scDesign

A statistical simulator for single-cell RNA sequencing experimental design

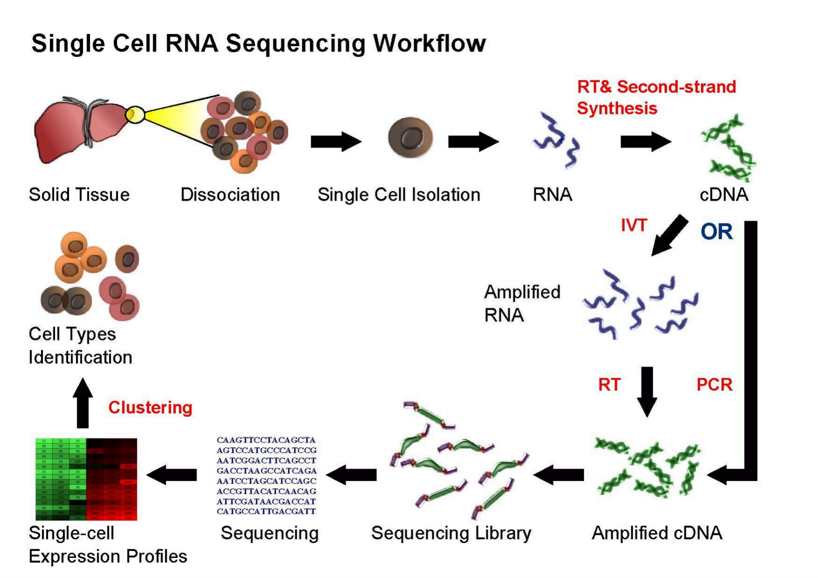
Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

joint work with Wei Vivian Li (UCLA)

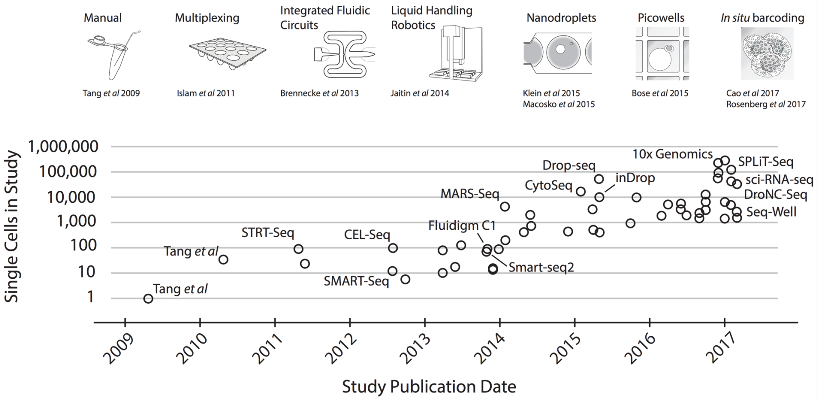
Introduction

Single Cell RNA Sequencing (scRNA-seq)



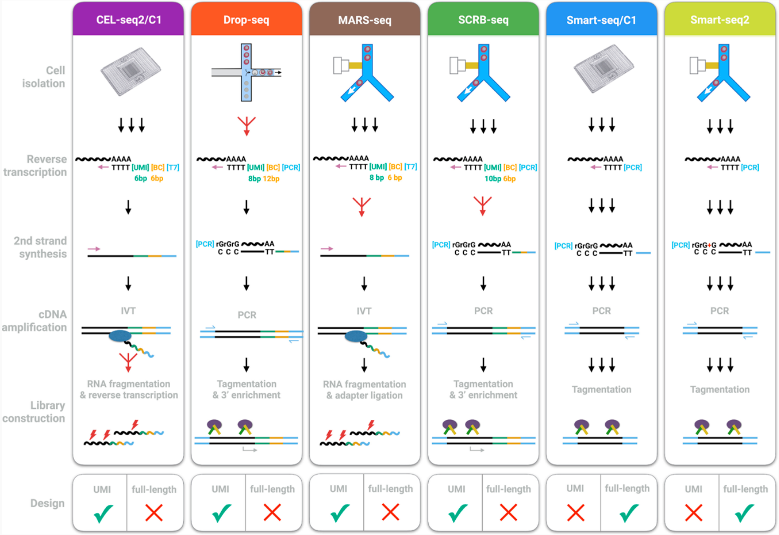
from *Wikipedia*, Single Cell Sequencing

Multiple scRNA-seq Protocols



[Svensson et al., 2018] *Nature Protocols*

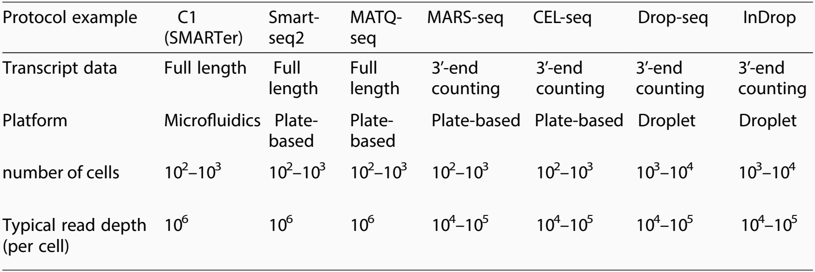
Full length vs. Tag-based Protocols



Design	UMI	full-length
CEL-seq2/C1	✓	✗
Drop-seq	✓	✗
MARS-seq	✓	✗
SCRB-seq	✓	✗
Smart-seq/C1	✗	✓
Smart-seq2	✗	✓

[Ziegenhain et al., 2017] *Molecular Cell*

Exploring the Depth vs. Breadth of Transcriptomes



Protocol example	C1 (SMARTer)	Smart-seq2	MATQ-seq	MARS-seq	CEL-seq	Drop-seq	InDrop
Transcript data	Full length	Full length	Full length	3′-end counting	3′-end counting	3′-end counting	3′-end counting
Platform	Microfluidics	Plate-based	Plate-based	Plate-based	Plate-based	Droplet	Droplet
number of cells	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^3–10^4	10^3–10^4
Typical read depth (per cell)	10^6	10^6	10^6	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5

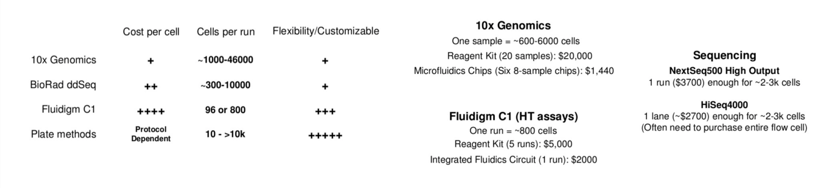
[Haque et al., 2017] *Genome Medicine*

Challenge: Budget vs. Cell number vs. Dropout

The frequency of dropout events depends on scRNA-seq protocols

- Fluidigm C1 platform: ~ 100 cells, ~ 1 million reads per cell
- Droplet microfluidics: $\sim 10,000$ cells, ~ 100K reads per cell [Zilionis et al., 2017]

