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.)

Motivations

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

Importance of benchmarking and in silico negative control

Teaser: false discoveries of DESeq2 and edgeR on population RNA-seq samples

Short Report | Open Access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li ✉ & Jingyi Jessica Li ✉

Genome Biology **23**, Article number: 79 (2022) | Cite this article

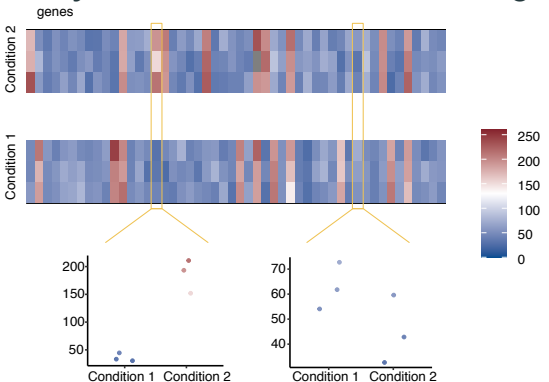
14k Accesses | **185** Altmetric | Metrics

— collaboration with Dr. Yumei Li in Dr. Wei Li's lab (UC Irvine)

Teaser: identifying differentially expressed genes (DEGs)

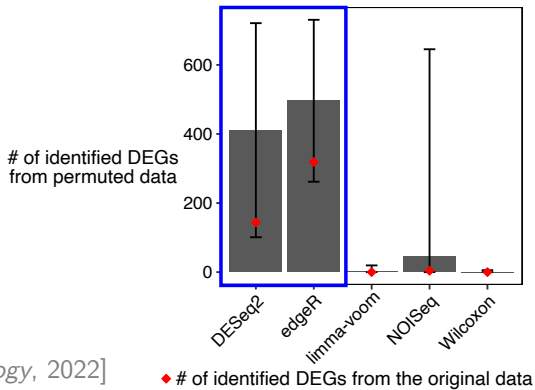
- Popular software (originally designed for **small** sample sizes):
  - `edgeR` [Robinson *et al.*, *Bioinformatics*, 2014]; cited $\sim$ 24K times
  - `DESeq2` [Love *et al.*, *Genome Biol*, 2014]; cited $>$ 33K times

  both assume a **negative binomial** distribution per gene and condition & use **empirical Bayes** to borrow information across genes



Teaser: in silico negative control by permutation

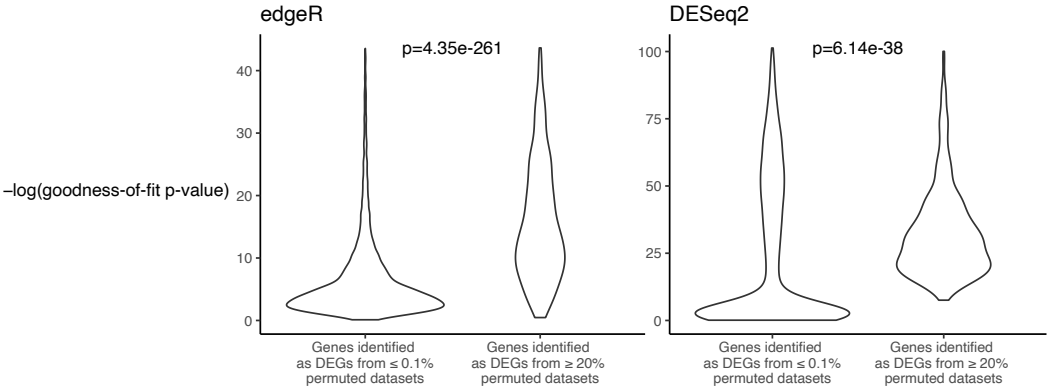
- 51 pre-nivolumab and 58 on-nivolumab anti-PD-1 therapy patients [Riaz *et al.*, *Cell*, 2017]
- Permute samples between conditions (no true DEGs)



[Li et al., *Genome Biology*, 2022]

