UCLA Statistics

A unified framework of realistic in silico data generation and statistical model inference for single-cell and spatial omics

Dongyuan Song

November 10, 2022

Bioinformatics IDP
University of California, Los Angeles

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

About Myself

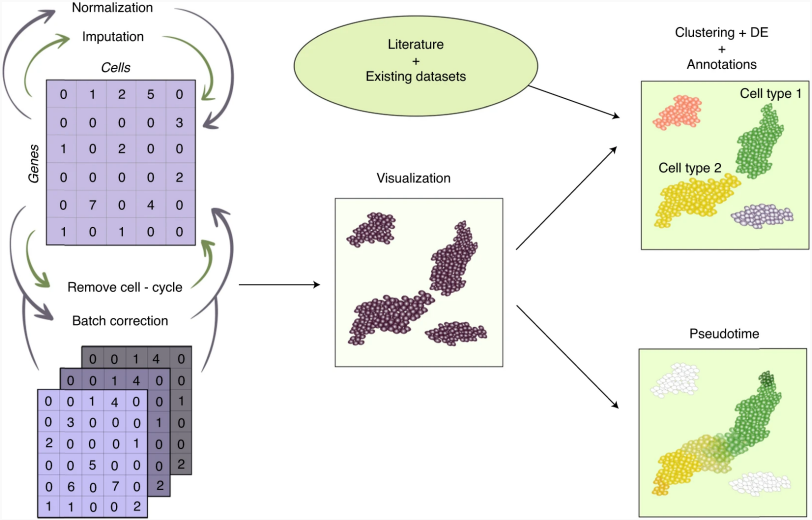
- I am a 4th year PhD student in Bioinformatics, UCLA
- I obtained my Master in Computational Biology from HSPH
- My advisor, Jingyi Jessica Li, is this year's Fellow of Harvard Radcliffe Institute
- I will be a visiting PhD student at Harvard Stats for one year
- Research interests: single-cell genomics, spatial genomics, applications of regression models, etc.
- Email: `dongyuansong@ucla.edu`

Introduction

Single-cell RNA Sequencing

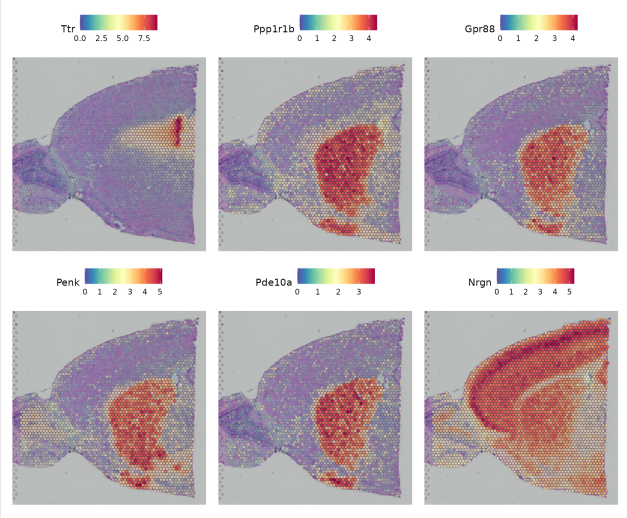
- 10 years ago, we can only measure the gene expression on **tissue** level
- Now we can measure the gene expression of each cell (scRNA-seq)
- Modelling of scRNA-seq data is challenging:
  - **High-dimensional**: $10^2 - 10^6$ cells $\times$ 10^4 genes
  - Complex (latent) structures: discrete cell types, continuous cell trajectories
  - Many genes are **correlated** (gene regulatory network)
  - Confounding covariates: batch effects, cell sizes, etc.

