UCLA Statistics

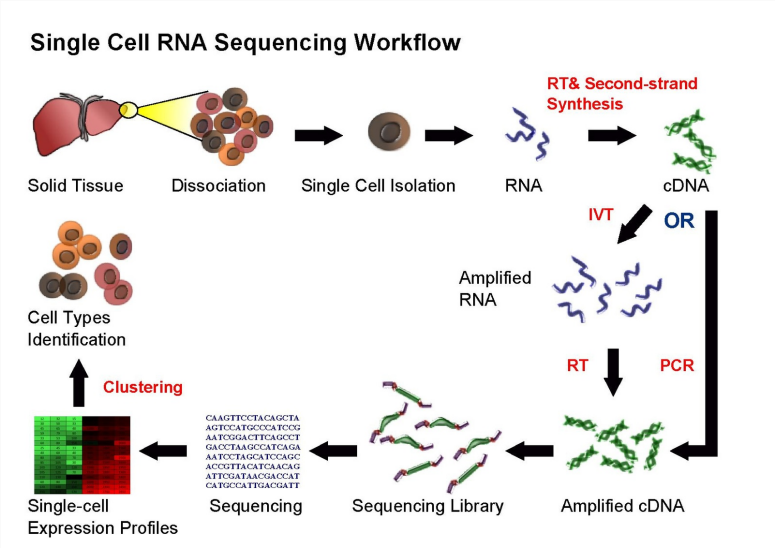
Applications of Statistical Models to Single-cell RNA-seq Data Analysis

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Single-cell RNA-sequencing (scRNA-seq)



from Wikipedia

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 (2nd-year Bioinformatics PhD student)
   - in press at *Genome Biology*
   - bioRxiv: `https://doi.org/10.1101/2020.11.17.387779`

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

PseudotimeDE

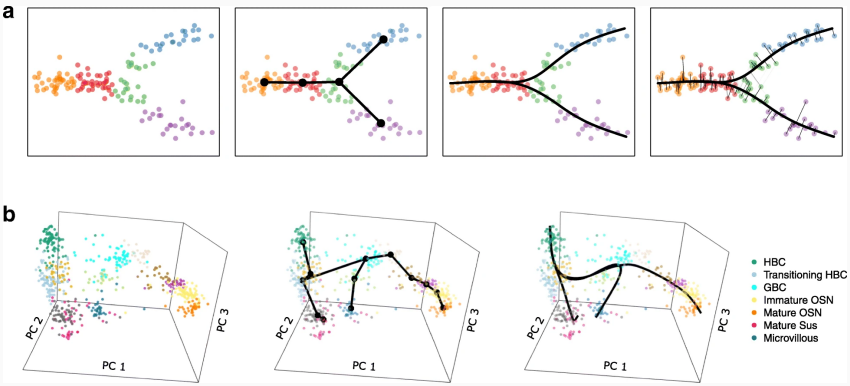
Pseudotime inference

- 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

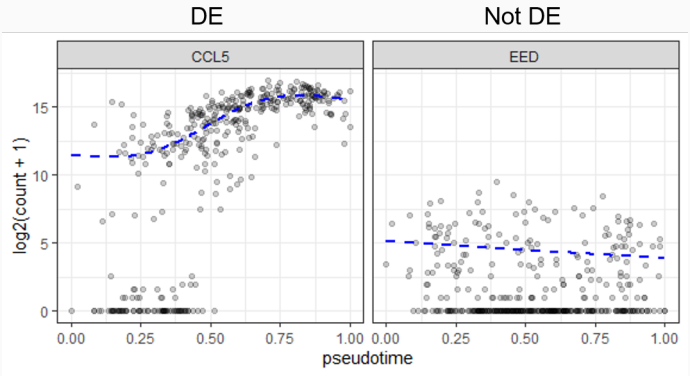
- Popular methods:
  - `Monocle3` (Trapnell *et al.* 2014)
  - `TSCAN` (Ji *et al.* 2016)
  - `Slingshot` (Street *et al.* 2018)

Example: pseudotime inference by Slingshot



Differential gene expression along cell pseudotime

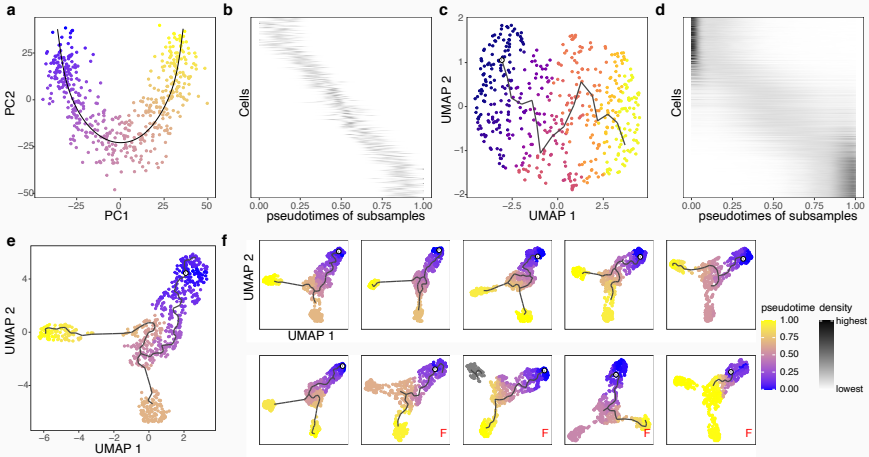
- Differentially expressed (DE) gene: a gene whose expected expression changes along cell pseudotime
- Question: how to identify DE genes?



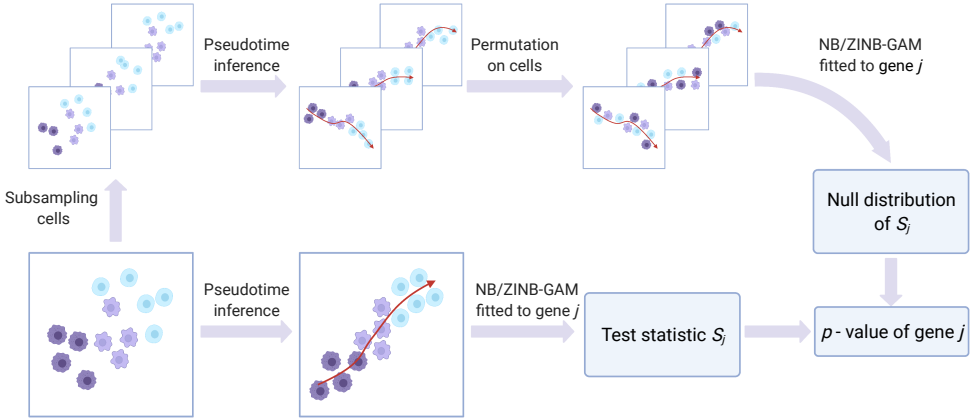
Limitations of existing methods

- `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)

