Analysis of single-cell RNA-seq data (III)

Hao Wu
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

Ziyi Li
Department of Biostatistics
The University of Texas MD
Anderson Cancer Center

ENAR 2021 short course
March 2021

Course outline

- 8-9:15: Intro and data preprocessing.
- 9:15-9:45: Lab: preprocessing and visualization.
- 10-11:15: Normalization, batch effect, imputation, DE, simulator.
- 11:15-12: Lab: Normalization, batch effect, imputation, DE, simulator

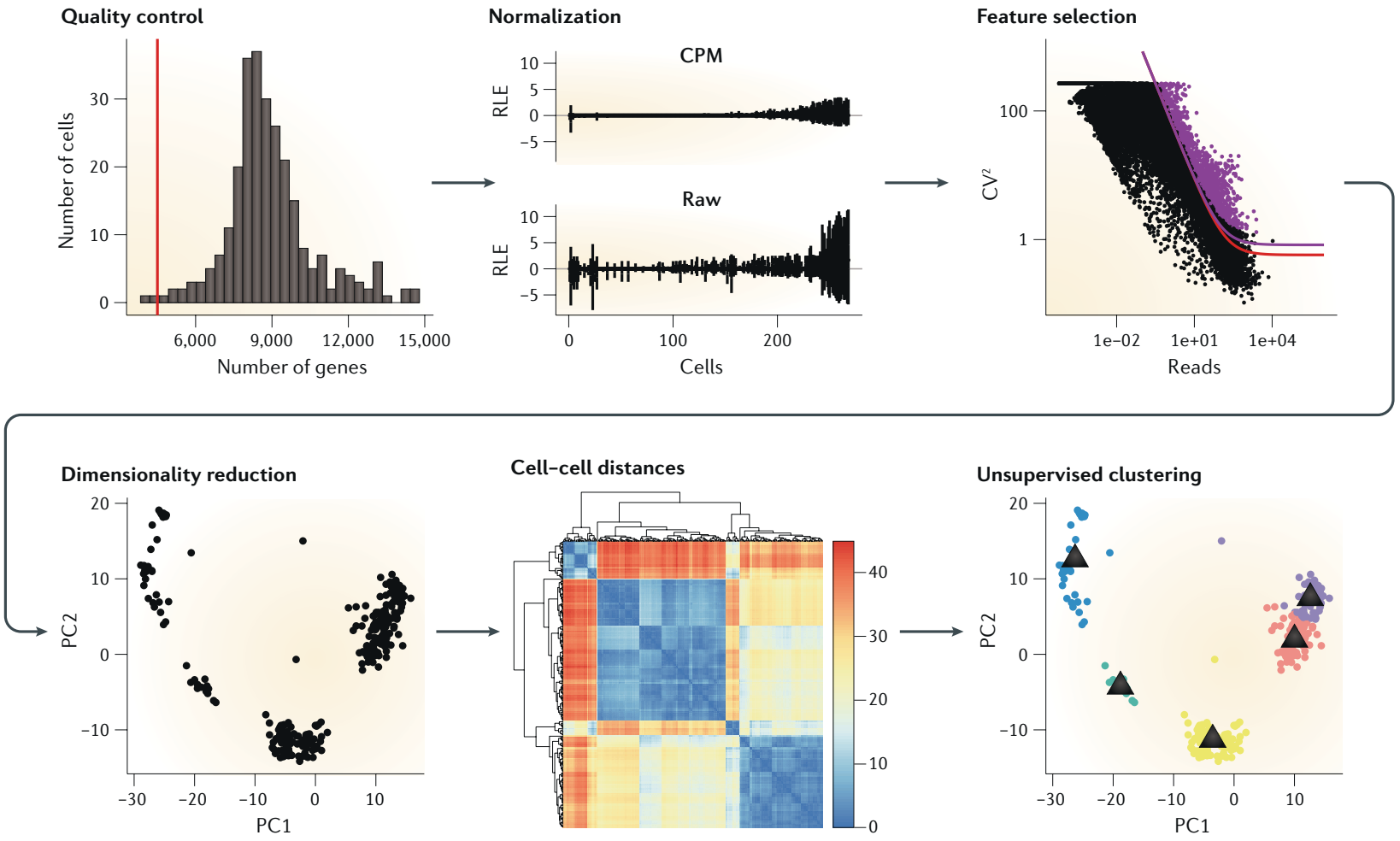
- 12-1: Lunch break

- **1-2: Clustering and pseudotime construction**
- 2-2:30: Lab: Clustering and pseudotime construction
- 2:45–3:30: Supervised cell typing & related single cell data sources
- 3:30-4: Lab: supervised cell typing.
- 4:15-5: scRNA-seq in cancer

Outline for this session

- **Background**
  - Scientific motivation
  - Assumptions and challenges
- **Clustering**
  - Existing methods
  - Performance comparisons and considerations
- **Pseudotime construction**
  - Existing methods
  - Pros and Cons
- **Future directions**

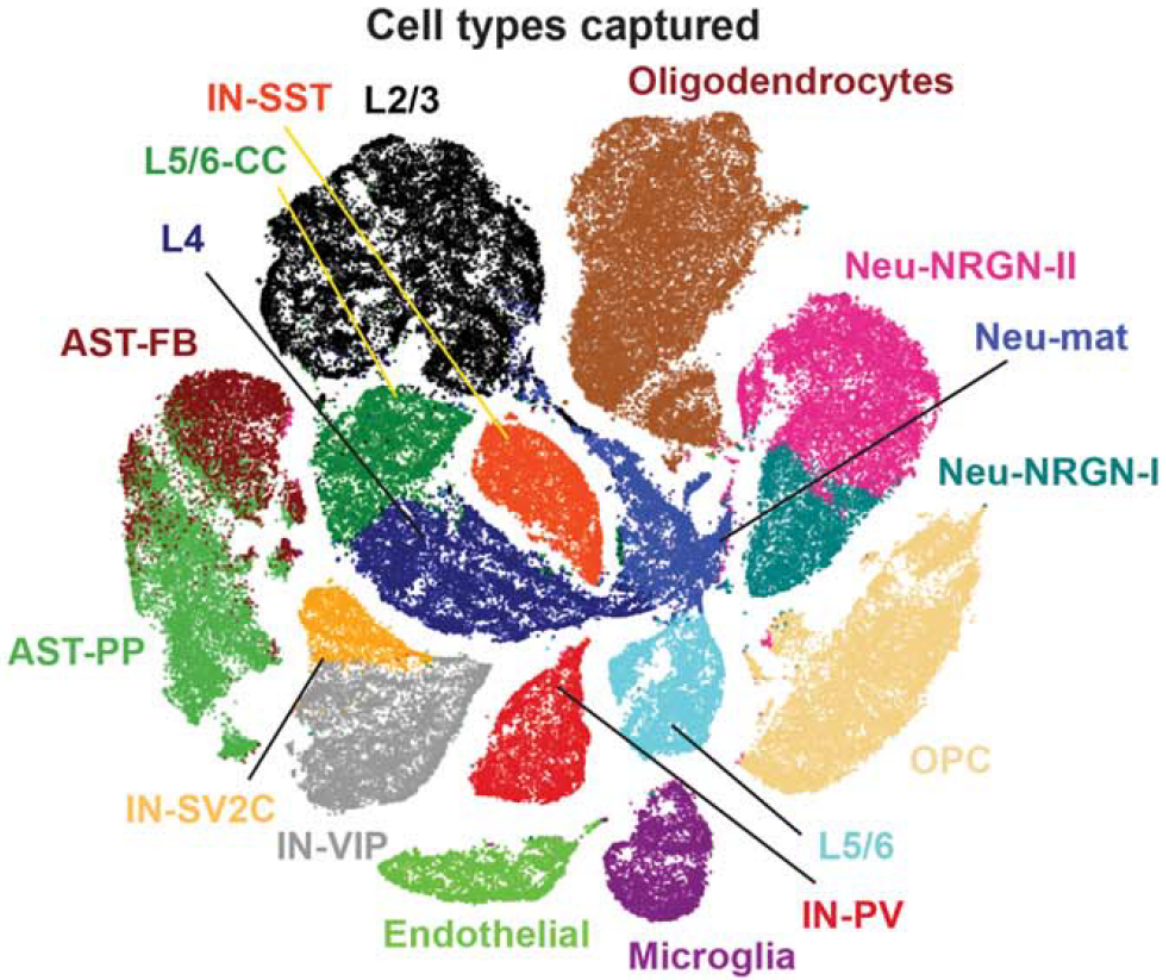
Example scRNA-seq analysis workflow



Kiselev et al. (2019) Nat. Rev. Genet.

