



scDesign3: single-cell and spatial omics simulator

benchmarking, inference & in silico controlled experiments

Jingyi Jessica Li

Associate Professor

Junction of Statistics and Biology (<http://jsb.ucla.edu>)

Department of Statistics

University of California, Los Angeles

Single-cell and spatial omics data: statistical characteristics

Processed data: a cell-by-feature matrix + cell covariates



Single-cell and spatial omics data: statistical characteristics

Processed data: a cell-by-feature matrix + cell covariates

Cell heterogeneity structures

- discrete cell types (known or latent)
- continuous trajectories (usually latent)
- spatial locations (known for spatial data)



Single-cell and spatial omics data: statistical characteristics

Processed data: a cell-by-feature matrix + cell covariates

Cell heterogeneity structures

- discrete cell types (known or latent)
- continuous trajectories (usually latent)
- spatial locations (known for spatial data)

Experimental designs

- batches (unwanted effects)
- conditions (biological signals)



Single-cell and spatial omics data: statistical characteristics

Processed data: a cell-by-feature matrix + cell covariates

Cell heterogeneity structures

- discrete cell types (known or latent)
- continuous trajectories (usually latent)
- spatial locations (known for spatial data)

Experimental designs

- batches (unwanted effects)
- conditions (biological signals)

Features

- gene expression (scRNA-seq, spatial transcriptomics, etc.)
- chromatin accessibility (scATAC-seq, SNARE-seq, etc.)
- protein abundance (CITE-seq, etc.)



Computational benchmarking

- > 1000 computational tools at www.scrna-tools.org
- how to choose among competing computational tools?



Motivations

Computational benchmarking

- > 1000 computational tools at www.scrna-tools.org
- how to choose among competing computational tools?

Inference

Conditional on a cell covariate (type, pseudotime, or spatial location)

- every gene's distribution
- every gene pair's correlation



Motivations

Computational benchmarking

- > 1000 computational tools at www.scrna-tools.org
- how to choose among competing computational tools?

Inference

Conditional on a cell covariate (type, pseudotime, or spatial location)

- every gene's distribution
- every gene pair's correlation

In silico controlled experiments

- negative control: to evaluate a pipeline's **false discoveries**
- positive control: to evaluate a pipeline's **discovery power**



Motivations

Computational benchmarking

- > 1000 computational tools at www.scrna-tools.org
- how to choose among competing computational tools?

Inference

Conditional on a cell covariate (type, pseudotime, or spatial location)

- every gene's distribution
- every gene pair's correlation

In silico controlled experiments

- negative control: to evaluate a pipeline's **false discoveries**
- positive control: to evaluate a pipeline's **discovery power**

A realistic simulator with interpretable parameters



A statistical simulator scDesign for rational scRNA-seq experimental design

Wei Vivian Li, Jingyi Jessica Li 

Bioinformatics, Volume 35, Issue 14, July 2019, Pages i41–i50,

<https://doi.org/10.1093/bioinformatics/btz321>

Published: 05 July 2019

scDesign pros:

- interpretable parameters
- variable cell number
- variable sequencing depth



Cell Systems

Volume 12, Issue 2, 17 February 2021, Pages 176-194.e6



Article

Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data

Nan Miles Xi ¹, Jingyi Jessica Li ^{1, 2, 3, 4}  



Cell Systems

Volume 12, Issue 2, 17 February 2021, Pages 176-194.e6



Article

Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data

Nan Miles Xi ¹, Jingyi Jessica Li ^{1, 2, 3, 4}  

scDesign cons:

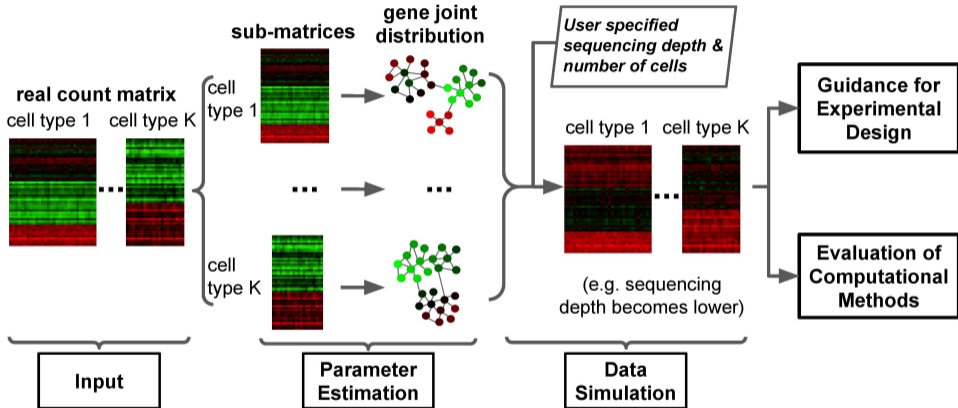
- cannot capture gene correlations
- does not directly model count data