Trade-off: given the same budget, more cells, more dropouts



	Cost per cell	Cells per run	Flexibility/Customizable
10x Genomics	**+**	**~1000-46000**	**+**
BioRad ddSeq	**++**	**~300-10000**	**+**
Fluidigm C1	**++++**	**96 or 800**	**+++**
Plate methods	**Protocol Dependent**	**10 - >10k**	**+++++**

10x Genomics
One sample = ~600-6000 cells
Reagent Kit (20 samples): $20,000
Microfluidics Chips (Six 8-sample chips): $1,440

Fluidigm C1 (HT assays)
One run = ~800 cells
Reagent Kit (5 runs): $5,000
Integrated Fluidics Circuit (1 run): $2000

Sequencing
NextSeq500 High Output
1 run ($3700) enough for ~2-3k cells

HiSeq4000
1 lane (~$2700) enough for ~2-3k cells
(Often need to purchase entire flow cell)

from David Cook, Ottawa Hospital Research Institute

Experimental Design for scRNA-seq

1. Model-based / theoretical analysis

   - Negative Binomial model to estimate the number of cells to sequence [Baran-Gale et al., 2017] `http://www.satijalab.org/howmanycells`
   - A Good-Toulmin like estimator of number of cells of each tissue to maximize cell type discovery across tissues [Dumitrascu et al., 2018]

Experimental Design for scRNA-seq

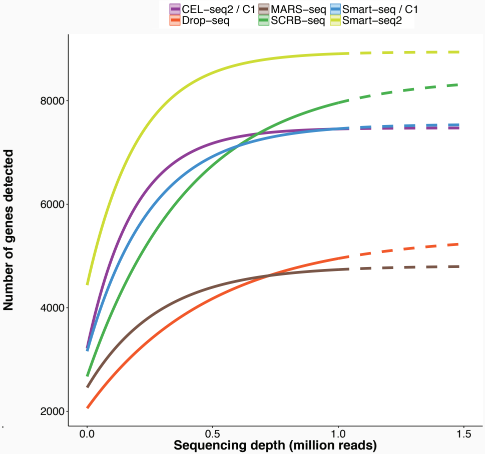
1. Model-based / theoretical analysis
    - Negative Binomial model to estimate the number of cells to sequence [Baran-Gale et al., 2017] `http://www.satijalab.org/howmanycells`
    - A Good-Toulmin like estimator of number of cells of each tissue to maximize cell type discovery across tissues [Dumitrascu et al., 2018]

Remarks:

- Requires prior knowledge of
    - Cell type composition
    - Cell type gene expression profiles
- Difficult to "visualize" the design

Experimental Design for scRNA-seq

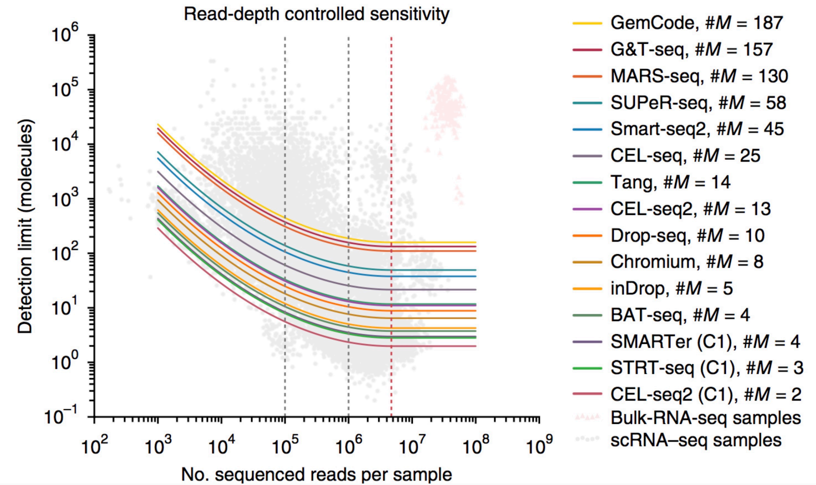
2. Descriptive statistics
    - Sensitivity of most protocols saturates at ∼one million reads per cell



[Ziegenhain et al., 2017]

Experimental Design for scRNA-seq

2. Descriptive statistics
    - Detection rate saturates at ∼4.5 million reads per cell



[Svensson et al., 2017]

Experimental Design for scRNA-seq

2. Descriptive statistics

    - Sensitivity of most protocols saturates at ∼one million reads per cell [Ziegenhain et al., 2017].
    - Detection rate saturates at ∼4.5 million reads per cell [Svensson et al., 2017].

Experimental Design for scRNA-seq

2. Descriptive statistics

   - Sensitivity of most protocols saturates at ∼one million reads per cell [Ziegenhain et al., 2017].
   - Detection rate saturates at ∼4.5 million reads per cell [Svensson et al., 2017].

Remarks:

- Difficult to unify to guide practices
- Not specific to the biological condition under study

scDesign: Statistical Simulator and Experimental Design

Simulation-based scRNA-seq experimental design

Advantages of scDesign:

- Protocol-adaptive and data-adaptive: learn from
  - public real data [Abugessaisa et al., 2017, Cao et al., 2017]
  - pilot-study data
- Generate synthetic data that well mimic real data under a pre-specified experimental setting
  - Assist experimental design & method development
- Flexible in accommodating user-specific analysis needs
- No experimental cost

Methods

scDesign for scRNA-seq Data Simulation

Key features:

1. Leverage existing real scRNA-seq data
2. Construct a Gamma-Normal mixture model to account for dropouts
   - Adapted from scImpute [Li and Li, 2018].
   - Estimate key gene expression parameters from real data
3. Two flexible modes:
   - One-state mode: a group of cells from a single cell state are sequenced
   - Two-state mode: two groups of cells from different cell states are sequenced together

The Generative Framework of scDesign: One-state Mode

Given an experimental setting

- total sequencing depth
- number of cells

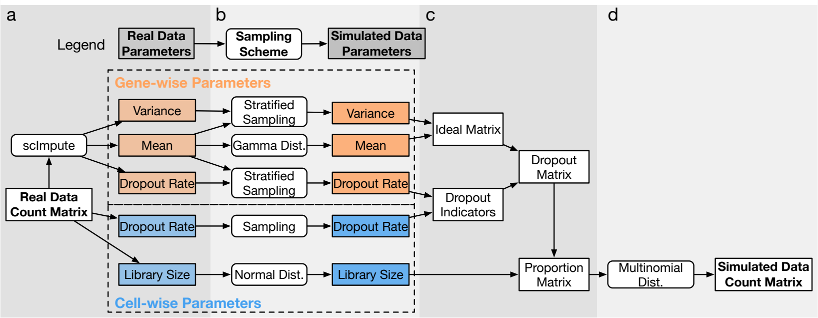
a real scRNA-seq dataset from one cell state

$$\rightarrow$$

a single scRNA-seq dataset

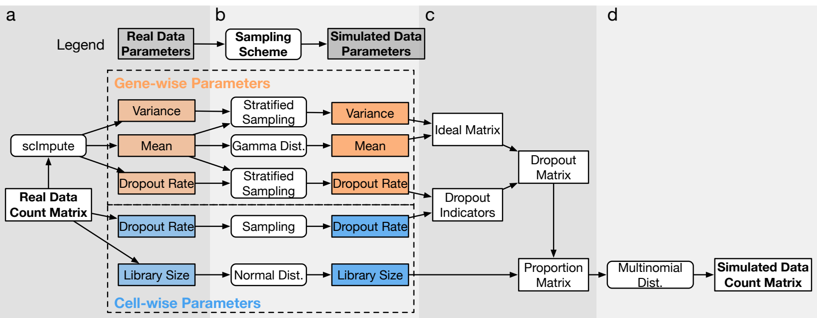
The Generative Framework of scDesign: One-state Mode

a. Estimate five parameters from the real scRNA-seq dataset: three gene-wise and two cell-wise parameters



