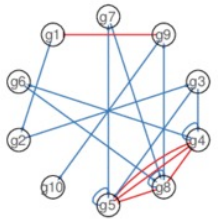
Cycle



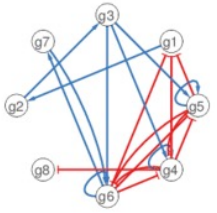
Bifurcating



Bifurcating Converging



Trifurcating

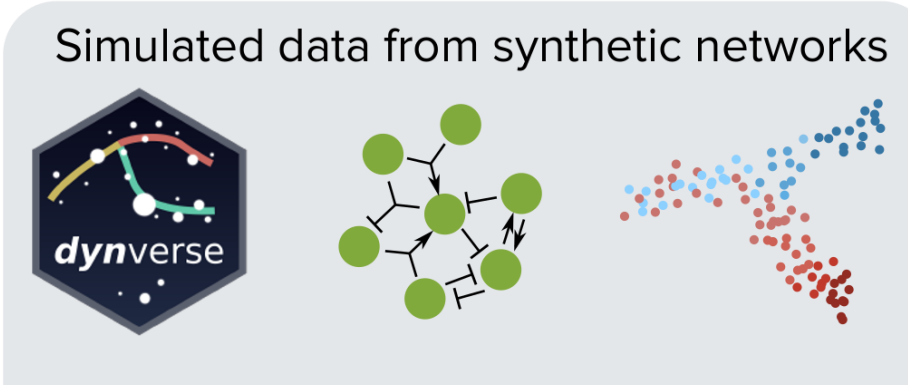


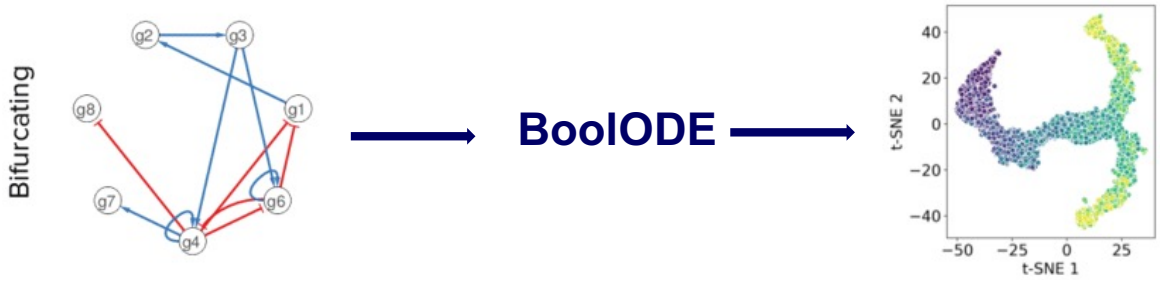
Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37,** 547–554 (2019). https://doi.org/10.1038/s41587-019-0071-9

Input type 1:
Simulated datasets from synthetic networks

Bifurcating trajectory

- The color of each ‘cell’ is determined by the timepoint at which it was sampled in the simulation
- Darker colors indicate earlier time points





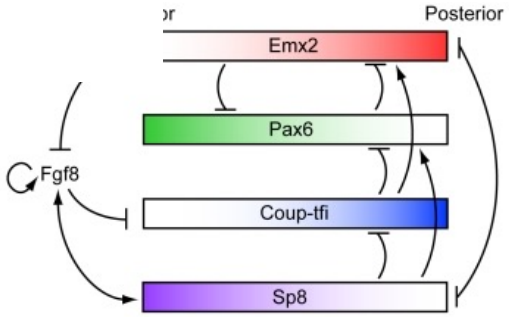
Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37,** 547–554 (2019). https://doi.org/10.1038/s41587-019-0071-9

Input type 2:
Simulated datasets from curated models (Boolean)

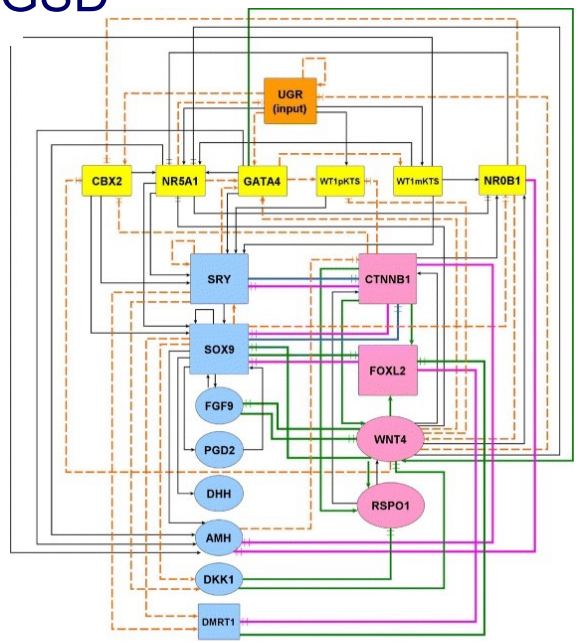
Four published Boolean models:

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

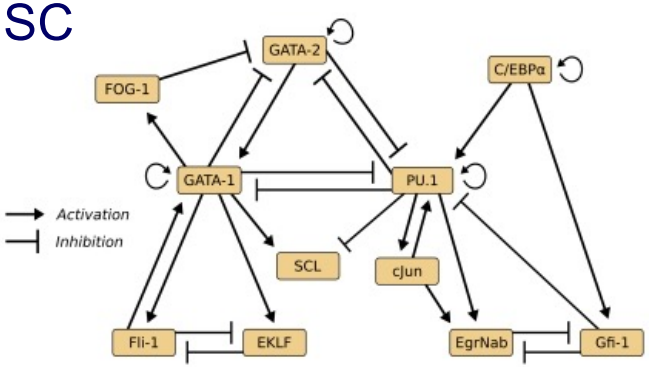
mCAD



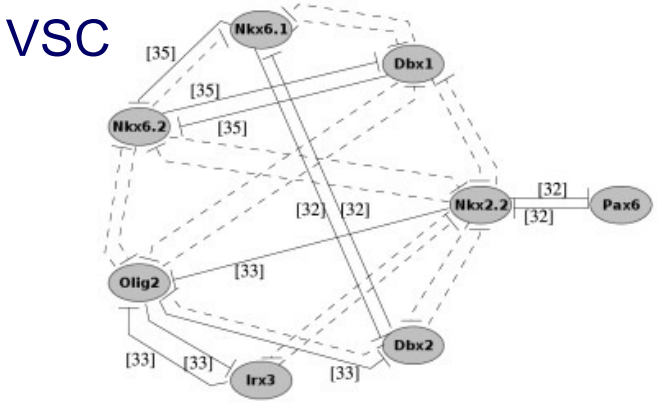
GSD



HSC



VSC

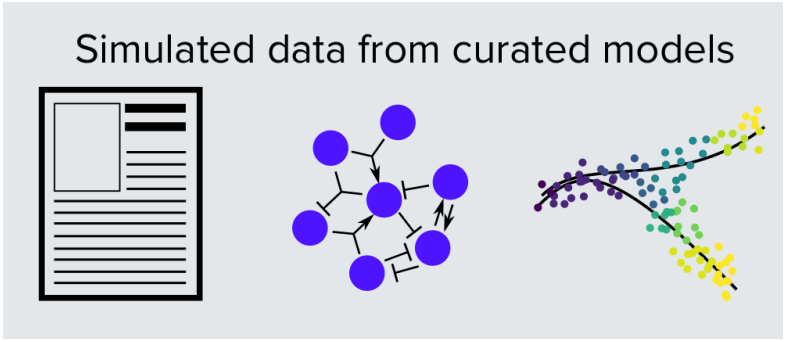


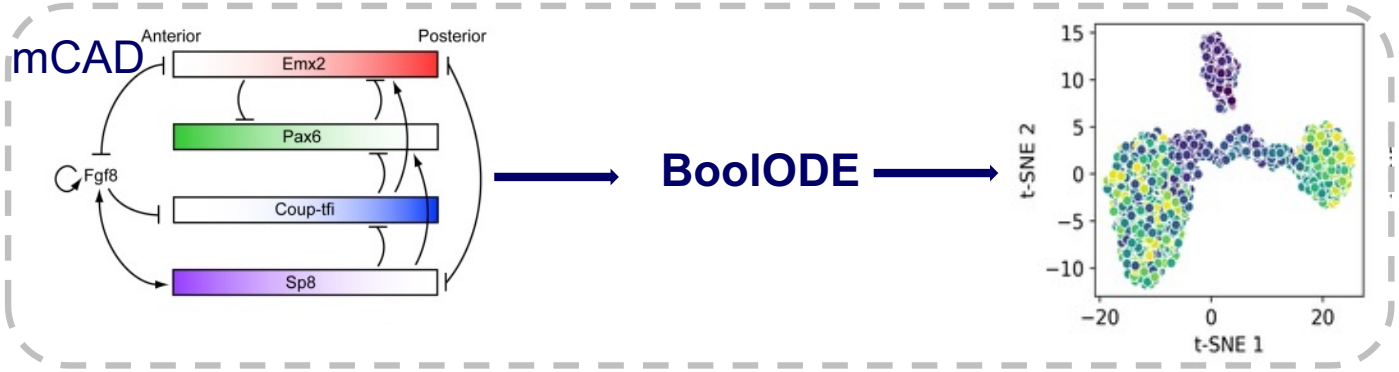
- Giacomantonio CE & Goodhill GJ A boolean model of the gene regulatory network underlying mammalian cortical area development. *PLoS Comput. Biol.* 6, e1000936 (2010).
- Lovrics A et al. Boolean Modelling reveals new regulatory connections between transcription factors orchestrating the development 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 Tactor Network. *PLoS One* 6, e22649 (2011).
- Ríos O et al. A Boolean network model of human gonadal sex determination. *Theor. Biol. Med. Model.* 12, 26 (2015).

