UCLA Statistics

scDesign3: in silico data generation for multimodal single-cell and spatial omics

Dongyuan Song, **Jingyi Jessica Li**

July 25, 2023

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Introduction

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 of scDesign3

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

Challenges in modeling single-cell and spatial multi-omics

- **High-dimensional**: 10^4 genes; for other features the number can be even larger
- **Correlation**: complex correlation structures between features
- **Diverse covariates**: cell types, continuous trajectories, spatial space
- **Multi-omics**: different omics may follow different distributions
- **Transparency**: not a black box

Our recent work: scDesign3

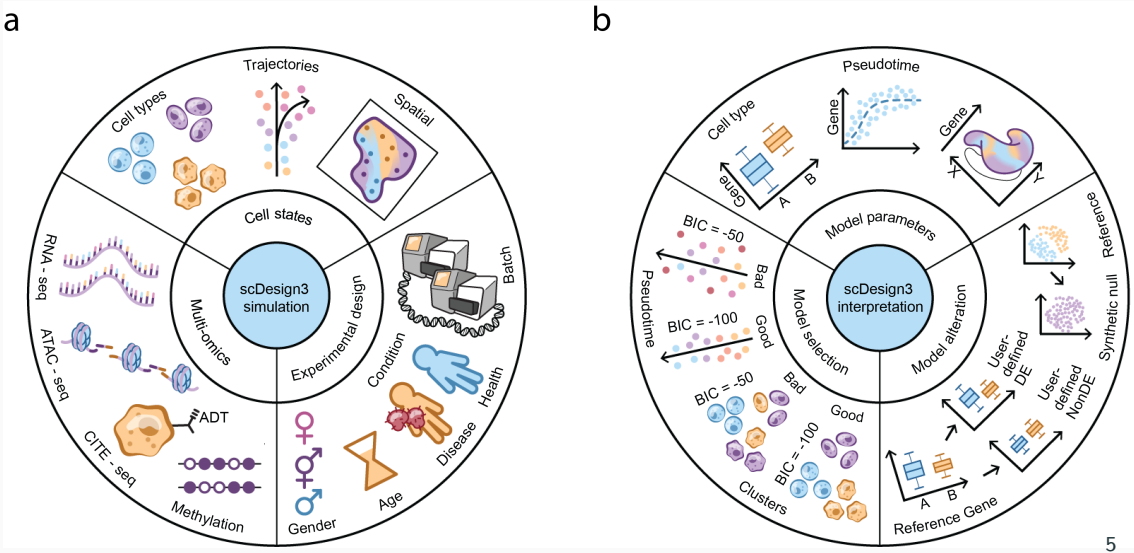
Brief Communication | Published: 11 May 2023

scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun & Jingyi Jessica Li

Nature Biotechnology (2023) | Cite this article

scDesign3's functionality



Methods

Mathematical notations of input data

- $\mathbf{Y} = [Y_{ij}]$ 2 $\mathrm{R}^{n \ \ m}$: cell-by-feature matrix
  - Y_{ij}: the measurement of feature j in cell i
  - $\mathbf{Y}$ is usually a **count matrix** (i.e., $\mathbf{Y}$ 2 $\mathrm{N}^{n \ \ m}$)
- $\mathbf{X} = [\mathbf{x}_1; \quad ; \mathbf{x}_n]^{\mathrm{T}}$ 2 $\mathrm{R}^{n \ \ p}$: cell-by-state-covariate matrix, such as
  - Cell type ($p = 1$ categorical variable)
  - Cell pseudotime in p lineage trajectories (p continuous variables)
  - 2-dimensional spatial coordinates ($p = 2$ continuous variables)
- $\mathbf{Z}$ 2 $\mathrm{R}^{n \ \ q}$: cell-by-design-covariate matrix
  - $\mathbf{Z} = [\mathbf{b}; \mathbf{c}]$
  - $\mathbf{b} = (b_1; : : : ; b_n)^{\mathrm{T}}$ has b_i 2 $f1; \quad ; Bg$ representing cell i's batch
  - $\mathbf{c} = (c_1; : : : ; c_n)^{\mathrm{T}}$ has c_i 2 $f1; \quad ; Cg$ representing cell i's condition

