UCLA Statistics

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

Dongyuan Song

June 28, 2023

Bioinformatics Program
University of California, Los Angeles

Advisor: Dr. Jingyi Jessica Li
http://jsb.ucla.edu

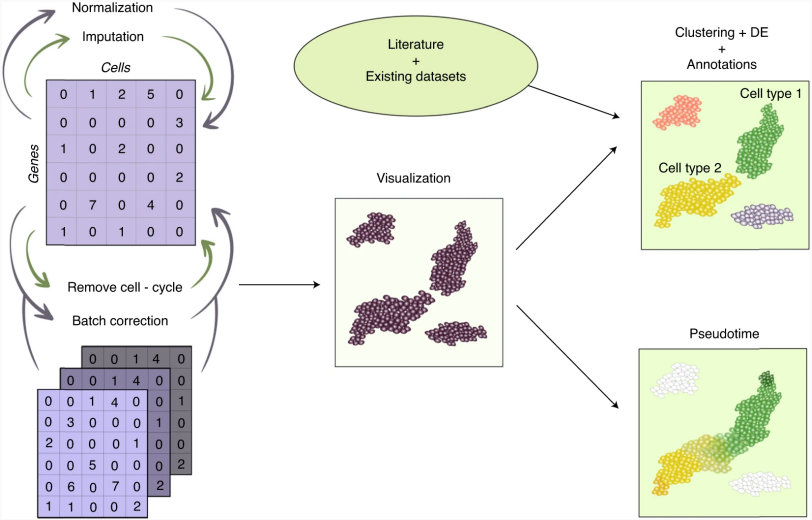
About me

- Ph.D. candidate in Bioinformatics, UCLA
- Currently, visiting student at Harvard Stats
- My advisor, Jingyi Jessica Li, is a Professor in Statistics at UCLA

Dr. Jingyi Jessica Li

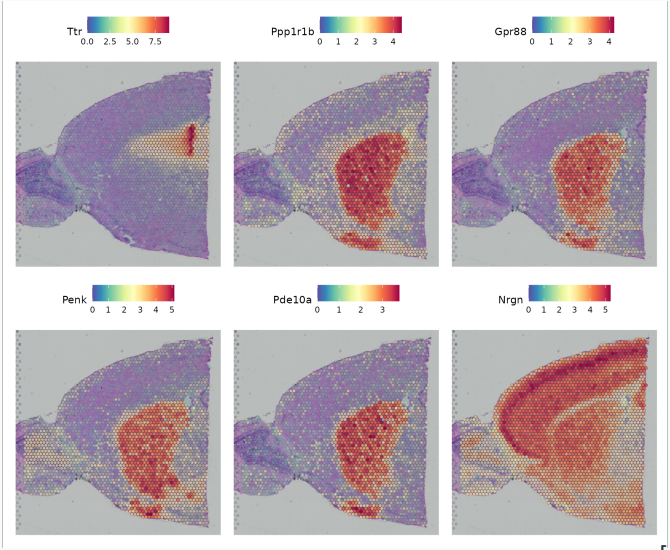
Introduction

Single-cell RNA sequencing (scRNA-seq)



[Andrews et al., 2020]

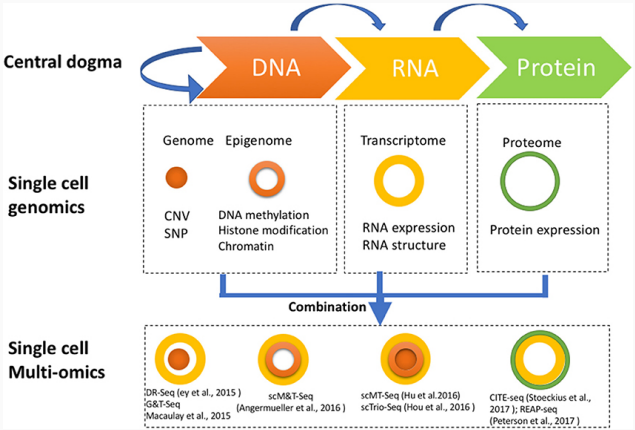
Spatial transcriptomics: gene expression in space