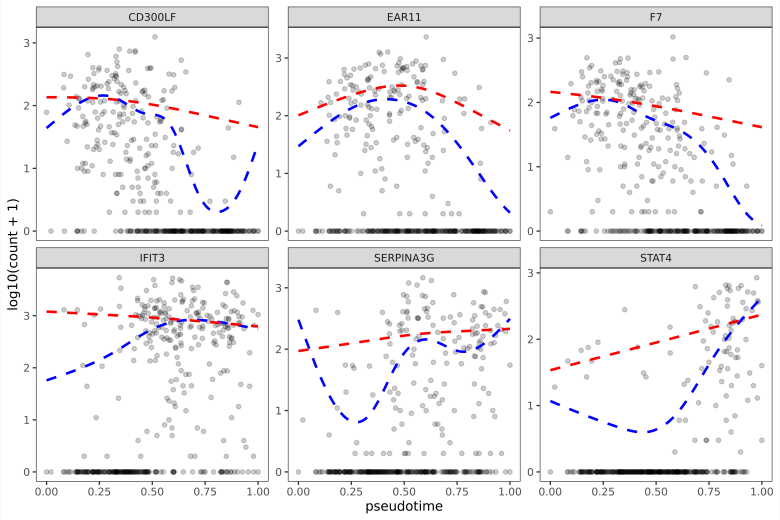
$$\begin{cases} Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_j) \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i) \end{cases}$$

- 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

PseudotimeDE: p-value calculation

- **Empirical p-value**:

$$p_j^{\text{emp}} = \frac{\sum_{b=1}^{B} \mathbb{I}(s_j^b \geq s_j) + 1}{B + 1}$$

The resolution of p_j^{emp} is $1/(B+1)$ - not enough for false discovery rate (FDR) control if B is not too large

- **Parametric p-value**:

Fit $\{s_j^1, \ldots, s_j^B\}$ by

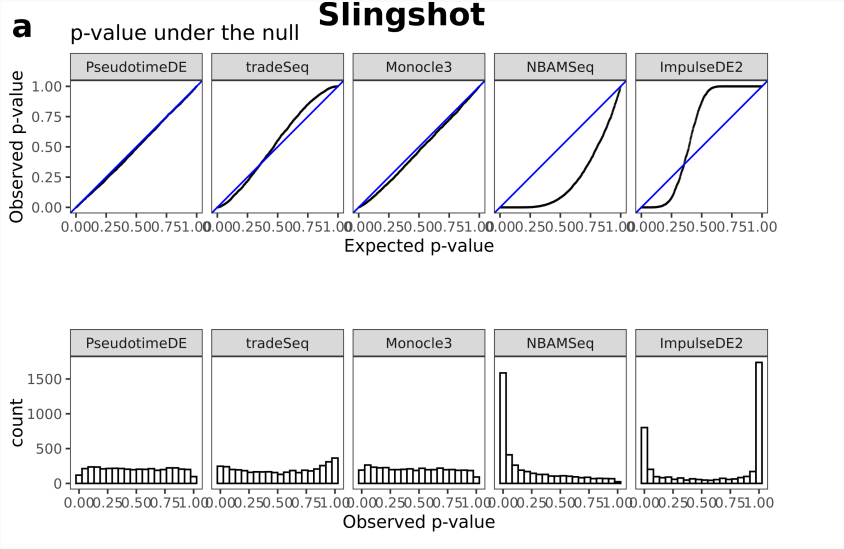
1. a gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha, \beta > 0$
2. a two-component gamma mixture model $\gamma\Gamma(\alpha_1, \beta_1) + (1-\gamma)\Gamma(\alpha_2, \beta_2)$ with $0 < \gamma < 1$ and $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$

Choose between the two distributions by the likelihood-ratio test

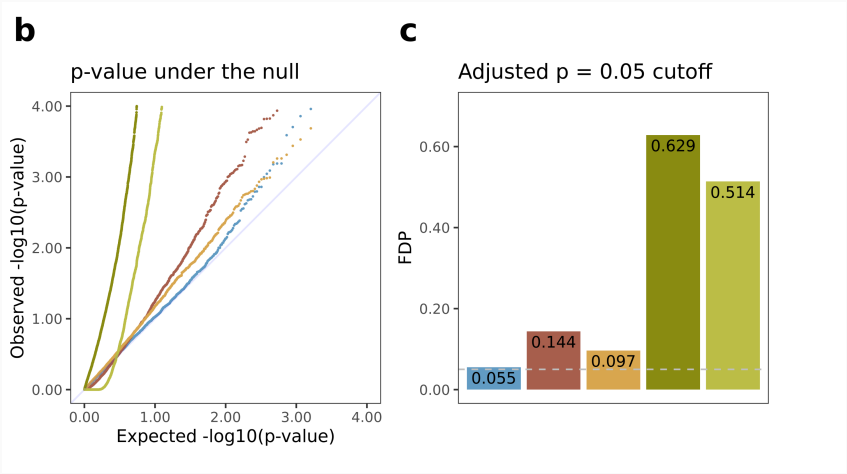
Parametric null distribution's cumulative distribution function: $\hat{F}_j(\cdot)$

$$p_j^{\text{param}} = 1 - \hat{F}_j(s_j)$$

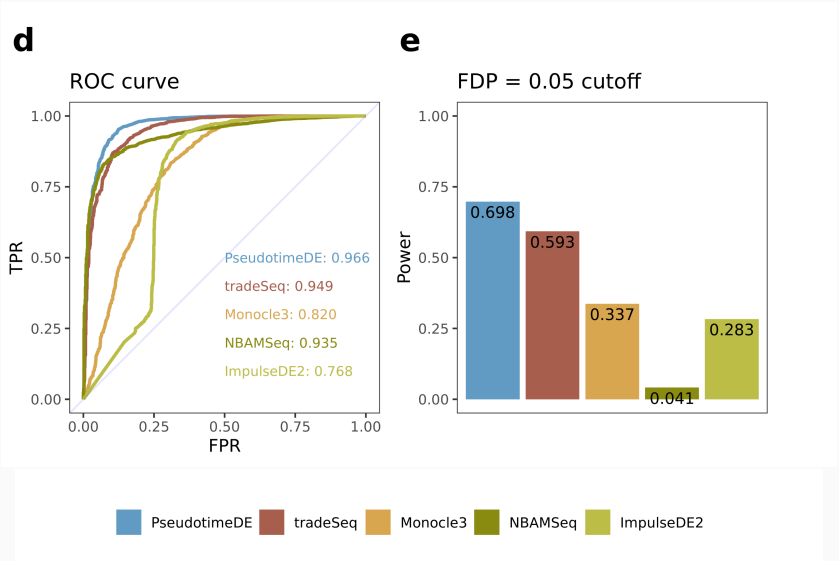
Simulation: PseudotimeDE generates well-calibrated *p*-values