Exemplar scRNA-seq simulators and properties

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		✓	✓	✓	✓	✓	✓





Related work:

SPsimSeq [Assefa *et al.*, *Bioinformatics*, 2020]; ESCO [Tian *et al.*, *Bioinformatics*, 2021]



scDesign2: notations

- Denote the scRNA-seq count matrix as $\mathbf{X} \in \mathbb{N}^{p \times n}$, with p genes and n cells
- Assume that \mathbf{X} contains K cell types and the cell memberships are known in advance
- Suppose there are $n^{(k)}$ cells in cell type k , $k = 1, \dots, K$, and denote the count matrix for cell type k as $\mathbf{X}^{(k)}$
- 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 $X_j = (X_{1j}, \dots, X_{pj}) \in \mathbb{N}^p$ as the gene expression vector for cell j , $j = 1, \dots, n$. We assume that the X_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 highly-expressed genes

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

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

- 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 $\text{Uniform}[0, 1]$ independently for $i = 1, \dots, p$ and $j = 1, \dots, n$
 - 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_j as

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

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



scDesign2: joint distribution fitting

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

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

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



scDesign2: data simulation

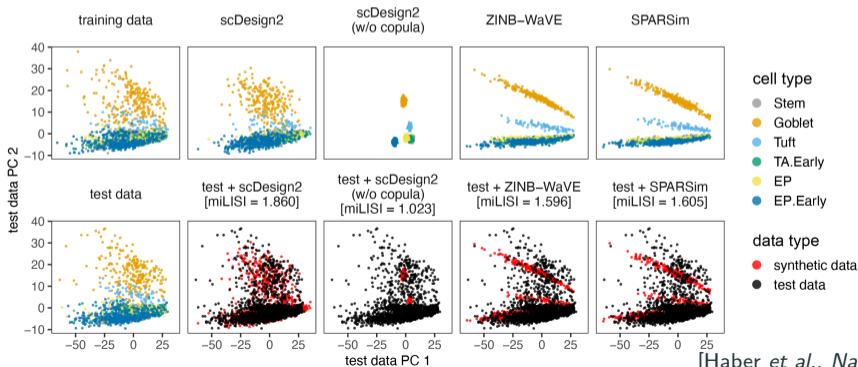
- **Input from previous step:**
 - fitted joint gene distributions (one per cell type)
 - cell type proportions
- **User-specified input:**
 - number of cells to simulate
 - total sequencing depth
- **Output:**
 - a synthetic gene-by-cell count matrix with K cell types
 - fitted model parameters



scDesign2: summary

A multi-gene probabilistic model **per cell type**

- Each gene \sim count distribution $\in \{\text{Poisson, negative binomial, ZIP, ZINB}\}$
- Gene correlations estimated via **Gaussian copula**



A multi-gene probabilistic model **per cell type**

- Each gene \sim count distribution \in {Poisson, negative binomial, ZIP, ZINB}
- Gene correlations estimated via **Gaussian copula**

Method | [Open Access](#) | [Published: 25 May 2021](#)

scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

[Tianyi Sun](#), [Dongyuan Song](#), [Wei Vivian Li](#)  & [Jingyi Jessica Li](#) 

[Genome Biology](#) **22**, Article number: 163 (2021) | [Cite this article](#)

5654 Accesses | **8** Citations | **31** Altmetric | [Metrics](#)

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 1, 2022
© Mary Ann Liebert, Inc.
Pp. 1–4
DOI: [10.1089/cmb.2021.0440](https://doi.org/10.1089/cmb.2021.0440)

RECOMB 2021

Simulating Single-Cell Gene Expression Count Data with Preserved Gene Correlations by scDesign2