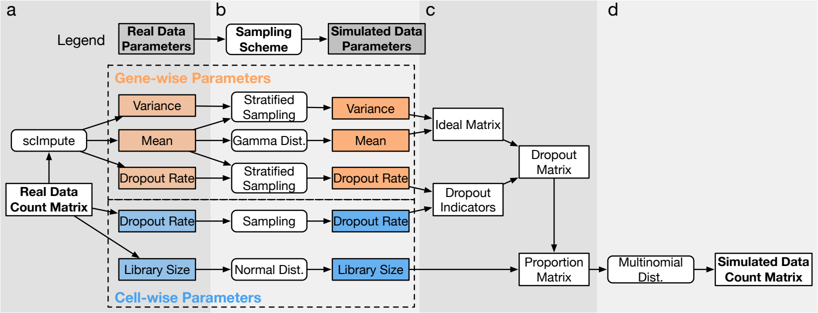
The Generative Framework of scDesign: One-state Mode

b. Simulate gene- and cell-wise parameters for new cells



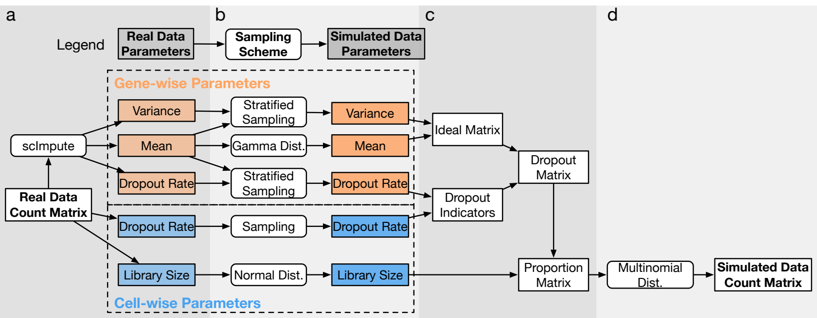
The Generative Framework of scDesign: One-state Mode

c. Simulate ideal gene expression levels for new cells, then introduce dropout values based on the estimated dropout parameters



The Generative Framework of scDesign: One-state Mode

d. Output a synthetic gene expression matrix with entries as read counts



The Generative Framework of scDesign: Two-state Mode

Mimics an experiment where two groups of cells from two cell states are sequenced together

Given an experimental setting

- total sequencing depth
- cell numbers of the two states

two real scRNA-seq dataset from two cell states

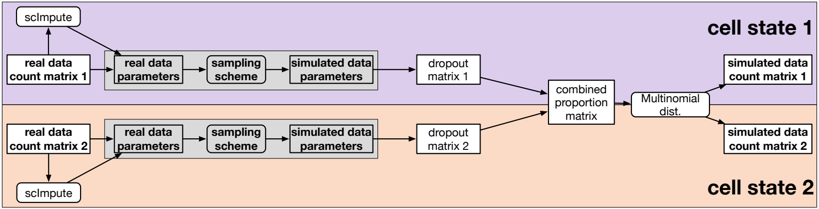
$$\rightarrow$$

two scRNA-seq datasets

Can generalize to multiple cell states

The Generative Framework of scDesign: Two-state Mode

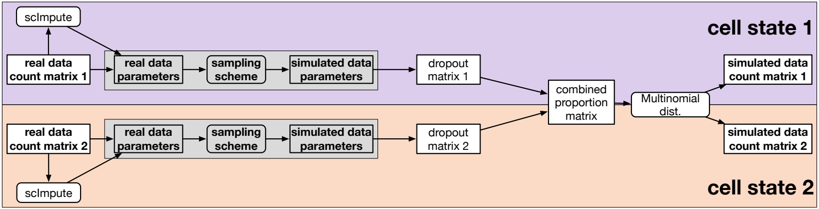
a. Independently estimate real data parameters for the two cell states



The Generative Framework of scDesign: Two-state Mode

a. Independently estimate real data parameters for the two cell states

b. Independently simulate ideal gene expression levels for new cells of the two cell states

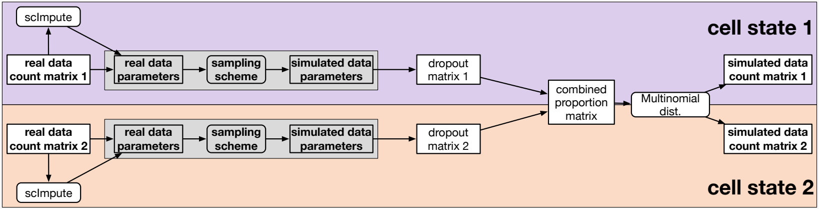


The Generative Framework of scDesign: Two-state Mode

a. Independently estimate real data parameters for the two cell states

b. Independently simulate ideal gene expression levels for new cells of the two cell states

c. Introduce dropout values based on the estimated dropout parameters of each state



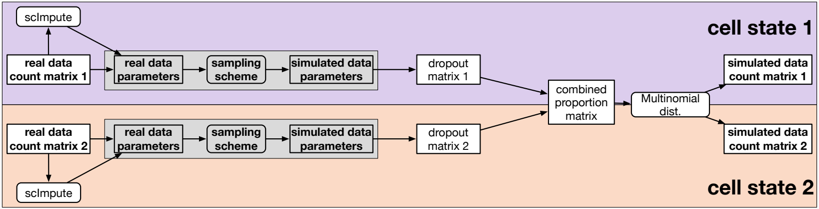
The Generative Framework of scDesign: Two-state Mode

a. Independently estimate real data parameters for the two cell states

b. Independently simulate ideal gene expression levels for new cells of the two cell states

c. Introduce dropout values based on the estimated dropout parameters of each state

d. Generate observed read counts by accounting for the fact that RNA fragments from the two batches of cells compete for the total sequencing depth



Problem Formulation

For one-state mode, given

- a real single-cell count matrix $\boldsymbol{X}^{\text{real}}$
- I genes and J_0 cells
- constraints on the sequencing depth and cell number

simulate a new count matrix with I genes and J cells

Problem Formulation

For one-state mode, given

- a real single-cell count matrix $\boldsymbol{X}^{\text{real}}$
- I genes and J_0 cells
- constraints on the sequencing depth and cell number

simulate a new count matrix with I genes and J cells

For two-state mode, given

- two real single-cell count matrices $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$
- cell state 1 with I genes and J_{01} cells
- cell state 2 with I genes and J_{02} cells
- constraints on sequencing depth and cell numbers

simulate a new count matrix for each cell state.