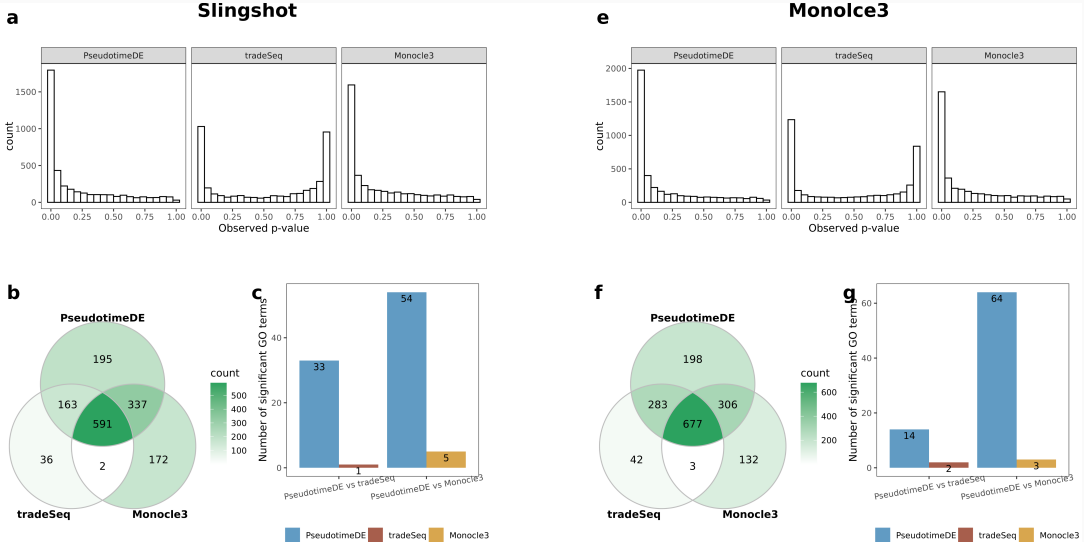
Simulation: PseudotimeDE leads to the best FDR control



Simulation: PseudotimeDE achieves the highest power



Real data example 1: dendritic cells stimulated with LPS



Real data example 1: dendritic cells stimulated with LPS

d

PseudotimeDE vs tradeSeq

ID	description	p
GO:0050729	positive regulation of inflammatory response	0.00424
GO:0071222	cellular responsee to lipopolysaccharide	0.00609

PseudotimeDE vs Monocle3

ID	description	p
GO:0006954	inflammatory responsee	0.00032
GO:0050729	positive regulation of inflammatory response	0.00479
GO:0006955	immune responsee	0.00541
GO:0042742	defense responsee to bacterium	0.00573

h

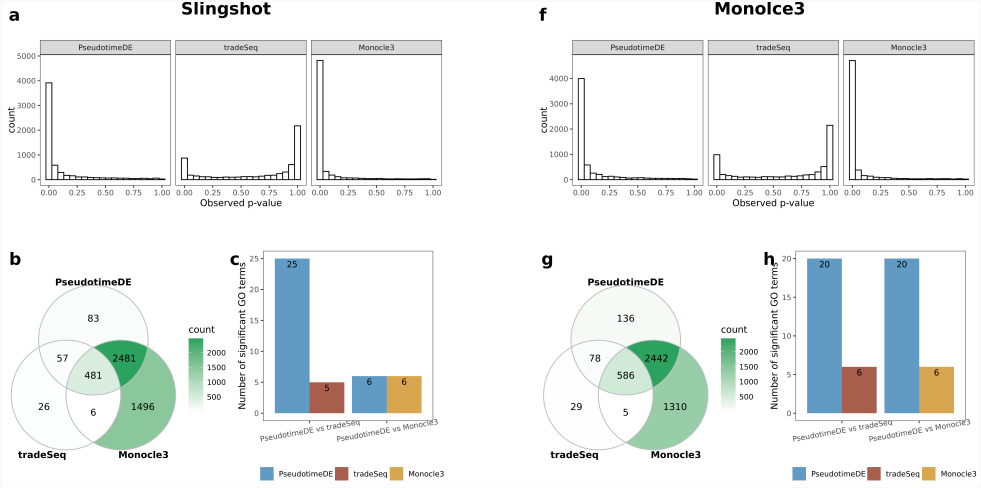
PseudotimeDE vs tradeSeq

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

PseudotimeDE vs Monocle3

ID	description	p
GO:0045087	innate immune response	0.00016
GO:0009617	response to bacterium	0.00039
GO:0032496	response to lipopolysaccharide	0.00049
GO:0006955	immune response	0.00054
GO:0061844	antimicrobial humoral immune response	0.00060
GO:0071222	cellular response to lipopolysaccharide	0.00075
GO:0002250	adaptive immune response	0.00179
GO:0042742	defense response to bacterium	0.00391
GO:0050829	defense response to Gram-negative bacterium	0.00418

Real data example 2: pancreatic beta cell maturation



Real data example 2: pancreatic beta cell maturation

d

PseudotimeDE vs tradeSeq

ID	description	p
GO:0045747	positive regulation of Notch signaling pathway	0.00162
GO:0006486	protein glycosylation	0.00221
GO:0045746	negative regulation of Notch signaling pathway	0.00326
GO:0009791	post-embryonic development	0.00478
GO:0031018	endocrine pancreas development	0.00490
GO:0098609	cell-cell adhesion	0.00975

PseudotimeDE vs Monocle3

ID	description	p
GO:0035773	insulin secretion involved in cellular response	0.0056

i

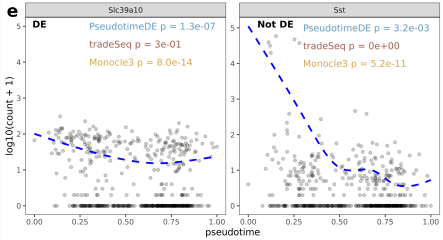
PseudotimeDE vs tradeSeq

ID	description	p
GO:0009791	post-embryonic development	0.00082
GO:0098609	cell-cell adhesion	0.00351