Illustration of scRNA-seq Analysis



[Andrews et al., 2020]

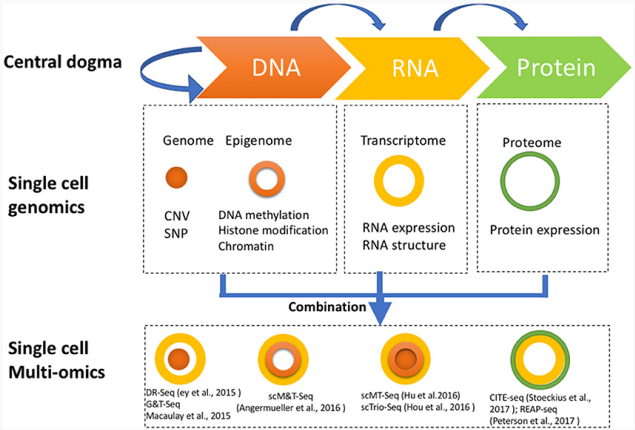
Spatial Transcriptomics: Measuring Cells In Situ



[Seurat tutorial]

More than just Genes: Other Omics and Multi-Omics

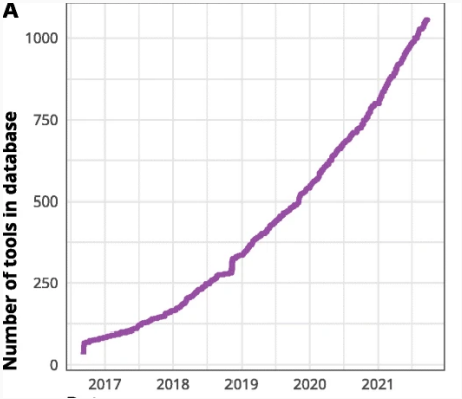
- Measure other types of features (omics) rather than genes
- Measure several types of features simultaneously (multi-omics)



[Hu et al., 2018]

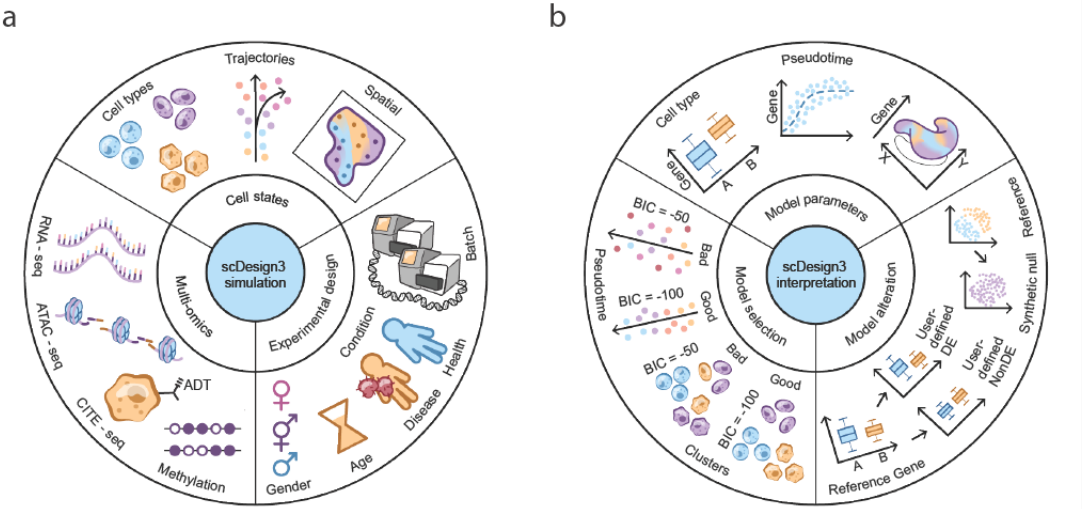
Massive Data and Numerous Computational Tools

- We have so many datasets and available tools! Can we have:
  - **A unified probabilistic model** for interpretating single-cell and spatial data?
  - **An all-in-one simulator** for comparing various computational tools?