Input type 2:
Simulated datasets from curated models (Boolean)

mCAD

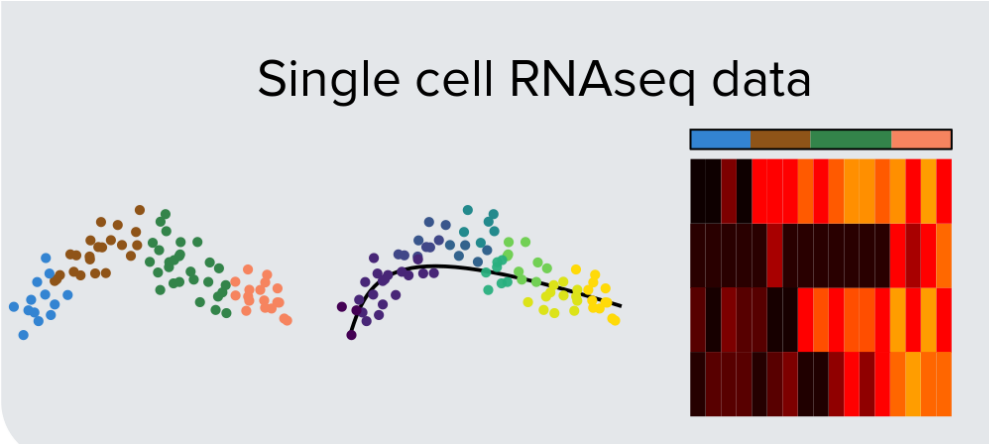
- Visualizations of t-SNE from the BoolODE output
- The color of each point indicates the corresponding simulation time.



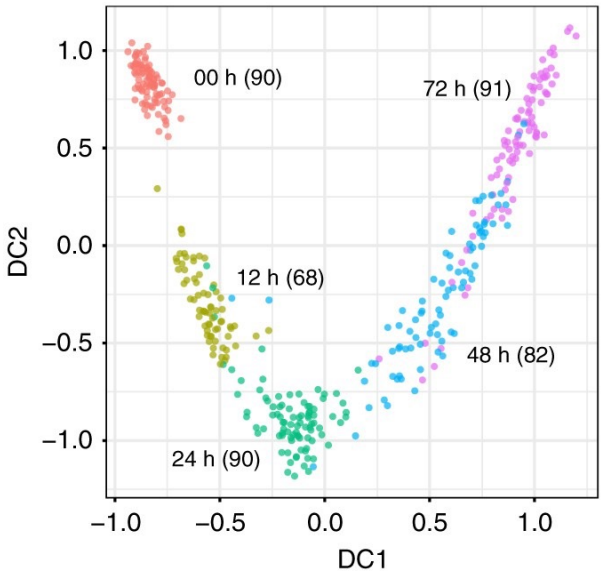


Giacomantonio CE & Goodhill GJ A boolean model of the gene regulatory network underlying mammalian cortical area development. *PLoS Comput. Biol.* 6, e1000936 (2010).
Lovrics A et al. Boolean Modelling reveals new regulatory connections between transcription factors orchestrating the development of the 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 Tactor Network. *PLoS One* 6, e22649 (2011).
Ríos O et al. A Boolean network model of human gonadal sex determination. *Theor. Biol. Med. Model.* 12, 26 (2015).

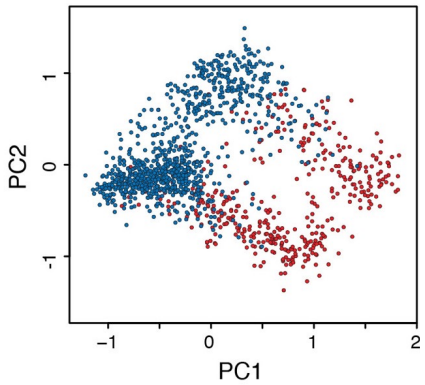
Input type 3:
Experimental scRNA-seq datasets (Mouse)



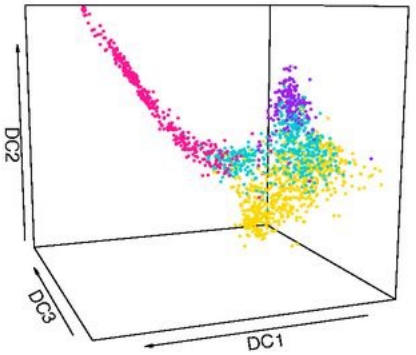
Embryonic stem cells



dendritic cells

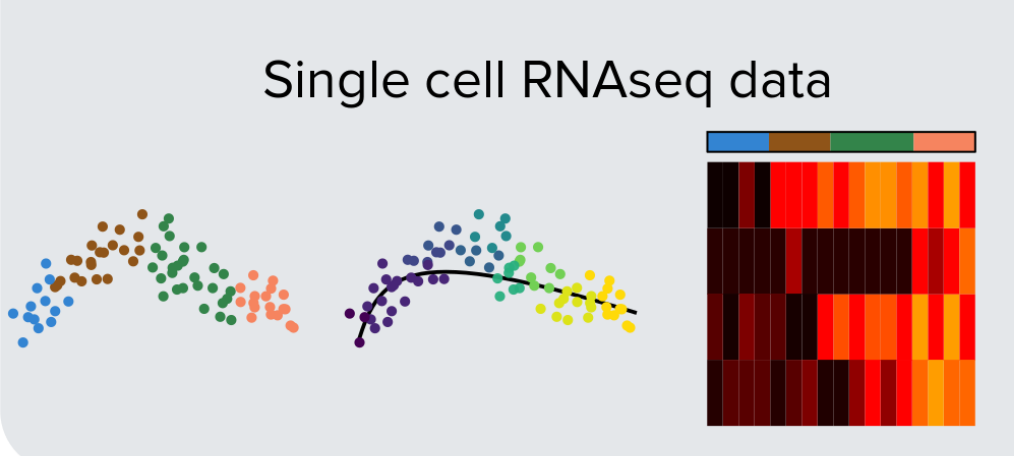


hematopoietic stem cells

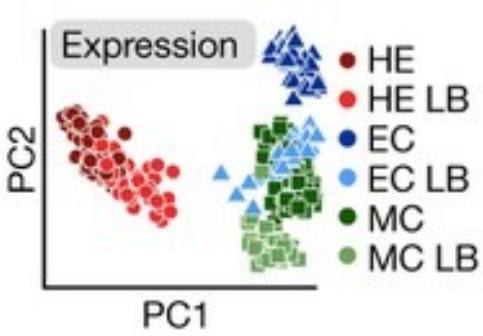


- hematopoietic stem and progenitor cell differentiation. Blood 128, 20–31 (2016).
- Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. 9, 619 (2018).
- Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine 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).

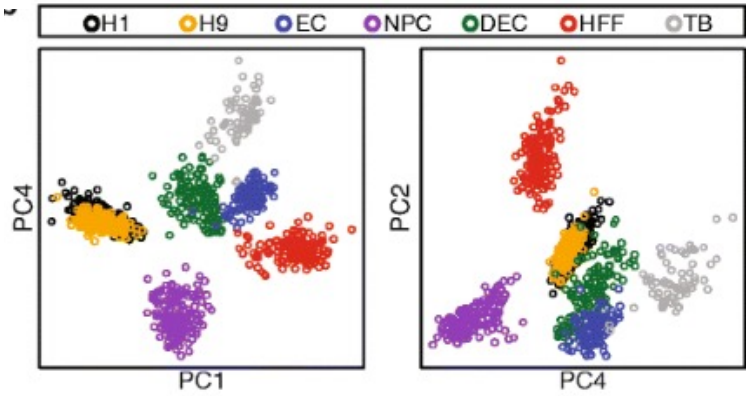
Input type 3:
Experimental scRNA-seq datasets (Human)