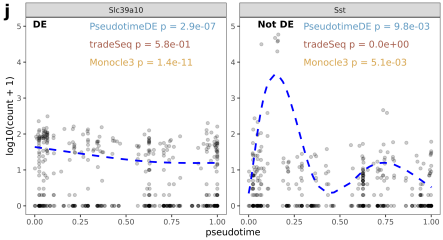
PseudotimeDE vs Monocle3

ID	description	p
GO:0046503	glycerolipid catabolic process	0.00076
GO:0006639	acylglycerol metabolic process	0.00489
GO:0001953	negative regulation of cell-matrix adhesion	0.00623

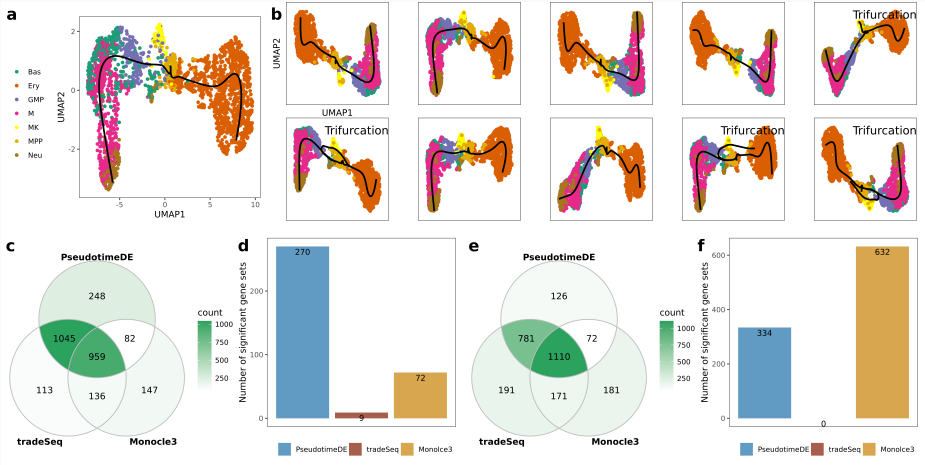
e



j



Real data example 3: bone marrow differentiation



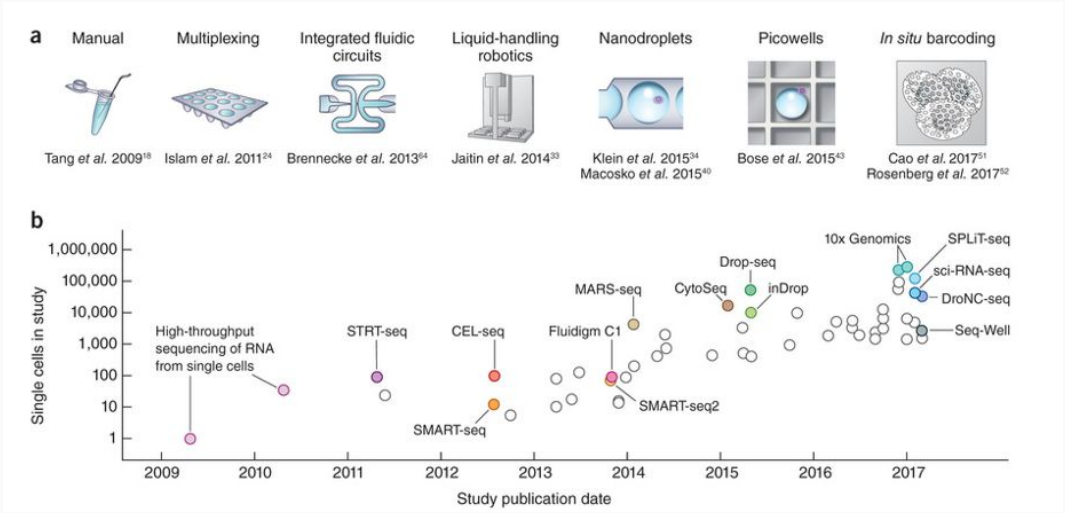
Learning more about GAMs

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

- R package `mgcv` by Dr. Simon Woods

scDesign2

Development of scRNA-seq Protocols



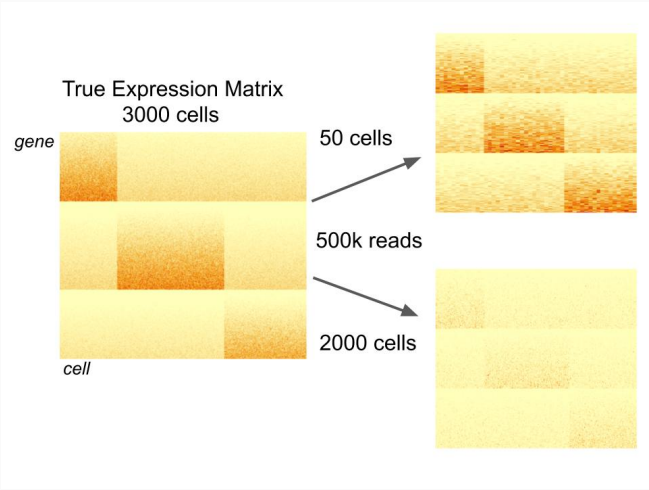
Scaling of scRNA-seq experiments. [Svensson et al 2018]

Experimental design questions

- How to choose among existing experimental protocols?
  - Tag-based vs full-length: more low-resolution cells or fewer high-resolution cells?

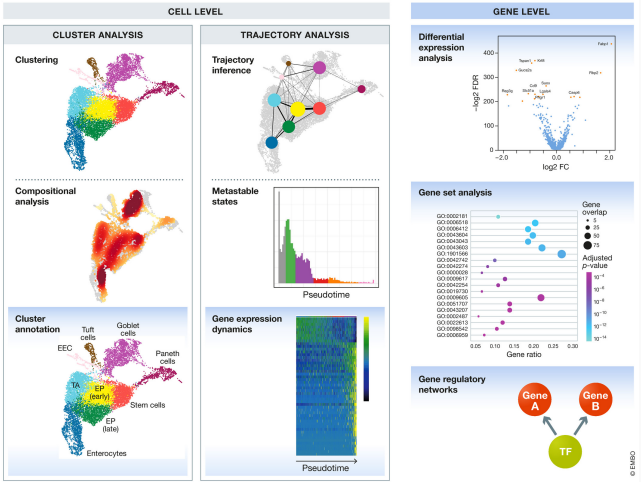
- Given a chosen protocol, how to determine the optimal cell number and sequencing depth for the experiment?
  - Under a fixed budget: breadth vs depth trade-off

Breadth vs. depth trade-off



A toy example of the breadth vs. depth trade-off under a **fixed budget**

Typical scRNA-seq data analysis



dim. reduction (for visualization) → clustering (& rare cell type detection) / trajectory inference → gene level analysis (e.g. DE gene identification). [Luecken and Theis 2019]