[Zippa et al., 2021]

Our Recent Work: scDesign3



Methods

Mathematical Notations of Training Data

- $\mathbf{Y} = [Y_{ij}] \in \mathbb{R}^{n \times m}$: the cell-by-feature matrix
  - Y_{ij}: the measurement of feature j in cell i
  - $\mathbf{Y}$ is often a count matrix (i.e., $\mathbf{Y} \in \mathbb{N}^{n \times m}$)
- $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$: 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
  - $\mathbf{Z} = [\mathbf{b}, \mathbf{c}]$,
  - $\mathbf{b} = (b_1, \ldots, b_n)^\mathsf{T}$ 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

- 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:

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

  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 Distribution

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

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

- We denote cell i's measurements of the m features as: a random vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\mathsf{T}$
- We denote the joint CDF as: $F(\cdot \mid \mathbf{x}_i, \mathbf{z}_i) : \mathbb{R}^m \to [0, 1]$
- Modeling the joint CDF can be challenging, thus we use **copula**
- We denote the conditional copula as $C(\cdot \mid \mathbf{x}_i, \mathbf{z}_i) : [0, 1]^m \to [0, 1]$:

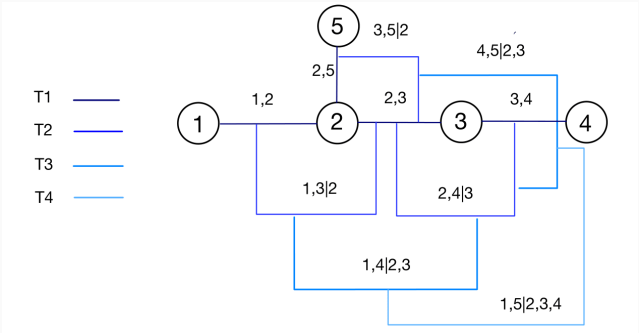
$$F(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i) = C\left(F_1(y_{i1} \mid \mathbf{x}_i, \mathbf{z}_i), \cdots, F_m(y_{im} \mid \mathbf{x}_i, \mathbf{z}_i) \mid \mathbf{x}_i, \mathbf{z}_i\right),$$

  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

- However, since we often have $m > n$, Gaussian copula can be problematic
- How to model high-dimensional correlation? One solution is **Vine copula** [Czado et al., 2009]
- "Decompose" a high-dimensional copula into a sequence of bivariate copulas
- It can be described by a graph:



The Estimation of Copula

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

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

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

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

- When $\hat{F}_j(\cdot \mid \mathbf{x}_i, \mathbf{z}_i)$ 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 $\mathbf{u}_i = (u_{i1}, \ldots, u_{im})^{\mathsf{T}}$

Synthetic Data Generation

- 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 its m-dimensional probability 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

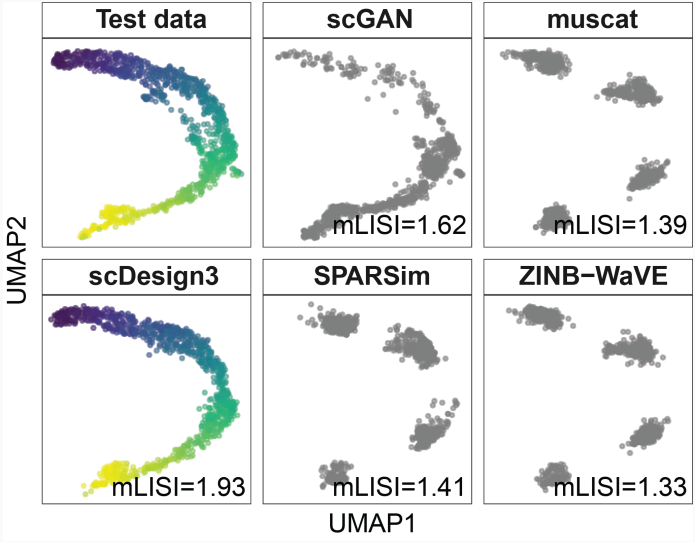
$$\begin{cases} \theta(\hat{\mu}_{i'j}) & = \hat{\alpha}_{j0} + \hat{\alpha}_{jb_{i'}} + \hat{\alpha}_{jc_{i'}} + \hat{f}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \log(\hat{\sigma}_{i'j}) & = \hat{\beta}_{j0} + \hat{\beta}_{jb_{i'}} + \hat{\beta}_{jc_{i'}} + \hat{g}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \operatorname{logit}(\hat{p}_{i'j})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