TIANYI SUN¹, DONGYUAN SONG², WEI VIVIAN LI³ and JINGYI JESSICA LI^{1,3}



A multi-gene probabilistic model **per cell type**

- Each gene \sim count distribution \in {Poisson, negative binomial, ZIP, ZINB}
- Gene correlations estimated via **Gaussian copula**

Method | [Open Access](#) | [Published: 25 May 2021](#)

scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

[Tianyi Sun](#), [Dongyuan Song](#), [Wei Vivian Li](#)  & [Jingyi Jessica Li](#) 

[Genome Biology](#) 22, Article number: 163 (2021) | [Cite this article](#)

5654 Accesses | 8 Citations | 31 Altmetric | [Metrics](#)

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 1, 2022
© Mary Ann Liebert, Inc.
Pp. 1–4
DOI: 10.1089/cmb.2021.0440

RECOMB 2021

Simulating Single-Cell Gene Expression Count Data
with Preserved Gene Correlations by scDesign2

TIANYI SUN¹, DONGYUAN SONG², WEI VIVIAN LI³ and JINGYI JESSICA LI^{1,3}

Generalization: the marginal distribution of gene i in cell $j \sim$ cell j 's covariates
Example covariates: cell type, cell pseudotime, spatial coordinates, condition, batch, ...

Generalized additive model + negative binomial distribution



A multi-gene probabilistic model **per cell type**

- Each gene \sim count distribution \in {Poisson, negative binomial, ZIP, ZINB}
- Gene correlations estimated via **Gaussian copula**

Method | [Open Access](#) | [Published: 25 May 2021](#)

scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

[Tianyi Sun](#), [Dongyuan Song](#), [Wei Vivian Li](#)  & [Jingyi Jessica Li](#) 

[Genome Biology](#) 22, Article number: 163 (2021) | [Cite this article](#)

5654 Accesses | 8 Citations | 31 Altmetric | [Metrics](#)

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 1, 2022
© Mary Ann Liebert, Inc.
Pp. 1–4
DOI: 10.1089/cmb.2021.0440

RECOMB 2021

Simulating Single-Cell Gene Expression Count Data
with Preserved Gene Correlations by scDesign2

TIANYI SUN¹, DONGYUAN SONG², WEI VIVIAN LI³ and JINGYI JESSICA LI^{1,3}

Generalization: the marginal distribution of gene i in cell $j \sim$ cell j 's covariates
Example covariates: cell type, cell pseudotime, spatial coordinates, condition, batch, ...

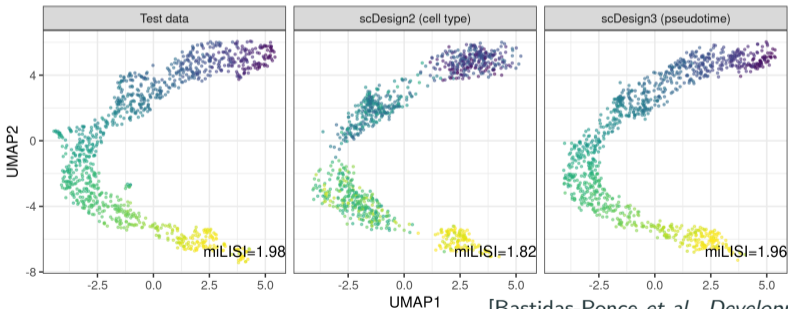
Generalized additive model + negative binomial distribution



scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states:** continuous trajectory & discrete cell types
- **Feature modalities:** RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood:** vine copula [Joe and Kurowicka, 2011]

Example: continuous trajectory (pancreatic cell differentiation)



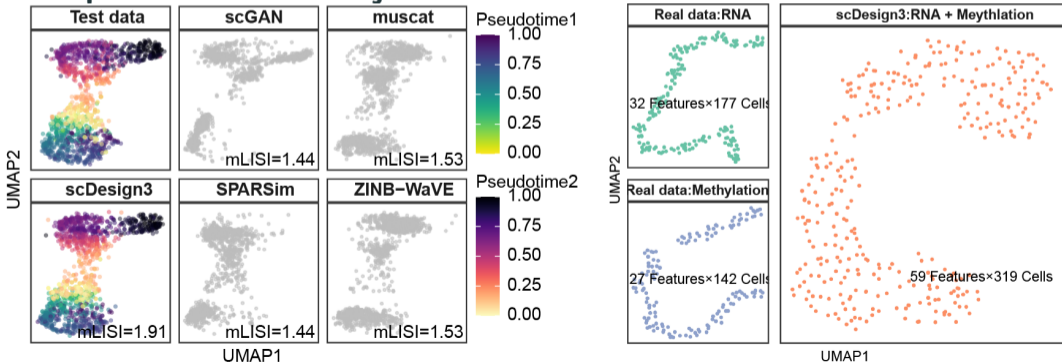
[Bastidas-Ponce *et al.*, *Development*, 2019]



scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states:** continuous trajectory & discrete cell types
- **Feature modalities:** RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood:** vine copula [Joe and Kurowicka, 2011]

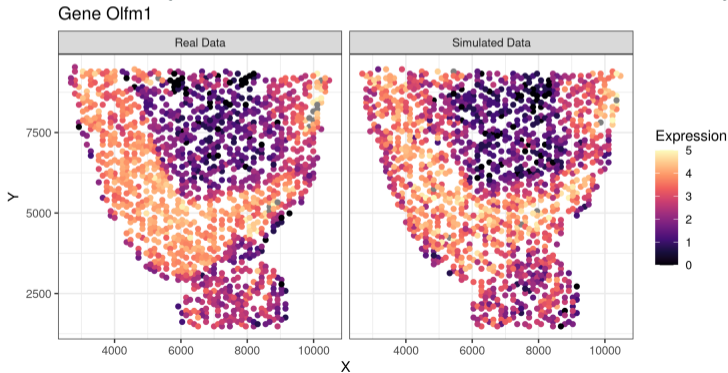
Examples: bifurcation trajectories & multiomics



scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states:** continuous trajectory & discrete cell types
- **Feature modalities:** RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood:** vine copula [Joe and Kurowicka, 2011]

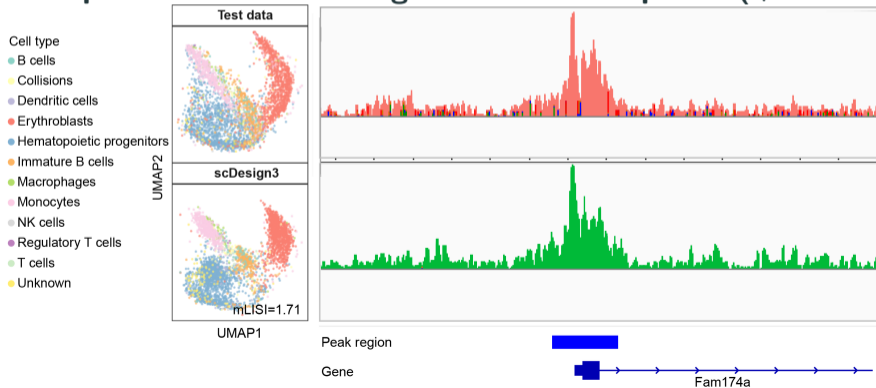
Example: spatial data (brain region measured by 10X Visium)



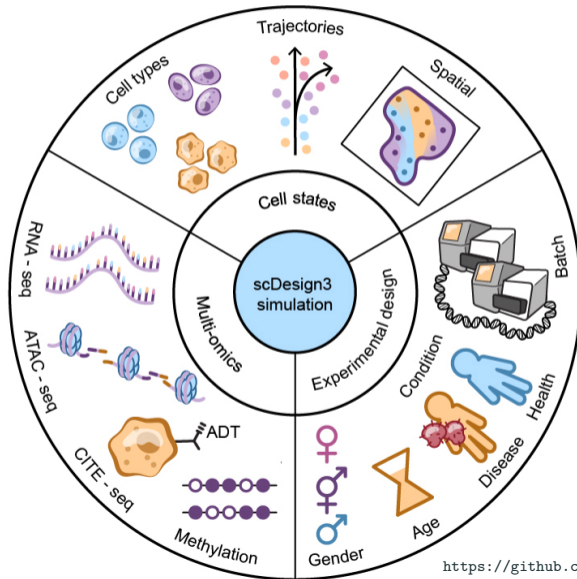
scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states:** continuous trajectory & discrete cell types
- **Feature modalities:** RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood:** vine copula [Joe and Kurowicka, 2011]