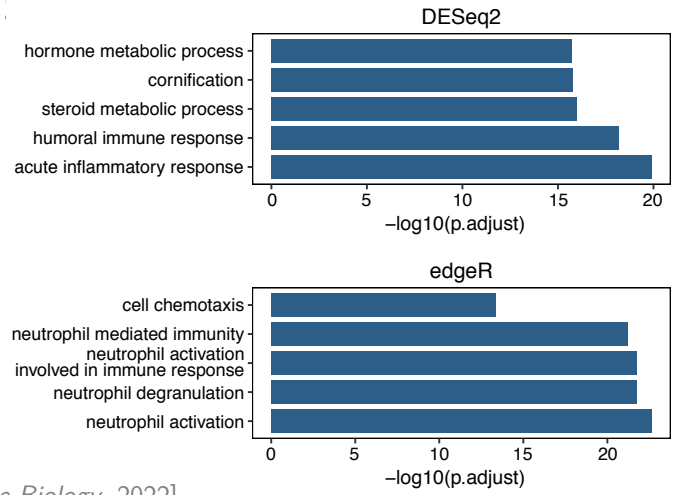
Teaser: model mis-specification

- Poor fit of **negative binomial model** $\longleftrightarrow$ false positive DEGs



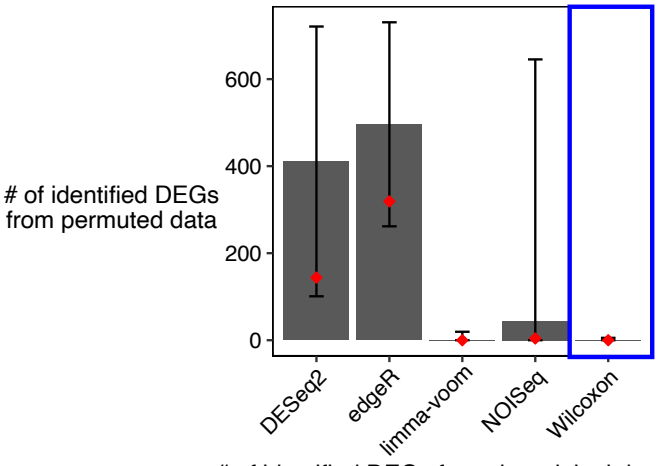
[Li et al., *Genome Biology*, 2022]

Teaser: false positive DEGs mislead scientific discoveries



[Li et al., *Genome Biology*, 2022]

Teaser: popular bioinformatics tools vs. classic statistical methods



[Li et al., *Genome Biology*, 2022]

@jsb_ucla

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

Use scDesign to benchmark doublet-detection methods

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]

Use scDesign to benchmark doublet-detection methods

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]

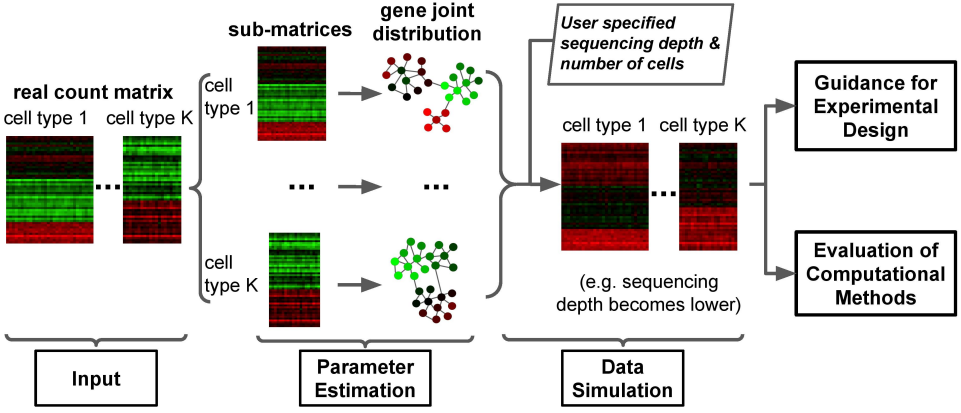
scDesign cons:

- 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	✓	✓	✓	✓	✓	✓

scDesign2



Related work:

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

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 highly-expressed 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