  3. Get $(Y_{i'1}, \ldots, Y_{i'm})^{\mathsf{T}}$, where

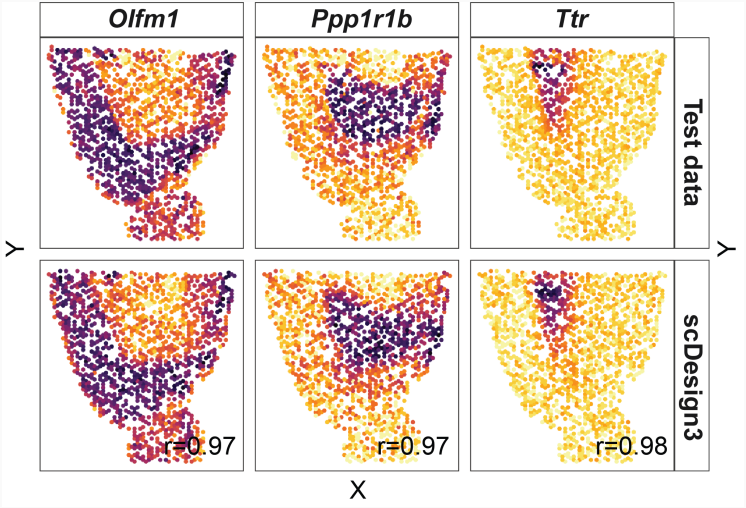
$$Y_{i'j} = \hat{F}_j^{-1}(U_{i'j} \mid \mathbf{x}_{i'}, \mathbf{z}_{i'}), \; j = 1, \ldots, m.$$

Results

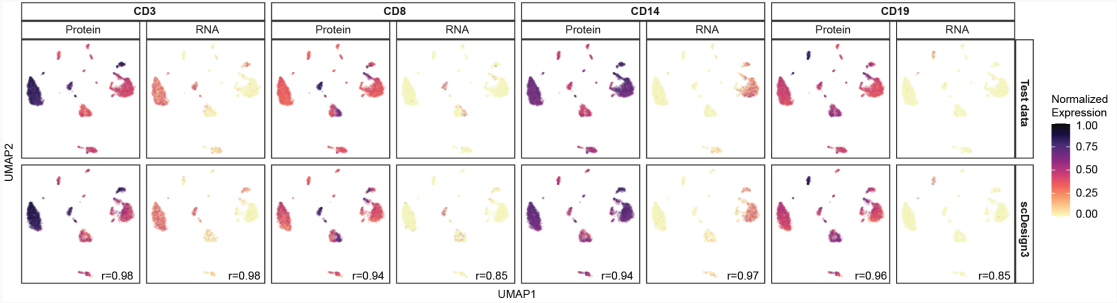
scDesign3 Simulates Continuous Cell Differentiation