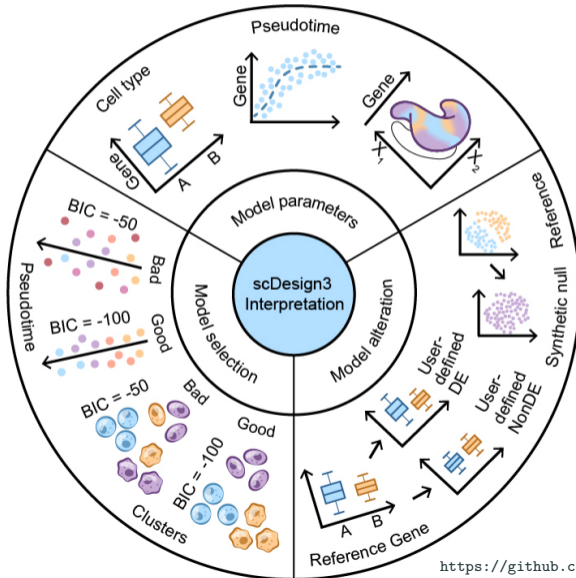
Example: bone marrow single-cell ATAC-seq data (+ scReadSim)



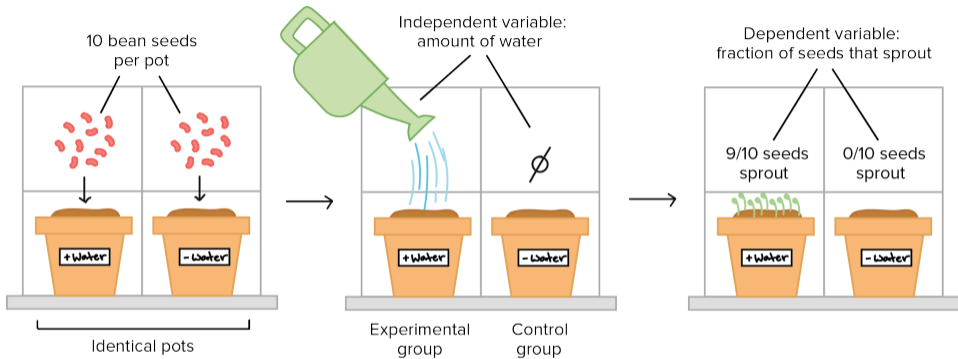
scDesign3 functionalities (simulation)



scDesign3 functionalities (interpretation)



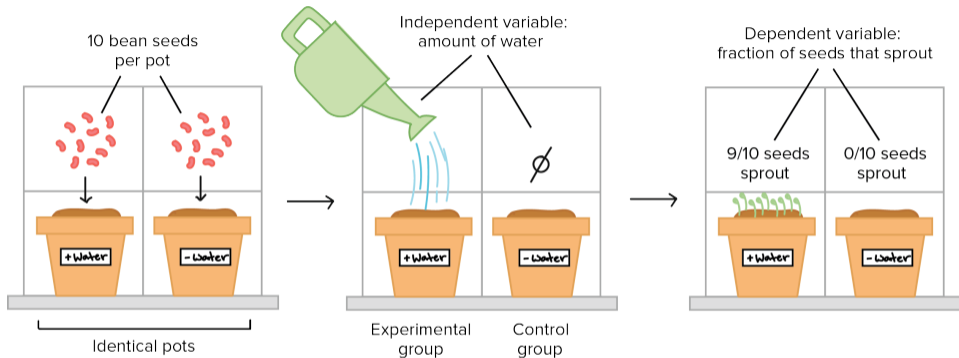
Why need in silico controlled experiments?



<https://www.khanacademy.org/science/biology/intro-to-biology/science-of-biology/a/experiments-and-observations>



Why need in silico controlled experiments?