[Seurat tutorial]

Not only genes: other omics and multi-omics

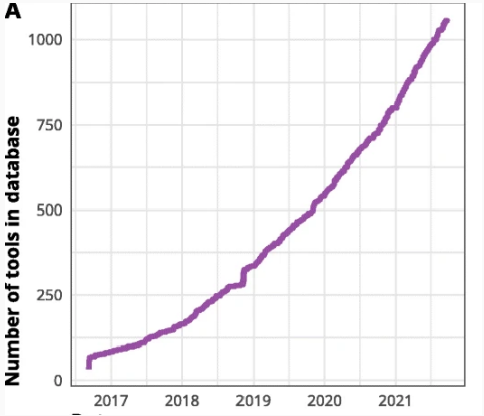
- Measure other types of features (omics) rather than genes
- Measure several types of features simultaneously (multi-omics)



[Hu et al., 2018]

Massive data and numerous tools

- So many datasets and computational tools! Therefore, people want:
  - **A unified probabilistic model** for interpreting single-cell and spatial data?
  - **An all-in-one simulator** for comparing various computational tools?



[Zippa et al., 2021]

Challenges in modeling single-cell and spatial multi-omics

- **High-dimensional**: $\sim 10^4$ 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

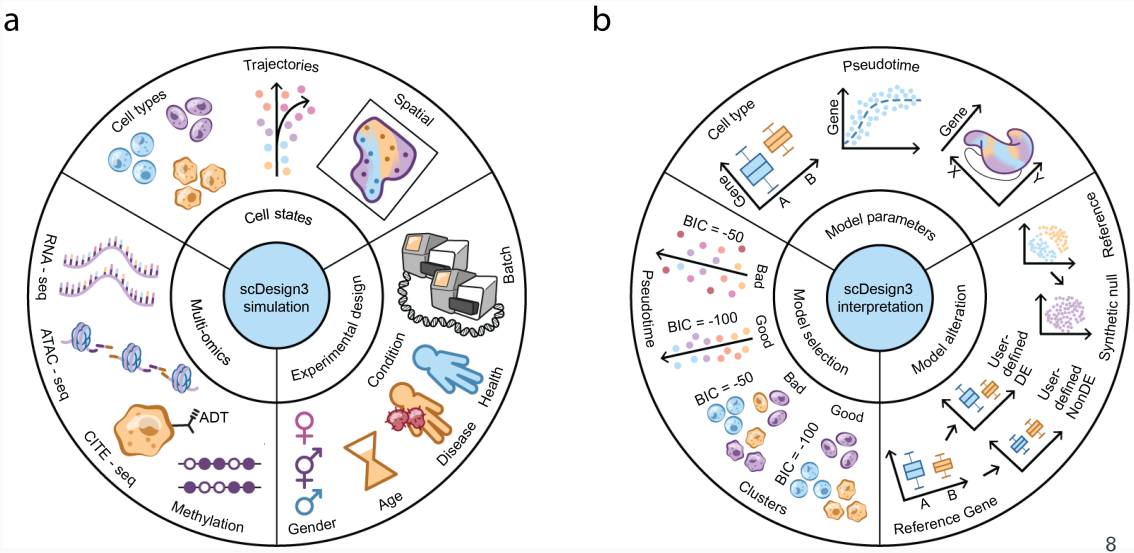
Brief Communication | Published: 11 May 2023

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

- $\mathbf{Y} = [Y_{ij}] \in \mathbb{R}^{n \times m}$: cell-by-feature matrix
  - Y_{ij}: the measurement of feature j in cell i
  - $\mathbf{Y}$ is often a count matrix (i.e., $\mathbf{Y} \in \mathbb{N}^{n \times m}$)
- $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$: 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)
- $\mathbf{Z} \in \mathbb{R}^{n \times q}$: cell-by-design-covariate matrix
  - $\mathbf{Z} = [\mathbf{b}, \mathbf{c}]$
  - $\mathbf{b} = (b_1, \ldots, b_n)^\mathsf{T}$ has $b_i \in \{1, \cdots, B\}$ representing cell i's batch
  - $\mathbf{c} = (c_1, \ldots, c_n)^\mathsf{T}$ has $c_i \in \{1, \cdots, C\}$ 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 regression model is:

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

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

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	$f(x) = \mu^x(1-\mu)^{1-x};\ x \in \{0,1\}$
Poisson	$f(x) = \frac{\mu^x e^{-\mu}}{x!};\ x \in \{0,1,2,\cdots\}$
Negative Binomial	$f(x) = \frac{\Gamma(x+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(x+1)} \left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^x;\ x \in \{0,1,2,\cdots\}$
Zero-inflated Poisson	$f(x) = \begin{cases} p + (1-p)e^{-\mu}; & x = 0 \\ \frac{(1-p)\mu^x e^{-\mu}}{x!}; & x = 1,2,3,\cdots \end{cases}$
Zero-inflated NB	$f(x) = \begin{cases} p + (1-p)(1+\sigma\mu)^{\frac{-1}{\sigma}}; & x = 0 \\ \frac{(1-p)\Gamma(x+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(x+1)} \left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^x; & x = 1,2,3,\cdots \end{cases}$

Functions of modeling cell states

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

[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 measurements of the m features as: a random vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\mathsf{T}$
- Denote the joint CDF as: $F(\cdot \mid \mathbf{x}_i, \mathbf{z}_i) : \mathbb{R}^m \to [0, 1]$
- Modeling the joint CDF is challenging; thus we use **copula**
- Denote the conditional copula as $C(\cdot \mid \mathbf{x}_i, \mathbf{z}_i) : [0, 1]^m \to [0, 1]$:

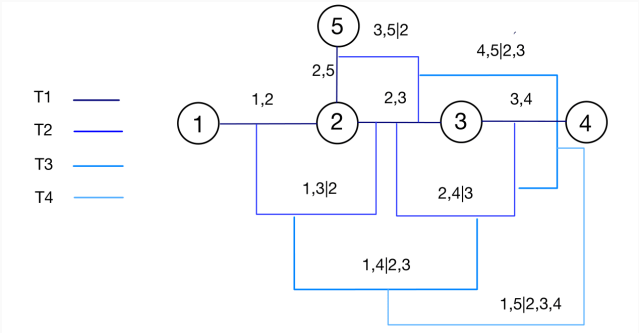
$$F(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i) = C\left(F_1(y_{i1} \mid \mathbf{x}_i, \mathbf{z}_i), \cdots, F_m(y_{im} \mid \mathbf{x}_i, \mathbf{z}_i) \mid \mathbf{x}_i, \mathbf{z}_i\right),$$

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

$$\begin{aligned} &C\left(F_1(y_{i1} \mid \mathbf{x}_i, \mathbf{z}_i), \cdots, F_m(y_{im} \mid \mathbf{x}_i, \mathbf{z}_i) \mid \mathbf{x}_i, \mathbf{z}_i\right) \\ &= \Phi_m\left(\Phi^{-1}(F_1(y_{i1} \mid \mathbf{x}_i, \mathbf{z}_i)), \cdots, \Phi^{-1}(F_m(y_{i1} \mid \mathbf{x}_i, \mathbf{z}_i)); \mathbf{R}(\mathbf{x}_i, \mathbf{z}_i)\right) \end{aligned}$$

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
- A graph can describe it:



The plug-in estimation of copula

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

$$\hat{F}_1(\cdot \mid \mathbf{x}_i, \mathbf{z}_i), \ldots, \hat{F}_m(\cdot \mid \mathbf{x}_i, \mathbf{z}_i)$$

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

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

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

$$u_{ij} = v_{ij}\hat{F}_j(y_{ij} - 1 \mid \mathbf{x}_i, \mathbf{z}_i) + (1 - v_{ij})\hat{F}_j(y_{ij} \mid \mathbf{x}_i, \mathbf{z}_i)\,,\ y_{ij} = 1, 2, \ldots\,,$$