Scientific motivations

- Subpopulation (cell type) identification is a fundamental step for many scRNA-seq data analyses
- Consistent and rigorous definition of cell type is elusive:
  - Early days, physical appearance, e.g. size, shape
  - Later, presence or absence of surface proteins
  - scRNAseq: define cell type based on transcriptome similarities
- Goal: discover the natural groupings of measured cells, discrete or continuous



AST-FB	Fibrous astrocytes
AST-PP	Protoplasmic astrocytes
OPC	Oligodendrocyte precursor cells
IN-PV	Parvalbumin interneurons
IN-SST	Somatostatin interneurons
IN-SV2C	SV2C interneurons
IN-VIP	VIP interneurons

L2/3	Layer 2/3 excitatory neurons
L4	Layer 4 excitatory neurons
L5/6	Layer 5/6 corticofugal projection neurons
L5/6-CC	Layer 5/6 cortico-cortical projection neurons
Neu-mat	Maturing neurons
Neu-NRGN-I	NRGN-expressing neurons
Neu-NRGN-II	NRGN-expressing neurons

Velmeshev et al., (2019) Science

Assumptions

- Clustering: discrete groups of cells present in the data.
- If assumption not hold, clustering methods still partition the data, and thus mistake random noise for true structure (!)
- Pseudotime construction:
  - place cells on a continuum connecting two or more end states
  - useful for understanding development or disease progression
- Strategies bridging the two approaches: soft or fuzzy clustering
- When assumptions are not clear, explore both

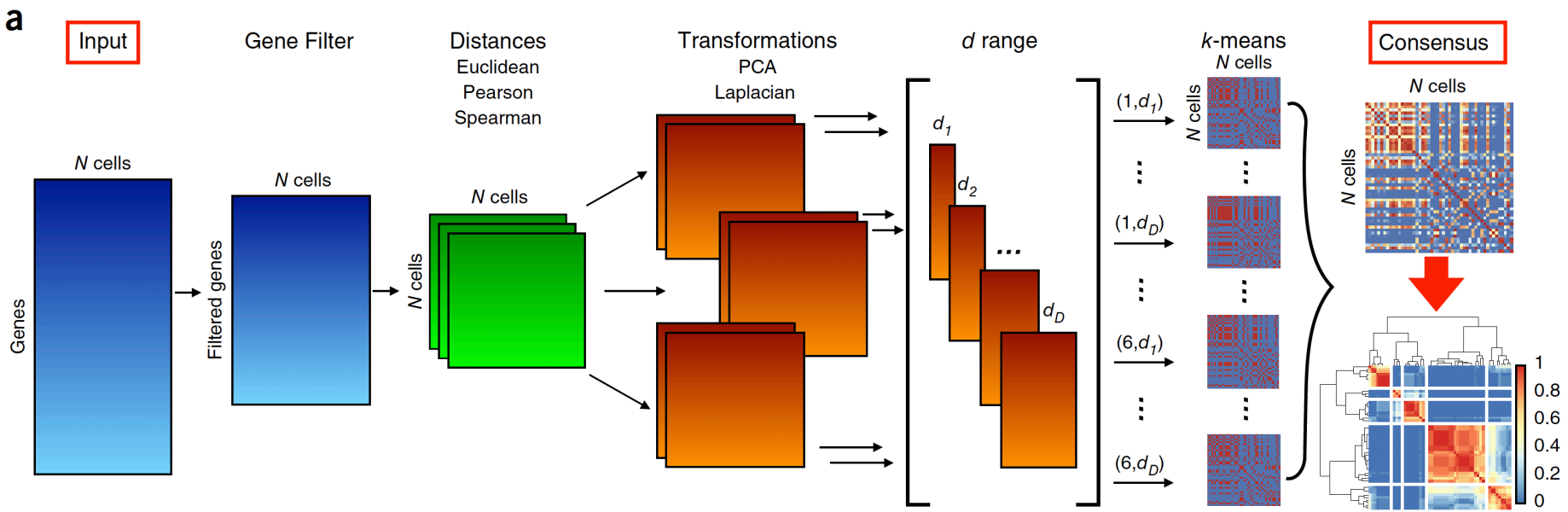
Existing clustering methods

- **K-means based**
  - RaceID (Grun et al. 2015 Nature)
  - SC3 (Kiselev et al. 2017 Nat methods)
  - SIMLR (Wang et al. 2017 Nat Methods)
- **Hierarchical clustering based**
  - CIDR (Lin et al. 2017 Genome Biology)
  - pcaReduce (Zeisel et al. 2016 BMC Bioinfo)
  - Ascend (Senabouth et al. 2019 Gigascience)
  - SINCERA (Guo et al. 2015 Plot Comp Biology)
  - BackSPIN (Zeiselet al. 2015 Science)
- **Graph or community-detection based**
  - Seurat (Macosko et al. 2015 Cell)
  - Scanpy (Wolf et al. 2018 Genome Biology)
  - PhenoGraph (Levin et al. 2015 Cell)
- **Model based clustering**
  - TSCAN (Ji and Ji, 2015 NAR)
  - monocle (Trapnell, 2014 Nat Biotech)

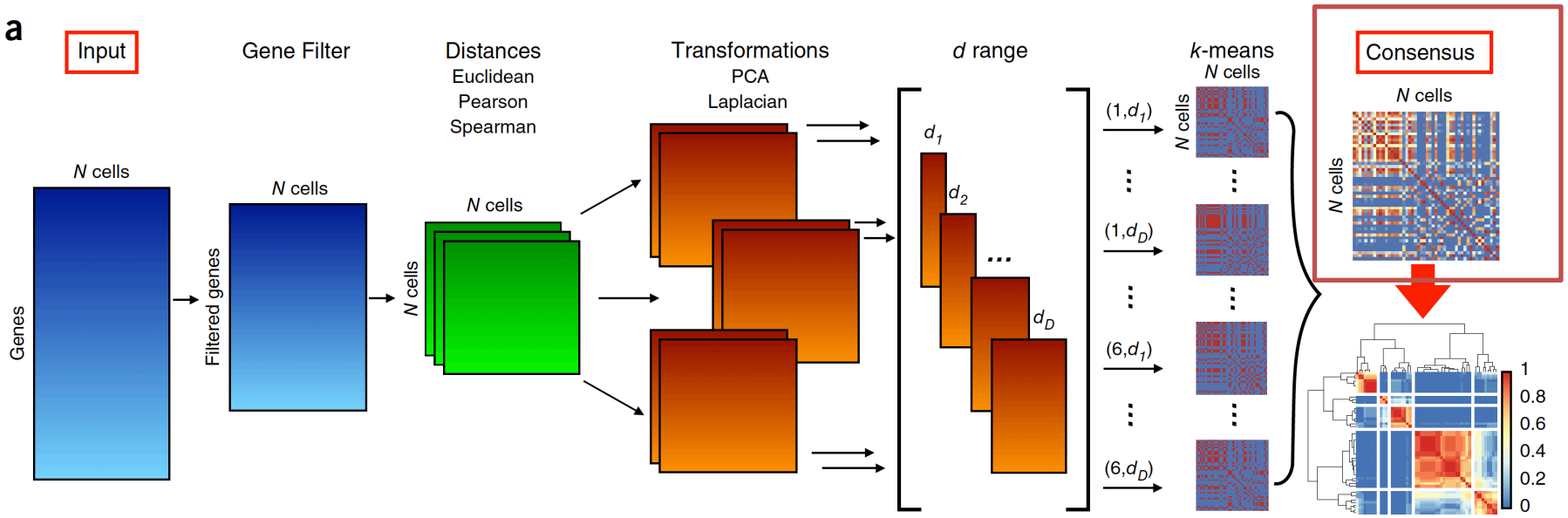
Existing cell clustering methods

- K-means based clustering methods
  - Iteratively identifies k cluster centroids, and each cell is assigned to the closest centroid
  - **Advantage**: scaling linearly with the number of points, can be applied to large datasets
  - **Drawback 1**: the algorithm is greedy. Global minimum is not guaranteed.
    - A solution: repeat the process using different initialization and find consensus result, SC3
  - **Drawback 2**: bias towards identifying equal-sized clusters. Rare cell types could be hidden in large clusters.
    - A solution: combine K-means cluster with outlier detection, RaceID

SC3


a
Input
Gene Filter
Distances
Euclidean
Pearson
Spearman
Transformations
PCA
Laplacian
d range
k-means
Consensus
N cells
Genes
Filtered genes
d_1
d_2
d_D
$(1,d_1)$
$(1,d_D)$
$(6,d_1)$
$(6,d_D)$
1
0.8
0.6
0.4
0.2
0

SC3



- Most important step: combines all the different clustering outcomes into a **consensus matrix** that summarizes how often each pair of cells is in the same cluster.
- The final result is determined by complete-linkage hierarchical clustering of the consensus matrix into *k* groups.