<https://www.khanacademy.org/science/biology/intro-to-biology/science-of-biology/a/experiments-and-observations>

Double-dipping challenges in single-cell inference

- Cell clustering + DEG identification



DEGs between inferred cell clusters from single-cell RNA-seq data

ClusterDE (cell clustering + DEG identification between cell clusters)

– existing methods assume **Gaussian** distributions

 TN test [Zhang, Kamath, and Tse, *Cell Syst*, 2019]

 clusterpval [Gao, Bien, and Witten, *arXiv*, 2020]

or require **count splitting**

 countsplit [Neufeld, Gao, Popp, Battle, and Witten, *arXiv*, 2022]



DEGs between inferred cell clusters from single-cell RNA-seq data

ClusterDE (cell clustering + DEG identification between cell clusters)

- existing methods assume **Gaussian** distributions

 TN test [Zhang, Kamath, and Tse, *Cell Syst*, 2019]

 clusterpval [Gao, Bien, and Witten, *arXiv*, 2020]

or require **count splitting**

 countsplit [Neufeld, Gao, Popp, Battle, and Witten, *arXiv*, 2022]

Our proposal: scDesign3 + Clipper

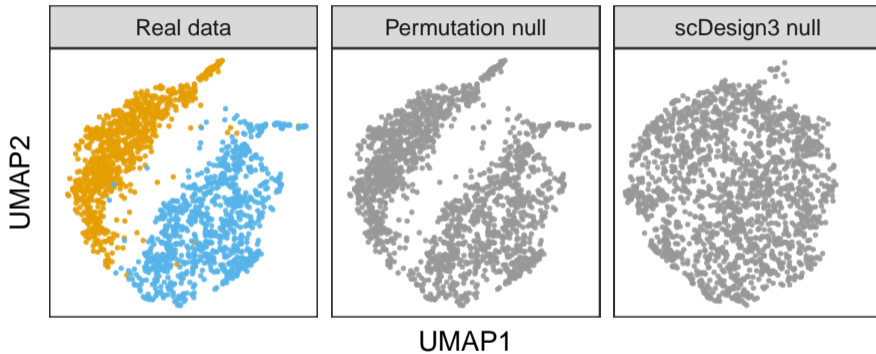
- inspired by

 gap statistic [Hastie, Tibshirani, and Walther, *JRSSB*, 2002]

 knockoffs [Barber and Candès, *Ann Stat*, 2015]



scDesign3: in silico negative control



Cell type ● Naive cytotoxic T cell ● Regulatory T cell ● Null



Clipper: a p-value-free FDR control framework

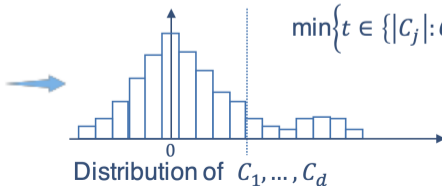
- **NO requirement of**
 - high-resolution p-values
 - parametric distributions
 - large sample sizes
- Foundation: knockoffs
- Two components:
 - **contrast scores**
 - **cutoff**



Goal: marginal screening for **interesting** features
 d features FDR threshold q

Contrast scores

C_1
 \vdots
 C_d



Contrast score cutoff

$$\min \left\{ t \in \{ |C_j| : C_j \neq 0 \} : \frac{1 + \#\{j : C_j \leq -t\}}{\#\{j : C_j \geq t\} \vee 1} \leq q \right\}$$



Method | [Open Access](#) | [Published: 11 October 2021](#)

Clipper: p -value-free FDR control on high-throughput data from two conditions

[Xinzhou Ge](#), [Yiling Elaine Chen](#), [Dongyuan Song](#), [MeiLu McDermott](#), [Kyla Woysner](#), [Antigoni Manousopoulou](#), [Ning Wang](#), [Wei Li](#), [Leo D. Wang](#) & [Jingyi Jessica Li](#) 

[Genome Biology](#) **22**, Article number: 288 (2021) | [Cite this article](#)

6865 Accesses | **5** Citations | **52** Altmetric | [Metrics](#)



Clipper offers a general p-value-free FDR control solution

Key: contrast score construction

example	target data (experiment)	null data (negative control)
ClusterDE	actual data	scDesign3 simulated data

Contrast score of feature $j = 1, \dots, d$, the

$$C_j := t(\mathbf{target\ data}) - t(\mathbf{null\ data}),$$

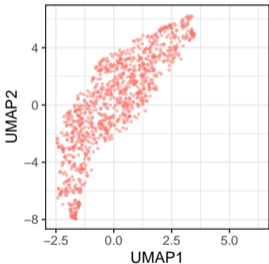
where $t(\cdot)$ is a summary statistic — can be a **complex pipeline**



ClusterDE: scDesign3 + Clipper (preliminary)