hepatocytes



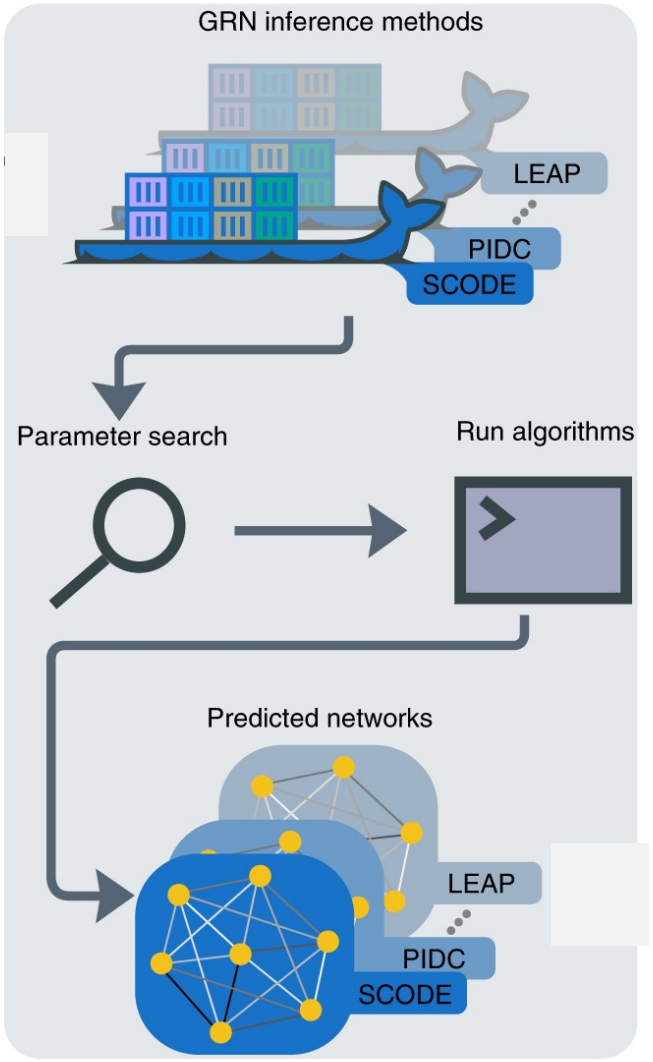
Embryonic stem cells

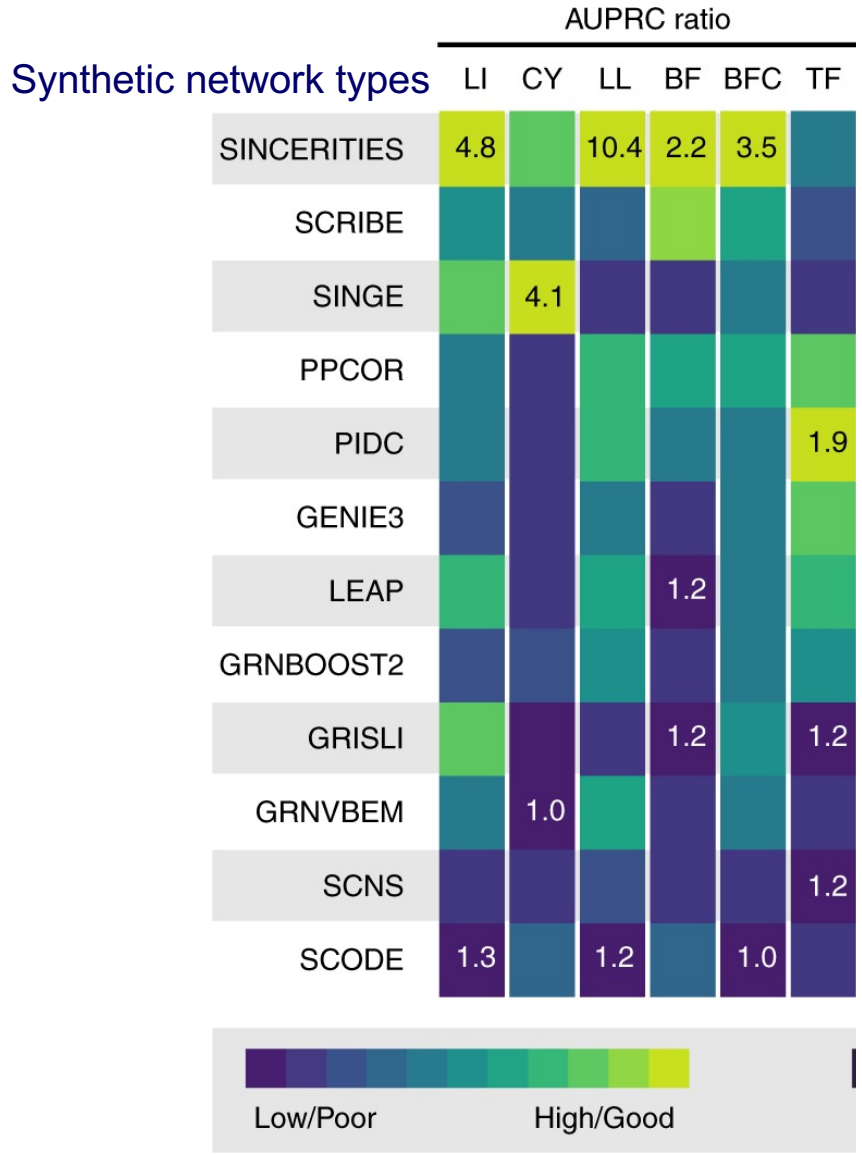


- Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128, 20–31 (2016).
- Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. 9, 619 (2018).
- Shalek, 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).

Run GRN inference algorithms

- Incorporation of 12 diverse GRN inference algorithms





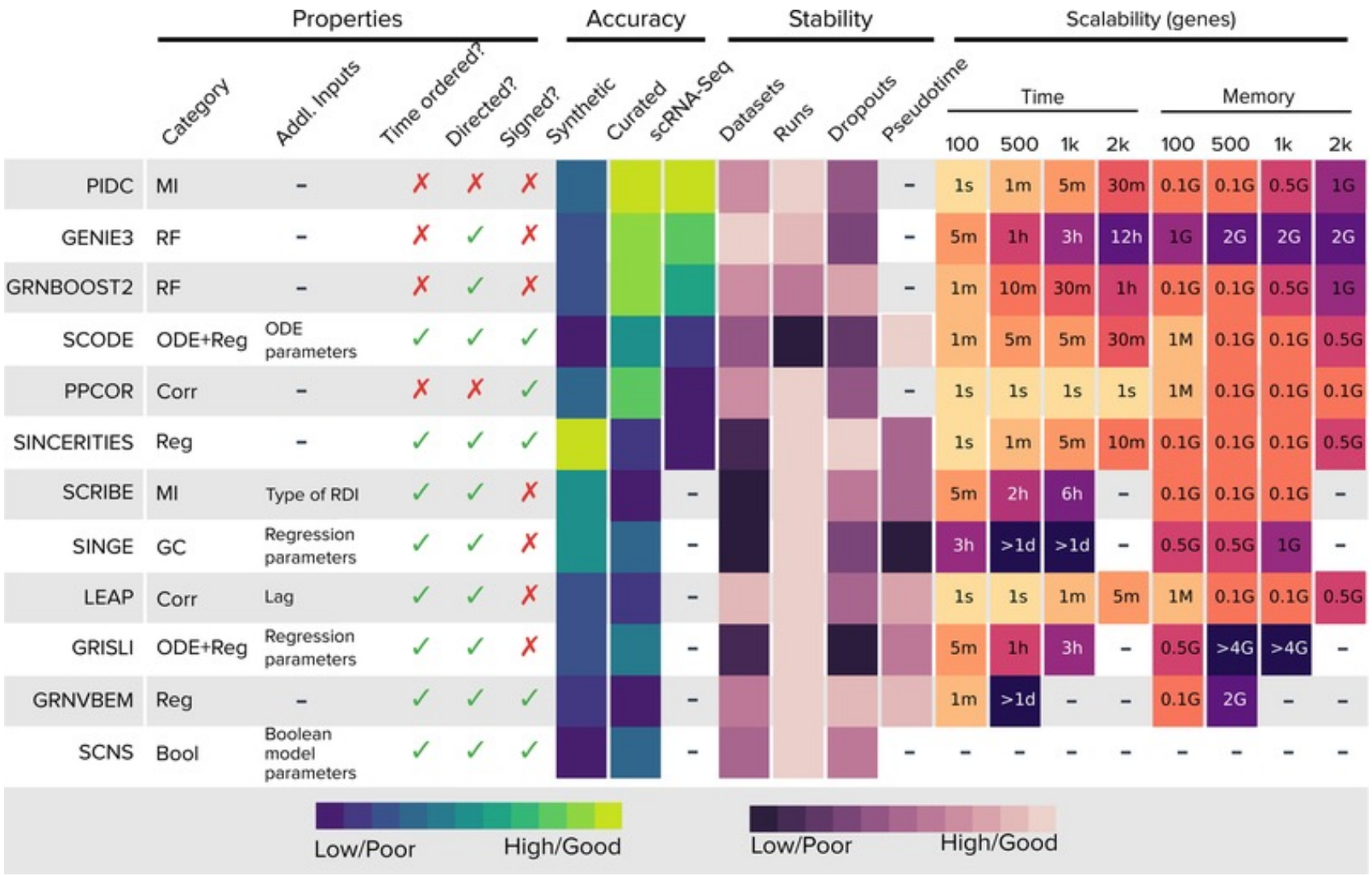
Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6

Evaluation

- Accuracy (AUPRC and early precision)
- Stability of results (across simulations, in the presence of dropouts and across algorithms)
- Analysis of network motifs
- Scalability