Modeling features' marginal distributions

- First model the distribution of each feature j
- Use the generalized additive model for location, scale, and shape (**GAMLSS**) [Stasinopoulos and Rigby, 2008]
- The feature j's regression model is:

$$\begin{cases} Y_{ij} \mid \mathbf{x}_i, \mathbf{z}_i & \overset{\text{ind}}{\sim} F_j(\mid \mathbf{x}_i, \mathbf{z}_i ; \quad ij ; \quad ij, p_{ij}) \\ {}_j(\quad ij) & = \quad j0 + \quad jb_i + \quad jc_i + f_{jc_i}(\mathbf{x}_i) \\ \log(\quad ij) & = \quad j0 + \quad jb_i + \quad jc_i + g_{jc_i}(\mathbf{x}_i) \\ \text{logit}(p_{ij}) & = \quad j0 + \quad jb_i + \quad jc_i + h_{jc_i}(\mathbf{x}_i) \end{cases} ; \tag{1}$$

  where ${}_j(\ )$ denotes feature j's specific link function $\quad ij$, depending on F_j
- The fitted distribution is denoted as $\hat{F}_j(\mid \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$; $j = 1, \ldots, m$

Choices of marginal distributions

Distribution	PDF or PMF
Gaussian	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2};\ x \in \mathbb{R}$
Bernoulli	$f(x) = p^x(1-p)^{1-x};\ x \in \{0,1\}$
Poisson	$f(x) = \frac{\mu^x e^{-\mu}}{x!};\ x \in \{0,1,2,\ldots\}$
Negative Binomial	$f(x) = \frac{\Gamma(x+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(x+1)} \left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^x;\ x \in \{0,1,2,\ldots\}$
Zero-inflated Poisson	$f(x) = \begin{cases} p + (1-p)e^{-\mu}; & x = 0 \\ (1-p)\frac{\mu^x e^{-\mu}}{x!}; & x = 1,2,3,\ldots \end{cases}$
Zero-inflated NB	$f(x) = \begin{cases} p + (1-p)(1+\alpha)^{-\frac{1}{\sigma}}; & x = 0 \\ (1-p)\frac{\Gamma(x+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(x+1)} \left(\frac{1}{1+\alpha}\right)^{\frac{1}{\sigma}} \left(\frac{\alpha}{1+\alpha}\right)^x; & x = 1,2,3,\ldots \end{cases}$

Functions of modeling cell states

Covariate type	Covariate form	Function form[1]
Discrete cell type	$x_i \ 2 \ f1; \quad ; K_C g$	$f_{jc_i}(x_i) = \ ^{jc_i x_i}$
One lineage	$x_i \ 2 \ [0; \ 1 \ )$	$f_{jc_i}(x_i) = \ \sum_{k=1}^{K} b_{jc_i k}(x_i) \ _{jc_i k}$
p lineages	$\mathbf{x}_i = (x_{i1}; \quad ; x_{ip})^{\mathsf{T}} \ 2 \ [0; \ 1 \ )^p$	$f_{jc_i}(\mathbf{x}_i) = \ \sum_{l=1}^{p} \sum_{k=1}^{K} b_{jc_i kl}(x_{il}) \ _{jc_i lk}$
Spatial location	$\mathbf{x}_i = (x_{i1}; x_{i2})^{\mathsf{T}} \ 2 \ \mathrm{R}^2$	$f_{jc_i}(\mathbf{x}_i) = f_{jc_i}^{\mathrm{GP}}(x_{i1}; x_{i2}; K)$

[1] For simplicity, we only show the form of $f_{jc_i}(\cdot)$ because $g_{jc_i}(\cdot)$ and $h_{jc_i}(\cdot)$ have the same form.