Example codes for SC3

```
sce = SingleCellExperiment(
    assays = list(
        counts = as.matrix(counts),
        logcounts = log2(as.matrix(counts) + 1)
    )
)
sce = sc3_prepare(sce)
if( missing(K) ) { ## estimate number of clusters
    sce = sc3_estimate_k(sce)
    K = metadata(sce)$sc3$k_estimation
}

sce <- sc3(sce, ks = K, biology = TRUE, n_cores = 4)
head(col_data[ , grep("sc3_", colnames(col_data))])
sc3clusters <- col_data$sc3_5_clusters
```

Existing cell clustering methods

- Hierarchical clustering-based methods
  - Sequentially combines individual cells into larger clusters (agglomerative) or divides clusters into smaller groups (divisive)
  - **Drawback**: both time and memory requirements scale at least quadratically with the number of data points. Does not scale well with large dataset.
  - CIDR adapts hierarchical clustering for scRNA-seq by adding an implicit imputation of zeros into the distance calculation – more stable

Existing cell clustering methods

- Graph or community-detection based
  - Instead of identifying groups of points that are close together, community detection identifies groups of nodes that are densely connected
  - Construct a k-nearest-neighbour graph first, then apply community-detection algorithm on the graph. The most popular one is the Louvain algorithm

  - **Disadvantage**: selection of k impacts the number and size of the final clusters
  - **Advantage**: users don't need to specify number of clusters

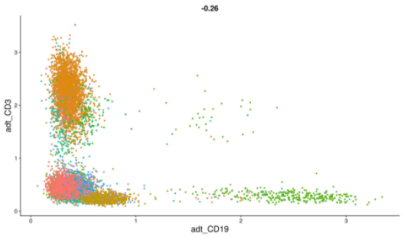
Seurat

Guided tutorial — 2,700 PBMCs

A basic overview of Seurat that includes an introduction to common analytical workflows.

GO

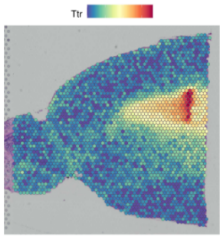
Multimodal analysis



An introduction to working with multi-modal datasets in Seurat.

GO

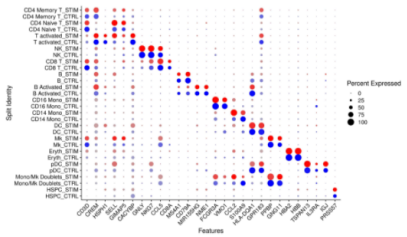
Analysis of spatial datasets



Learn to explore spatially-resolved transcriptomic data with examples from 10x Visium and Slide-seq v2.

GO

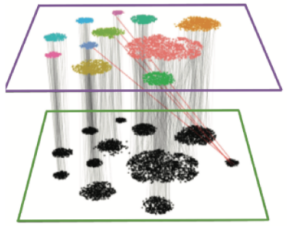
Introduction to scRNA-seq integration



An introduction to integrating scRNA-seq datasets in order to identify and compare shared cell types across experiments

GO

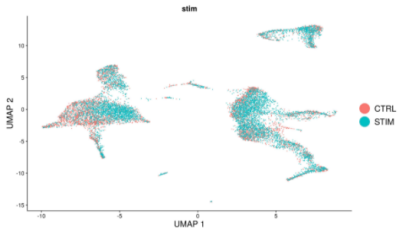
Mapping and annotating query datasets



Learn how to map a query scRNA-seq dataset onto a reference in order to automate the annotation and visualization of query cells

GO

Fast integration using reciprocal PCA (RPCA)



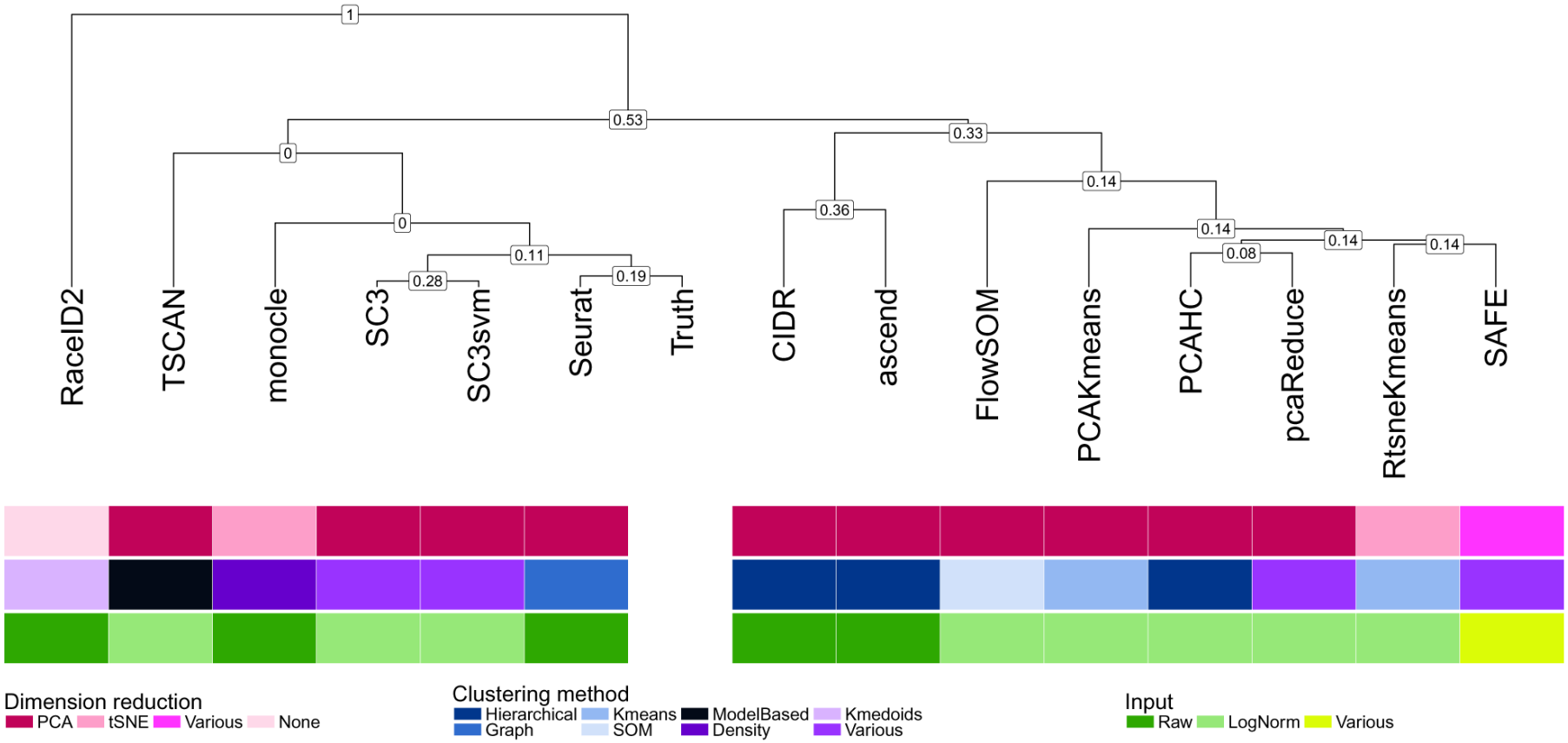
Identify anchors using the reciprocal PCA (rPCA) workflow, which performs a faster and more conservative integration

GO

Example code for Seurat

```
seuset = CreateSeuratObject( counts )
seuset = NormalizeData(object = seuset)
seuset = FindVariableFeatures(object = seuset)
seuset = ScaleData(object = seuset)
seuset = RunPCA(object = seuset)
seuset = FindNeighbors(object = seuset)
seuset = FindClusters(object = seuset)
Result = seuset@active.ident
```