Estimation the Five Parameters

- $\boldsymbol{X}^{\text{real}}$: real single-cell gene count matrix
- I: number of genes (rows)
- J_0: number of cells (columns)
- $\hat{s}_{0j}$: cell-wise library size

$$\hat{s}_{0j} = \sum_{i=1}^{I} X_{ij}^{\text{real}},\ j = 1, \ldots, J_0$$

Estimation the Five Parameters

- $\boldsymbol{X}^{\text{real}}$: real single-cell count matrix
- I: number of genes (rows)
- J_0: number of cells (columns)
- $\hat{q}_{0j}$: cell-wise dropout rate

$$\hat{q}_{0j} = \frac{1}{I}\sum_{i=1}^{I} \mathbb{I}\{X_{ij}^{\text{real}} = 0\},\ j = 1, \ldots, J_0$$

Estimation the Five Parameters

Normalization and transformation:

$$X_{ij}^{\log} = \log_{10}\left(\frac{\text{median}\{\hat{s}_{01},\ldots,\hat{s}_{0J}\}}{\hat{s}_{0j}} X_{ij}^{\text{real}} + 1.01\right)$$

Estimation the Five Parameters

Normalization and transformation:

$$X_{ij}^{\log} = \log_{10}\left(\frac{\mathrm{median}\{\hat{s}_{01}, \ldots, \hat{s}_{0J}\}}{\hat{s}_{0j}} X_{ij}^{\mathrm{real}} + 1.01\right)$$

Assume $X_{i1}^{\log}, \ldots, X_{iJ_0}^{\log}$ are i.i.d. following the density function

$$f_i(x) = \lambda_{0i}\,\mathrm{Gamma}\left(x; \alpha_{0i}, \beta_{0i}\right) + (1 - \lambda_{0i})\,\mathrm{Normal}\left(x; \mu_{0i}, \sigma_{0i}^2\right)\ ,\ x \in \mathbb{R}\,,$$

- λ_{0i}: gene-wise dropout rate
- μ_{0i}: gene expression mean
- σ_{0i}: gene expression standard deviation

Estimation the Five Parameters

Normalization and transformation:

$$X_{ij}^{\log} = \log_{10}\left(\frac{\text{median}\{\hat{s}_{01}, \ldots, \hat{s}_{0J}\}}{\hat{s}_{0j}} X_{ij}^{\text{real}} + 1.01\right)$$

Assume $X_{i1}^{\log}, \ldots, X_{iJ_0}^{\log}$ are i.i.d. following the density function

$$f_i(x) = \lambda_{0i}\,\text{Gamma}\left(x; \alpha_{0i}, \beta_{0i}\right) + (1 - \lambda_{0i})\,\text{Normal}\left(x; \mu_{0i}, \sigma_{0i}^2\right)\ ,\ x \in \mathbb{R}\,,$$

- λ_{0i}: gene-wise dropout rate
- μ_{0i}: gene expression mean
- σ_{0i}: gene expression standard deviation
- EM algorithm → $\hat{\lambda}_{0i}$, $\hat{\mu}_{0i}$, $\hat{\sigma}_{0i}$

Experimental Design for Differential Gene Expression

Scenario 1:

- Cells from the two cell states are prepared as two separate libraries and sequenced independently
    - Cells collected at two differentiating time points
    - Cells of the same tissue type from patients and healthy subjects
    - Cells of the same type but exposed to different experimental treatments
- Select the optimal cell numbers simultaneously for two libraries to optimize the subsequent DE analysis
- Constraints are the total sequencing depths of the two cell states

Experimental Design for Differential Gene Expression

In Scenario 1, given

- Two real count matrices $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$ from two cell states
- Pre-determined total sequencing depths for the two cell states

For each pair of candidate cell numbers,

1. Simulate one synthetic real count matrix for each cell state
2. Perform DE analysis
3. Calculate the target criterion (e.g., FDR)
    - “True” DE genes are top N genes ranked by

$$\frac{|\hat{\mu}_{0i}^{1} - \hat{\mu}_{0i}^{2}|}{\sqrt{(\hat{\sigma}_{0i}^{1})^2 + (\hat{\sigma}_{0i}^{2})^2}}.$$

Experimental Design for Differential Gene Expression

In Scenario 1, given

- Two real count matrices $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$ from two cell states
- Pre-determined total sequencing depths for the two cell states

For each pair of candidate cell numbers,

1. Simulate one synthetic real count matrix for each cell state
2. Perform DE analysis
3. Calculate the target criterion (e.g., FDR)
   - "True" DE genes are top N genes ranked by

$$\frac{|\hat{\mu}_{0i}^{1} - \hat{\mu}_{0i}^{2}|}{\sqrt{(\hat{\sigma}_{0i}^{1})^2 + (\hat{\sigma}_{0i}^{2})^2}}\,.$$

⇒ Select the best pair of cell numbers based on the target criterion

Experimental Design for Differential Gene Expression

In Scenario 2, given

- Two real count matrices $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$ from two cell states
- Pre-determined total sequencing depth
- Proportions of the two cell states
  - Collected from domain knowledge or estimated from real data

For each candidate total cell number,

1. Simulate one synthetic real count matrix for each cell state
2. Perform DE analysis
3. Calculate the target criterion (e.g., FDR)

Experimental Design for Differential Gene Expression

In Scenario 2, given

- Two real count matrices $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$ from two cell states
- Pre-determined total sequencing depth
- Proportions of the two cell states
    - Collected from domain knowledge or estimated from real data

For each candidate total cell number,

1. Simulate one synthetic real count matrix for each cell state
2. Perform DE analysis
3. Calculate the target criterion (e.g., FDR)

⇒ Select the best cell number based on the target criterion

Results

Evaluation of the Simulated Data by scDesign

Six scRNA-seq protocols, each having three cell types:

- Smart-seq2 [Picelli et al., 2013]
- Fluidigm C1 [Pollen et al., 2014]
- Drop-seq [Macosko et al., 2015]
- 10x Genomics [Zheng et al., 2017]
- inDrop [Klein et al., 2015]
- Seq-Well [Gierahn et al., 2017]

Evaluation of the Simulated Data by scDesign

Six scRNA-seq protocols, each having three cell types:

- Smart-seq2 [Picelli et al., 2013]
- Fluidigm C1 [Pollen et al., 2014]
- Drop-seq [Macosko et al., 2015]
- 10x Genomics [Zheng et al., 2017]
- inDrop [Klein et al., 2015]
- Seq-Well [Gierahn et al., 2017]

Five simulation methods:

- scDesign
- splat [Zappia et al., 2017]
- powsimR [Vieth et al., 2017]
- scDD [Korthauer et al., 2016]
- Lun [Lun et al., 2016]

Evaluation of the Simulated Data by scDesign

For each real count matrix, randomly split the cells into two subsets:

- one half used to estimate gene expression parameters and simulate new count matrices
- the other half used to evaluate the simulation results

Compare methods using six summary statistics

- Per-gene count mean, count variance, count coefficient of variation, and zero proportion
- Per-cell library size and zero proportion

Evaluation of the Simulated Data by scDesign

For each real count matrix, randomly split the cells into two subsets:

- one half used to estimate gene expression parameters and simulate new count matrices
- the other half used to evaluate the simulation results

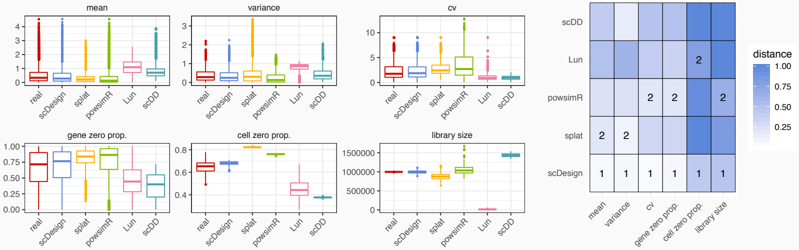
Compare methods using six summary statistics

- Per-gene count mean, count variance, count coefficient of variation, and zero proportion
- Per-cell library size and zero proportion

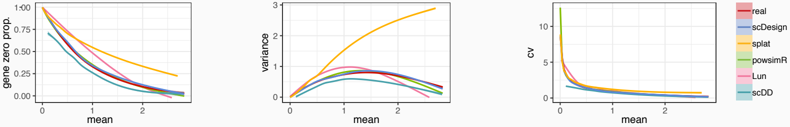
scDesign is ranked the best in 84 comparisons and the second best in 20 comparisons, among all the 108 comparisons.

Evaluation Results based on Smart-seq2

six summary statistics

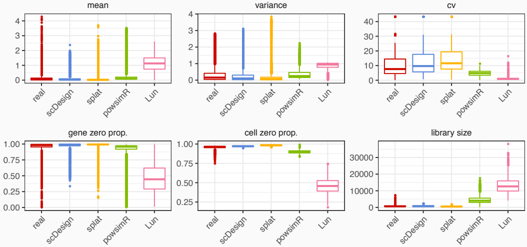


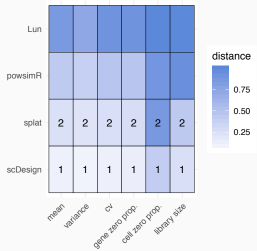
relationships between statistics



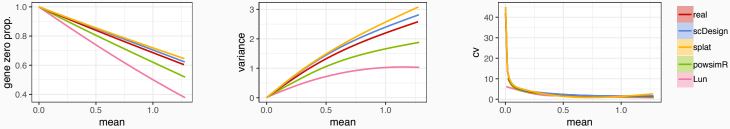
Evaluation Results based on Drop-seq

six summary statistics





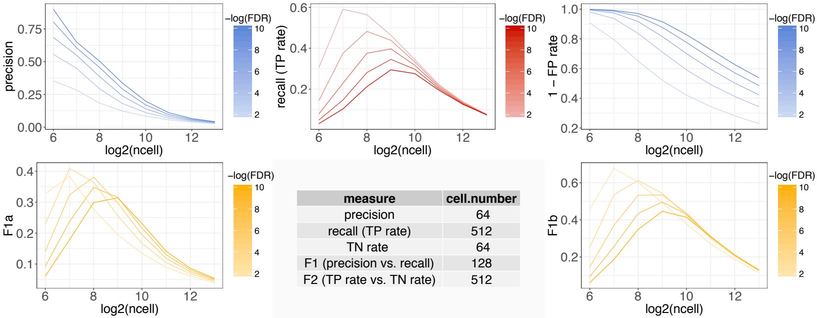
relationships between statistics



Gene Differential Expression (Scenario 1)

Same cell types but different protocols:

Astrocytes vs. Oligodendrocytes (Fluidigm C1)

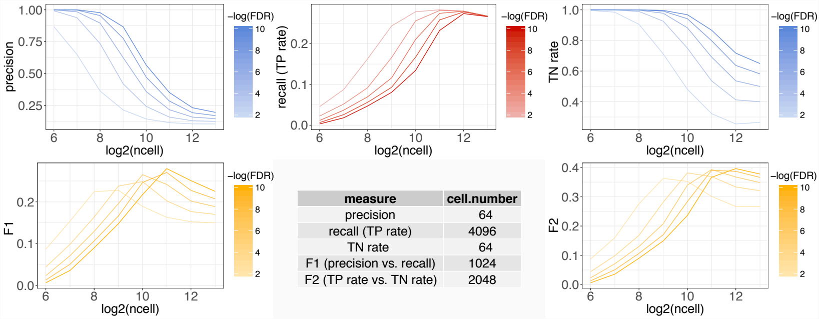


measure	cell.number
precision	64
recall (TP rate)	512
TN rate	64
F1 (precision vs. recall)	128
F2 (TP rate vs. TN rate)	512

Gene Differential Expression (Scenario 1)

Same cell types but different protocols:

Astrocytes vs. Oligodendrocytes (inDrop)

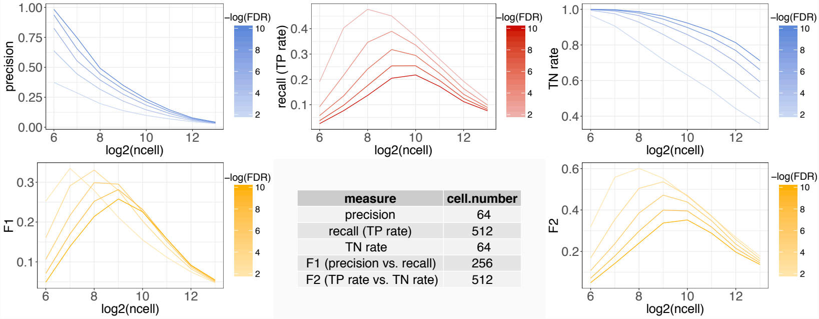


measure	cell.number
precision	64
recall (TP rate)	4096
TN rate	64
F1 (precision vs. recall)	1024
F2 (TP rate vs. TN rate)	2048

Gene Differential Expression (Scenario 1)

Different cell types but same protocol:

Dendrocytes subtype 1 vs. Dendrocytes subtype 2 (Smart-seq2)

