UCLA Statistics

Applications of generalized additive models and copulas to single-cell RNA-seq computational method development

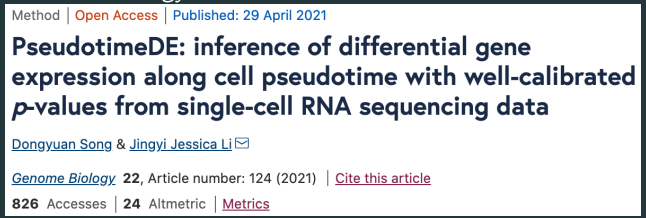
Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Our recent work

1. **PseudotimeDE**: inference of differential gene expression along cell pseudotime with valid p-values from single-cell RNA-seq data
   - by **Dongyuan Song** (宋东源; 2nd-year Bioinformatics PhD student)
   - *Genome Biology*


Method | Open Access | Published: 29 April 2021

PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated *p*-values from single-cell RNA sequencing data

Dongyuan Song & Jingyi Jessica Li

Genome Biology **22**, Article number: 124 (2021) | Cite this article

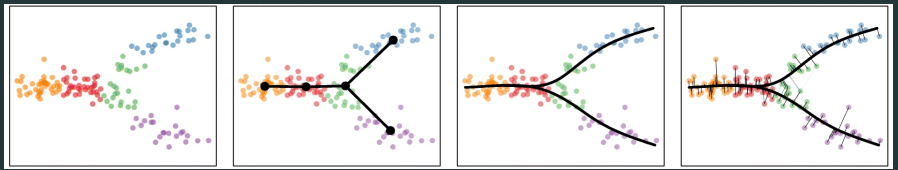
826 Accesses | **24** Altmetric | Metrics

2. **scDesign2**: a transparent simulator that generates realistic single-cell gene expression count data with gene correlations captured
   - by **Tianyi Sun** (孙天毅; 4th-year Statistics PhD student) *et al.*
   - accepted by *RECOMB* and in press at *Genome Biology*
   - bioRxiv: `https://doi.org/10.1101/2020.11.17.387795`

PseudotimeDE

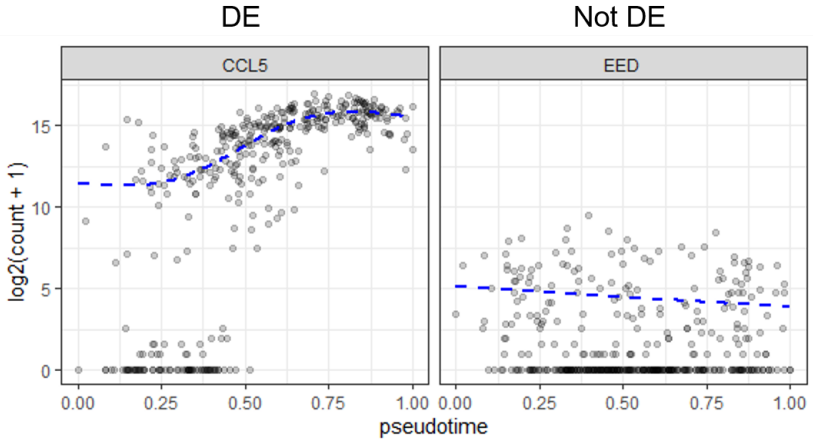
Pseudotime inference

- Pseudotime: a latent "temporal" variable that reflects a cell's relative transcriptome status among all cells
- Pseudotime inference (trajectory inference): estimate the pseudotime of cells, i.e., order cells along a trajectory (lineage) based on transcriptome similarities
- Popular methods:
  - `Monocle3` (Trapnell *et al.* 2014)
  - `TSCAN` (Ji *et al.* 2016)
  - `Slingshot` (Street *et al.* 2018)