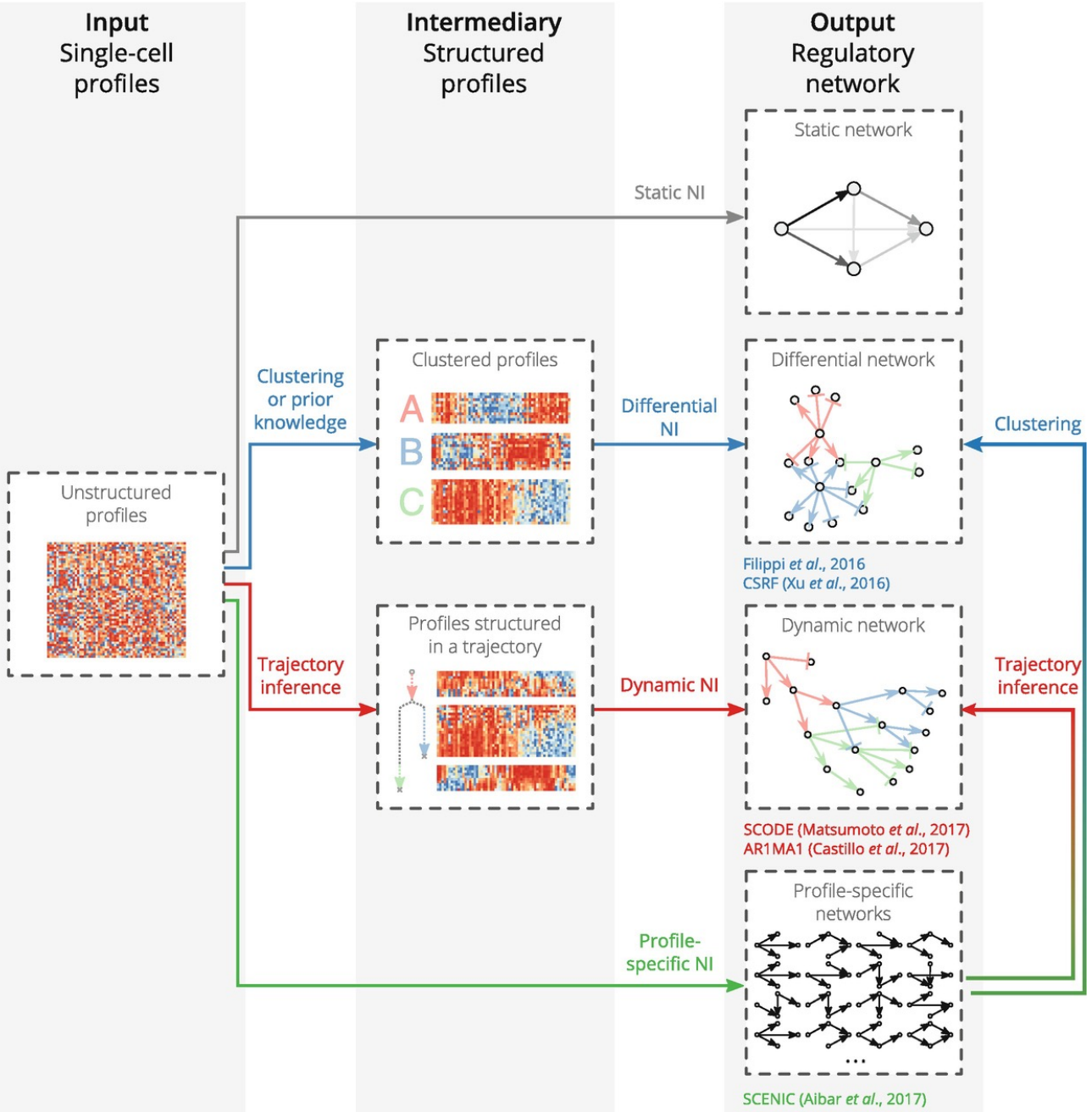
Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6

BEELINE summary



Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6

Network inference algorithms summary



Todorov H., Cannoodt R., Saelens W., Saeys Y. (2019) Network Inference from Single-Cell Transcriptomic Data. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-8882-2_10

Outline

- scRNA-seq data analysis
  - Cell type annotation
    - SingleR
  - Cell type markers identification
  - Pseudo timing
    - Monocle
  - Cell-type gene regulatory networks
    - SCENIC
    - BEELINE
  - **Single cell deconvolution**
    - **CIBERSORTx**

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

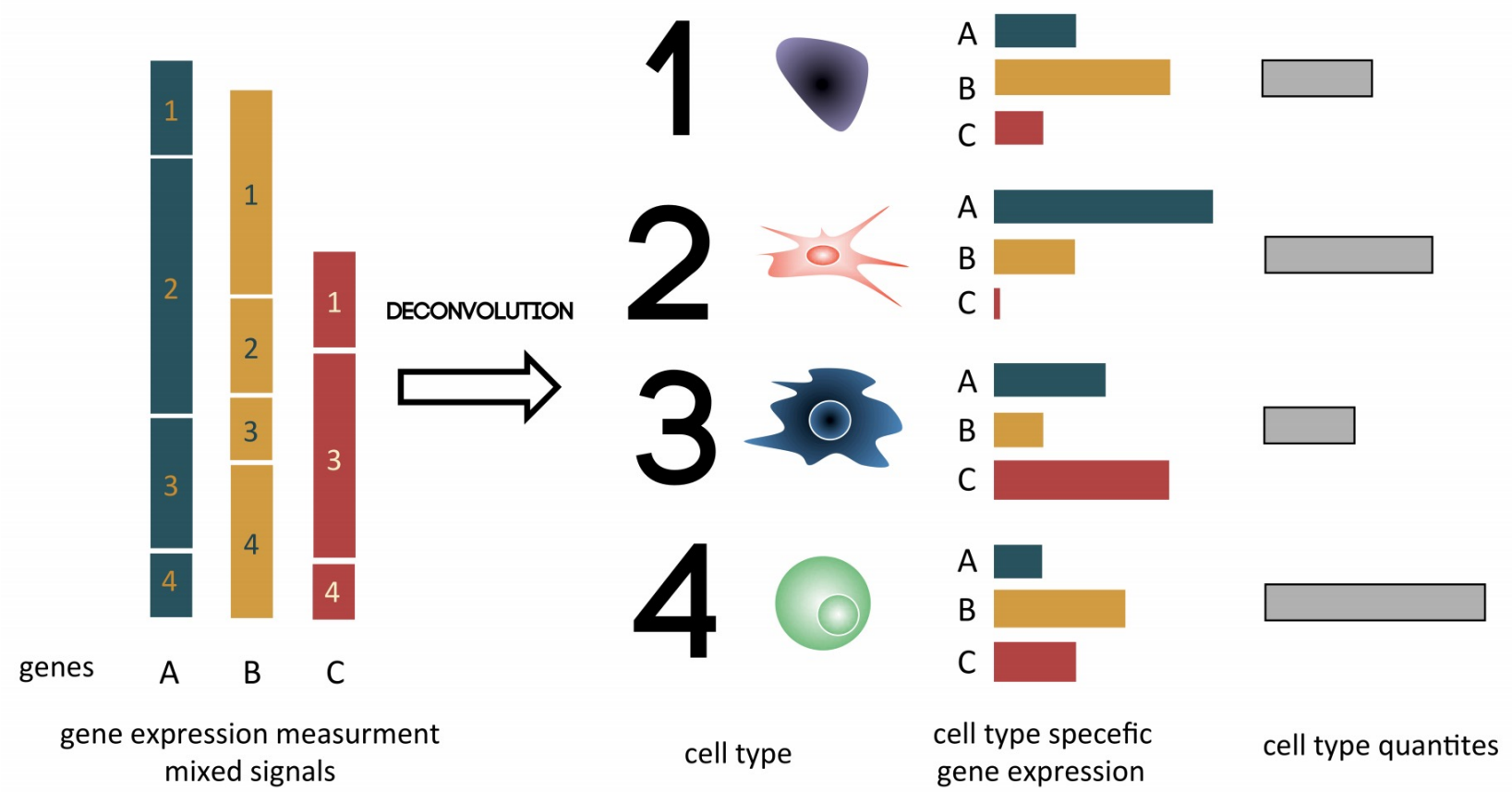
Single cell transcriptomic analyses retain single cell

https://projects.iq.harvard.edu/files/chanbioinformatics/files/cell_type_deconvolution.pdf

ASHG 2019 scRNAseq HiPlex oral presentation https://www.youtube.com/watch?v=YlRemO_TE3Y

