scDesign2: a high-fidelity scRNA-seq simulator that captures gene correlations

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu
Why Do We Need a High-fidelity scRNA-seq Simulator?

- Design of experiments:
  - How to choose among existing experimental protocols?
    - Full-length (e.g., Smart-seq2): fewer cells & more genes
    - Tag-based (e.g., 10X Genomics): more cells & fewer genes
  - Given a protocol, how to determine the optimal experimental parameters?
    - Number of cells to sequence
    - Sequencing depth

- Benchmarking of computational methods
  - Cell clustering
  - Rare cell type detection
  - Cell trajectory inference
  - Differentially expressed gene identification
## Summary of Simulators

<table>
<thead>
<tr>
<th>Simulator</th>
<th>Property</th>
<th>protocol</th>
<th>gene preserved</th>
<th>gene cor. captured</th>
<th>cell num. seq. dep. 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>
<td>✓</td>
</tr>
<tr>
<td>Lun2</td>
<td>✓</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>
<td>✓</td>
</tr>
<tr>
<td>PROSST</td>
<td>×</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>
<td>✓</td>
</tr>
<tr>
<td>scDesign</td>
<td>✓</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>
<td>×</td>
</tr>
<tr>
<td>splat simple</td>
<td>✓</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>
<td>✓</td>
</tr>
<tr>
<td>kersplat</td>
<td>✓</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>
<td>✓</td>
</tr>
<tr>
<td>SymSim</td>
<td>✓</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>
<td>✓</td>
</tr>
<tr>
<td>scDesign2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Diagram of scDesign2

- Gene expression count
  - Low
  - High

- Real count matrix
  - Cell type 1
  - Cell type K

- Sub-matrices
  - Gene joint distribution

- User-specified sequencing depth & number of cells

- Guidance for Experimental Design
- Evaluation of Computational Methods

Input

Parameter Estimation

Data Simulation

(e.g. sequencing depth becomes lower)
Comparison to Existing scRNA-seq Simulators

Kendall's tau

Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
Comparison to Existing scRNA-seq Simulators

Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
Comparison to Existing scRNA-seq Simulators

Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
Comparison to Existing scRNA-seq Simulators

Data: dendrocytes subtype 1 of human blood by Smart-Seq2 [Villani et al., Science (2017)]
## Comparison to Existing scRNA-seq Simulators

<table>
<thead>
<tr>
<th>Training Data</th>
<th>scDesign2</th>
<th>scDesign2 (w/o copula)</th>
<th>ZINB-WaVE</th>
<th>SPARSim</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Training Data" /></td>
<td><img src="image2" alt="scDesign2" /></td>
<td><img src="image3" alt="scDesign2 (w/o copula)" /></td>
<td><img src="image4" alt="ZINB-WaVE" /></td>
<td><img src="image5" alt="SPARSim" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Data</th>
<th>test + scDesign2</th>
<th>test + scDesign2 (w/o copula)</th>
<th>test + ZINB-WaVE</th>
<th>test + SPARSim</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image6" alt="Test Data" /></td>
<td><img src="image7" alt="test + scDesign2" /></td>
<td><img src="image8" alt="test + scDesign2 (w/o copula)" /></td>
<td><img src="image9" alt="test + ZINB-WaVE" /></td>
<td><img src="image10" alt="test + SPARSim" /></td>
</tr>
</tbody>
</table>

Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]

8
Comparison to Existing scRNA-seq Simulators

Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
Application 1: Other Single-cell Technologies

Top data: mouse hypothalamic preoptic region by MERFISH [Moffitt et al., Science (2018)]
Bottom data: mouse hippocampal area CA1 by pciSeq [Qian et al., Nature Methods (2020)]
### Application 2: Clustering

#### Data

Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]

#### Performance of Seurat and SC3

<table>
<thead>
<tr>
<th>Method</th>
<th>AMI</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seurat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Total number of UMI

<table>
<thead>
<tr>
<th>Total UMI</th>
<th>Seurat</th>
<th>SC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,2M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28,57M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>113,05M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>457,23M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Cell number

<table>
<thead>
<tr>
<th>Cell Number</th>
<th>Seurat</th>
<th>SC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>474</td>
<td></td>
<td></td>
</tr>
<tr>
<td>948</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15,172</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Clustering Results

- **Seurat**
  - cluster 1 / Stem
  - cluster 2 / Goblet
  - cluster 3 / Tuft
  - cluster 4 / TA.Early
  - cluster 5 / EP
  - cluster 6 / EP.Early
  - cluster 7

- **SC3**
  - cluster 1 / Stem
  - cluster 2 / Goblet
  - cluster 3 / Tuft
  - cluster 4 / TA.Early
  - cluster 5 / EP
  - cluster 6 / EP.Early
  - cluster 7
  - cluster 8

---

Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
Application 3: Rare Cell Type Detection

Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]
• **scDesign2**: an interpretable simulator that generates realistic single-cell gene expression count data with gene correlations

  - Motivated by our previous work scDesign (Li and Li, Bioinformatics 2019)
  - A multi-gene generative model (probabilistic, transparent, interpretable)
  - Guidance for scRNA-seq experimental design
  - Benchmarking of computational methods

• R package: [https://github.com/JSB-UCLA/scDesign2](https://github.com/JSB-UCLA/scDesign2)

• Future work
  - Extend the current model to accommodate continuous cell trajectories
Acknowledgements

Tianyi Sun
(Ph.D. student, UCLA)

Dongyuan Song
(Ph.D. student, UCLA)

Dr. Wei Vivian Li
(former Ph.D. student; assistant professor, Rutgers)
PseudotimeDE: Identification of DEGs along Pseudotime with Valid p-values

Dongyuan Song

R package: https://github.com/SONGDONGYUAN1994/PseudotimeDE