Differential gene expression along cell pseudotime

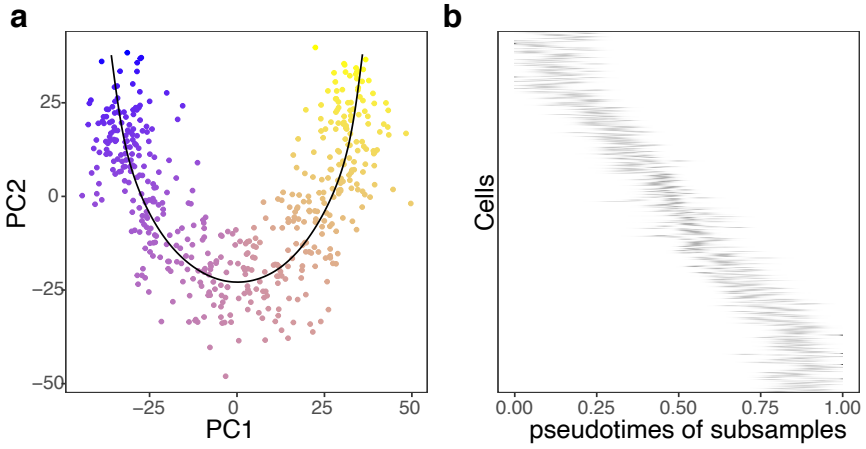
- Differentially expressed (DE) gene: a gene whose expected expression changes along cell pseudotime
- Question: how to identify DE genes?



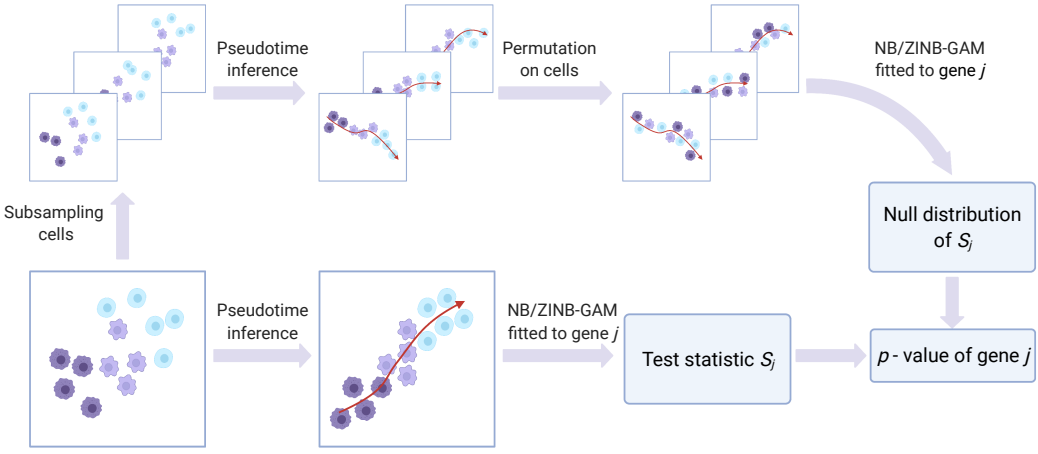
Limitations of existing methods

- `tradeSeq` (Van den Berge *et al.* 2020)
  - Test if a gene is DE based on a generalized additive model (GAM) (Hastie and Tibshirani, 1986, 1990)
- `Monocle3` (Trapnell *et al.* 2014)
  - Test if a gene is DE based on a generalized linear model (GLM) (McCullagh, 1983)
- Both methods are regression-based:
  - response: a gene's expression level in a cell
  - predictor/covariate: a cell's pseudotime
- Limitation: cell pseudotime is treated as fixed with uncertainty ignored
- Why is cell pseudotime random?
  - pseudotime is not observed but inferred; inference involves uncertainty
- This ignorance of pseudotime uncertainty may result in invalid *p*-values

Pseudotime inference uncertainty



Our proposal: PseudotimeDE



PseudotimeDE: notations

- $\boldsymbol{Y} = (Y_{ij})$: an $n \times m$ gene expression count matrix (n cells and m genes)
- $\boldsymbol{T} = (T_1, \ldots, T_i, \ldots, T_n)^\mathsf{T}$: cell pseudotime inferred from $\boldsymbol{Y}$
- To capture the uncertainty of pseudotime $\boldsymbol{T}$, we subsample 80% cells in $\boldsymbol{Y}$ for B times (default $B = 1000$); in the b-th subsample:
  - $\boldsymbol{Y}^b = (Y_{ij}^b)$, an $n' \times m$ matrix where $n' = \lfloor .8n \rfloor$
  - $\boldsymbol{T}^b = (T_1^b, \ldots, T_{n'}^b)^\mathsf{T}$: cell pseudotime inferred from $\boldsymbol{Y}^b$
  - $\boldsymbol{T}^{*b} = (T_1^{*b}, \ldots, T_{n'}^{*b})^\mathsf{T}$: permuted cell pseudotime