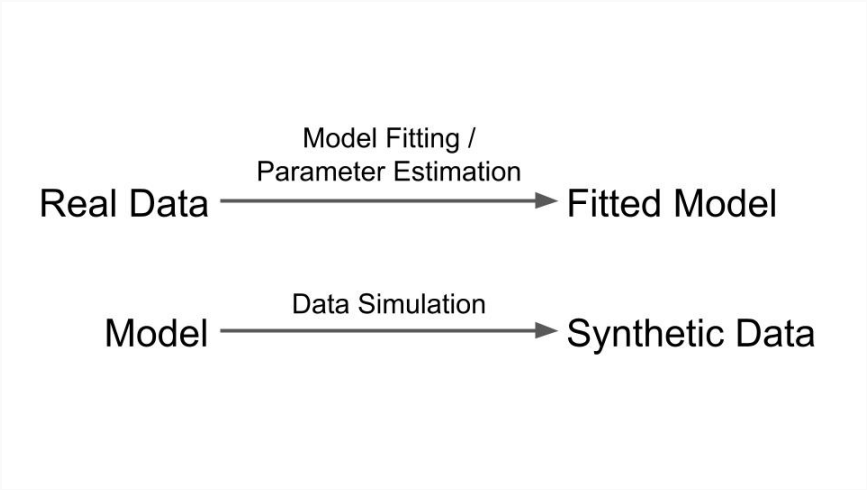
Computational benchmarking question

- How to choose among available computational methods?
  - Dimensionality reduction: PCA / t-SNE / UMAP / ZIFA / ...
  - Cell clustering: K-means / CIDR / SC3 / Seurat / ...
  - Rare cell type detection: RaceID / FiRE / GiniClust2 / GiniClust3 / ...
  - Trajectory inference: Slingshot / TSCAN / Monocle2 / destiny / ...
  - Identification of DE genes between cell types: SCDE / MAST / scDD / D3E / ...

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!**

Model fitting and data simulation

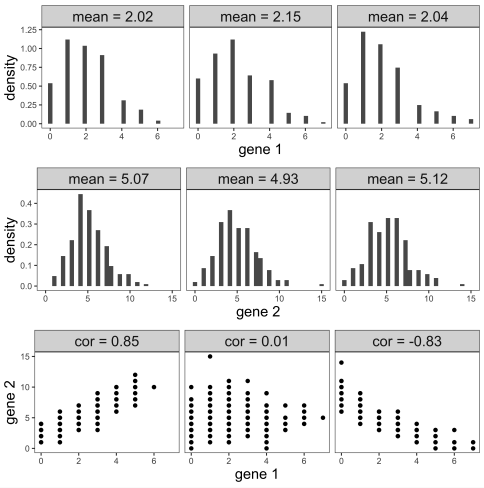


Model fitting and data simulation. A good model should capture key characteristics of real data.

Existing scRNA-seq simulators

Simulator \ Property	protocol adaptive	gene preserved	gene cor. captured	cell num. seq. dep. flexible	easy to interpret	comp. & sample efficient
dyngen	✓✗	×	×	✓✗	✓	✓
Lun2	✓✗	✓	×	✓	✓✗	✓
powsimR	✓	✓	×	✓	✓	✓
PROSST	✓✗	✓	×	✓✗	✓	✓
scDD	✓	×	×	✓✗	✓✗	✓
scDesign	✓	✓✗	×	✓	✓	✓
scGAN	✓	✓	✓✗	✓✗	×	×
splat simple	✓	×	×	×	✓	✓
splat	✓	×	×	×	✓	✓
kersplat	✓	×	✓✗	×	✓	✓
SPARSim	✓	✓	✓✗	×	✓	✓
SymSim	✓	×	×	×	✓	✓
ZINB-WaVE	✓	✓✗	✓✗	×	✓	✓
scDesign2	✓	✓	✓	✓	✓	✓

Why do gene correlations matter?



Correlation affect the joint distribution of genes.

Our proposal: scDesign2

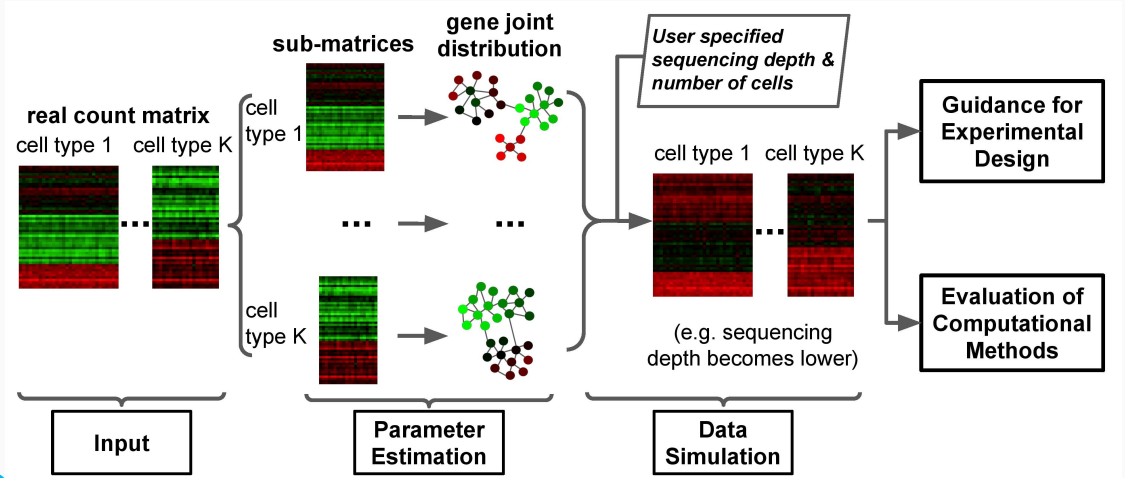
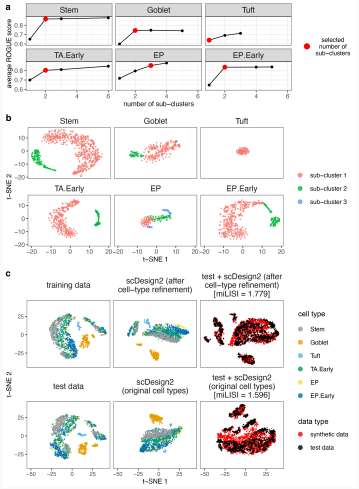


Diagram of scDesign2.