Single cell deconvolution

- Qualify cell types in mixture

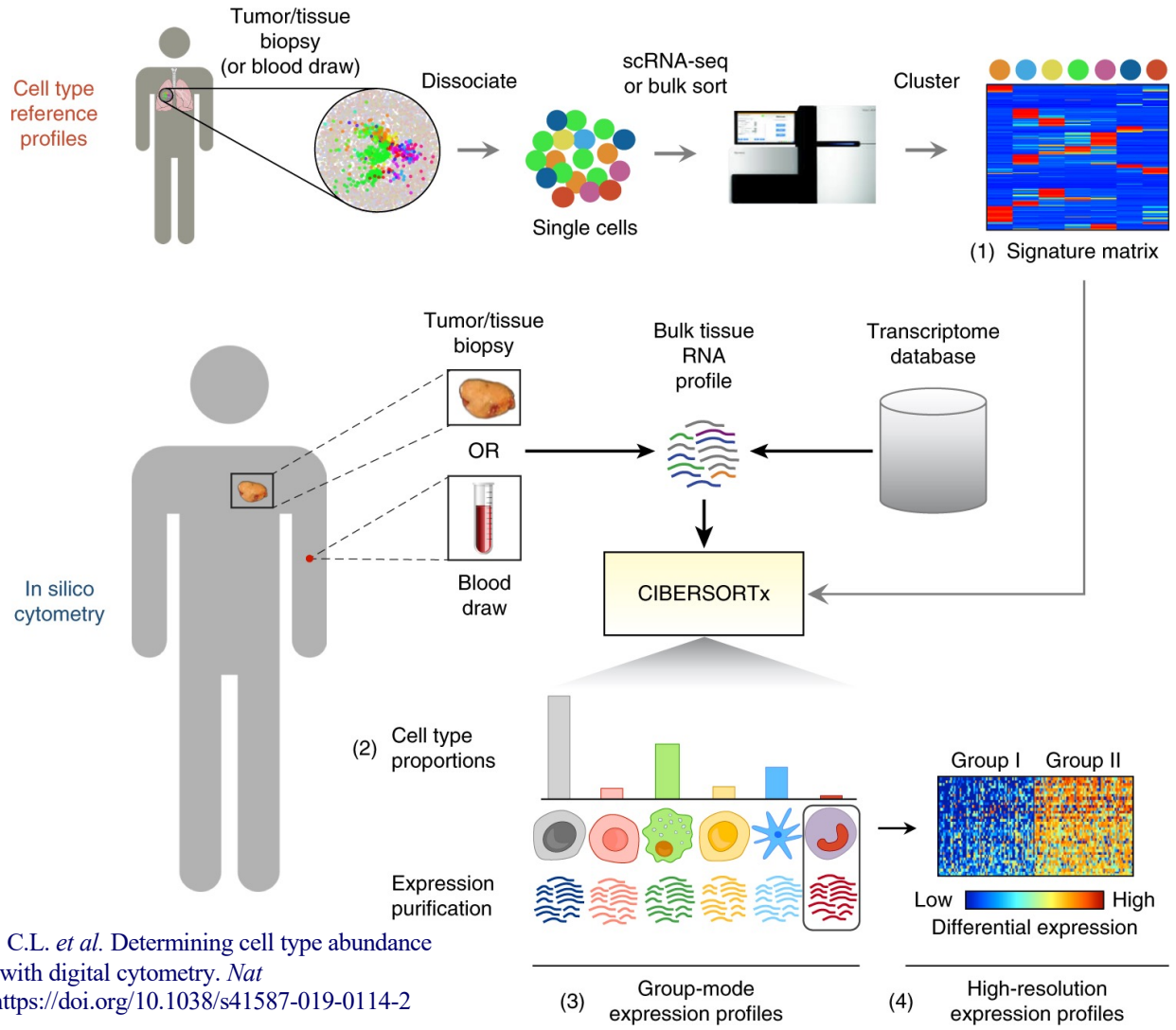


Finotello, F. & Trajanoski, Z. Quantifying tumorinfiltrating immune cells from transcriptomics data. Cancer Immunol. Immunother. 67, 1031–1040 (2018).

CIBERSORTx

cell-type **i**dentification **b**y **e**stimating **r**elative **s**ubsets **o**f **R**NA **t**ranscripts

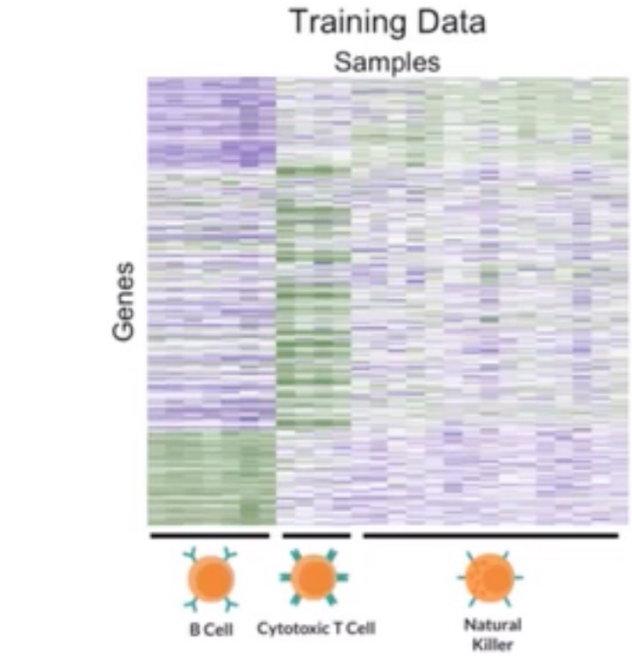
- Infer cell-type-specific gene expression profiles without physical cell isolation



Newman, A.M., Steen, C.B., Liu, C.L. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37,** 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2

CIBERSORTx

cell-type **i**dentification **b**y **e**stimating **r**elative **s**ubsets **o**f **R**NA **t**ranscripts



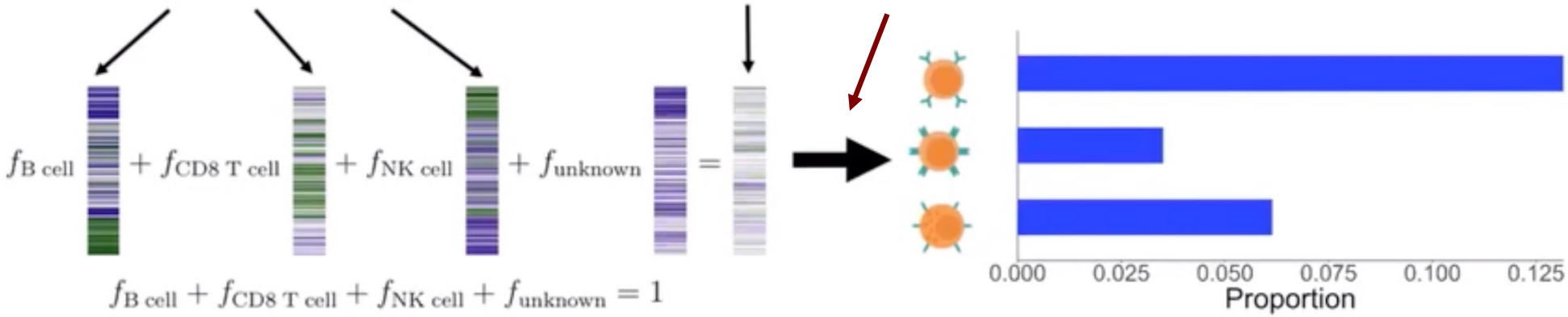
Unknown fractions of each cell type in the mixture

mRNA mixture

Gene expression profiles signature matrix

$$\mathrm{m} = f \times B$$

Support vector regression (SVR)

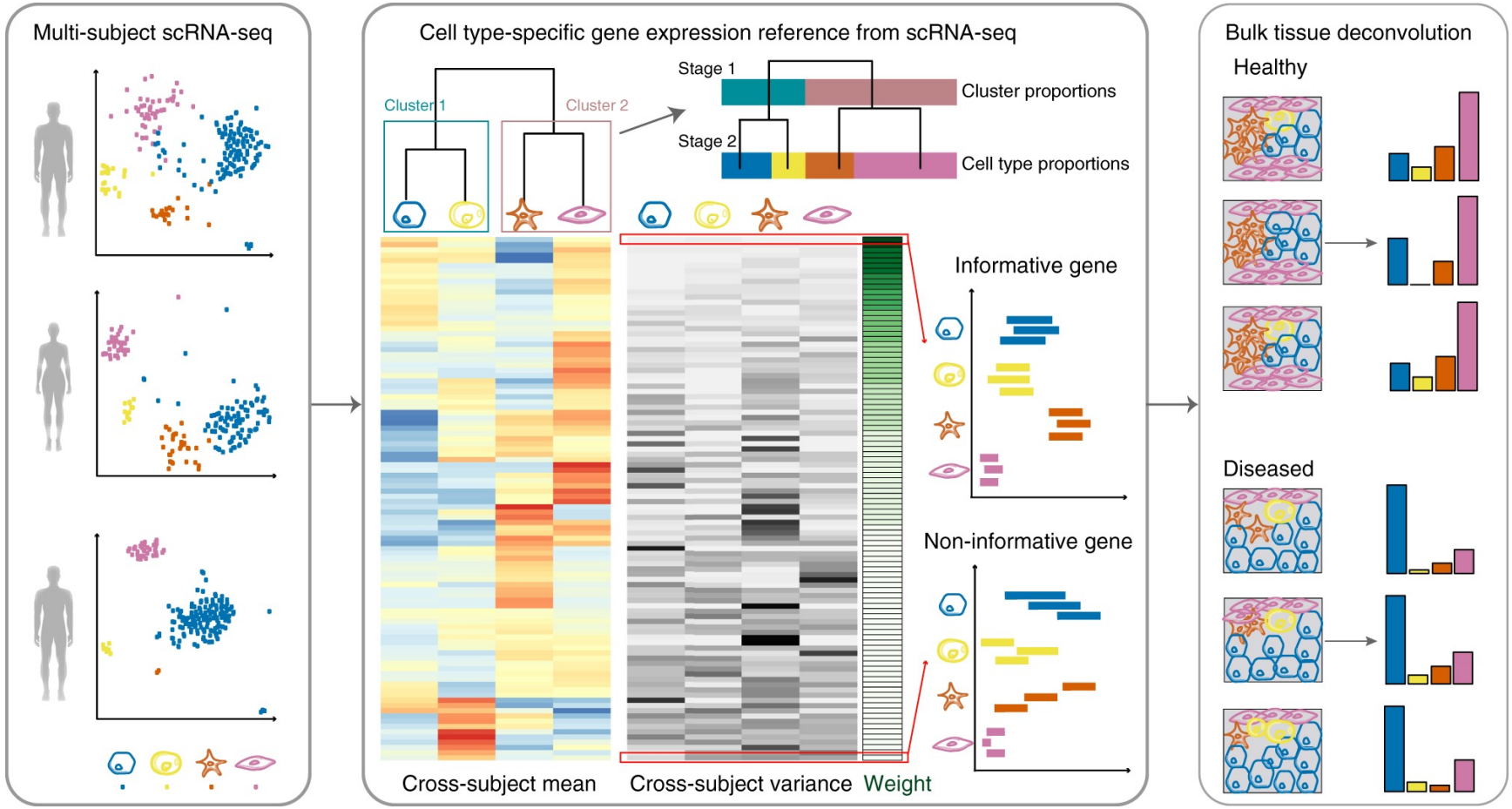


Newman, A.M., Steen, C.B., Liu, C.L. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37,** 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2