PseudotimeDE: GAM

- Negative-Binomial Generalized Additive Model (NB-GAM)

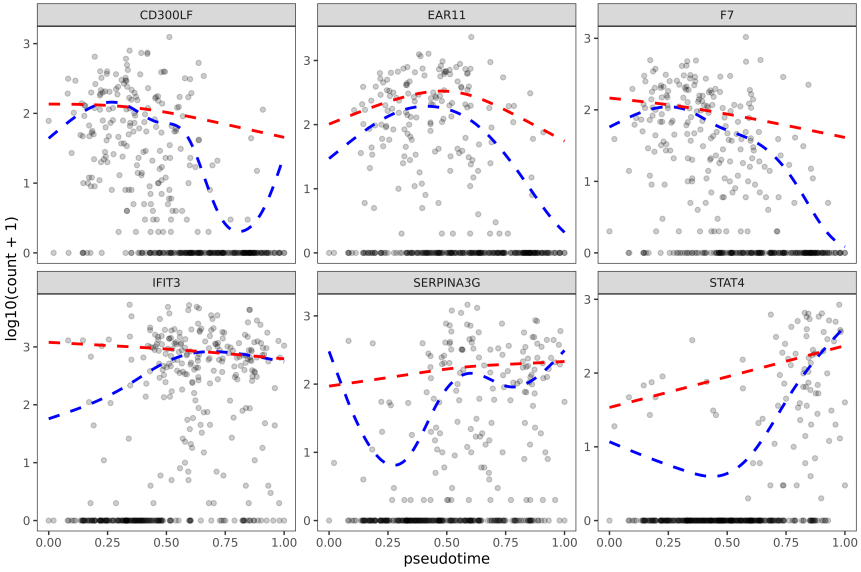
$$\begin{cases} Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_j) \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i) \end{cases}$$

- Zero-Inflated Negative-Binomial Generalized Additive Model (ZINB-GAM)

$$\begin{cases} Z_{ij} \sim \mathrm{Ber}(p_{ij}) \\ Y_{ij} | Z_{ij} \sim Z_{ij} \cdot \mathrm{NB}(\mu_{ij}, \phi_j) + (1 - Z_{ij}) \cdot 0 \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i) \\ \mathrm{logit}(p_{ij}) = \alpha_{j0} + \alpha_{j1} \log(\mu_{ij}) \end{cases}$$

where $f_j(T_i) = \sum_{k=1}^{K} b_k(T_i)\beta_{jk}$ is a cubic spline function

NB-GAM (blue) v.s. ZINB-GAM (red)



PseudotimeDE: statistical test

- Null and alternative hypotheses for gene j:

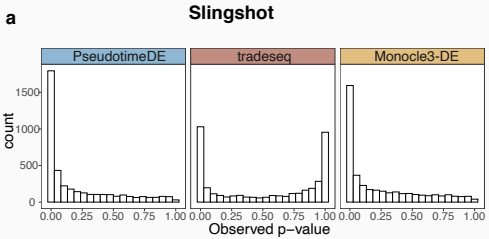
$$H_0 : f_j(\cdot) = 0 \quad \text{vs.} \quad H_1 : f_j(\cdot) \neq 0$$

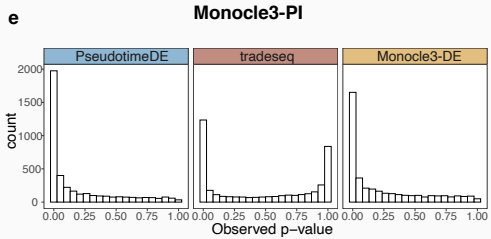
- Fit GAM to $\boldsymbol{Y}$ and $\boldsymbol{T}$. Denote the estimate of $(f_j(T_1), \ldots, f_j(T_n))^\mathsf{T}$ by $\hat{\boldsymbol{f}}_j$ and estimated covariance matrix of $\hat{\boldsymbol{f}}_j$ by $\widehat{\mathrm{V}}_{\boldsymbol{f}_j}$
- Test statistic:

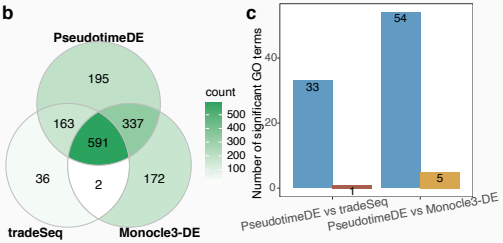
$$S_j = \hat{\boldsymbol{f}}_j^\mathsf{T} \widehat{\mathrm{V}}_{\boldsymbol{f}_j}^{r-} \hat{\boldsymbol{f}}_j$$

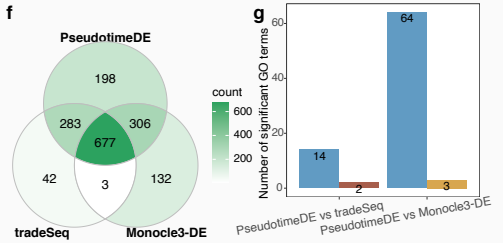
  where $\widehat{\mathrm{V}}_{\boldsymbol{f}_j}^{r-}$ is the rank-r pseudo-inverse of $\widehat{\mathrm{V}}_{\boldsymbol{f}_j}$
- Observed value of S_j denoted by s_j
- For $b = 1, \ldots, B$, fit GAM to $\boldsymbol{Y}^b$ and $\boldsymbol{T}^{*b}$; calculate the test statistic s_j^b
- $\{s_j^1, \ldots, s_j^B\}$: null values of the test statistic S_j
- **Gene j's p-value** $p_j \Longleftarrow s_j, \{s_j^1, \ldots, s_j^B\}$

Real data example: dendritic cells stimulated with LPS









PseudotimeDE tradeSeq Monocle3-DE

Real data example: dendritic cells stimulated with LPS

d

PseudotimeDE vs tradeSeq

ID	description	p
GO:0043410	positive regulation of MAPK cascade	0.00331
GO:0050729	positive regulation of inflammatory response	0.00424
GO:0071222	cellular response to lipopolysaccharide	0.00609

PseudotimeDE vs Monocle3−DE

ID	description	p
GO:0006954	inflammatory response	0.00032
GO:0043410	positive regulation of MAPK cascade	0.00079
GO:0042742	defense response to bacterium	0.00573

h

PseudotimeDE vs tradeSeq

ID	description	p
GO:0002250	adaptive immune response	0.00812
GO:0002673	regulation of acute inflammatory response	0.00955
GO:0017001	antibiotic catabolic process	0.00955

PseudotimeDE vs Monocle3−DE

ID	description	p
GO:0045087	innate immune response	0.00016
GO:0009617	response to bacterium	0.00039
GO:0050829	defense response to Gram−negative bacterium	0.00418