Cell clustering methods

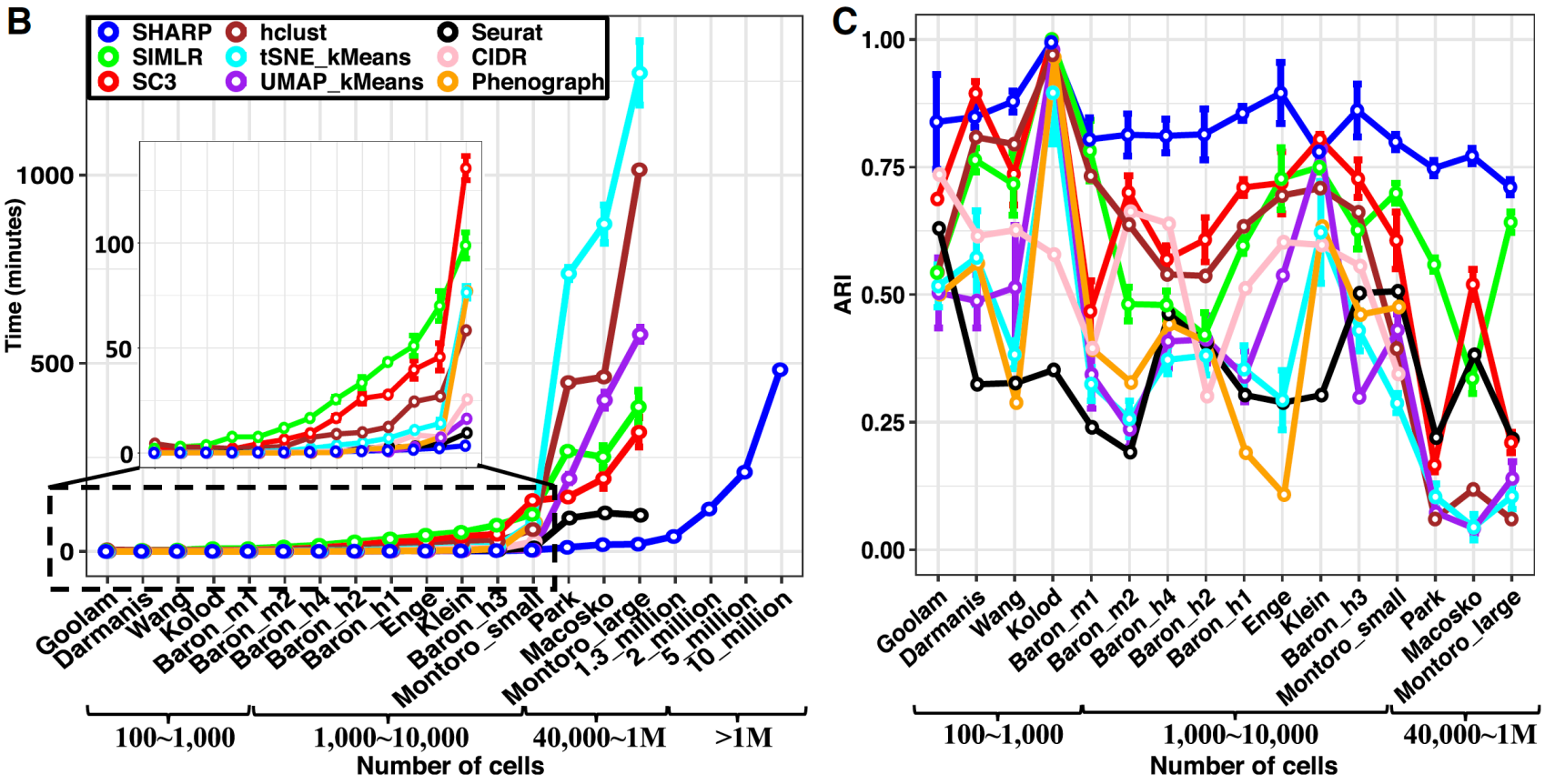


Duo et. al (2018) F1000 Research

SHARP (Genome Research, 2020)

- Hyperfast and accurate clustering of scRNA-seq data
- Based on ensemble random projection
- RP (Bingham and Mannila 2001) is a powerful dimension-reduction method that reduces the dimension while the distances between the points are approximately preserved
- RP is very fast because it does not require calculation of pairwise cell-to-cell distances or principal components

SHARP (Genome Research, 2020)



Performance comparisons and considerations

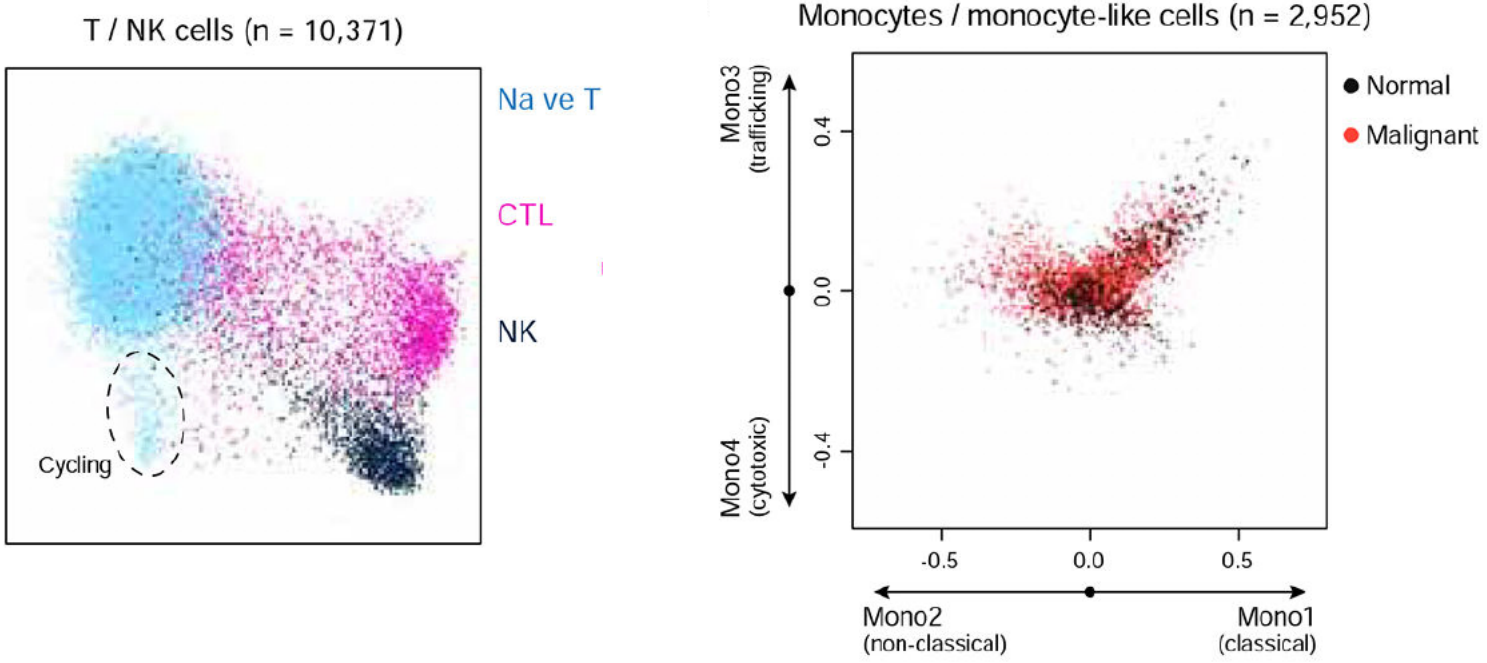
- Due to speed and scalability, SHARP or Seurat or scanpy is the top choice for large data set
- Louvain-based method does not have good accuracy in smaller data set
- SC3 has been shown to have the highest clustering accuracy, but is the slowest
- For rare cell type detection, RaceID and GiniClust should be considered. But they perform poorly if no rare cell type exists.

and robust [73]. Due to the heavy time consuming nature of consensus clustering, a rule of thumb for unsupervised single cell clustering is to use single-cell consensus clustering (SC3, integrated in Scater [52]) when the number of cells is < 5000 but use Seurat instead when there are more than 5000 cells.

Mu et al. Genomics Proteomics Bioinformatics (2019)

Sub-cluster identification

- After major clusters or cell types are identified, recursive clustering could be applied to define finer cell types



Galen et al. 2019. Cell

Challenges of clustering methods

- Technical challenges
  - Dropout: imputation methods available, but all rely on existing observations
- Technical noise: spike-in RNA can be used for normalization.
  - Batch effect is especially hard to correct in scRNA-seq. Batch effects can have a large impact on clustering. Balanced experimental design is hard to implement for perishable samples used in scRNA-seq.
- Multiple sample clustering