Modeling features' joint distribution

- Denote cell i's m features as a random vector $\mathbf{Y}_i = (Y_{i1}; : : : ; Y_{im})^{\mathsf{T}}$
- Denote the joint CDF as: $F(\ j \ \mathbf{x}_i; \mathbf{z}_i) : \mathrm{R}^m \ ! \ [0; 1]$
- Modeling the joint CDF is challenging; thus we use **copula**
- Denote the conditional copula as $C(\ j \ \mathbf{x}_i; \mathbf{z}_i) : [0; 1]^m \ ! \ [0; 1]$:

$$F(\mathbf{y}_i \ j \ \mathbf{x}_i; \mathbf{z}_i) = C \ (F_1(y_{i1} \ j \ \mathbf{x}_i; \mathbf{z}_i); \quad ; F_m(y_{im} \ j \ \mathbf{x}_i; \mathbf{z}_i) \ j \ \mathbf{x}_i; \mathbf{z}_i) \ ;$$

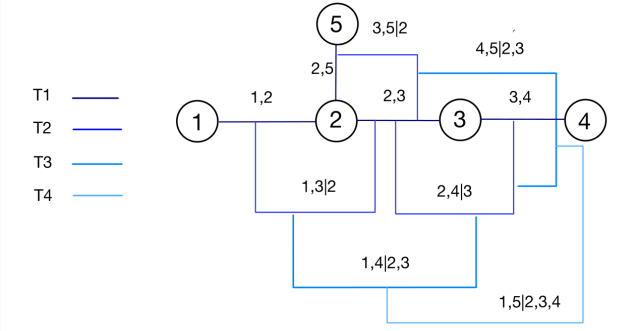
  where $\mathbf{y}_i = (y_{i1}; : : : ; y_{im})^{\mathsf{T}}$ is a realization of $\mathbf{Y}_i = (Y_{i1}; : : : ; Y_{im})^{\mathsf{T}}$
- The simplest choice is a **Gaussian copula**:

$$\begin{aligned} & C \ (F_1(y_{i1} \ j \ \mathbf{x}_i; \mathbf{z}_i); \quad ; F_m(y_{im} \ j \ \mathbf{x}_i; \mathbf{z}_i) \ j \ \mathbf{x}_i; \mathbf{z}_i) \\ = \ & \Phi_m \ \ \Phi \ ^1(F_1(y_{i1} \ j \ \mathbf{x}_i; \mathbf{z}_i)); \quad ; \Phi \ ^1(F_m(y_{i1} \ j \ \mathbf{x}_i; \mathbf{z}_i)); \ \mathbf{R}(\mathbf{x}_i; \mathbf{z}_i) \end{aligned}$$

  where $\mathbf{\Phi}_m(\ j\boldsymbol{R})$ is a joint Gaussian CDF with a zero mean vector and a covariance matrix that is equal to the correlation matrix $\boldsymbol{R}$

From Gaussian copula to vine copula

- Since we often have m n, Gaussian copula can be problematic
- One solution to model high-dimensional correlation is **Vine copula** [Czado et al., 2009]
- "Decompose" a high-dimensional copula into a sequence of bivariate copulas
- Graphic illustration of a 5-dimensional vine copula:



The plug-in estimation of copula

- To estimate $C(\ j\ \mathbf{x}_i; \mathbf{z}_i)$, we use the plug-in approach:

$$\hat{F}_1(\ j\ \mathbf{x}_i; \mathbf{z}_i); \dots; \hat{F}_m(\ j\ \mathbf{x}_i; \mathbf{z}_i)$$

- If $\hat{F}_j(\ j\ \mathbf{x}_i; \mathbf{z}_i)$ is a continuous distribution, each observed y_{ij} is transformed as:

$$u_{ij} = \hat{F}_j(y_{ij}\ j\ \mathbf{x}_i; \mathbf{z}_i)$$

- If $\hat{F}_j(\ j\ \mathbf{x}_i; \mathbf{z}_i)$ is a discrete distribution, we use the **distributional transformation** to make it "continuous":

$$u_{ij} = v_{ij}\hat{F}_j(y_{ij} \quad 1\ j\ \mathbf{x}_i; \mathbf{z}_i) + (1 \quad v_{ij})\hat{F}_j(y_{ij}\ j\ \mathbf{x}_i; \mathbf{z}_i);\ y_{ij} = 1; 2; \dots;$$