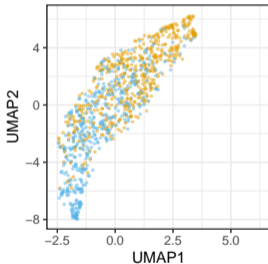
Complete null case: no cell clusters

Real Data



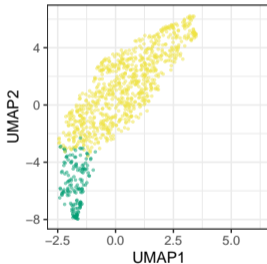
Cell_Type ● naive.cytotoxic

Seurat Clustering



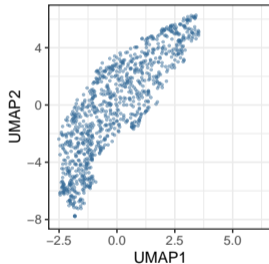
Seurat_Clusters ● 0 ● 1

Kmeans Clustering



Kmeans_Clusters ● 0 ● 1

Null Data by scDesign3

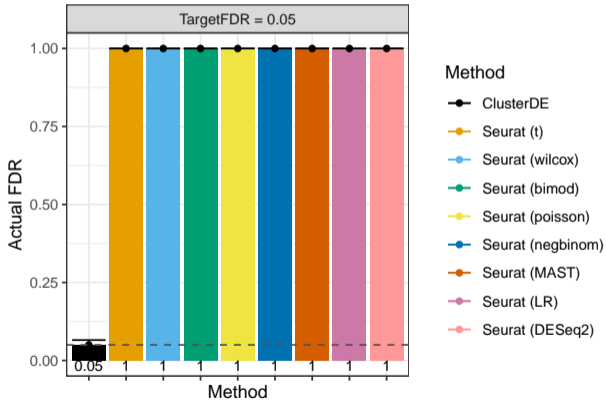
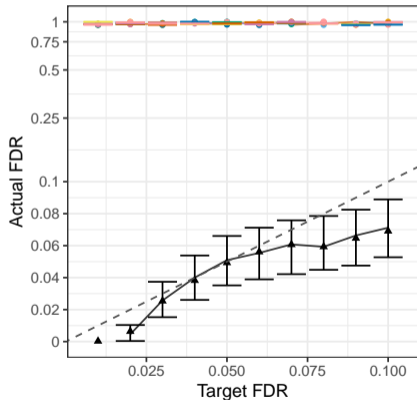


[Zheng et al., Nat Commun, 2017]



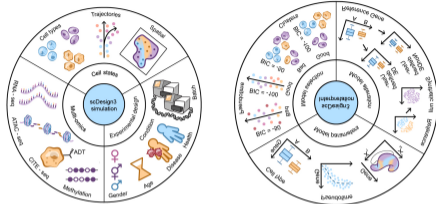
ClusterDE: scDesign3 + Clipper (preliminary)

Complete null case: no cell clusters



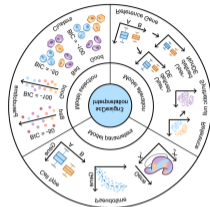
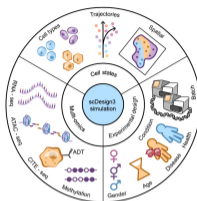
Take-home messages

- **scDesign3 usages**
 - Method benchmarking
 - Parameter inference
 - In silico controlled data generation



Take-home messages

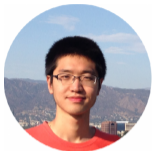
- **scDesign3 usages**
 - Method benchmarking
 - Parameter inference
 - In silico controlled data generation
- **Double dipping** is ubiquitous in genomic data science
Statistical inference is often NOT the first step of a pipeline



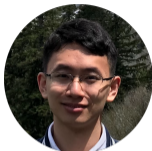
Acknowledgements



Wei Vivian Li
(former Ph.D. student
Assist. Prof. @ Rutgers)
scDesign



Tianyi Sun
(Ph.D. student)
scDesign2



Dongyuan Song
(Ph.D. student)
scDesign3



Guan'ao Yan
(Ph.D. student)
scReadSim



Xinzhou Ge
(Postdoc)
Clipper



Kexin Li
(Ph.D. student)
scDesign3+
Clipper



Alfred P. Sloan
FOUNDATION