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

where $\boldsymbol{\Phi}_p(\cdot | \boldsymbol{R})$ 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 $\widehat{F}_i$ 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

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

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

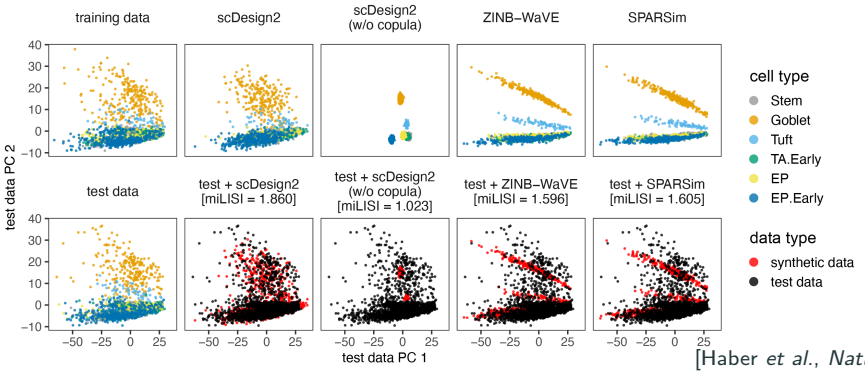
scDesign2: data simulation

- 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

scDesign2: summary

A multi-gene probabilistic model per cell type

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



[Haber *et al.*, *Nature*, 2017]

scDesign2: summary

A multi-gene probabilistic model per cell type

- Each gene $\sim$ count distribution $\in \{\text{Poisson}, \text{negative binomial}, \text{ZIP}, \text{ZINB}\}$
- 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

5144 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

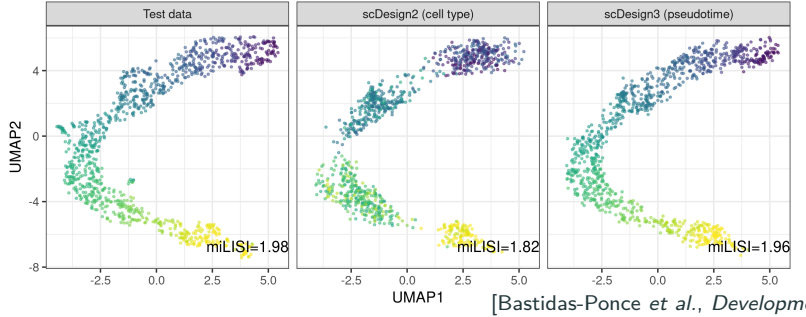
Simulating Single-Cell Gene Expression Count Data with Preserved Gene Correlations by scDesign2

TIANYI SUN,[1] DONGYUAN SONG,[2] WEI VIVIAN LI,[3] and JINGYI JESSICA LI[1,4]

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)

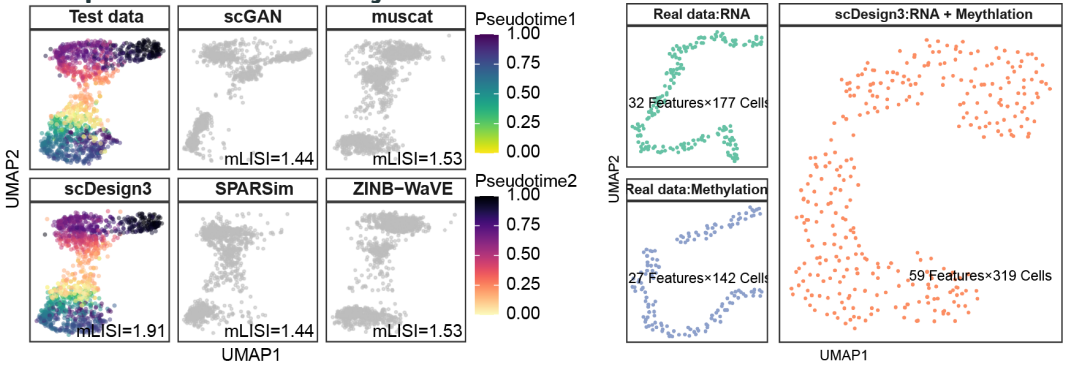


[Bastidas-Ponce *et al.*, *Development*, 2019]