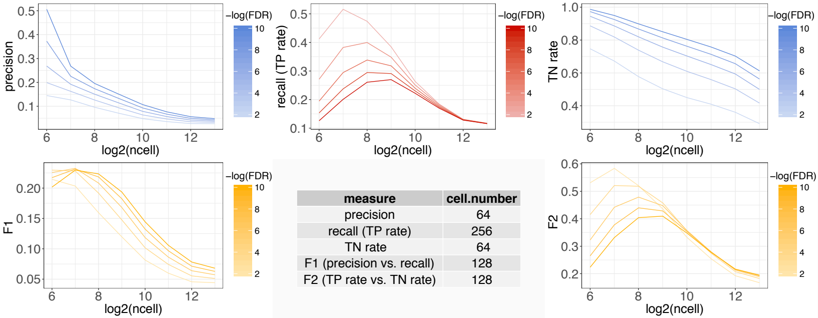


measure	cell.number
precision	64
recall (TP rate)	512
TN rate	64
F1 (precision vs. recall)	256
F2 (TP rate vs. TN rate)	512

Gene Differential Expression (Scenario 1)

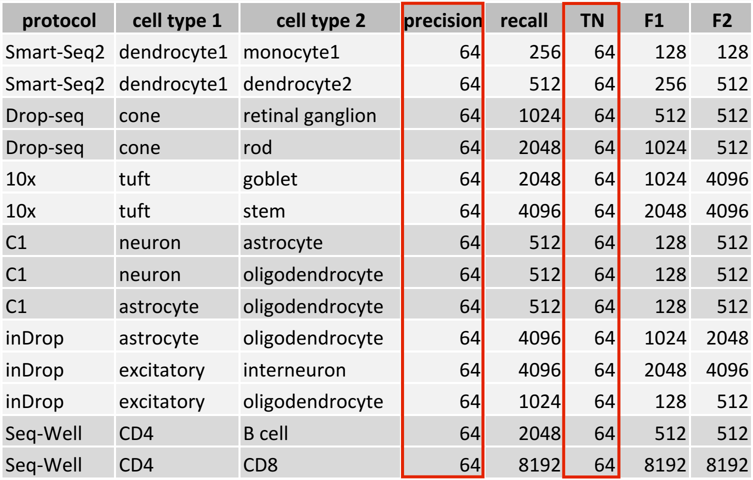
Different cell types but same protocol:

Dendrocytes vs. Monocytes (Smart-seq2)



measure	cell.number
precision	64
recall (TP rate)	256
TN rate	64
F1 (precision vs. recall)	128
F2 (TP rate vs. TN rate)	128

Gene Differential Expression (Scenario 1)

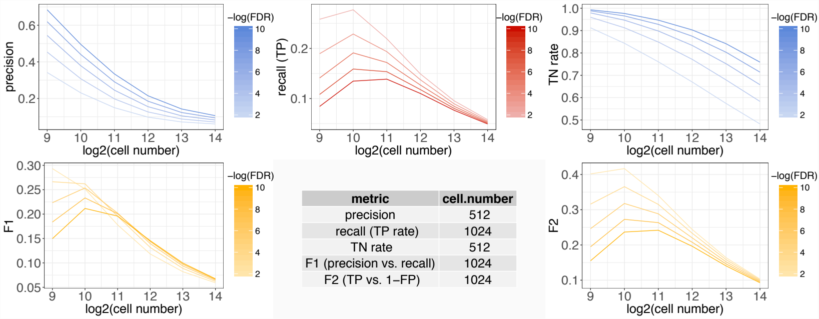


protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1	monocyte1	64	256	64	128	128
Smart-Seq2	dendrocyte1	dendrocyte2	64	512	64	256	512
Drop-seq	cone	retinal ganglion	64	1024	64	512	512
Drop-seq	cone	rod	64	2048	64	1024	512
10x	tuft	goblet	64	2048	64	1024	4096
10x	tuft	stem	64	4096	64	2048	4096
C1	neuron	astrocyte	64	512	64	128	512
C1	neuron	oligodendrocyte	64	512	64	128	512
C1	astrocyte	oligodendrocyte	64	512	64	128	512
inDrop	astrocyte	oligodendrocyte	64	4096	64	1024	2048
inDrop	excitatory	interneuron	64	4096	64	2048	4096
inDrop	excitatory	oligodendrocyte	64	1024	64	128	512
Seq-Well	CD4	B cell	64	2048	64	512	512
Seq-Well	CD4	CD8	64	8192	64	8192	8192

Gene Differential Expression (Scenario 2)

Same cell types but different protocols and cell proportions:

Human Astrocytes (19%) vs. Oligodendrocytes (15%) (Fluidigm C1)

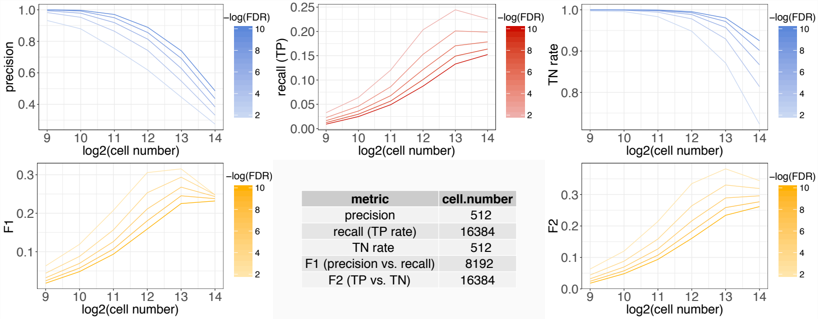


metric	cell.number
precision	512
recall (TP rate)	1024
TN rate	512
F1 (precision vs. recall)	1024
F2 (TP vs. 1−FP)	1024

Gene Differential Expression (Scenario 2)

Same cell types but different protocols and cell proportions:

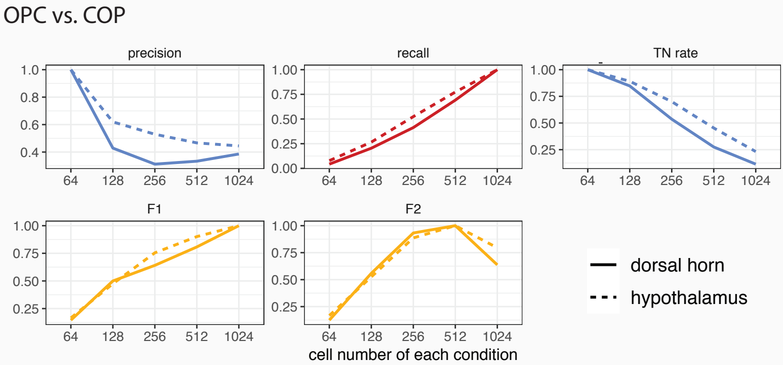
Human Astrocytes (8%) vs. Oligodendrocytes (11%) (inDrop)



metric	cell.number
precision	512
recall (TP rate)	16384
TN rate	512
F1 (precision vs. recall)	8192
F2 (TP vs. TN)	16384