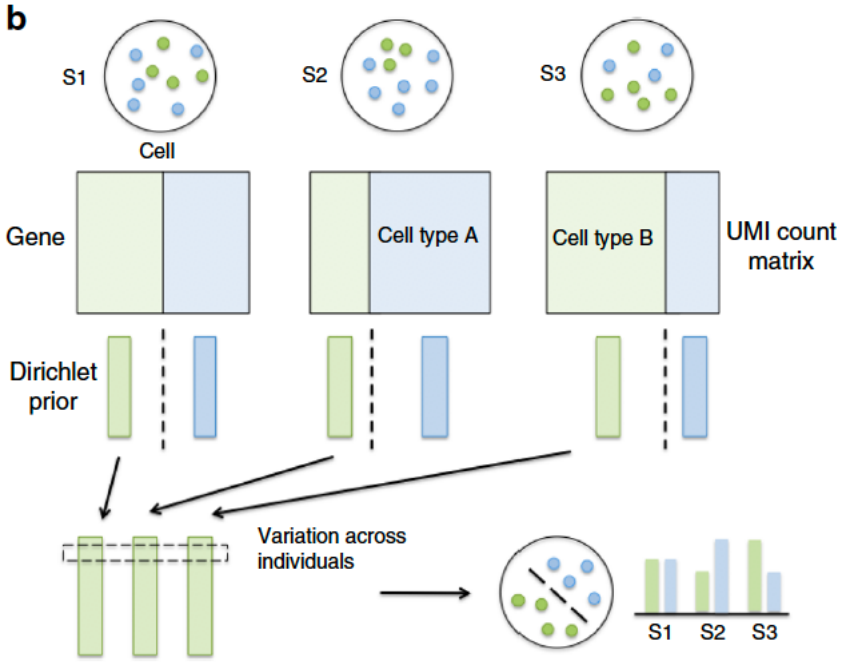
Cell clustering for multiple samples

- When scRNA-seq data are from multiple samples, batch/subject effects could have significant impact on the results.
- Cells from the same sample, instead of the same cell type form different sample, can cluster together.
- Possible solution:
  - Remove batch effect then cluster: MNN + SC3
  - Jointly model cell type and sample effect: BAMM- SC (Sun et al. 2019, Nat. Comm), BUSseq (Song et al. 2020, Nat. Comm), DESC (Li et al. 2020, Nat. Comm), CarDEC (Lakkis et al. 2021+)
- Is an active research field

Multiple subject clustering

A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies

Zhe Sun[1], Li Chen[2], Hongyi Xin[3], Yale Jiang[3,4], Qianhui Huang[5], Anthony R. Cillo[6], Tracy Tabib[7], Jay K. Kolls[8], Tullia C. Bruno[6,9], Robert Lafyatis[7], Dario A.A. Vignali[6,9,10], Kong Chen[11], Ying Ding[1], Ming Hu[12] & Wei Chen[1,3]



BAMM-SC

- Bayesian hierarchical Dirichlet multinomial mixture model
- Impose cell type-specific Dirichlet prior when modeling each individual
- Solve by Gibbs Sampler

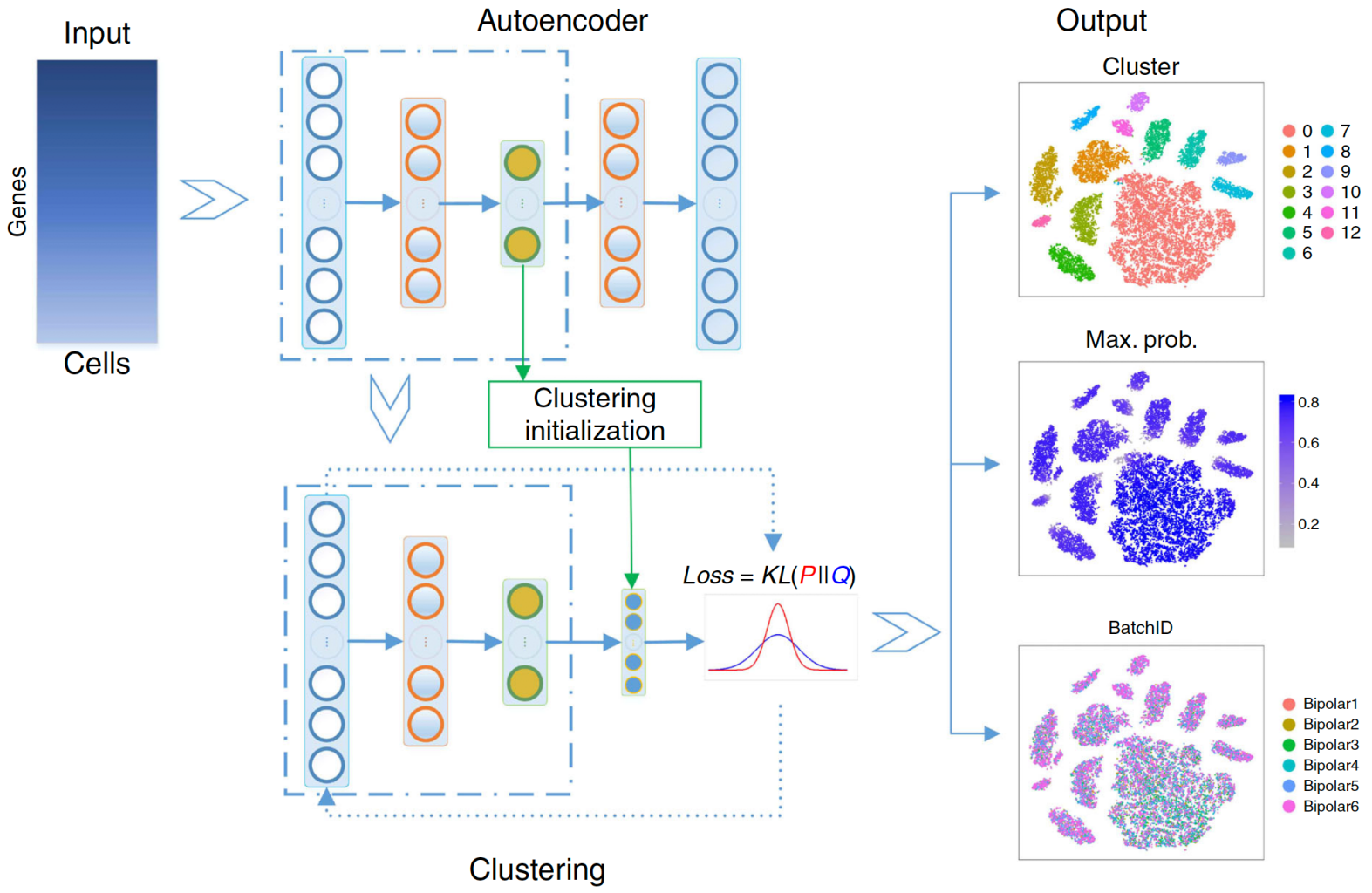
Multiple subject clustering

Table 1 Performance of clustering across ten times analyses for three real datasets

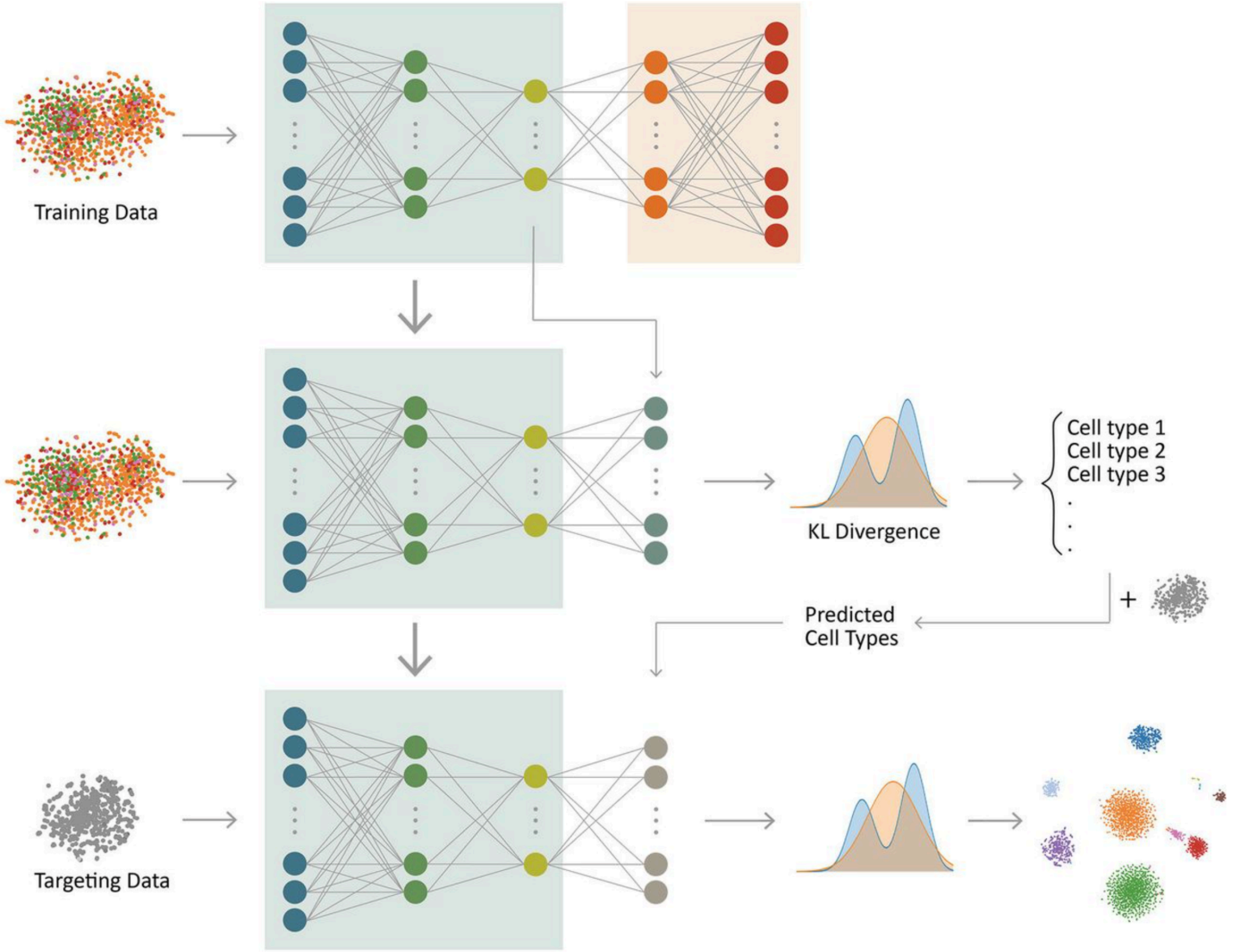
Method	Mean_P	SD_P	Range_P	Mean_L	SD_L	Range_L	Mean_S	SD_S	Range_S
MNN+K-means	0.379	0.083	(0.283–0.485)	0.662	0.066	(0.596–0.815)	0.597	0.075	(0.461–0.676)
MNN+TSCAN	0.373	NA	NA	0.720	NA	NA	0.553	NA	NA
MNN+SC3	0.348	0.084	(0.266–0.511)	0.640	0.061	(0.556–0.687)	0.517	0.034	(0.436–0.557)
MNN+Seurat	0.325	NA	NA	0.749	NA	NA	0.647	NA	NA
CCA+K-means	0.414	0.056	(0.307–0.464)	0.695	0.114	(0.505–0.883)	0.619	0.129	(0.424–0.737)
CCA+TSCAN	0.210	NA	NA	0.611	NA	NA	0.398	NA	NA
CCA+SC3	0.145	0.052	(0.051–0.215)	0.610	0.068	(0.531–0.708)	0.369	0.071	(0.277–0.488)
CCA+Seurat	0.468	NA	NA	0.729	NA	NA	0.702	NA	NA
DIMM-SC	0.333	0.071	(0.302–0.529)	0.809	0.030	(0.742–0.868)	0.715	0.045	(0.671–0.779)
BAMM-SC	0.487	0.056	(0.362–0.532)	0.882	0.042	(0.764–0.910)	0.762	0.032	(0.717–0.843)