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 one parametric model of all genes' expression for each cell type** k
- For simplicity of notation, we drop the subscript k in the following discussion

scDesign2: cell type/cluster refinement by ROGUE scores



Application of ROGUE scores [Liu *et al.*, Nat Comm (2020)] combined with dimensionality reduction plots to refine cell types before training scDesign2

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.
- We assume that $X_{ij} \sim \mathrm{ZINB}(p_i, \mu_i, \psi_i)$, for gene $i = 1, \ldots, p$. That is, $Z_{ij} \sim \mathrm{Ber}(p_i)$, and $X_{ij} = 0$, if $Z_{ij} = 1$; $X_{ij} \sim \mathrm{NB}(\psi_i, \mu_i)$, if $Z_{ij} = 0$.

$$\mathbb{E}(X_{ij}|Z_{ij} = 0) = \mu_i$$

$$\mathrm{Var}(X_{ij}|Z_{ij} = 0) = \mu_i + \frac{(\mu_i)^2}{\psi_i}$$

- The Z_{ij}'s are unobserved
- The ZINB distribution is a general model that also includes Poisson, zero-inflated Poisson and NB

scDesign2: marginal distribution fitting for gene i

- Denote $\boldsymbol{X} \in \mathbb{N}^{p \times n}$ as the count matrix with p genes and n cells. (k dropped for simplicity).
- For $X_{i\cdot}$, if $\text{mean}(X_{i\cdot}) \geq \text{var}(X_{i\cdot})$,
  - Fit a Poisson distribution and ZIP distribution by maximum likelihood estimation (MLE) and perform a χ^2_1 likelihood-ratio test to determine if zero-inflation is significant
- Otherwise
  - Fit a NB distribution and ZINB distribution by MLE and perform a χ^2_1 likelihood-ratio test to determine if zero-inflation is significant
- The default p-value cutoff for the χ^2_1 test is 0.05

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 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 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 (DT): 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

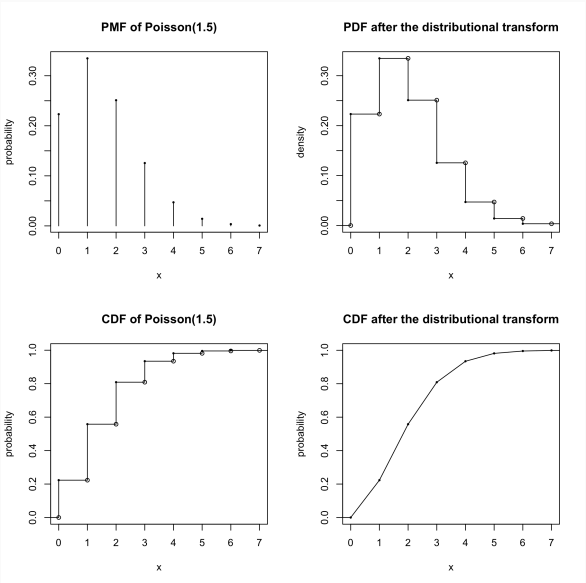
$$u_{ij} = v_{ij} F_i(X_{ij} - 1) + (1 - v_{ij}) F_i(X_{ij})$$

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

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

  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 $\boldsymbol{R}$

Effect of distributional transform



Gaussian copula correlation and gene Kendall's tau

- If we denote R_{hl} as the Gaussian copula correlation between genes h and l, i.e., the (h, l)-th entry of R, and τ_{hl} as the Kendall's tau between the same two genes on the original scale, i.e., $\tau_{hl} = \tau(X_{hj}, X_{lj})$, then we have the following relationship

$$R_{hl} = \sin\left(\frac{\pi}{2}\tau_{hl}\right)$$

- This relationship links the copula correlation with the Kendall's tau of the two original variables, thus providing an interpretation of the copula correlation
- It also suggests that $\boldsymbol{R}$ can be estimated by plugging the sample tau matrix into the above formula
- However, this estimate of $\boldsymbol{R}$ may not always be positive semidefinite. Therefore, we use another procedure to estimate $\boldsymbol{R}$

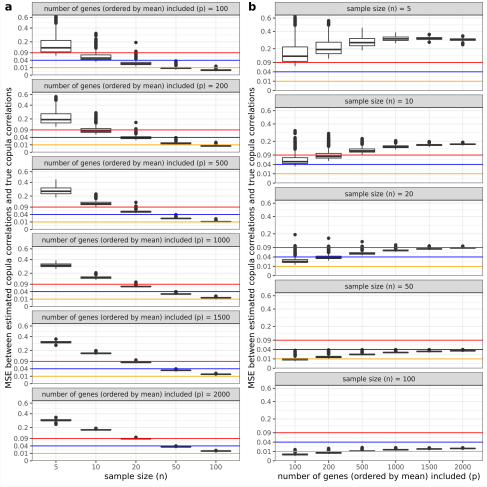
scDesign2: joint distribution fitting

- Denote $(\hat{p}_i, \hat{\mu}_i, \hat{\psi}_i)$ as the fitted marginal parameters for gene i, which also specifies the fitted CDF $\hat{F}_i$
- 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 $(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}))^{\mathsf{T}}$, $j = 1, \ldots, n$

Estimation of R - a high-dimensional problem



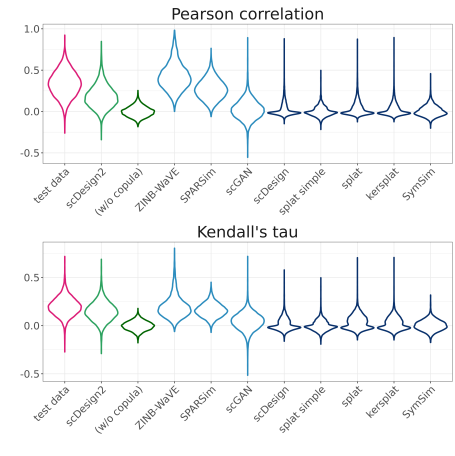
The estimation of ***R*** is a high-dimensional problem. We partially avoid it by only estimating for the top highly to moderately expressed genes

scDesign2: data simulation

- Input: model parameters (one for each cell type), cell type proportions, number of cells to simulate, total number of reads in the simulated data