scDesign Demonstrates Reproducibility across Datasets

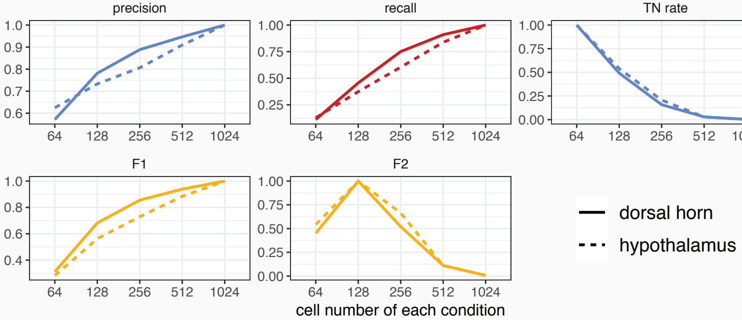
Experimental design based on datasets from two brain regions: dorsal horn and hypothalamus [Marques et al., 2016] (Fluidigm C1)



scDesign Demonstrates Reproducibility across Datasets

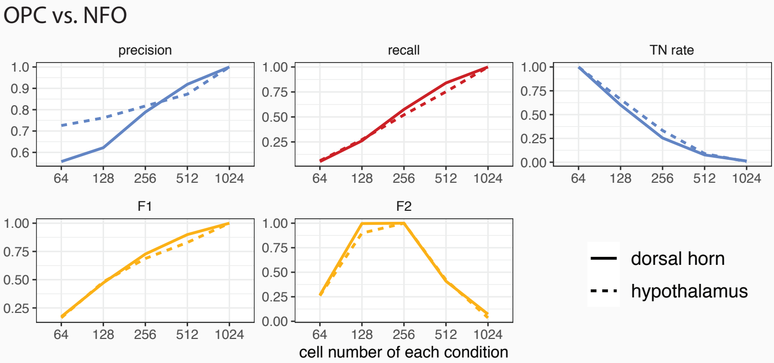
Experimental design based on datasets from two brain regions: dorsal horn and hypothalamus [Marques et al., 2016] (Fluidigm C1)

OPC vs. MFO



scDesign Demonstrates Reproducibility across Datasets

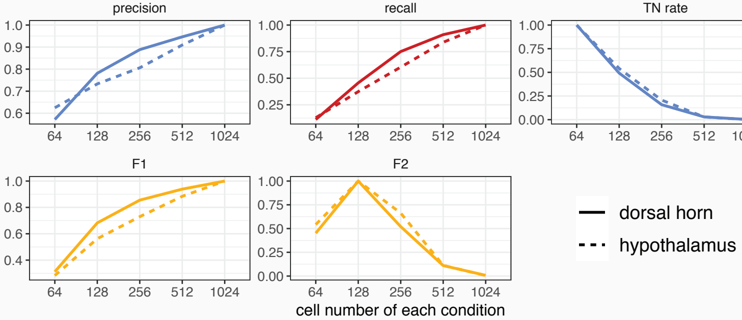
Experimental design based on datasets from two brain regions: dorsal horn and hypothalamus [Marques et al., 2016] (Fluidigm C1)



scDesign Demonstrates Reproducibility across Datasets

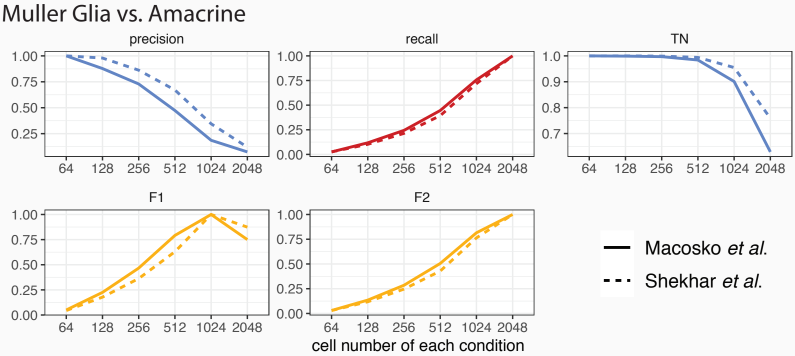
Experimental design based on datasets from two brain regions: dorsal horn and hypothalamus [Marques et al., 2016] (Fluidigm C1)

OPC vs. MFO



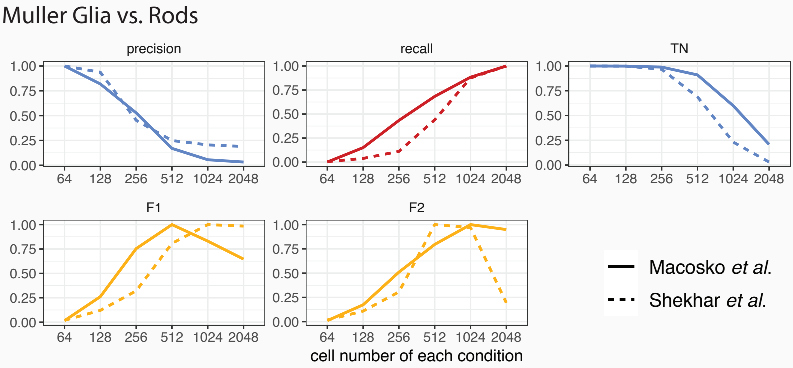
scDesign Demonstrates Reproducibility across Studies

Experimental design based on datasets from two independent studies: [Macosko et al., 2015] and [Shekhar et al., 2016] (Drop-seq)



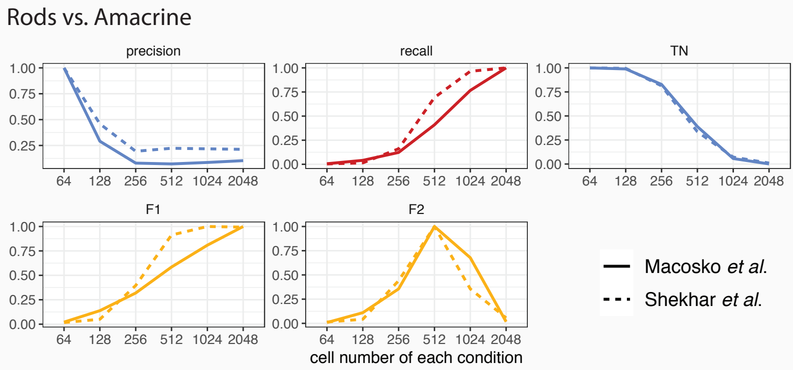
scDesign Demonstrates Reproducibility across Studies

Experimental design based on datasets from two independent studies: [Macosko et al., 2015] and [Shekhar et al., 2016] (Drop-seq)



scDesign Demonstrates Reproducibility across Studies

Experimental design based on datasets from two independent studies: [Macosko et al., 2015] and [Shekhar et al., 2016] (Drop-seq)



scDesign Assists Comparison of DE methods

Five DE methods:

- two-sample *t* test (baseline)
- MAST [Finak et al., 2015]
- SCDE [Kharchenko et al., 2014]
- DESeq2 + zingeR [Love et al., 2014, Van den Berge et al., 2017]
- edgeR + zingeR [Robinson et al., 2010]