MuSiC

Multi-**S**ubject **Si**ngle **C**ell deconvolution

- A method for characterizing the cell type composition of large amounts of RNA sequencing data in complex tissues using cell type-specific gene expression in single-cell RNA sequencing (RNA-seq) data

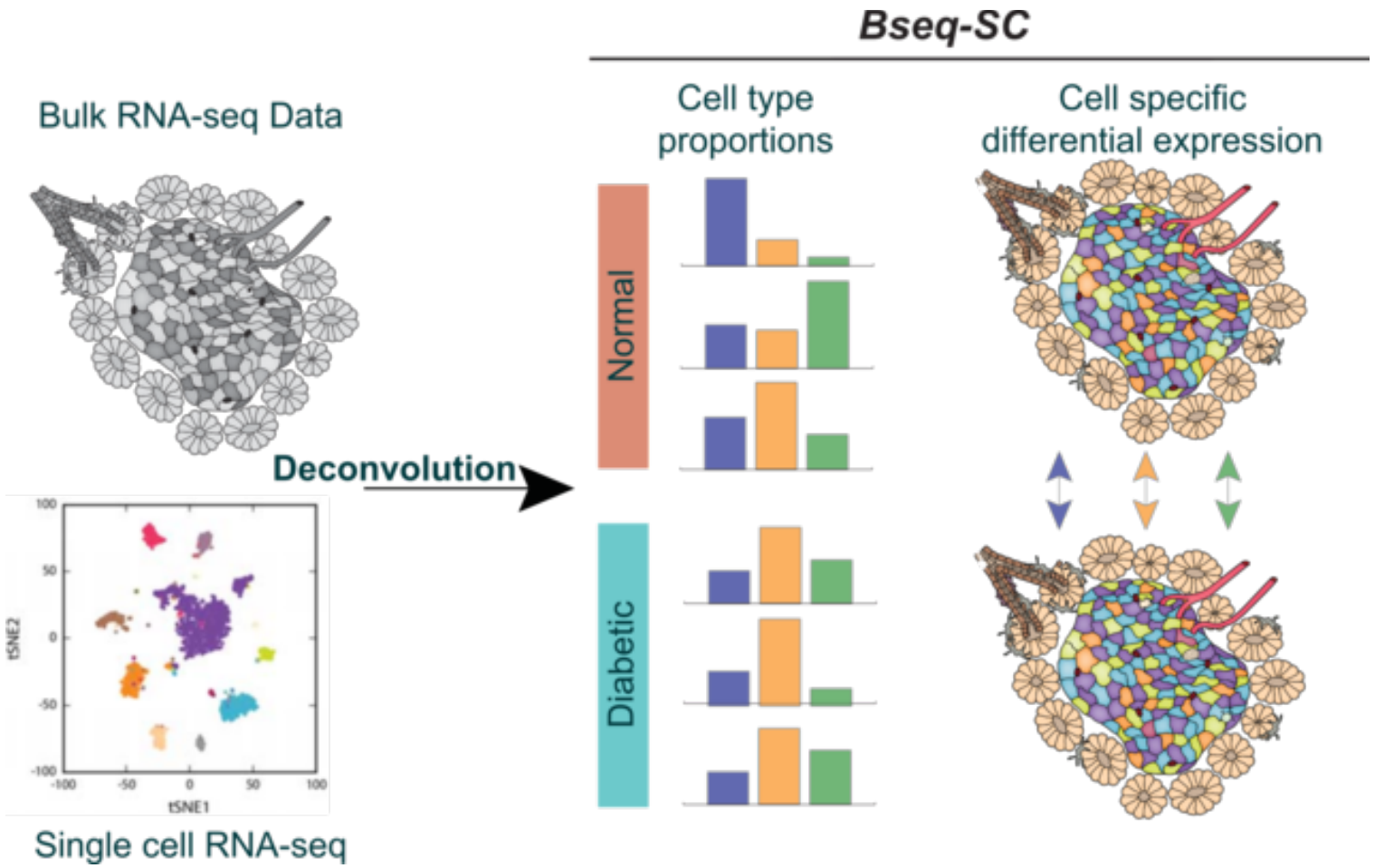


Wang, X., Park, J., Susztak, K. *et al.* Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10,** 380 (2019). https://doi.org/10.1038/s41467-018-08023-x

BSEQ-sc

Deconvolution of Bulk Sequencing Experiments using Single Cell Data

- Leverage single-cell sequencing data to estimate cell type proportion and cell type-specific gene expression differences from RNA-seq data from bulk tissue samples

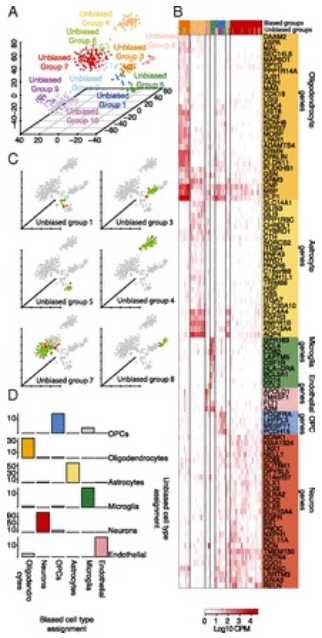


A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure M. Baron, A. Veres, S.L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. Hyoje Ryu, B. K. Wagner, S. Shen-Orr, A. M. Klein, D. A. Melton, I. Yanai Cell Systems. 2016 Oct 26 10.1016/j.cels.2016.08.011

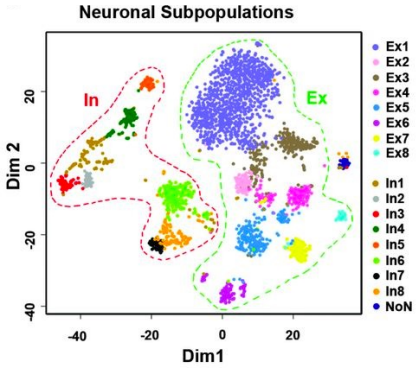
Single cell RNA-seq data for human brain

- 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)



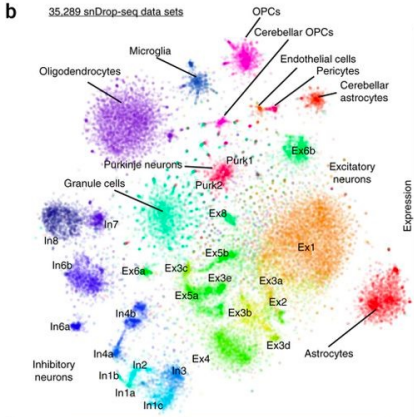
~400 cells (Darmanis et al., PNAS, 2015)



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

~ 900 cells (PsychENCODE)

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



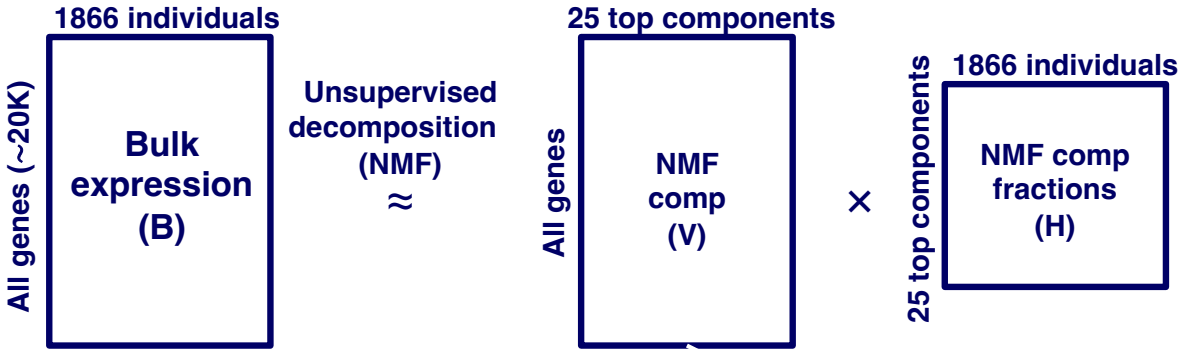
~10319 cells (Lake et al., Nature Biotech, 2018)

~ 17,093 cells (PsychENCODE)