Learning more about GAMs

- Book: *Generalized Additive Models: an introduction with R* by Dr. Simon Woods

- R package `mgcv` by Dr. Simon Woods

scDesign2

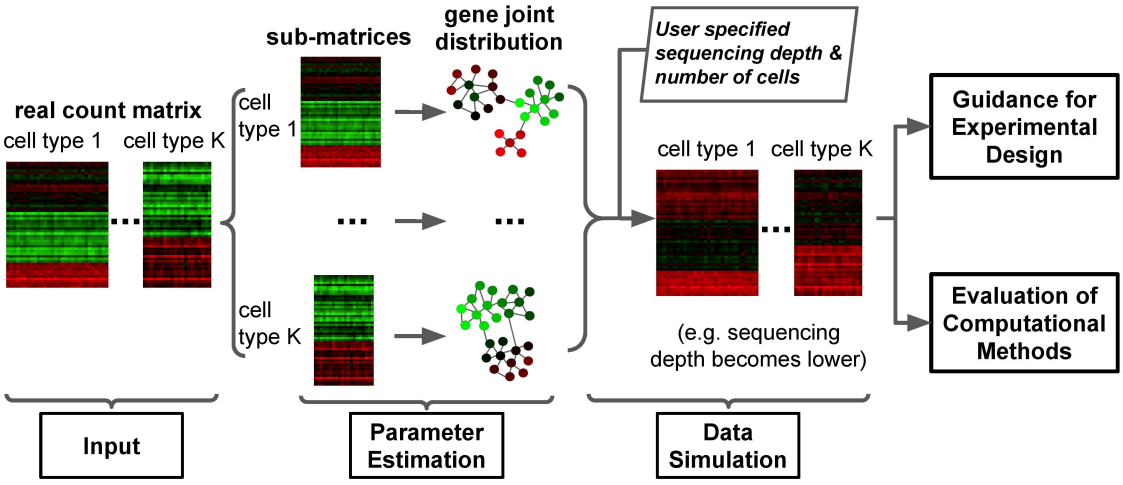
Motivation

- Experimental design:
  - How to choose among existing experimental protocols?
  - Given a chosen protocol, how to determine the optimal parameters for the experiment (cell number and seq. depth)?
- Computational benchmarking:
  - How to choose among available computational methods for data analysis?
- **Use a realistic simulator to answer these questions!**

Existing scRNA-seq simulators

Simulator \ Property	protocol adaptive	genes preserved	gene cor. captured	cell num. seq. depth flexible	easy to interpret	comp. & sample efficient
dyngen	✓✗	×	×	✓✗	✓	✓
Lun2	✓✗	✓	×	✓	✓✗	✓
powsimR	✓	✓	×	✓	✓	✓
PROSST	✓✗	✓	×	✓✗	✓	✓
scDD	✓	×	×	✓✗	✓✗	✓
scDesign	✓	✓✗	×	✓	✓	✓
scGAN	✓	✓	✓✗	✓✗	×	×
splat simple	✓	×	×	×	✓	✓
splat	✓	×	×	×	✓	✓
kersplat	✓	×	✓✗	×	✓	✓
SPARSim	✓	✓	✓✗	×	✓	✓
SymSim	✓	×	×	×	✓	✓
ZINB-WaVE	✓	✓✗	✓✗	×	✓	✓
SPsimSeq	✓	✓	✓	✓✗	✓	✓

Our proposal: scDesign2



scDesign2: notations

- Denote the scRNA-seq count matrix as $\boldsymbol{X} \in \mathbb{N}^{p \times n}$, with p genes and n cells
- Assume that $\boldsymbol{X}$ contains K cell types and the cell memberships are known in advance
- Suppose there are $n^{(k)}$ cells in cell type k, $k = 1, ..., K$, and denote the count matrix for cell type k as $\boldsymbol{X}^{(k)}$
- Our goal is to fit a parametric, probabilistic model of all genes' expression in each cell type k
- For simplicity of notation, we drop the subscript k in the following discussion

scDesign2: marginal distribution of each gene i

- Model counts directly
- Denote $X_{\cdot j} = (X_{1j}, \ldots, X_{pj}) \in \mathbb{N}^p$ as the gene expression vector for cell j, $j = 1, \ldots n$. We assume that the $X_{\cdot j}$'s are i.i.d. — p variables; n observations
- x_{ij}: observed count of gene i in cell j
- Select a marginal count distribution for gene i's count X_{ij} from Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial

scDesign2: joint distribution of all genes

- Use the copula framework
- Denote $F : \mathbb{N}^p \to [0,1]$ as the joint cumulative distribution function (CDF) of $X_{\cdot j} \in \mathbb{N}^p$ and $F_i : \mathbb{N} \to [0,1]$ as the marginal CDF of X_{ij}
- By Sklar's theorem [Sklar 1959], there exists a copula function $C : [0,1]^p \to [0,1]$ such that

$$F(x_{1j}, \ldots, x_{pj}) = C(F_1(x_{1j}), \ldots, F_p(x_{pj}))$$

- The copula function $C(\cdot)$ is unique for continuous distributions, but not for discrete distributions (unidentifiable) [Genest et al 2007]

scDesign2: distributional transform and the Gaussian copula

- Distributional transform: necessary for discrete variable [Rüschendorf 2013].
  - Sample v_{ij} from $\mathrm{Uniform}[0,1]$ independently for $i = 1, \ldots, p$ and $j = 1, \ldots, n$
  - Calculate u_{ij} as