  where v_{ij}'s are sampled independently from $\mathrm{Uniform}[0; 1]$
- $u_{ij} = \tilde{F}_j(y_{ij}\ j\ \mathbf{x}_i; \mathbf{z}_i)$, where $\tilde{F}_j(\ j\ \mathbf{x}_i; \mathbf{z}_i)$ is the CDF of a continuous distribution
- Then $C(\ j\ \mathbf{x}_i; \mathbf{z}_i)$ is estimated from $\mathbf{u}_1; \dots; \mathbf{u}_n$, where $\mathbf{u}_i = (u_{i1}; \dots; u_{im})^{\mathsf{T}}$

Generating the synthetic data

- Goal: generate $\mathbf{Y}^0 \; 2 \; \mathrm{R}^{n^0 \;\; m}$ (n^0 synthetic cells and the same m features as $\mathbf{Y}$)
- Given $\mathbf{X}^0 \; 2 \; \mathrm{R}^{n^0 \;\; p}$ and $\mathbf{Z}^0 \; 2 \; \mathrm{N}^{n^0 \;\; q}$,
  1. Sample a m-dimensional vector from the m-dimensional copula:

$$(U_{i'1}; : : : ; U_{i'm})^{\mathsf{T}} \quad \hat{C}(\;\; j \; \mathbf{x}_{i'}; \mathbf{z}_{i'}); \; i^0 = 1; : : : ; n^0$$

  2. Calculate the marginal distribution:

$$Y_{i'j} \; j \; \mathbf{x}_{i'}; \mathbf{z}_{i'} \quad \hat{F}_j(\;\; j \; \mathbf{x}_{i'}; \mathbf{z}_{i'}) = F_j(\;\; j \; \mathbf{x}_{i'}; \mathbf{z}_{i'} \; ; \; \hat{} _{i'j}; \; \hat{} _{i'j}; \; \hat{p}_{i'j}) ;$$

where

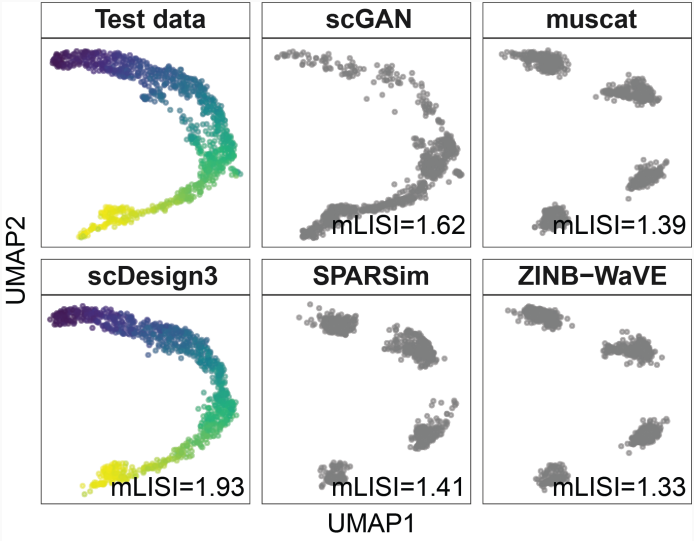
$$\begin{cases} (\hat{} _{i'j}) & = \hat{} _{j0} + \hat{} _{jb_{i'}} + \hat{} _{jc_{i'}} + \hat{f}_{jc_{i'}}(\mathbf{x}_{i'}) ; \\ \log(\hat{} _{i'j}) & = \hat{} _{j0} + \hat{} _{jb_{i'}} + \hat{} _{jc_{i'}} + \hat{g}_{jc_{i'}}(\mathbf{x}_{i'}) ; \\ \mathrm{logit}(\hat{p}_{i'j}) & = \hat{} _{j0} + \hat{} _{jb_{i'}} + \hat{} _{jc_{i'}} + \hat{h}_{jc_{i'}}(\mathbf{x}_{i'}) : \end{cases}$$