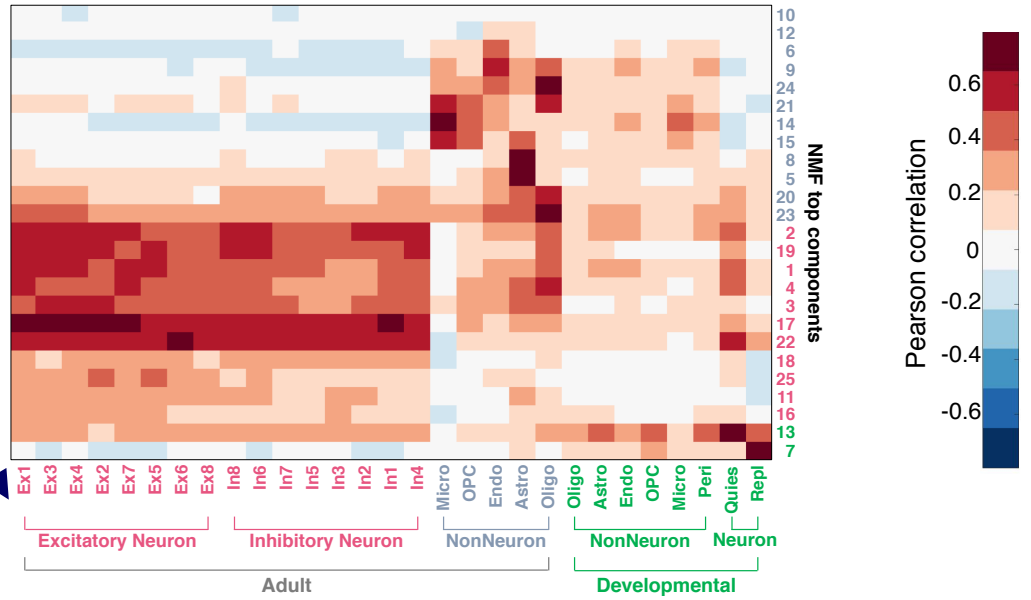
Wang, et al., *Science*, 2018

Single cell deconvolution
Step 1: unsupervised learning to see brain cell types

Non-negative matrix factorization (NMF)

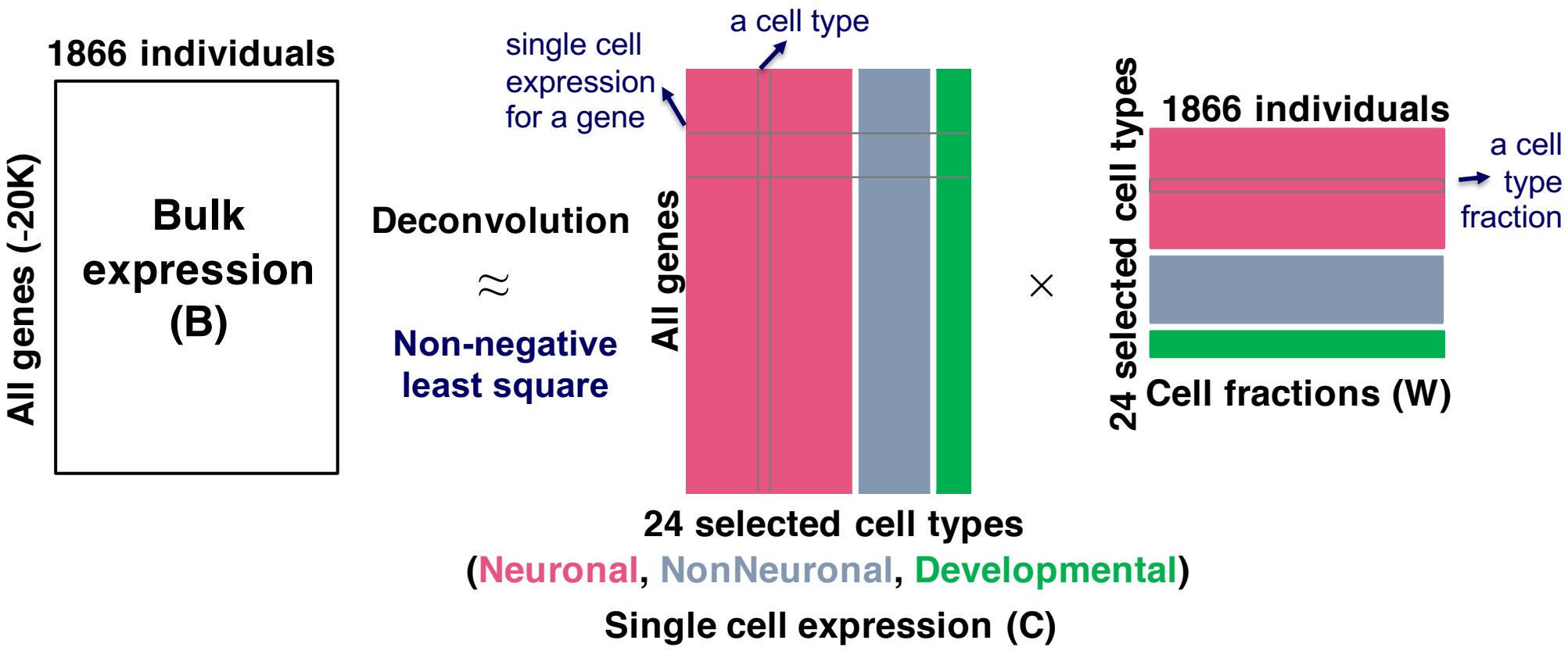


Single cell signatures
- ~14,000 cells (Lake et al., Science, 2016&2018)
- ~400 cells (Darmanis et al., PNAS, 2015)
- ~18,000 cells (PsychENCODE)



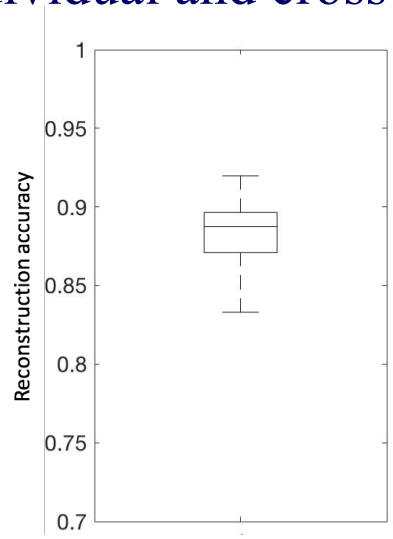
Wang, et al., *Science*, 2018

Single cell deconvolution
Step 2: supervised learning to estimate cell fractions



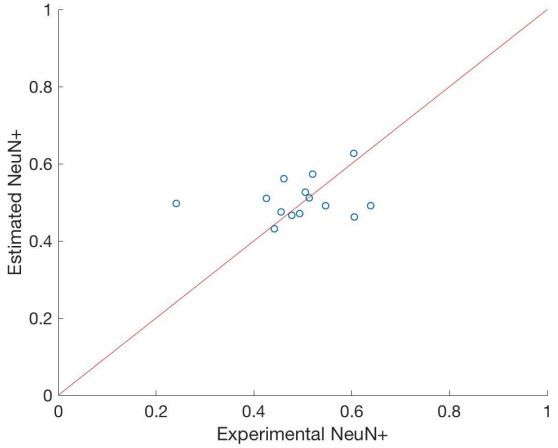
Cell fractions explain cross-population variation in human brain

Individual and cross-population reconstruction accuracy via deconvolution

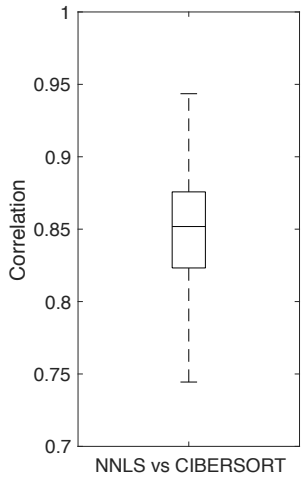


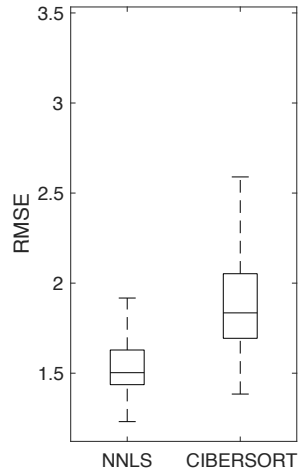
$$1 - \frac{\left\| B - CW \right\|^2}{\left\| B \right\|^2} > 85\%$$