$$u_{ij} = v_{ij} F_i(x_{ij} - 1) + (1 - v_{ij}) F_i(x_{ij})$$

- Gaussian copula: Denote Φ as the CDF of a standard Gaussian random variable, we can express the joint distribution of $X_{\cdot j}$ as

$$F(x_{1j}, \ldots, x_{pj}) = \Phi_p(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}) | \boldsymbol{R})$$

  where $\Phi_p(\cdot | \boldsymbol{R})$ is a joint Gaussian CDF with a zero mean vector and a covariance matrix that is equal to the correlation matrix $\boldsymbol{R}$

scDesign2: joint distribution fitting

- Denote $\hat{F}_i$ as the estimated marginal distribution of gene i
- Sample v_{ij} from $\mathrm{Uniform}[0,1]$ independently for $i = 1, ..., p$ and $j = 1, ..., n$
- Calculate u_{ij} as

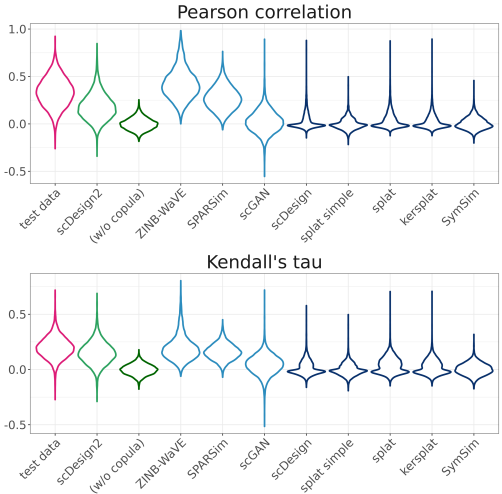
$$u_{ij} = v_{ij}\hat{F}_i(x_{ij} - 1) + (1 - v_{ij})\hat{F}_i(x_{ij})$$

- Calculate $\hat{\boldsymbol{R}}$ as the sample correlation matrix of $(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}))^{\mathsf{T}}$, $j = 1, \ldots, n$

scDesign2: data simulation

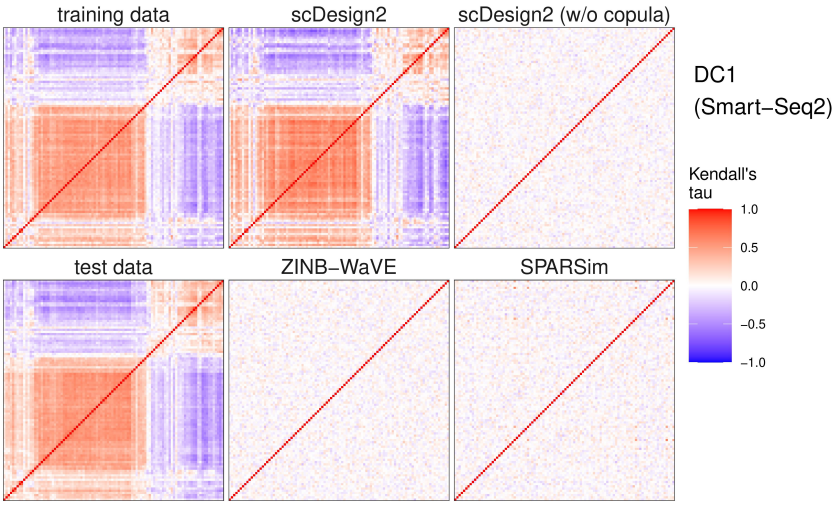
- Input from previous step:
  - fitted joint gene distributions (one per cell type)
  - cell type proportions
- User-specified input:
  - number of cells to simulate
  - total sequencing depth
- Output:
  - a synthetic gene-by-cell count matrix with K cell types
  - fitted model parameters (optional)

