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

Simulator	Property	protocol adaptive	gene preserved	gene cor. captured	cell num. seq. dep. flexible	easy to interpret	comp. & sample efficient
dyngen		×	×	×	\checkmark	\checkmark	\checkmark
Lun2		\mathbf{x}	\checkmark	×	\checkmark	\checkmark	\checkmark
powsimR		\checkmark	\checkmark	\times	\checkmark	\checkmark	\checkmark
PROSST		\checkmark	\checkmark	×	\checkmark	\checkmark	\checkmark
scDD		\checkmark	×	×	\checkmark	\checkmark	\checkmark
scDesign		\checkmark	\checkmark	×	\checkmark	\checkmark	\checkmark
scGAN		\checkmark	\checkmark	\checkmark	\checkmark	×	×
splat simple	2	\checkmark	×	×	×	\checkmark	\checkmark
splat		\checkmark	×	×	×	\checkmark	\checkmark
kersplat		\checkmark	×	\checkmark	×	\checkmark	\checkmark
SPARSim		\checkmark	\checkmark	×	×	\checkmark	\checkmark
SymSim		\checkmark	×	×	×	\checkmark	\checkmark
ZINB-WaV	E	\checkmark	×	×	×	\checkmark	\checkmark
scDesign2		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	✓ 2

Diagram of scDesign2





Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)] 4



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



Data: goblet cells of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)] ⁶



Data: dendrocytes subtype 1 of human blood by Smart-Seq2 [Villani et al., Science (2017)]



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

8



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

9

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



11Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)]

(4) cell number

1e+04 1e+05

- 15172

eluster 1 /

cluster 2 / Goblet

cluster 3 / Tuft

cluster 4 / TA.Early

cluster 5 / EP

cluster 6 / EP.Early

cluster 7

cluster 8

0.041

AR

; real data

; cell num.

Stem

Application 3: Rare Cell Type Detection





Data: six cell types of mouse small intestinal epithelium by 10x Genomics [Haber et al., Nature (2017)] 12

- **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
- 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/PseudotimeDE5