scDesign3: single-cell and spatial omics simulator
benchmarking, inference & in silico controlled experiments

Jingyi Jessica Li

Professor
Junction of Statistics and Biology (http://jsb.ucla.edu)
Department of Statistics
University of California, Los Angeles
Processed data: a cell-by-feature matrix + cell covariates
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](#)

24k Accesses  |  12 Citations  |  184 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)

![Graph showing the number of identified DEGs from permuted data and the original data for different methods.](image)

[Li et al., *Genome Biology*, 2022]
Teaser: model mis-specification

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

![Graph showing the distribution of gene expression levels | edgeR p=4.35\times10^{-261} | DESeq2 p=6.14\times10^{-38}](image)

[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, Jingyi Jessica Li
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 Xi¹, Jingyi Jessica Li¹, ², ³, ⁴, ⁸

scDesign cons:

- cannot capture gene correlations
- does not directly model count data
<table>
<thead>
<tr>
<th>Simulator</th>
<th>Protocol adaptive</th>
<th>Genes preserved</th>
<th>Gene cor. captured</th>
<th>Cell num. seq. depth flexible</th>
<th>Easy to interpret</th>
<th>Comp. &amp; sample efficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyngen</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lun2</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>powsimR</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROSST</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scDD</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scDesign</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scGAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>splat simple</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>splat</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>kersplat</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SPARSim</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SymSim</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ZINB-WaVE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SPsimSeq</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Related work:
SPsimSeq [Assefa et al., Bioinformatics, 2020]; ESCO [Tian et al., Bioinformatics, 2021]
scDesign2: notations

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

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

- The copula function $C(\cdot)$ is unique for continuous distributions, but not for discrete distributions (unidentifiable) [Genest et al 2007]
- **Distributional transform**: necessary for discrete variable [Ruschendorf 2013].
  - Sample $v_{ij}$ from Uniform[0, 1] independently for $i = 1, \ldots, p$ and $j = 1, \ldots, n$
  - Calculate $u_{ij}$ as $u_{ij} = v_{ij} F_i(x_{ij} - 1) + (1 - v_{ij}) F_i(x_{ij})$
- **Distributional transform**: necessary for discrete variable [Rüschendorf 2013].
  - Sample $v_{ij}$ from $\text{Uniform}[0, 1]$ independently for $i = 1, \ldots, p$ and $j = 1, \ldots, n$
  - Calculate $u_{ij}$ as $u_{ij} = v_{ij}F_i(x_{ij} - 1) + (1 - v_{ij})F_i(x_{ij})$
- **Distributional transform**: necessary for discrete variable [Rüschendorf 2013].
  - Sample $v_{ij}$ from 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 $\Phi$ as the CDF of a standard Gaussian random variable, we can express the joint distribution of $X_j$ as
  \[
  F(x_{1j}, \ldots, x_{pj}) = \Phi_p(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}) | R)
  \]
  where $\Phi_p(\cdot | R)$ is a joint Gaussian CDF with a zero mean vector and a covariance matrix that is equal to the correlation matrix $R$
Denote \( \hat{F}_i \) as the estimated marginal distribution of gene \( i \)

Sample \( v_{ij} \) from Uniform\([0, 1]\) independently for \( i = 1, \ldots, p \) and \( j = 1, \ldots, n \)

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 \( (\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}))^T, \)
\( j = 1, \ldots, n \)
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 $\in \{\text{Poisson, negative binomial, ZIP, ZINB}\}$
- Gene correlations estimated via Gaussian copula

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

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

**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](#)

7989 Accesses | 12 Citations | 30 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, DONGYUAN SONG, WEI VIVIAN LI, and JINGYI JESSICA LI
scDesign3 functionalities (simulation)

https://github.com/SONGDONGYUAN1994/scDesign3
From scDesign2 to scDesign3 (Modeling)

- \( \mathbf{Y} = [Y_{ij}] \in \mathbb{R}^{n \times m} \): the cell-by-feature matrix
  - \( Y_{ij} \): the measurement of feature \( j \) in cell \( i \)
  - \( \mathbf{Y} \) is often a count matrix (i.e., \( \mathbf{Y} \in \mathbb{N}^{n \times m} \))

- \( \mathbf{X} = [x_1, \cdots, x_n]^T \in \mathbb{R}^{n \times p} \): the cell-by-state-covariate matrix, such as
  - Cell type (\( p = 1 \) categorical variable)
  - Cell pseudotime in \( p \) lineage trajectories (\( p \) continuous variables)
  - 2-dimensional cell spatial coordinates (\( p = 2 \) continuous variables)

- \( \mathbf{Z} \in \mathbb{R}^{n \times q} \): the cell-by-design-covariate matrix
  - \( \mathbf{Z} = [\mathbf{b}, \mathbf{c}] \),
  - \( \mathbf{b} = (b_1, \ldots, b_n)^T \) has \( b_i \in \{1, \cdots, B\} \) representing cell \( i \)'s batch
  - \( \mathbf{c} = (c_1, \ldots, c_n)^T \) has \( c_i \in \{1, \cdots, C\} \) representing cell \( i \)'s condition
From scDesign2 to scDesign3 (Modeling)