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

Generating the synthetic data

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

$$(U_{i'1}, \ldots, U_{i'm})^\mathsf{T} \sim \hat{C}(\cdot \mid \mathbf{x}_{i'}, \mathbf{z}_{i'}), \; i' = 1, \ldots, n'$$

  2. Calculate the marginal distribution:

$$Y_{i'j} \mid \mathbf{x}_{i'}, \mathbf{z}_{i'} \sim \hat{F}_j(\cdot \mid \mathbf{x}_{i'}, \mathbf{z}_{i'}) = F_j(\cdot \mid \mathbf{x}_{i'}, \mathbf{z}_{i'} \; ; \; \hat{\mu}_{i'j}, \hat{\sigma}_{i'j}, \hat{p}_{i'j}),$$

     where

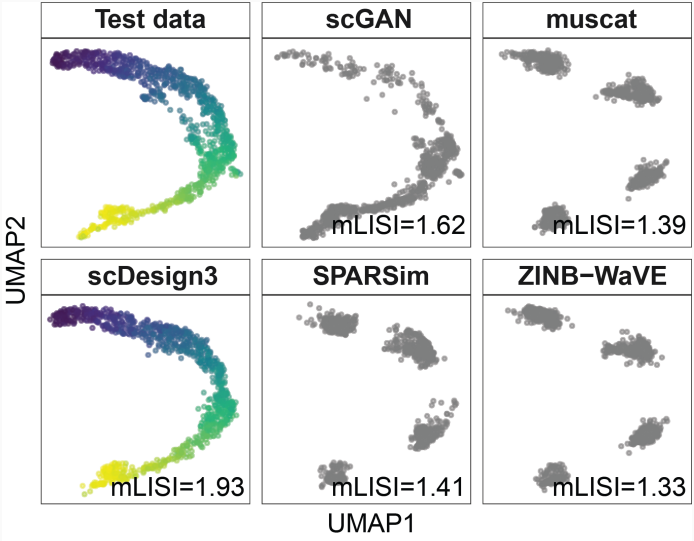
$$\begin{cases} \theta(\hat{\mu}_{i'j}) & = \hat{\alpha}_{j0} + \hat{\alpha}_{jb_{i'}} + \hat{\alpha}_{jc_{i'}} + \hat{f}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \log(\hat{\sigma}_{i'j}) & = \hat{\beta}_{j0} + \hat{\beta}_{jb_{i'}} + \hat{\beta}_{jc_{i'}} + \hat{g}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \mathrm{logit}(\hat{p}_{i'j}) & = \hat{\gamma}_{j0} + \hat{\gamma}_{jb_{i'}} + \hat{\gamma}_{jc_{i'}} + \hat{h}_{jc_{i'}}(\mathbf{x}_{i'}). \end{cases}$$

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

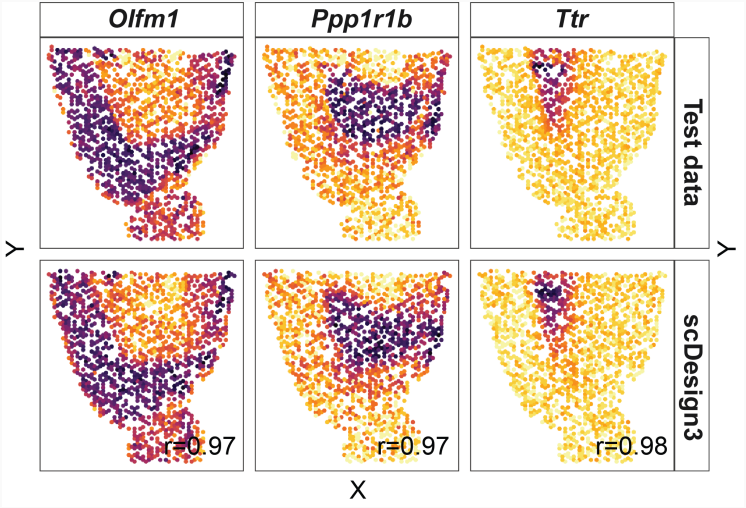
$$Y_{i'j} = \hat{F}_j^{-1}(U_{i'j} \mid \mathbf{x}_{i'}, \mathbf{z}_{i'}), \; j = 1, \ldots, m$$