  3. Get $(Y_{i'1}; : : : ; Y_{i'm})^{\mathsf{T}}$ by inverse CDF:

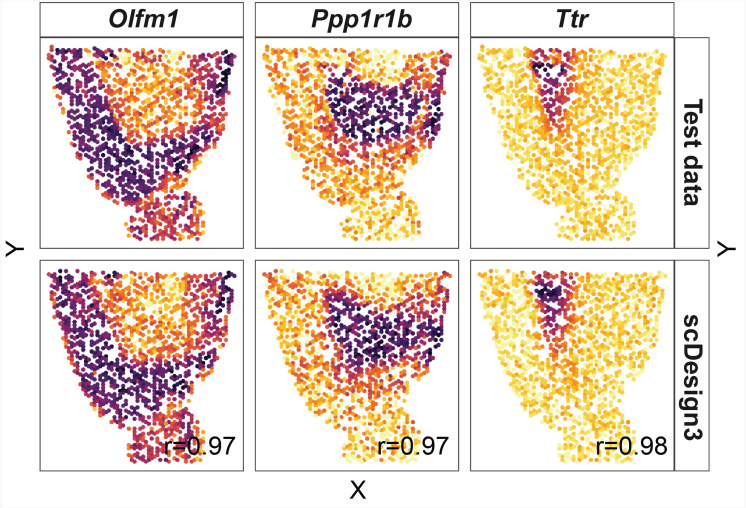
$$Y_{i'j} = \hat{F}_j \; ^1(U_{i'j} \; j \; \mathbf{x}_{i'}; \mathbf{z}_{i'}) ; \; j = 1; : : : ; m$$

Results

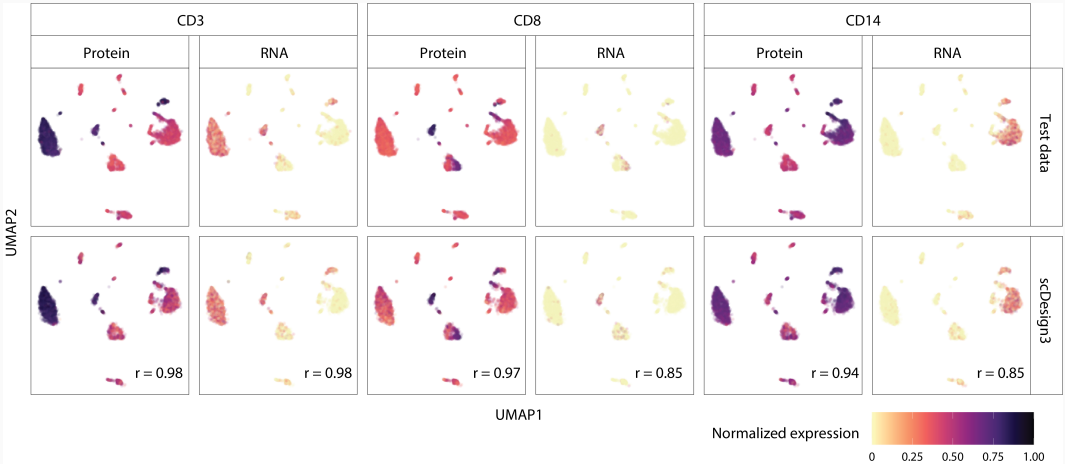
scDesign3 simulates continuous cell differentiation