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]

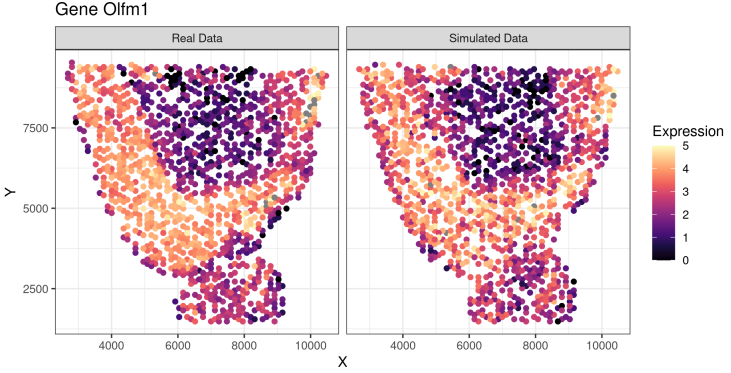
Examples: bifurcation trajectories & multiomics



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]

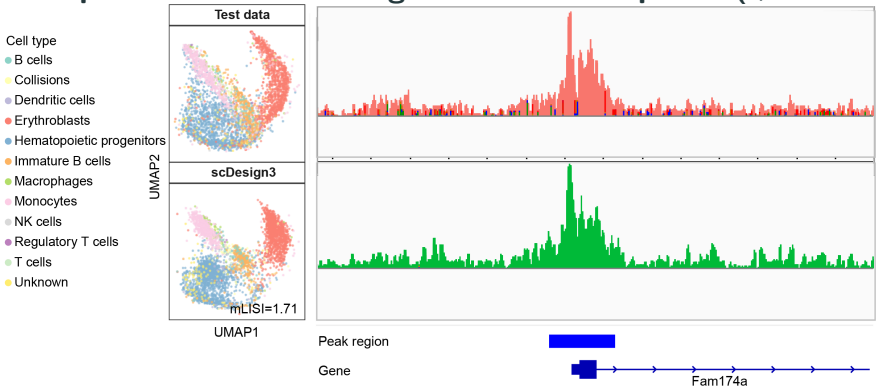
Example: spatial data (brain region measured by 10X Visium)



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)



by Guan'ao Yan

scDesign3 functionalities

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.)

scDesign3 usages

Computational benchmarking

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

Inference

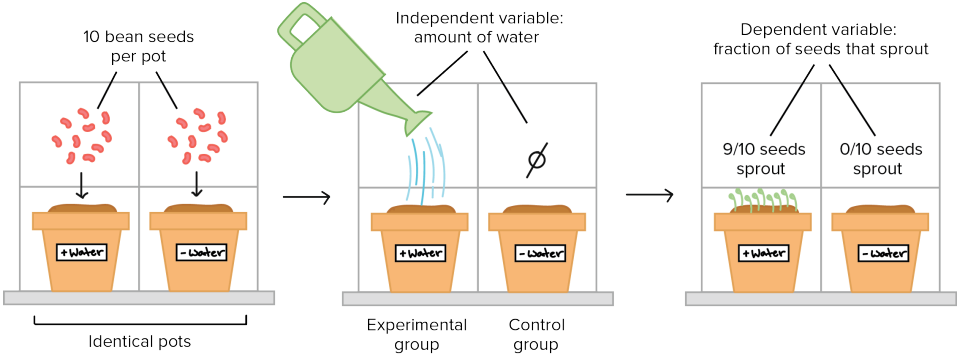
Conditional on a cell covariate (type, pseudotime, or spatial location)