Columns Mean_P, SD_P, and Range_P were calculated from human PBMC dataset. Columns Mean_L, SD_L, and Range_L were calculated from mouse lung dataset. Columns Mean_S, SD_S, and Range_S were calculated from human skin dataset.

DESC (Li et al. 2020, Nat. Comm)

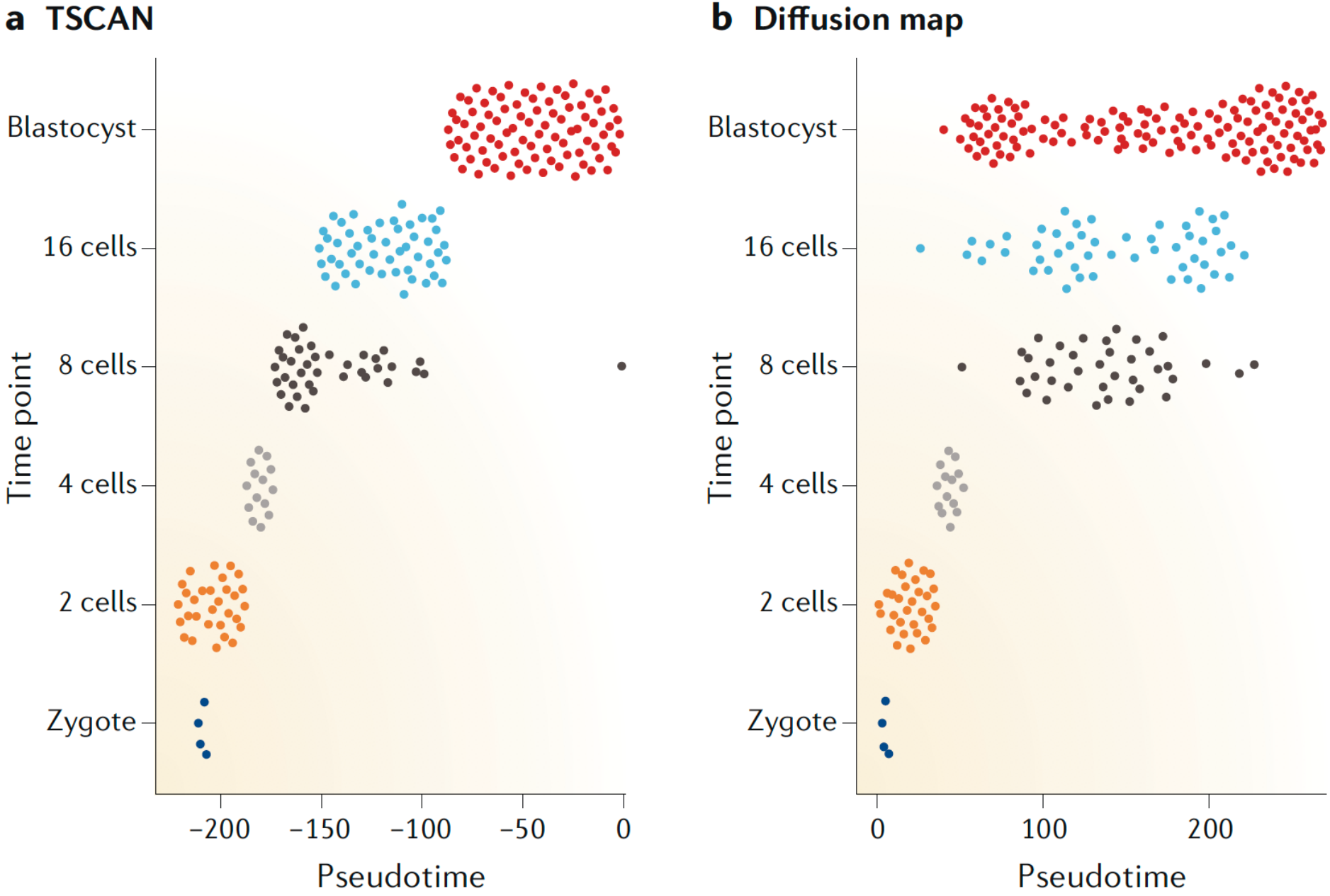


itClust (Hu et al. 2020, Nat. Machine Intelligence)



Pseudotime construction

- This belongs to the “clustering” category. Do not have discrete cluster numbers.
- Instead of putting cells into independent, exchangeable groups, it orders the cells by underlying temporal stage (estimated).
- Methods/tools:
  - Monocle/monocle2: Trapnell et al. (2014) Nat. Biotechnol; Qiu et al. (2017) Nat. Methods.
  - Waterfall: Shin et al. (2015) Cell Stem Cell
  - Wanderlust: Bendall et al. (2014) Cell
  - TSCAN: Ji et al. (2016) NAR