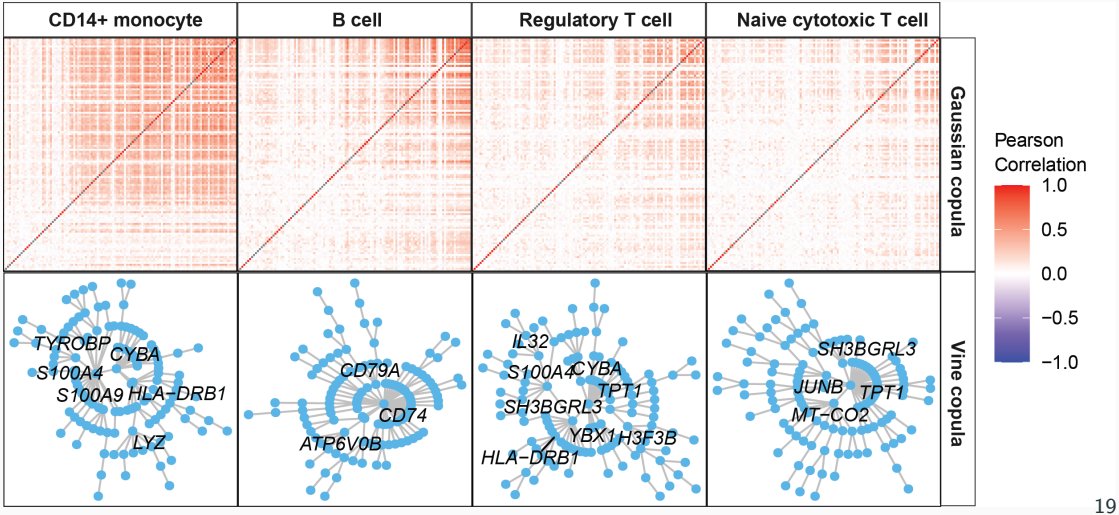
scDesign3 Simulates Brain Spatial Patterns



scDesign3 Simulates RNA and Protein Among Blood Cell Types



The Copula Reveals the Biological Differences between Immune Cell Types



Discussion

Open Questions and Potential Directions

- Better modeling of high-dimensional correlation
  - Scalable **sparse** matrix correlation estimation
  - Other possible statistical methods?
- Vine copula and gene regulatory network (GRN)
  - People know that gene regulation is hierarchical and can be treated as a network
  - Vine copula models the correlation in a hierarchical way
  - Can we use vine copula to infer the GRN?
- Interactions between cell states
  - For instance, spatial locations and cell types determine the gene expression pattern together
  - Hierarchical model?

Links and Acknowledgements

Qingyang Wang

Dr. Jingyi Jessica Li

- Paper link:
  `https://www.biorxiv.org/content/10.1101/2022.09.20.508796v2`
- Software: `https://github.com/SONGDONGYUAN1994/scDesign3`
- Thanks to all members of JSB lab

Supplementary materials

The Five-Dimension Vine Copula

The joint density of $\mathbf{X}$ can be written as

$$
\begin{aligned}
& f_{12345}(x_1, x_2, x_3, x_4, x_5) \\
=& f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \cdot f_4(x_4) \cdot f_5(x_5) \\
& \cdot c_{1,2}\left(F_1(x_1),\, F_2(x_2)\right) \cdot c_{2,3}\left(F_2(x_2),\, F_3(x_3)\right) \cdot c_{2,5}\left(F_2(x_2),\, F_5(x_5)\right) \cdot c_{3,4}\left(F_3(x_3),\, F_4(x_4)\right) \\
& \cdot c_{1,3|2}\left(F_{1|2}(x_1|x_2),\, F_{3|2}(x_3|x_2)\right) \cdot c_{2,4|3}\left(F_{2|3}(x_2|x_3),\, F_{4|3}(x_4|x_3)\right) \cdot c_{3,5|2}\left(F_{3|2}(x_3|x_2),\, F_{5|2}(x_5\right. \\
& \cdot c_{1,4|2,3}\left(F_{1|2,3}(x_1|x_2, x_3),\, F_{4|2,3}(x_4|x_2, x_3)\right) \cdot c_{4,5|2,3}\left(F_{4|2,3}(x_4|x_2, x_3),\, F_{5|2,3}(x_5|x_2, x_3)\right) \\
& \cdot c_{1,5|2,3,4}\left(F_{1|2,3,4}(x_1|x_2, x_3, x_4),\, F_{5|2,3,4}(x_5|x_2, x_3, x_4)\right) ,
\end{aligned}
$$

where $c_{i,j|D} : [0,1]^2 \to [0,\infty)$ is a bivariate copula density function of $F_{i|D}(X_i)$ and $F_{j|D}(X_j)$ conditional on the variable set $\{X_k : k \in D\}$, and $F_{i|D}$ is the conditional CDF of X_i given $\{X_k : k \in D\}$, $i = 1, \ldots, m$.