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

In silico controlled experiments

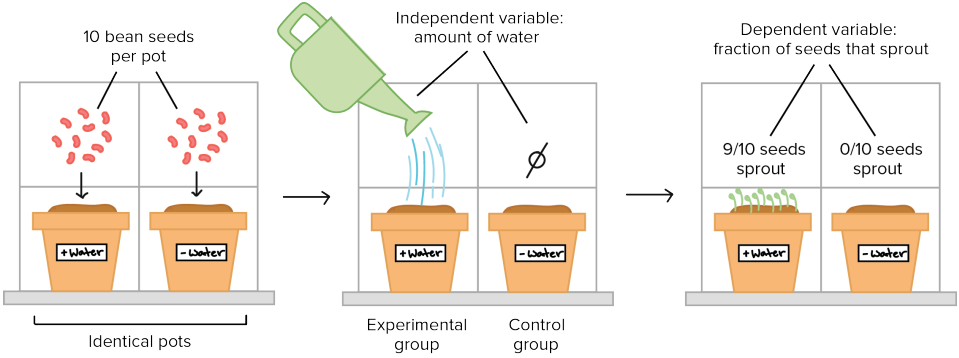
- negative control: to evaluate a pipeline's false discoveries
- positive control: to evaluate a pipeline's discovery power

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?



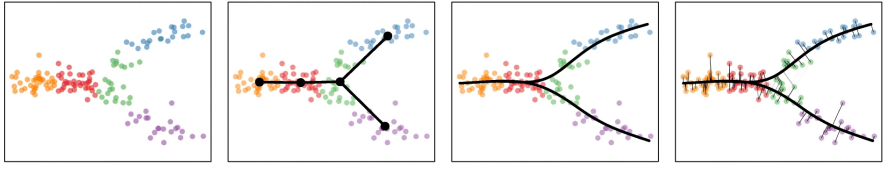
https://www.khanacademy.org/science/biology/intro-to-biology/science-of-biology/a/experiments-and-observations

Double-dipping challenges in single-cell inference

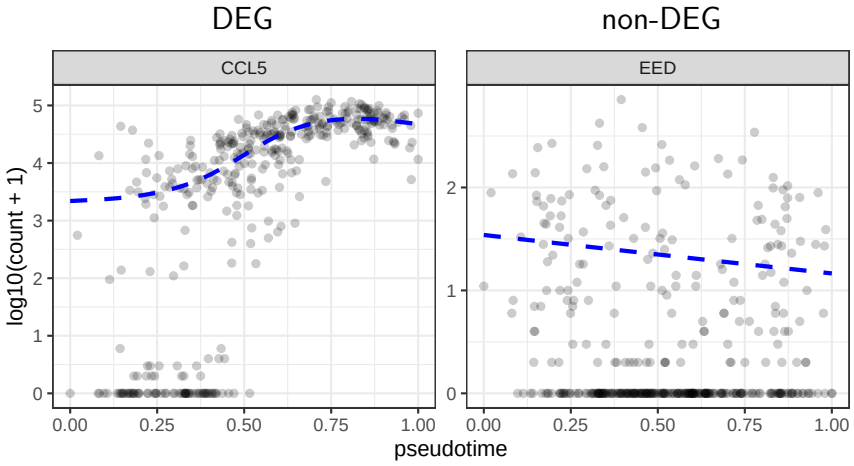
- **Cell pseudotime inference** + DEG identification
- **Cell clustering** + DEG identification

DEGs along inferred pseudotime from single-cell RNA-seq data

- **Cell 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 based on transcriptome similarities
- Popular software:
  - `Monocle3` [Trapnell *et al.*, *Nat Biotechnol*, 2014]; cited $>$ 2.8K times
  - `Slingshot` [Street *et al.*, *BMC Bioinform*, 2018]; cited 700 times

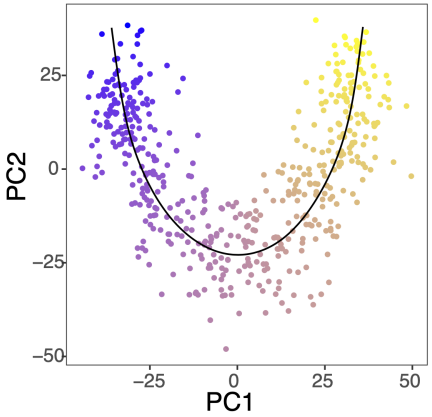


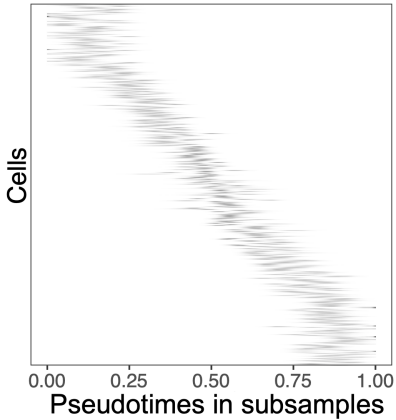
DEGs along inferred pseudotime from single-cell RNA-seq data