Results

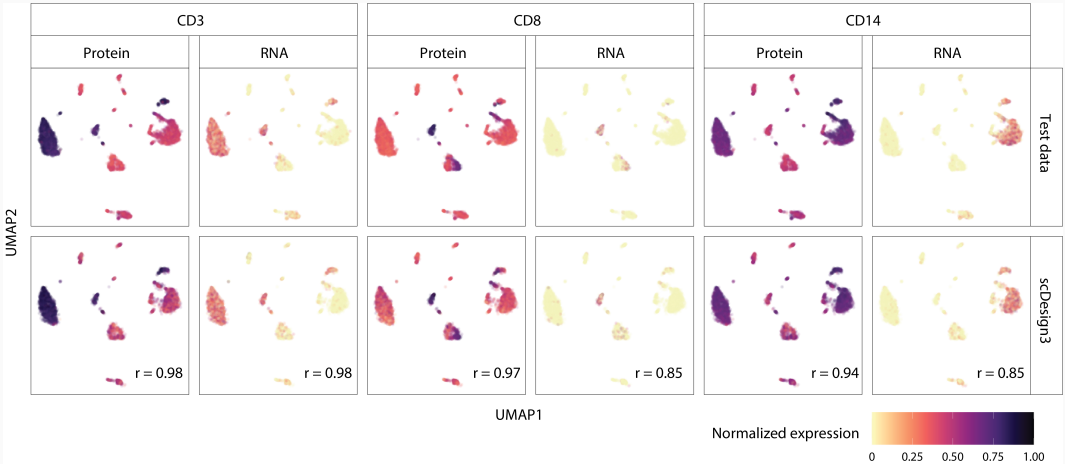
scDesign3 simulates continuous cell differentiation



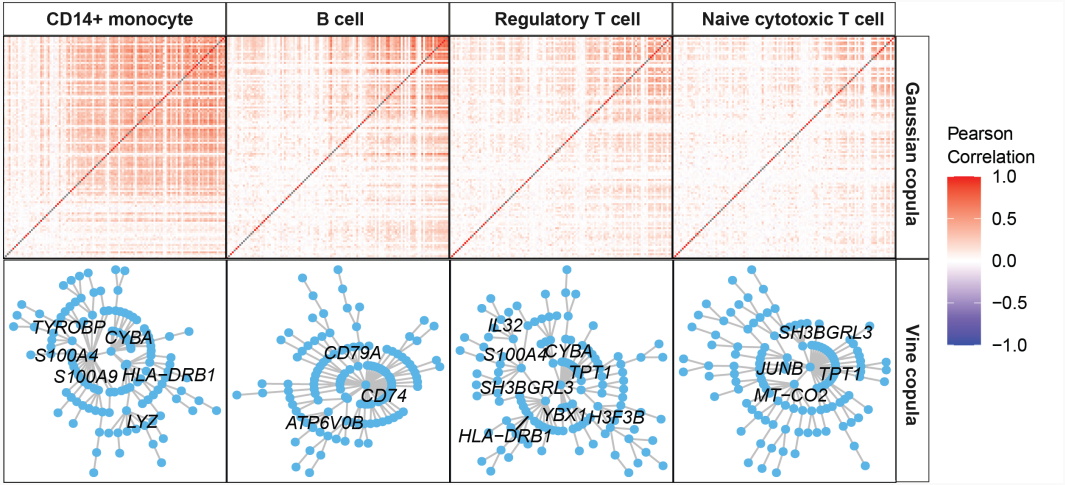
scDesign3 simulates brain spatial patterns



scDesign3 simulates RNA and protein co-expression in blood cells



Copula reveals biological differences between immune cell types



Links and Acknowledgements

Qingyang Wang

Dr. Jingyi Jessica Li

- Paper link: `https://www.nature.com/articles/s41587-023-01772-1`
- Software: `https://github.com/SONGDONGYUAN1994/scDesign3`
- My email: `dongyuansong@ucla.edu`
- Many thanks to the BKT workshop 2023!