Experimental validation



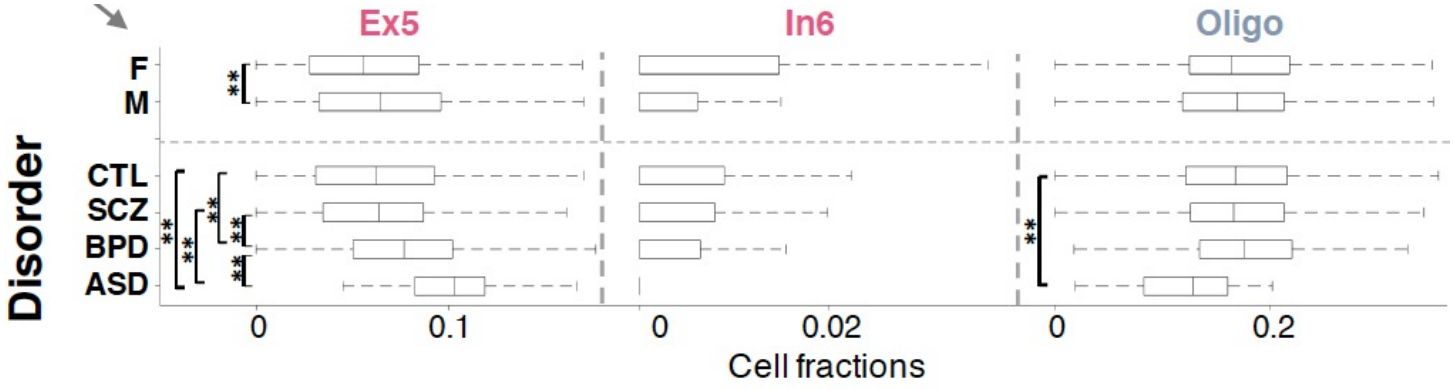
Comparison with existing methods





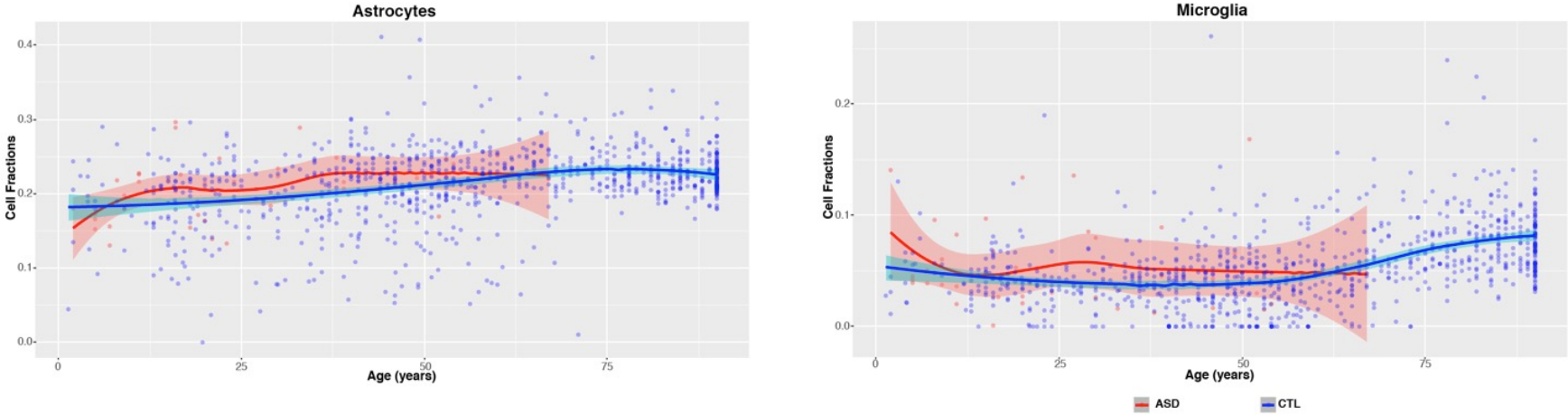
Wang, et al., *Science*, 2018

Neuronal and glial cell fraction changes in gender and disorders



Excitatory to Inhibitory imbalance at neuronal subtype level for ASD*

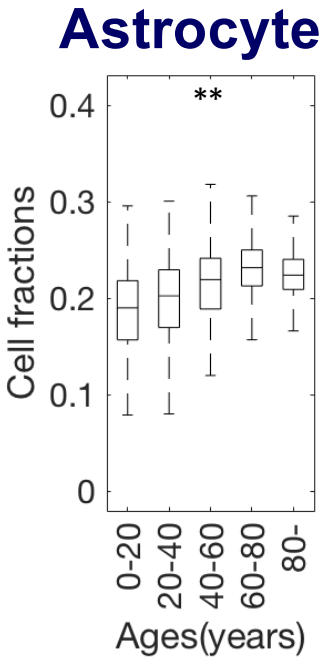
Astrocyte and Microglia increase in ASD**

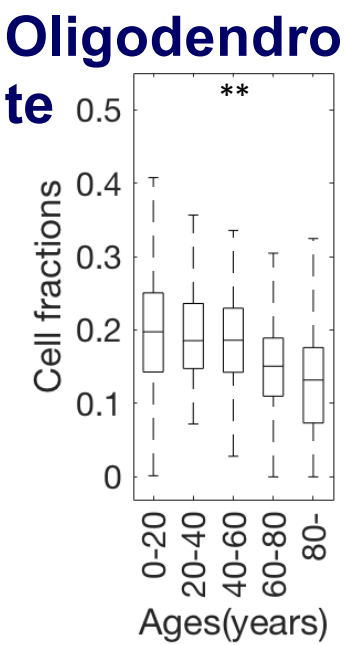


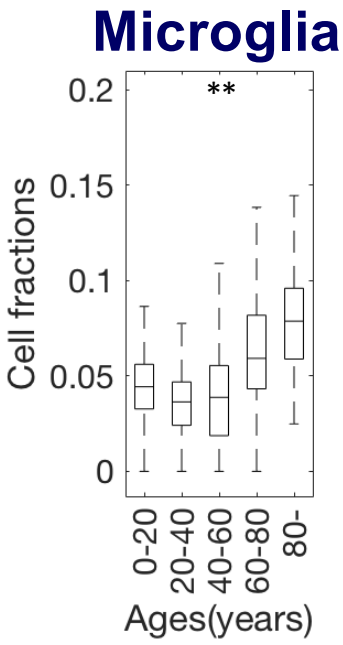
* Rubenstein et al., Model of autism: increased ratio of excitation/inhibition in key neural systems, Genes Brain Behav. 2003
** Gandal et al., Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap, Science 2018

Wang, et al., *Science*, 2018

Cell-type fraction changes in Age

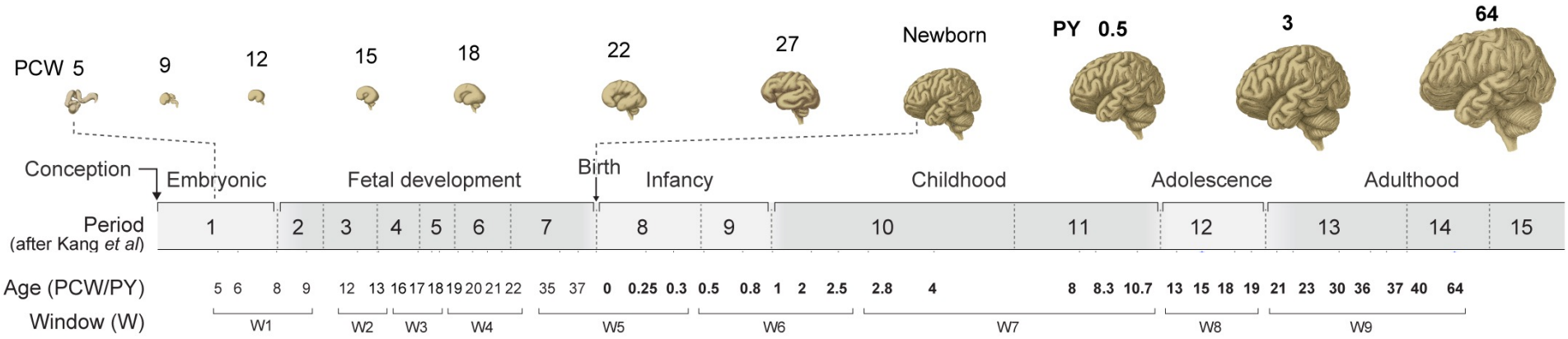


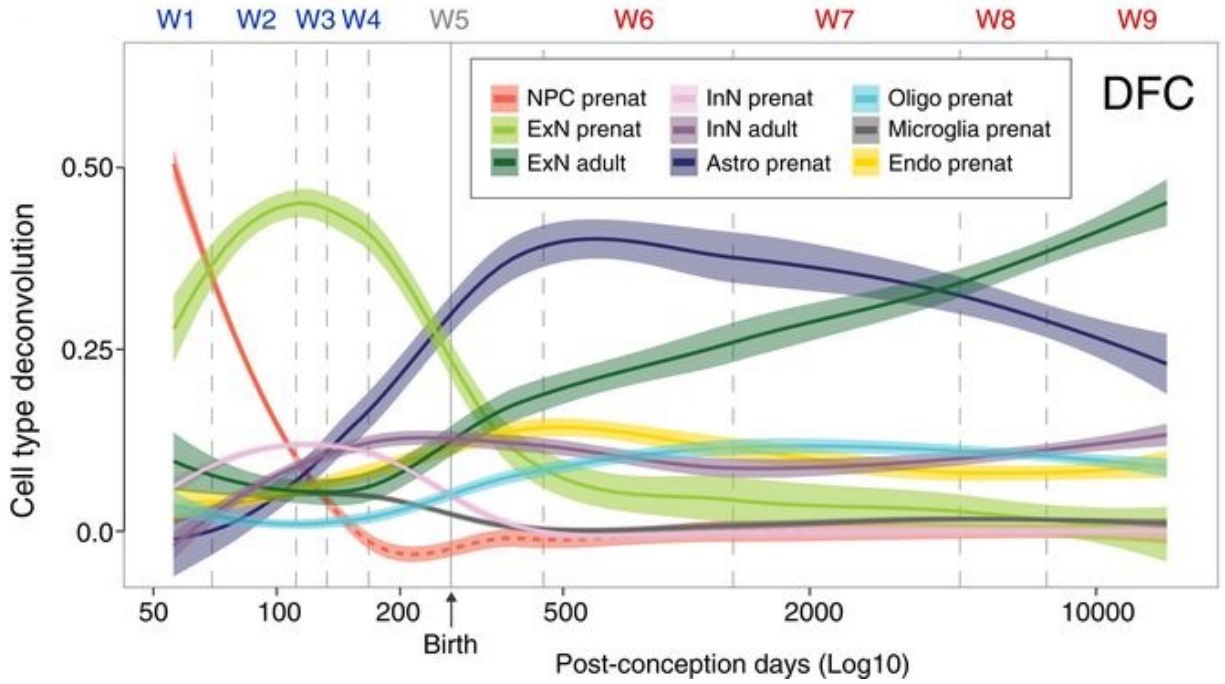




Wang, et al., *Science*, 2018

Cell-type fraction changes in human brain development





Li, …, **Wang**, ..., Sestan, *Science*, 2018

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