scDesign2 vs. existing scRNA-seq simulators



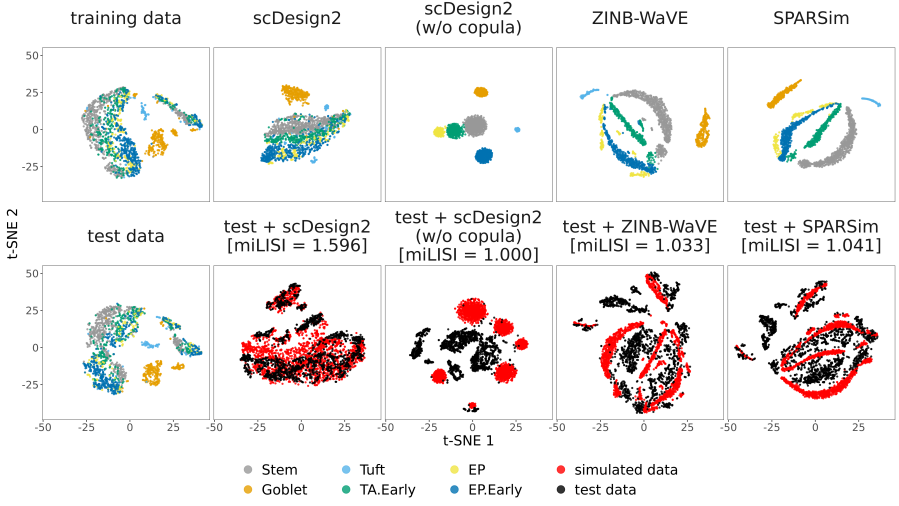
Data: mouse small intestinal goblet cells by 10x Genomics [Haber *et al.*, Nature (2017)]

scDesign2 vs. existing scRNA-seq simulators



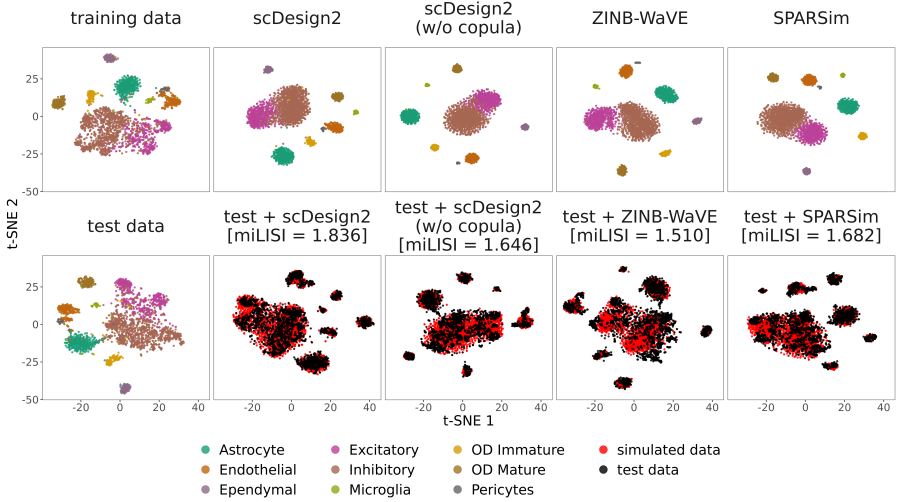
Data: dendrocytes subtype 1 of human blood by Smart-Seq2 [Villani *et al.*, Science (2017)]

scDesign2 vs. existing scRNA-seq simulators



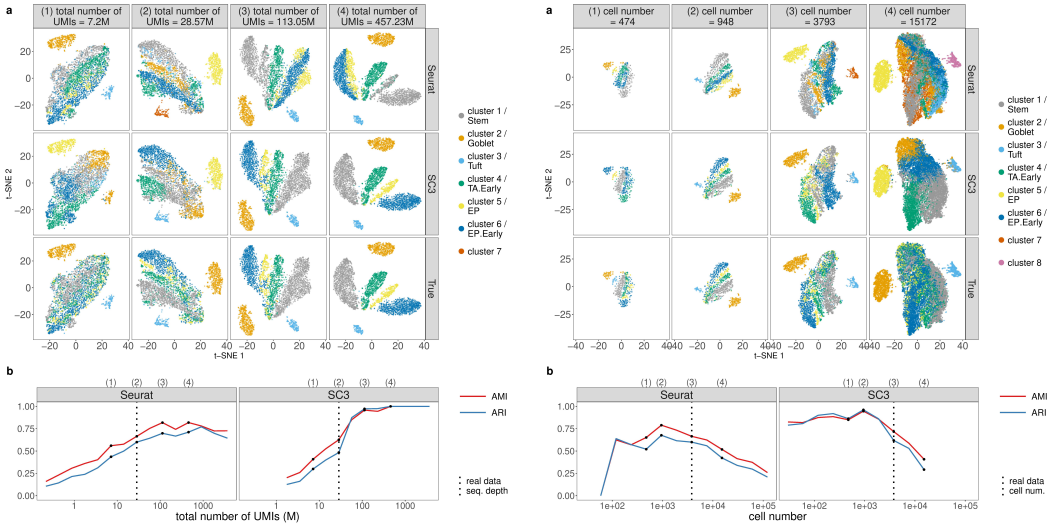
Data: mouse small intestinal epithelium cells by 10x Genomics [Haber *et al.*, Nature (2017)]