DEGs along inferred pseudotime from single-cell RNA-seq data

- Cell pseudotime is inferred from the same data and thus random





DEGs along inferred pseudotime from single-cell RNA-seq data

- However, existing methods treat cell pseudotime as an observed covariate

DEGs along inferred pseudotime from single-cell RNA-seq data

- However, existing methods treat cell pseudotime as an observed covariate
- Our solution: **PseudotimeDE** considers the uncertainty of pseudotime

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

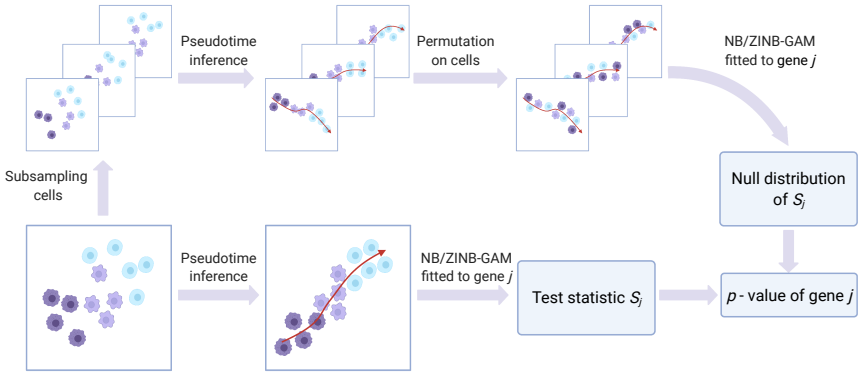
Dongyuan Song & Jingyi Jessica Li

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

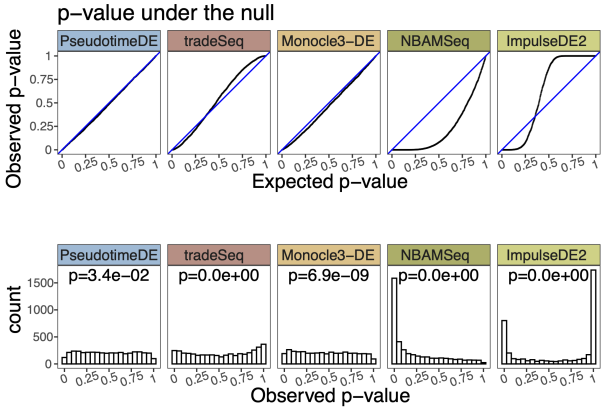
PseudotimeDE

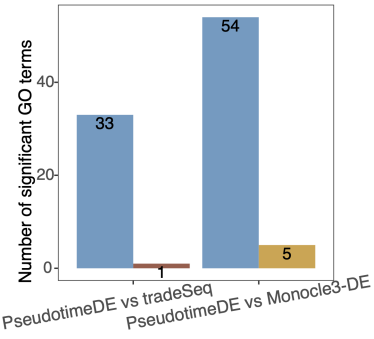
Generalized additive model (GAM): powerful test statistic

Subsampling + **pseudotime inference** + **permutation**: p-value calibration



PseudotimeDE performance





scRNA-seq methods:

`tradeSeq` [Van den Berge *et al.*, *Nat Comms*, 2020]

`Monocle3` [Trapnell *et al.*, *Nat Biotechnol*, 2014]

bulk RNA-seq methods:

`NBAMSeq` [Ren and Kuan, *BMC Bioinfo*, 2020]

`ImpulseDE2` [Fischer *et al.*, *NAR*, 2018]

PseudotimeDE limitations

- **Complete null**: what if cells do not follow a trajectory?