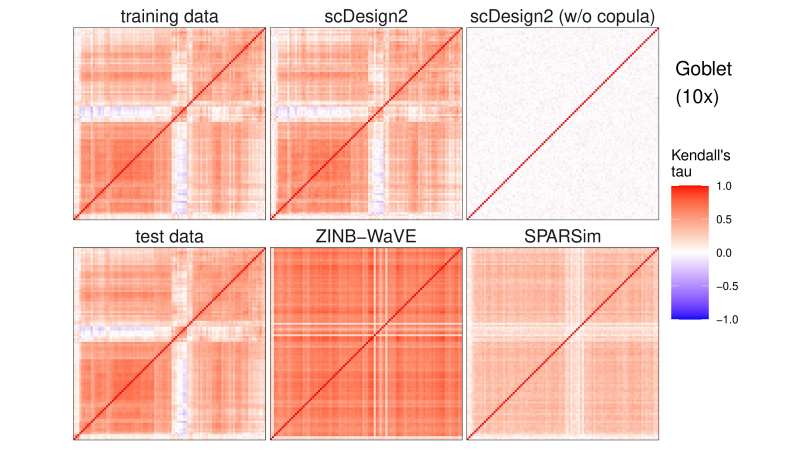
- Simulation:
  1. Determine the cell numbers for each cell type
  2. Compute a scaling factor r as the proportion of the average number of reads in each cell in the simulated data to the average number of reads in the original data
  3. Scale the mean parameters by r
  4. Simulate data for each cell type, and then combine the simulated data together as one data matrix

scDesign2 vs. existing scRNA-seq simulators



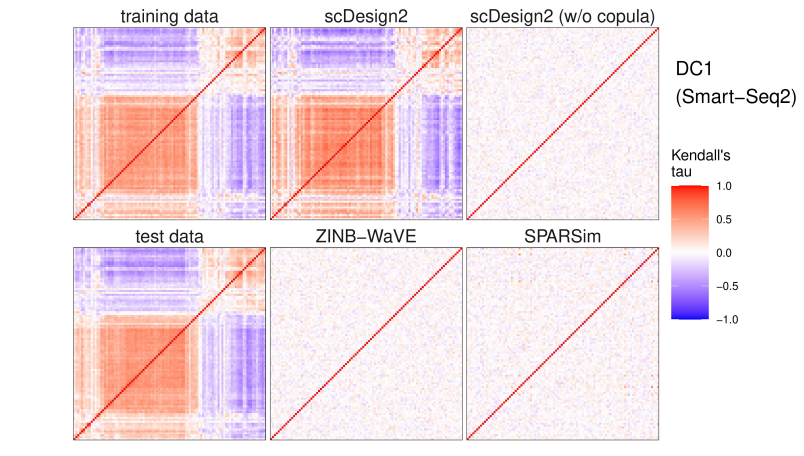
Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber *et al.*, Nature (2017)]

scDesign2 vs. existing scRNA-seq simulators



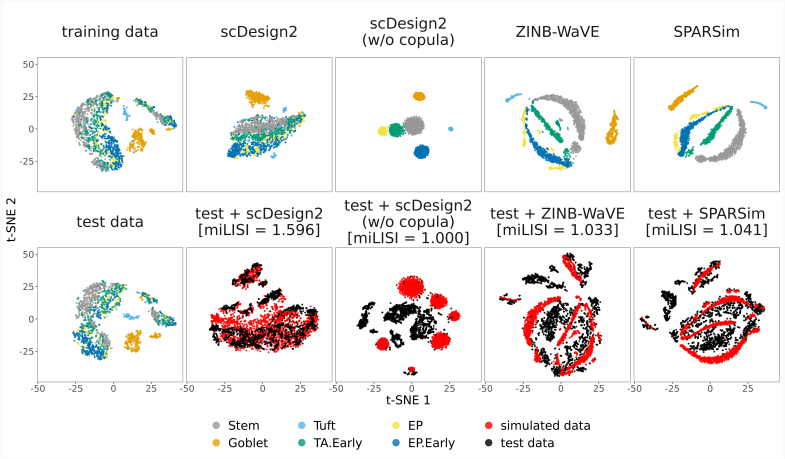
Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber *et al.*, Nature (2017)]

scDesign2 vs. existing scRNA-seq simulators



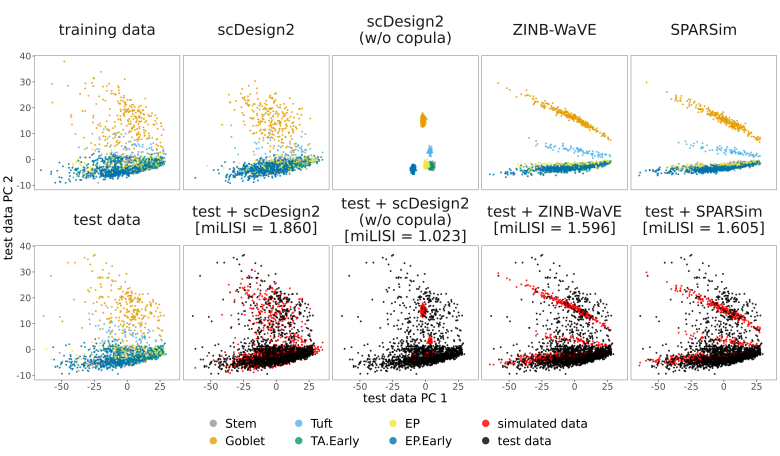
Data: dendrocytes subtype 1 of human blood by Smart-Seq2 [Villani *et al.*, Science (2017)]

scDesign2 vs. existing scRNA-seq simulators



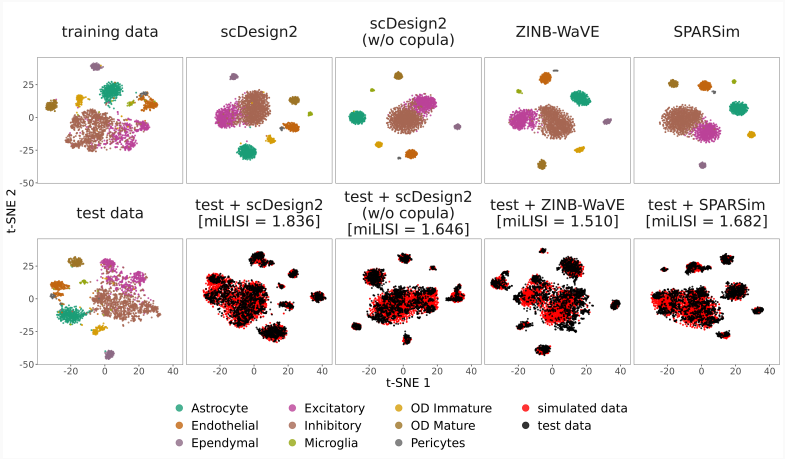
Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber *et al.*, Nature (2017)]

scDesign2 vs. existing scRNA-seq simulators



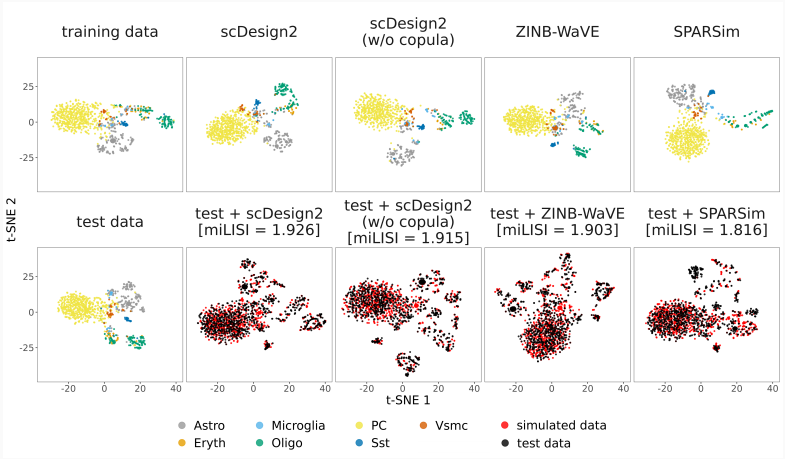
Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber *et al.*, Nature (2017)]