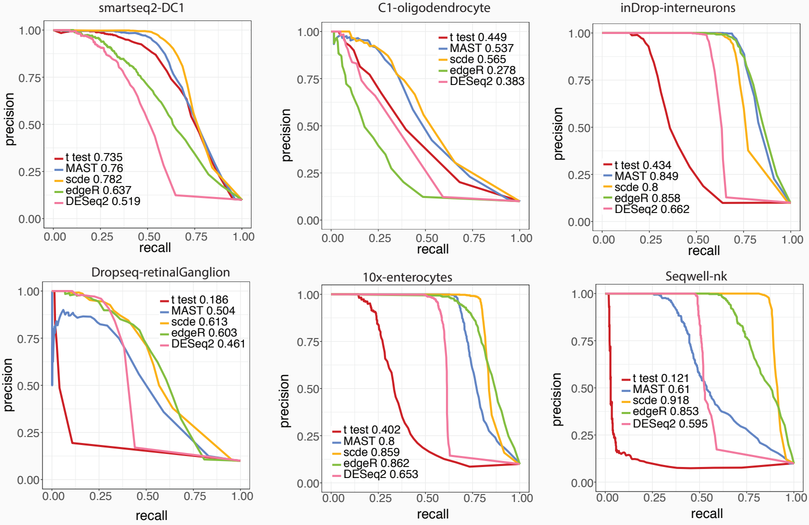
scDesign Assists Comparison of DE methods

Five DE methods:

- two-sample *t* test (baseline)
- MAST [Finak et al., 2015]
- SCDE [Kharchenko et al., 2014]
- DESeq2 + zingeR [Love et al., 2014, Van den Berge et al., 2017]
- edgeR + zingeR [Robinson et al., 2010]

For each of the six protocols, simulate a pair of read count matrices with 5% DE genes

scDesign Assists Comparison of DE methods



scDesign Assists Comparison of DE methods

Ranking the DE methods for each protocol:

- Smart-seq2: SCDE $>$ MAST $>$ t test $>$ edgeR $>$ DESeq2
- Fluidigm C1: SCDE $>$ MAST $>$ t test $>$ DESeq2 $>$ edgeR

scDesign Assists Comparison of DE methods

Ranking the DE methods for each protocol:

- Smart-seq2: SCDE $>$ MAST $>$ t test $>$ edgeR $>$ DESeq2
- Fluidigm C1: SCDE $>$ MAST $>$ t test $>$ DESeq2 $>$ edgeR

- inDrop: edgeR $>$ MAST $>$ SCDE $>$ DESeq2 $>$ t test
- 10x Genomics: edgeR $>$ SCDE $>$ MAST $>$ DESeq2 $>$ t test

scDesign Assists Comparison of DE methods

Ranking the DE methods for each protocol:

- Smart-seq2: SCDE > MAST > *t* test > edgeR > DESeq2
- Fluidigm C1: SCDE > MAST > *t* test > DESeq2 > edgeR

- inDrop: edgeR > MAST > SCDE > DESeq2 > *t* test
- 10x Genomics: edgeR > SCDE > MAST > DESeq2 > *t* test

- Drop-seq: SCDE > edgeR > MAST > DESeq2 > *t* test
- Seq-Well: SCDE > edgeR > MAST > DESeq2 > *t* test

scDesign Assists Comparison of Dimension Reduction methods

Four dimension reduction methods:

- principal component analysis (PCA)
- t-distributed stochastic neighbor embedding (tSNE)
- independent component analysis (ICA) [Hyvärinen and Oja, 2000]
- ZINB-WaVE [Risso et al., 2018]

scDesign Assists Comparison of Dimension Reduction methods

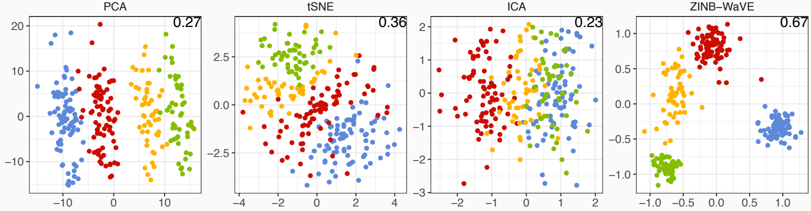
Four dimension reduction methods:

- principal component analysis (PCA)
- t-distributed stochastic neighbor embedding (tSNE)
- independent component analysis (ICA) [Hyvärinen and Oja, 2000]
- ZINB-WaVE [Risso et al., 2018]

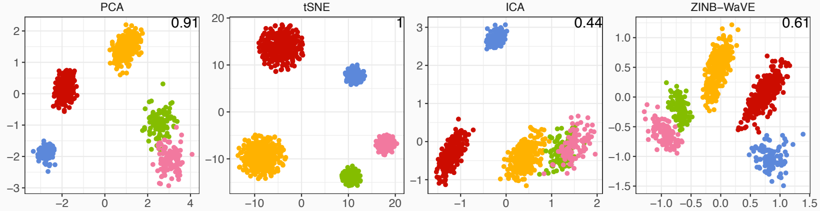
For each of the six protocols, simulate a set of read count matrices following a differentiation path

scDesign Assists Comparison of Dimension Reduction methods

Fluidigm C1



inDrop



scDesign Assists Comparison of Dimension Reduction methods

Ranking the dimension reduction methods for each protocol:

- Smart-seq2: PCA > ZINB-WaVE > ICA > tSNE
- Fluidigm C1: ZINB-WaVE > tSNE > PCA > ICA

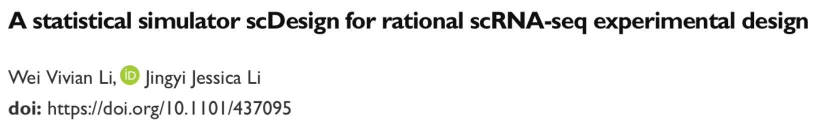
scDesign Assists Comparison of Dimension Reduction methods

Ranking the dimension reduction methods for each protocol:

- Smart-seq2: PCA $>$ ZINB-WaVE $>$ ICA $>$ tSNE
- Fluidigm C1: ZINB-WaVE $>$ tSNE $>$ PCA $>$ ICA

- inDrop: tSNE $>$ PCA $>$ ZINB-WaVE $>$ ICA
- 10x Genomics: tSNE $>$ PCA $>$ ZINB-WaVE $>$ ICA
- Drop-seq: tSNE $>$ PCA $>$ ICA $>$ ZINB-WaVE
- Seq-Well: tSNE $>$ PCA $>$ ZINB-WaVE $\approx$ ICA

Conlusions



A statistical simulator scDesign for rational scRNA-seq experimental design

Wei Vivian Li, Jingyi Jessica Li

doi: https://doi.org/10.1101/437095

- scDesign simulates synthetic scRNA-seq datasets that well capture gene expression characteristics in real data
- The experimental design depends on
  - analysis task
  - scRNA-seq protocol
  - cell population heterogeneity
- Cross-study comparisons verify that scDesign leads to reproducible experimental designs

`https://github.com/Vivianstats/scDesign`