PseudotimeDE limitations

- **Complete null**: what if cells do not follow a trajectory?

  Q: how to generate the in silico negative control under this complete null?
  — simulator **scDesign3**

PseudotimeDE limitations

- **Complete null**: what if cells do not follow a trajectory?

  Q: how to generate the in silico negative control under this complete null?
  — simulator **scDesign3**

- **Computational time**: high-resolution p-values require $> 10^3$ rounds of (subsampling + pseudotime inference + permutation)

PseudotimeDE limitations

- **Complete null**: what if cells do not follow a trajectory?

  Q: how to generate the in silico negative control under this complete null?
  — simulator **scDesign3**

- **Computational time**: high-resolution p-values require $> 10^3$ rounds of (subsampling + pseudotime inference + permutation)

  Q: how to reduce the number of rounds while still achieving FDR control?
  — contrast + FDR control framework **Clipper**

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]

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

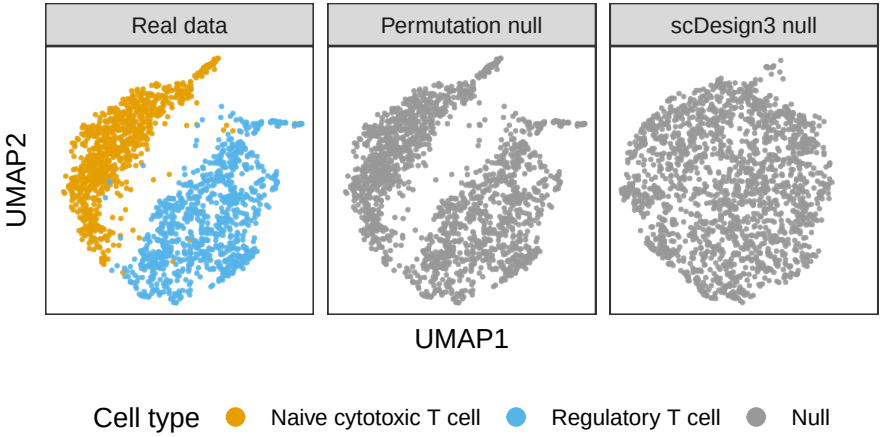
ClusterDE (cell clustering + DEG identification between cell clusters)

- existing methods assume **Gaussian** distributions
  - `TN test` [Zhang, Kamath, and Tse, *Cell Syst*, 2019]
  - `clusterpval` [Gao, Bien, and Witten, *arXiv*, 2020]

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



Clipper: a p-value-free FDR control framework

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

- Foundation: knockoffs
- Applications
  - RNA-seq DEG identification
  - PseudotimeDE
  - ClusterDE

Goal: marginal screening for **interesting** features

d features

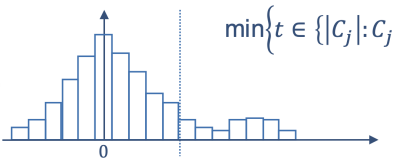
FDR threshold q

Contrast scores

C_1

$\vdots$

C_d



Contrast score cutoff

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

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

Key: **contrast score** construction

example	**target data** (experiment)	**null data** (negative control)
ChIP-seq peak calling (1 vs. 1)	experimental condition	background condition
RNA-seq DEG identification	actual data	permuted data
PseudotimeDE & **ClusterDE**	actual data	**scDesign3** simulated data