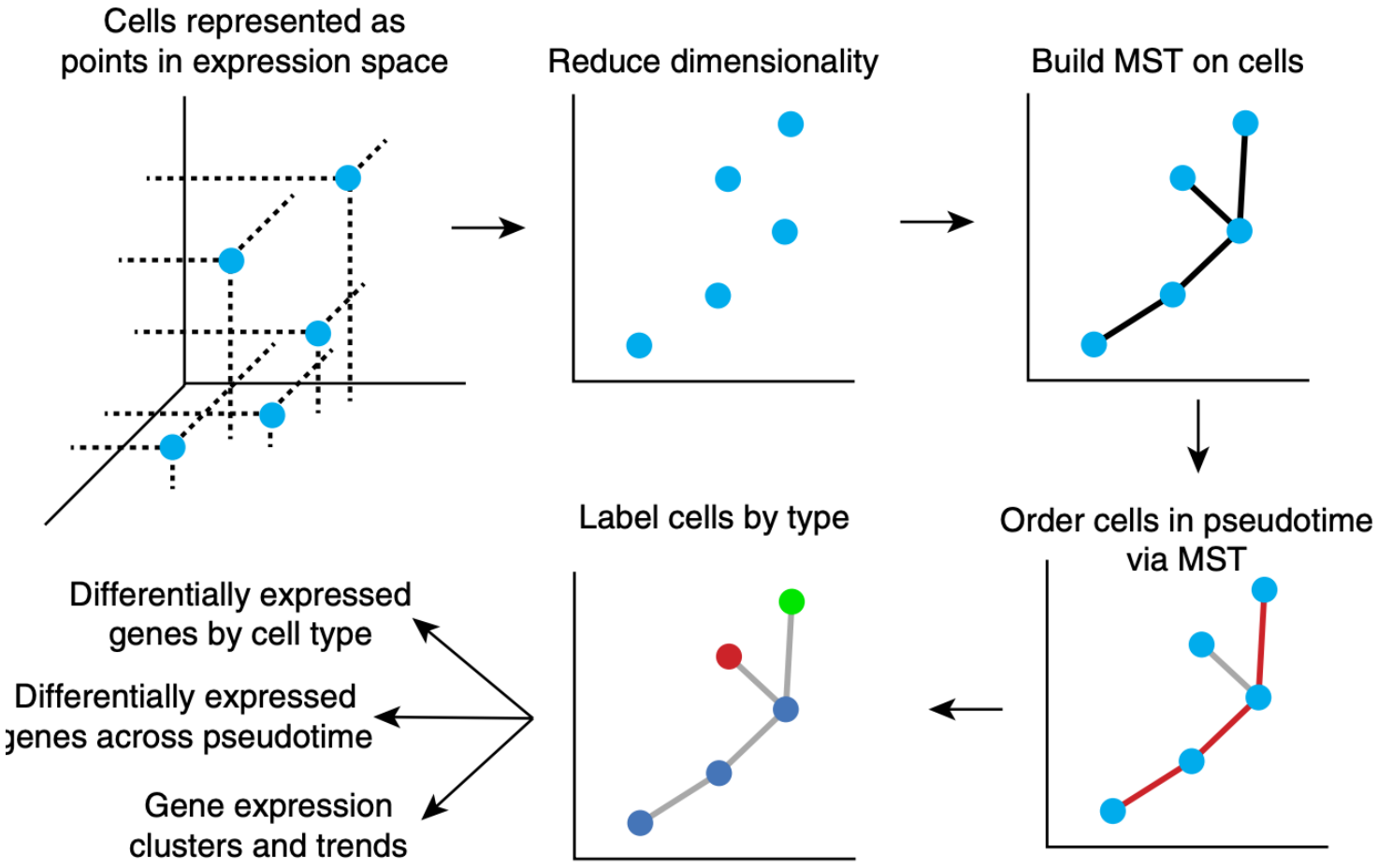
Kiselev et al. 2009. Nat Review Genetics

Pseudotime construction method

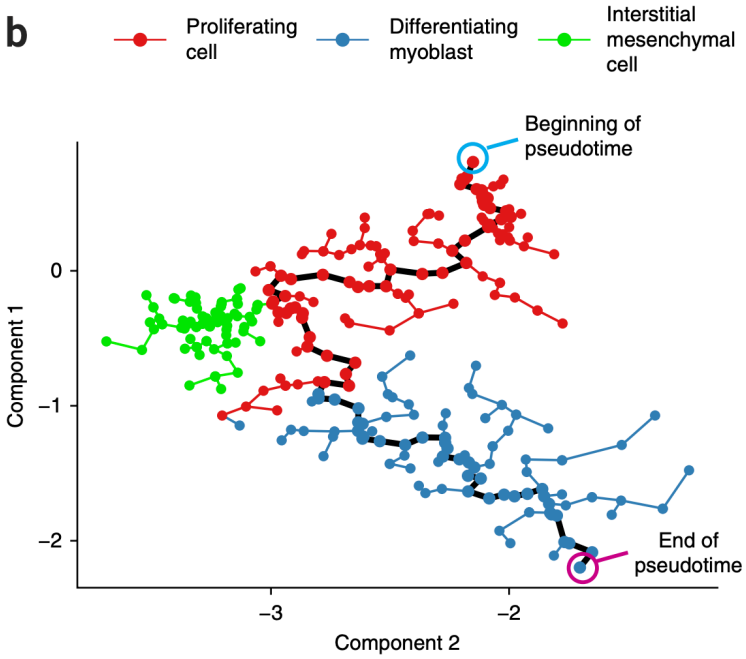
General steps:

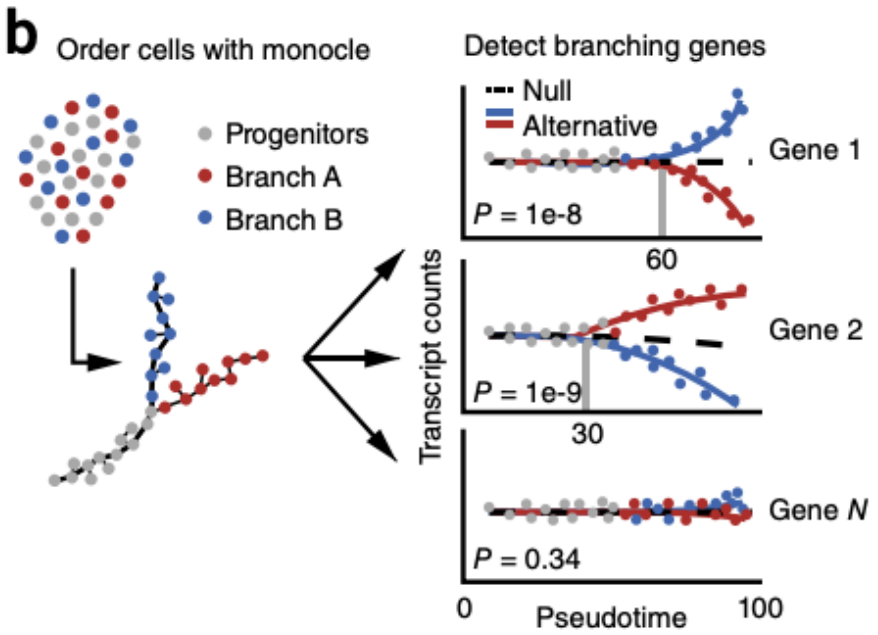
1. Select informative genes.
2. Dimension reduction of GE.
3. Cluster the cells based on reduced data. Often want to over-cluster them to have many groups.
4. Construct an MST (minimum spanning tree) from the clustering results.
5. Map cells to the MST.

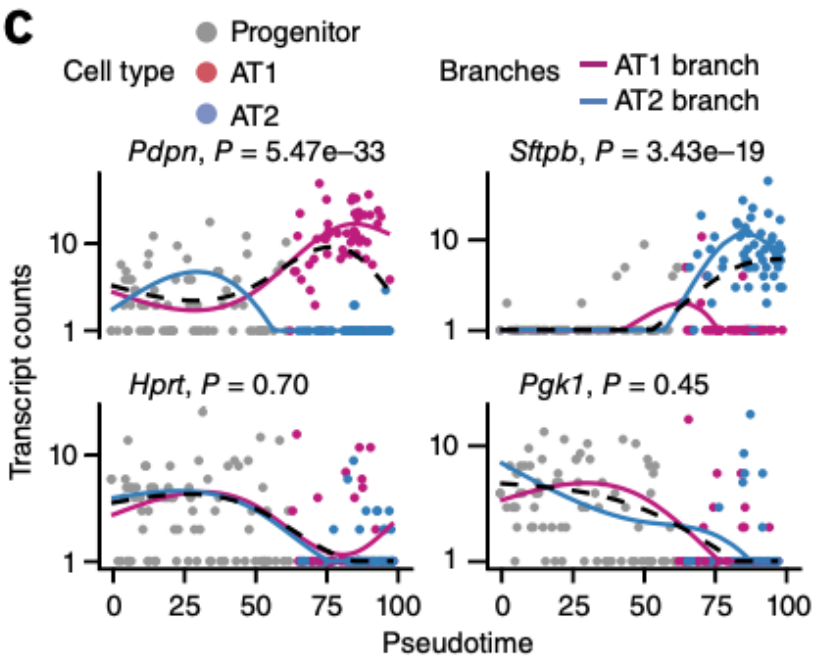
Monocle


Cells represented as
points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime
via MST
Label cells by type
Differentially expressed
genes by cell type
Differentially expressed
genes across pseudotime
Gene expression
clusters and trends

Monocle (v1: 2014, v2: 2017, v3: preprint)