scDesign3 simulates brain spatial patterns

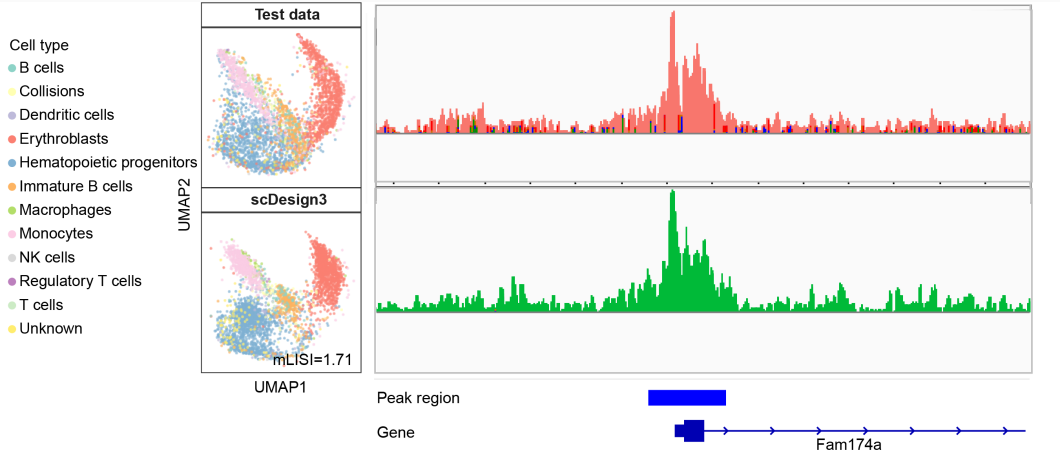


scDesign3 simulates RNA and protein co-expression in blood cells

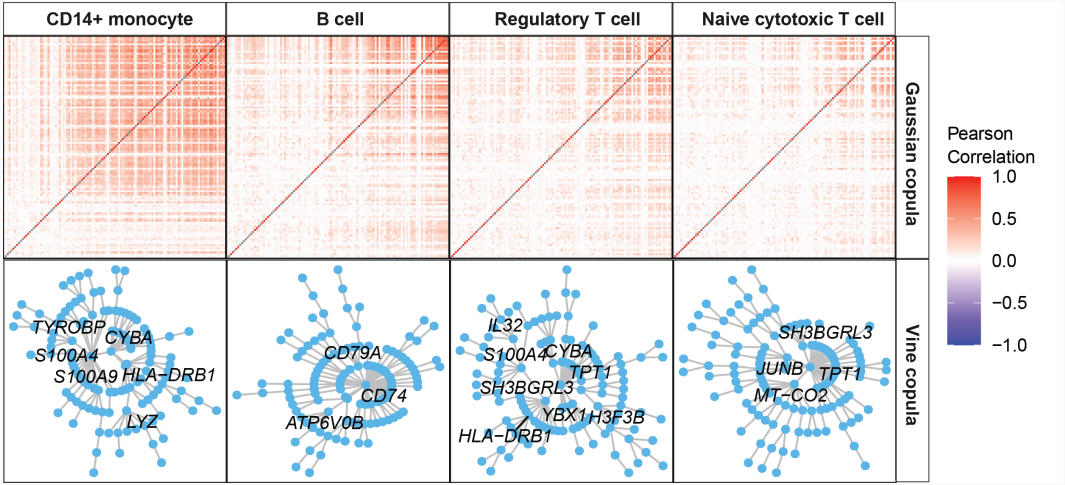


scDesign3 simulates counts and reads of bone marrow ATAC-seq data

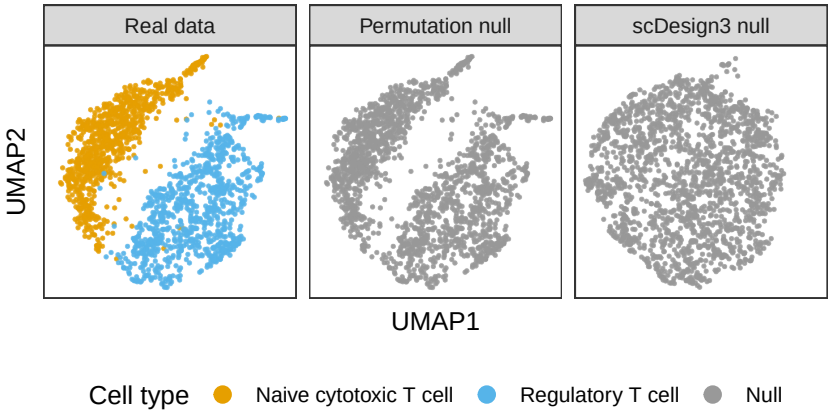
scDesign3 + scReadSim(https://www.biorxiv.org/content/10.1101/2022.05.29.493924v3)



Copula reveals biological differences between immune cell types



scDesign3 generates in silico negative control



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