- We first model the distribution of each gene $j$
- We use the generalized additive model for location, scale, and shape (GAMLSS) [Stasinopoulos and Rigby, 2008]
- The regression model is:

$$
\begin{align*}
Y_{ij} | x_i, z_i & \sim F_j(\cdot | x_i, z_i ; \mu_{ij}, \sigma_{ij}, p_{ij}) \\
\theta_j(\mu_{ij}) & = \alpha_{j0} + \alpha_{jb_i} + \alpha_{jc_i} + f_{jc_i}(x_i) \\
\log(\sigma_{ij}) & = \beta_{j0} + \beta_{jb_i} + \beta_{jc_i} + g_{jc_i}(x_i) \\
\logit(p_{ij}) & = \gamma_{j0} + \gamma_{jb_i} + \gamma_{jc_i} + h_{jc_i}(x_i)
\end{align*}
$$

where $\theta_j(\cdot)$ denotes feature $j$’s specific link function $\mu_{ij}$, depending on $F_j$
- The fitted distribution is denoted as $\hat{F}_j(\cdot | x_i, z_i), i = 1, \ldots, n; j = 1, \ldots, m$
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**

<table>
<thead>
<tr>
<th>Test data</th>
<th>scGAN</th>
<th>muscat</th>
<th>Real data:RNA</th>
<th>scDesign3:RNA + Methylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMAP2</td>
<td>mLISI=1.44</td>
<td>mLISI=1.53</td>
<td>32 Features×177 Cells</td>
<td>UMAP2</td>
</tr>
<tr>
<td>scDesign3</td>
<td>SPARSim</td>
<td>ZINB–WaVE</td>
<td>Real data: Methylation</td>
<td>27 Features×142 Cells</td>
</tr>
<tr>
<td>mLISI=1.91</td>
<td>mLISI=1.44</td>
<td>mLISI=1.53</td>
<td>69 Features×319 Cells</td>
<td>UMAP1</td>
</tr>
</tbody>
</table>
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: spot-resolution spatial data (mouse olfactory bulb 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 (interpretation)

https://github.com/SONGDONGYUAN1994/scDesign3
scDesign3: model inference
scDesign3: unsupervised trajectory / cluster quality assessment
scDesign3: model alteration
A unified framework of realistic in silico data generation and statistical model inference for single-cell and spatial omics

DOI: https://doi.org/10.1101/2022.09.20.508796
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.)
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?

Why need in silico controlled experiments?

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

Double-dipping challenges in single-cell inference

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

DEG

non-DEG

CCL5

EED

log10(count + 1)
pseudotime

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

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

12k Accesses | 11 Citations | 29 Altmetric | Metrics
**PseudotimeDE**

**Generalized additive model (GAM): powerful test statistic**

**Subsampling + pseudotime inference + permutation: p-value calibration**
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, *JASA*, 2022]
- or require *count splitting* and assume *Poisson* distribution
  - **countspli**t [Neufeld, Gao, Popp, Battle, and Witten, *arXiv*, 2022]
ClusterDE (cell clustering + DEG identification between cell clusters)
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, JASA, 2022]
- or require count splitting and assume Poisson distribution
  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
- Orange: Naive cytotoxic T cell
- Blue: Regulatory T cell
- Gray: Null

Real data
Permutation null
scDesign3 null
Clipper: p-value-free FDR control for genomics feature screening

- **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

\[ d \text{ features} \quad \text{FDR threshold } q \]

**Contrast scores**

\[ C_1 \quad \vdots \quad C_d \]

\[ \text{Contrast score cutoff} \]

\[ \min \left\{ t \in \{|C_j|: C_j \neq 0\} : \frac{1 + \#\{j:C_j \leq -t\}}{\#\{j:C_j \geq t\}} \leq q \right\} \]
Clipper offers a general p-value-free FDR control solution

**Key:** contrast score construction

<table>
<thead>
<tr>
<th>example</th>
<th>target data (experiment)</th>
<th>null data (negative control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-seq DEG identification</td>
<td>actual data</td>
<td>permuted data</td>
</tr>
<tr>
<td>PseudotimeDE &amp; ClusterDE</td>
<td>actual data</td>
<td>scDesign3 simulated data</td>
</tr>
</tbody>
</table>

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

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

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

8505 Accesses | 10 Citations | 50 Altmetric | [Metrics](#)
ClusterDE: scDesign3 + Clipper (preliminary)

Complete null case: no cell clusters

Real Data

Seurat Clustering

Kmeans Clustering

Null Data by scDesign3

Cell_Type  naive.cytotoxic
Seurat_Clusters  0  1
Kmeans_Clusters  0  1

[Zheng et al., Nat Commun, 2017]
ClusterDE: scDesign3 + Clipper (preliminary)

**Complete null case:** no cell clusters

Null Cases – nDE = 0

<table>
<thead>
<tr>
<th>Method</th>
<th>Target FDR = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClusterDE</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (t)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (wilcox)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (bimod)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (poisson)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (negbinom)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (MAST)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (LR)</td>
<td>1</td>
</tr>
<tr>
<td>Seurat (DESeq2)</td>
<td>1</td>
</tr>
</tbody>
</table>

![Graph showing actual FDR vs target FDR for different methods](image-url)
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
Patterns

Perspective

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

Jingyi Jessica Li\textsuperscript{1,*} and Xin Tong\textsuperscript{2}
\textsuperscript{1}Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA
\textsuperscript{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

NIH
NIGMS
NSF
Alfred P. Sloan FOUNDATION
Johnson & Johnson
PhRMA
W. M. Keck Foundation