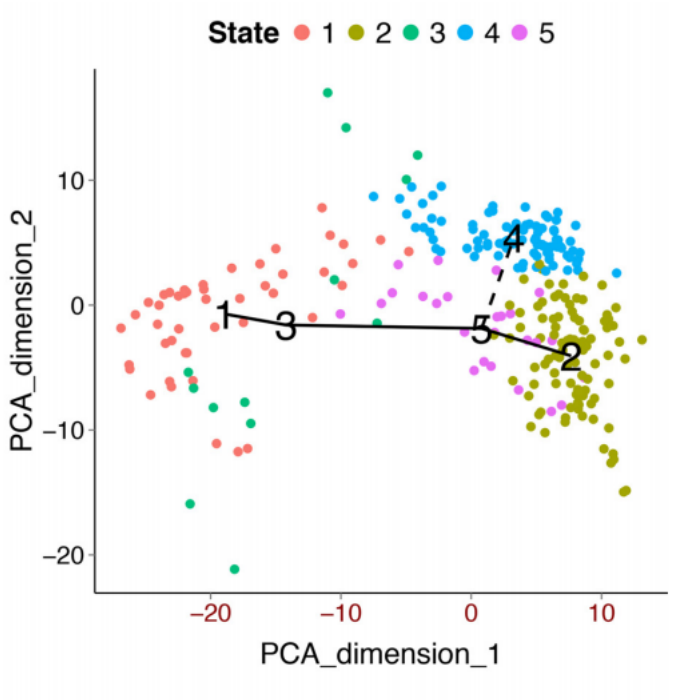


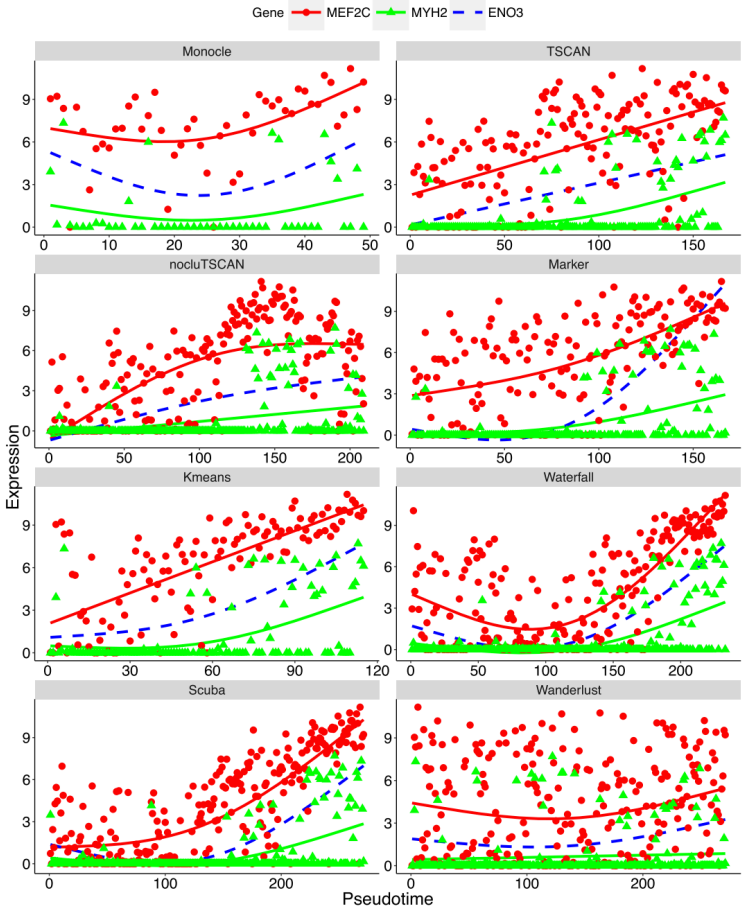




TSCAN (2016)

- In silico pseudo-Time reconstruction in Single-Cell RNA-seq ANalysis



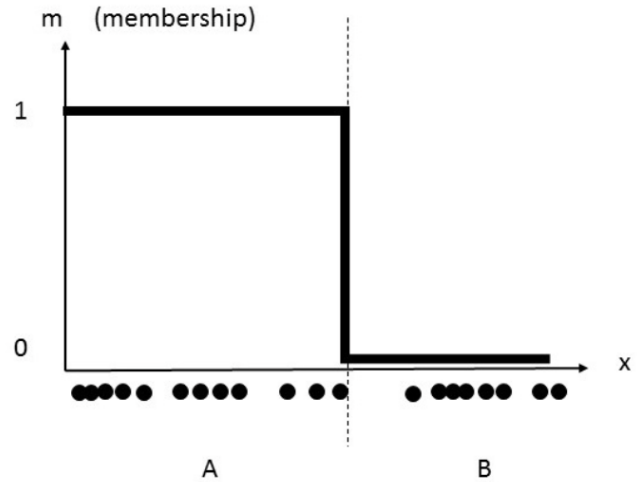


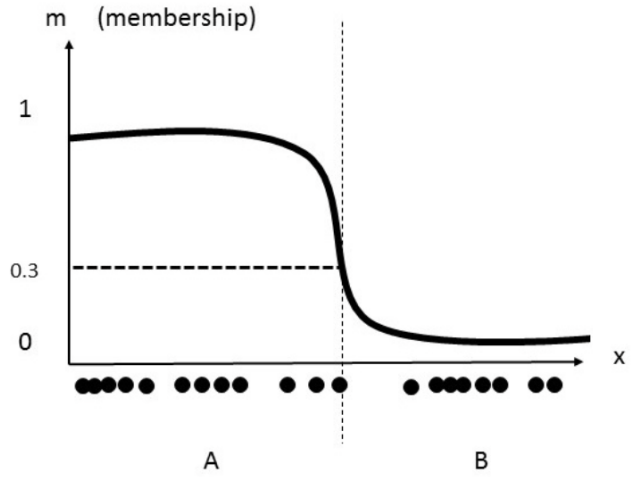
Example code for monocle

```
cds <- new_cell_data_set(expression_matrix,
                         cell_metadata = cell_metadata,
                         gene_metadata = gene_annotation)
cds <- preprocess_cds(cds, num_dim = 50)
cds <- align_cds(cds, alignment_group = "batch")
cds <- learn_graph(cds)
plot_cells(cds,
           color_cells_by = "cell.type",
           label_groups_by_cluster=FALSE,
           label_leaves=TRUE,
           label_branch_points=FALSE)
```

Soft-clustering

- Also named fuzzy clustering
- A form of clustering in which each data point can belong to more than one cluster
- Compared to hard clustering, cells in soft clustering have probabilities that belonging to each cluster
- Clusters are identified through similarity measures, e.g. distance, connectivity, intensity, etc.





SOUP (2019)

- Semisoft clustering: expect the existence of both **pure cells**, each belonging to a single cluster and requiring a hard cluster assignment, and **mixed cells** (transitional cells) that are transitioning between two or more cell types and hence should obtain soft assignments

1. Compute similarities:

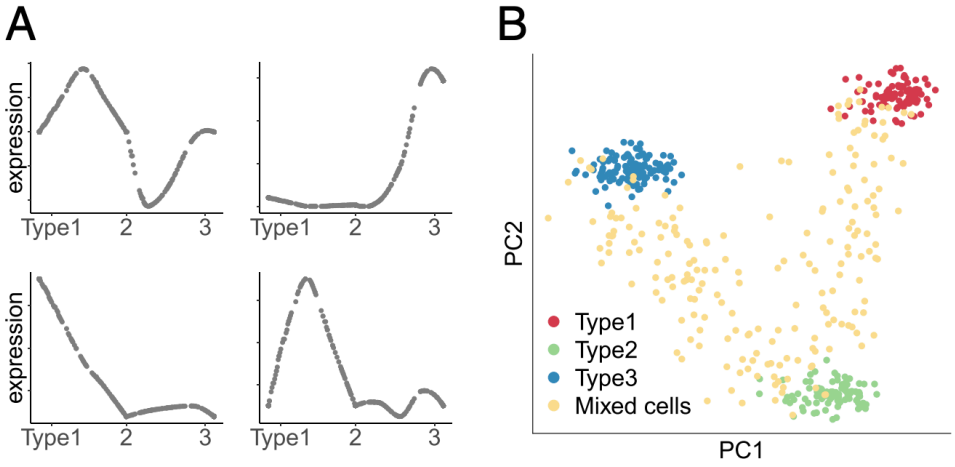
$$A := \mathbb{E}\left[XX^T\right] = \Theta Z \Theta^T + \sigma^2 I$$

2. Select pure cells:

$\hat{S}_i = \{j \neq i : \text{the top } \epsilon \text{ percent with the largest } |\hat{A}_{ij}|\}$,

$$\hat{p}_i = \frac{1}{|\hat{S}_i|} \sum_{j \in \hat{S}_i} \frac{|\hat{A}_{ij}|}{\hat{m}_j}, \text{ where } \hat{m}_i = \max_{j \neq i} |\hat{A}_{ij}|,$$

3. Obtain membership for all the cells



When does a cluster represent a new cell type?

- “For a new cell type to be accepted, it is necessary to go beyond characterization of the transcriptome.”
- A must: demonstrate that the newly identified cluster is also functionally distinct.
- A good example: Villani et al. (2017) discovered new cell types from blood – differences in morphology, stimulation by pathogens and ability to activate T cells.

Future directions

- Accurate and scalable clustering methods
- How to best choose the number of clusters or what quality of antibody is required for a validation experiment
- Determine how many marker genes are required to uniquely identify a specific cell type
- There is also a need for methods that will facilitate biological interpretation and annotation