Application 1: simulation for other single-cell technologies



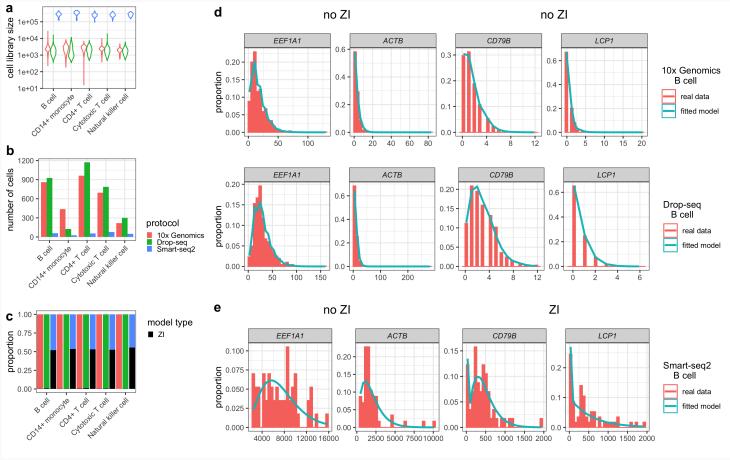
Data: mouse hypothalamic preoptic region by MERFISH [Moffitt *et al.*, Science (2018)]

Application 1: simulation for other single-cell technologies



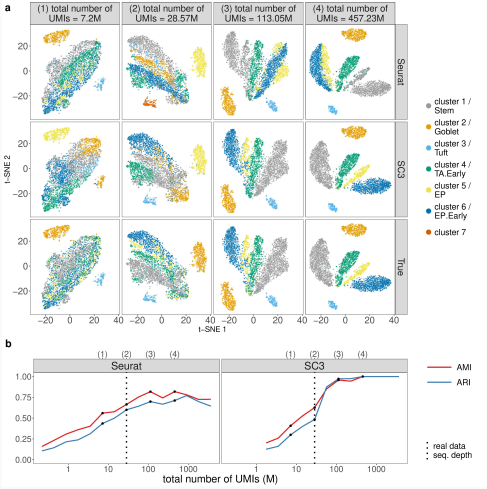
Data: mouse hippocampal area CA1 by pciSeq [Qian *et al.*, Nature Methods (2020)]

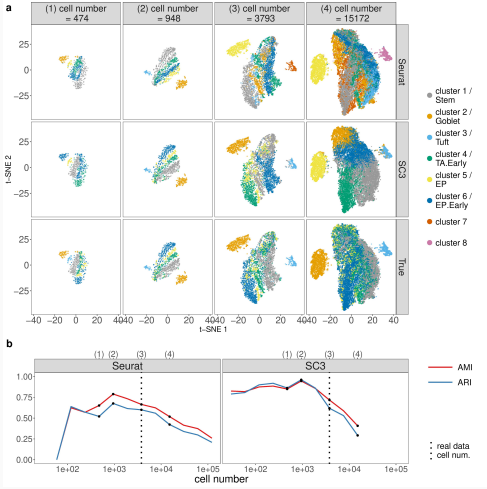
Application 2: evaluation of scRNA-seq protocols



Data: five cell types of PBMC by three different protocols [Ding *et al.*, Nature Biotechnology (2020)]

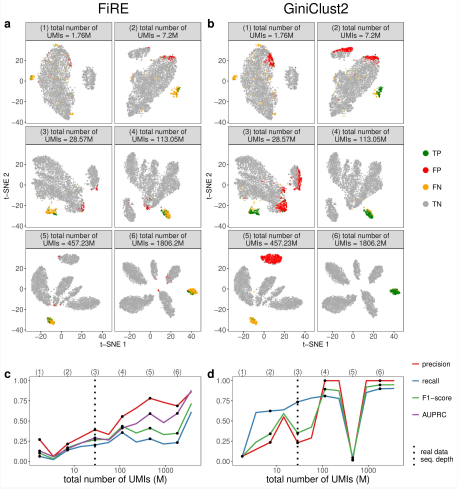
Application 3: clustering

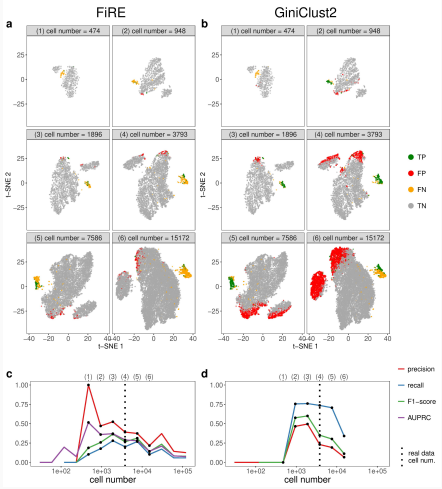




Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber *et al.*, Nature (2017)]

Application 4: rare cell type detection





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

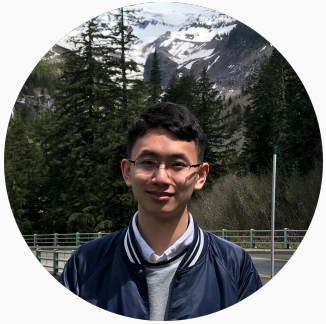
scDesign2 summary

- **scDesign2**: an interpretable simulator that generates realistic single-cell gene expression count data with gene correlations
  - Motivated by our previous work scDesign (Li and Li, Bioinformatics 2019)
  - A multi-gene generative model (probabilistic, transparent, interpretable)
  - Guidance for scRNA-seq experimental design
  - Benchmarking of computational methods
- R package: `https://github.com/JSB-UCLA/scDesign2`
- Future work
  - Extend the current model to accommodate continuous cell trajectories

Learning more about copulas

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

Acknowledgements



Dongyuan Song
(Ph.D. student, UCLA)

Tianyi Sun
(Ph.D. student, UCLA)

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