Application 1: simulation for other single-cell technologies



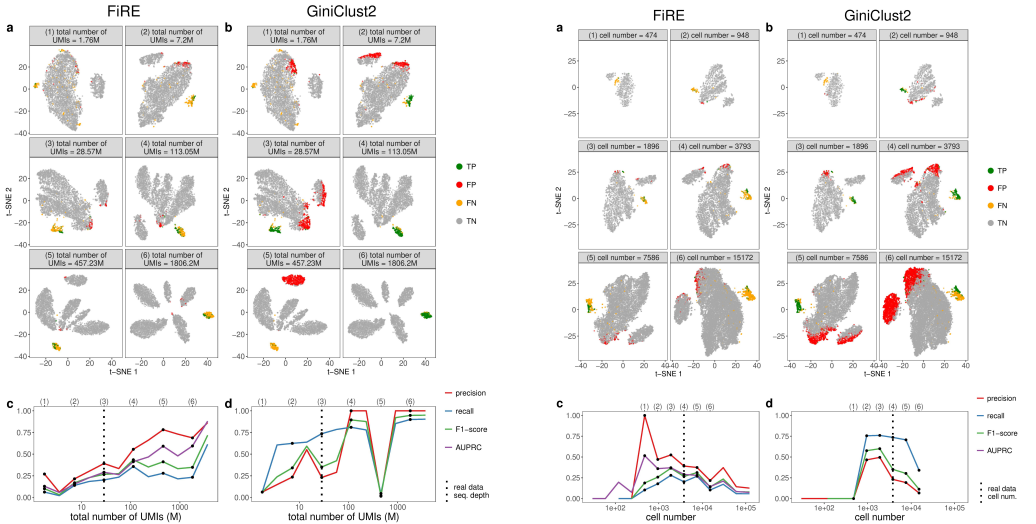
Data: mouse hypothalamic preoptic region by MERFISH [Moffitt *et al.*, Science (2018)]

Application 2: benchmarking cell clustering methods



Data: mouse small intestinal epithelium cells by 10x Genomics [Haber *et al.*, Nature (2017)]

Application 3: benchmarking rare cell type detection methods



Data: mouse small intestinal epithelium cells by 10x Genomics [Haber *et al.*, Nature (2017)]

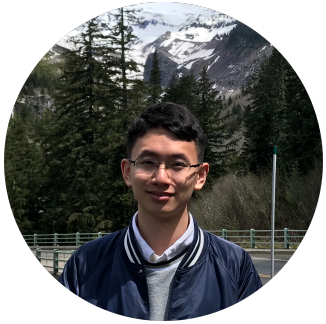
Learning more about copulas

- Book: *Introduction to copulas* by Dr. Roger B Nelson

Summary

- **PseudotimeDE**: finding DE genes along cell pseudotime
  - Well-calibrated *p*-values (essential for FDR control and GSEA)
  - Powerful (thanks to GAM)
  - R package: `https://github.com/SONGDONGYUAN1994/PseudotimeDE`
- **scDesign2**: generating realistic synthetic single-cell gene expression data
  - Gene correlations preserved (thanks to copula)
  - Probabilistic, transparent, interpretable
  - R package: `https://github.com/JSB-UCLA/scDesign2`

Acknowledgements



Dongyuan Song 宋东源
(Ph.D. student, UCLA)

Tianyi Sun 孙天毅
(Ph.D. student, UCLA)

Dr. Wei Vivian Li 李维
(former Ph.D. student;
assistant professor, Rutgers)