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

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

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

Clipper paper

Method | Open Access | Published: 11 October 2021

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

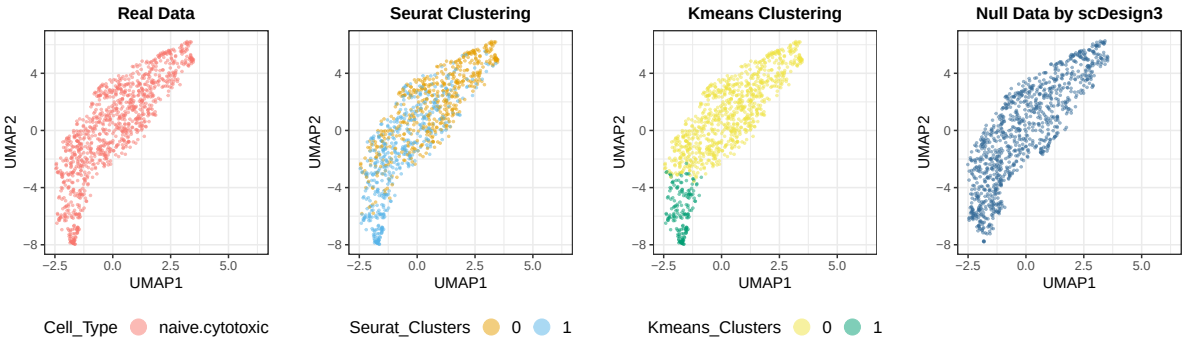
Xinzhou Ge, Yiling Elaine Chen, Dongyuan Song, MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Ning Wang, Wei Li, Leo D. Wang & Jingyi Jessica Li

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

6389 Accesses | **4** Citations | **51** Altmetric | Metrics

ClusterDE: scDesign3 + Clipper (preliminary)

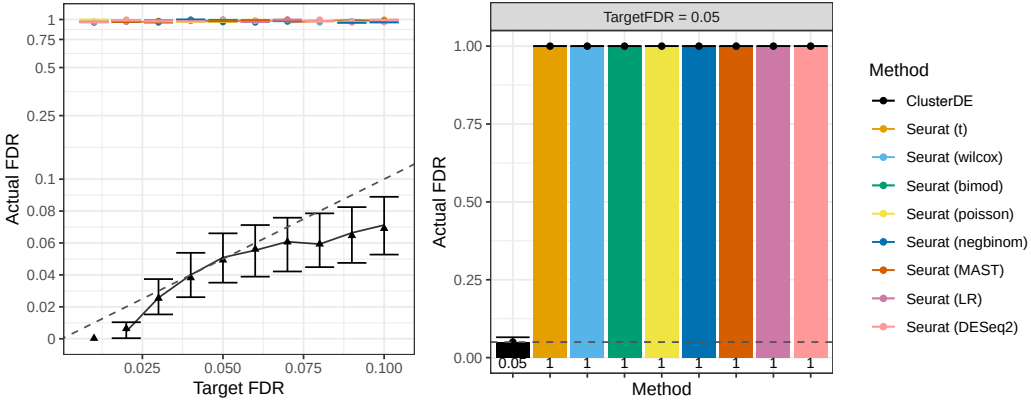
Complete null case: no cell clusters



[Zheng *et al.*, *Nat Commun*, 2017]

ClusterDE: scDesign3 + Clipper (preliminary)

Complete null case: no cell clusters



Take-home messages

- **Sanity check** is essential: popular methods do NOT always work
  Benchmarking against classic methods is crucial for method developers

Take-home messages

- **Sanity check** is essential: popular methods do NOT always work
  Benchmarking against classic methods is crucial for method developers
- **scDesign3 usages**
  - Method benchmarking
  - Parameter inference
  - In silico controlled data generation

Take-home messages

- **Sanity check** is essential: popular methods do NOT always work
  Benchmarking against classic methods is crucial for method developers
- **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

Take-home messages

- **Sanity check** is essential: popular methods do NOT always work
  Benchmarking against classic methods is crucial for method developers
- **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
- Our proposal for single-cell inference
  – **scDesign3**: generating data from the specified null
  – **Clipper**: FDR control that only requires null data generation for once

Final note

Patterns

CellPress
OPEN ACCESS

Perspective

Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines

Jingyi Jessica Li[1,*] and Xin Tong[2]
[1]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA
[2]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA
*Correspondence: jli@stat.ucla.edu
https://doi.org/10.1016/j.patter.2020.100115

Podcast with Glen Colopy @ YouTube

Acknowledgements

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

Tianyi Sun
(Ph.D. student)
scDesign2

Dongyuan Song
(Ph.D. student)
scDesign3
PseudotimeDE

Xinzhou Ge
(Postdoc)
Clipper

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