The shaky foundations of simulating single-cell RNA sequencing data

Seonmi Choi

Chung-Ang University

Abstract

Backgrounds

- Single-cell RNA sequencing (scRNA-seq) dataset의 데이터를 분석하는 방법의 성능 평가가 반복적으로 이루어짐
- 벤치마크 연구: 매개변수의 변화에 따른 분석 방법, 계산 비용 측면에서의 확장성, 다양한 시나리오에서의 성능, 방법들의 경쟁력 평가

Results

- Synthetic scRNA-seq 데이터 생성 방법 평가

1. 세포, 배치, 클러스터 수준에서 품질 관리 요약 지표를 1차원 및 2차원 설정에서 비교, 정량화
2. 시뮬레이터가 클러스터링 및 배치 교정 방법 비교에 미치는 영향 조사
3. Quality control summaries의 참조 데이터와 합성 데이터간의 유사성 포착을 분석

Conclusions

- 대부분의 시뮬레이터는 복잡한 실험 설계 반영하지 못함
- Integration 방법의 성능을 과대평가
- 시뮬레이션 기반 방법 비교에서 summary의 정의 필요

Background

- 시뮬레이션 분석 방법
  - Differential expression analysis
  - Trajectory inference
  - Data integration

- 일반적인 시뮬레이션 방법은 제한적이고 편향된 벤치마크 평가 제시

  ⇒ scRNA-seq 데이터의 주요 특성들이 잘 재현되는지 중립적인 평가 필요

  counts, sample-effect, subpopulation-effect 등

Background

- 시뮬레이션 방법 구분
  - De novo: 사용자가 정의한 매개변수를 기반으로 데이터 생성,
    서로 다른 세포 그룹 또는 샘플 간의 인위적 차이 도입
  - Reference 데이터셋 활용: 실제 참조 데이터에서 관찰된 유전자 발현 패턴을
    재현하도록 매개변수를 추정하는 접근법

- 16가지 scRNA-seq data의 시뮬레이션 방법을 평가

Results

Benchmark design

- 12개의 데이터셋

  Type n: 하나의 배치와 하나의 클러스터

  Type b: 여러 개의 배치 포함

  Type k: 여러 개의 클러스터 포함

Dataset	Subset(s)	Type	Batch(es)	Cluster(s)
CellBench	✗	b,k	3	3
	H2228	b	3	H2228
	celseq	k	sc_celseq	3
Ding20	✗	b,k	4	8
	10x.InhibNeuron	n	10x Chromium	Inhibitory neuron
	ExcitNeuron	b	4	Excitatory neuron
	DroNcSeq	k	DroNc-seq	5
Gierahn17	✓	n	0	0
Kang18	✗	b,k	8	8
	1015	k	1015	6
	B	n	1015	B cells
	NK	n	1015	NK cells
Koh16	✓	k	0	7
MCA20	✗	b,k	13	9
	gland.AT2	b	4	T cell_Cd8b1 high
	lung.AT2	b	4	AT2 Cell
Mereu20	✗	b,k	13	9
	CD4T	b	13	CD4 T cells
	ddSeq	k	ddSeq	9
Oetjen18	✓	b	18	0
	R	n	R	0
panc8	✗	b,k	5	9
	inDrop1.beta	n	indrop1	beta
	inDrop.ductal	b	indrop1-4	ductal
	SmartSeq2	k	smartseq2	7
TabulaMuris	✗	b,k	10	31
	limb.MSCs	n	Limb_Muscle	mesenchymal stem cell
	spleen	k	Spleen	4
Tung17	✓	b	3	0
	NA19101	n	NA19101	0
Zheng17	✓	k	0	7
	HSCs	n	0	HSCs CD34+
	Monocytes	n	0	Monocytes CD14+

Results

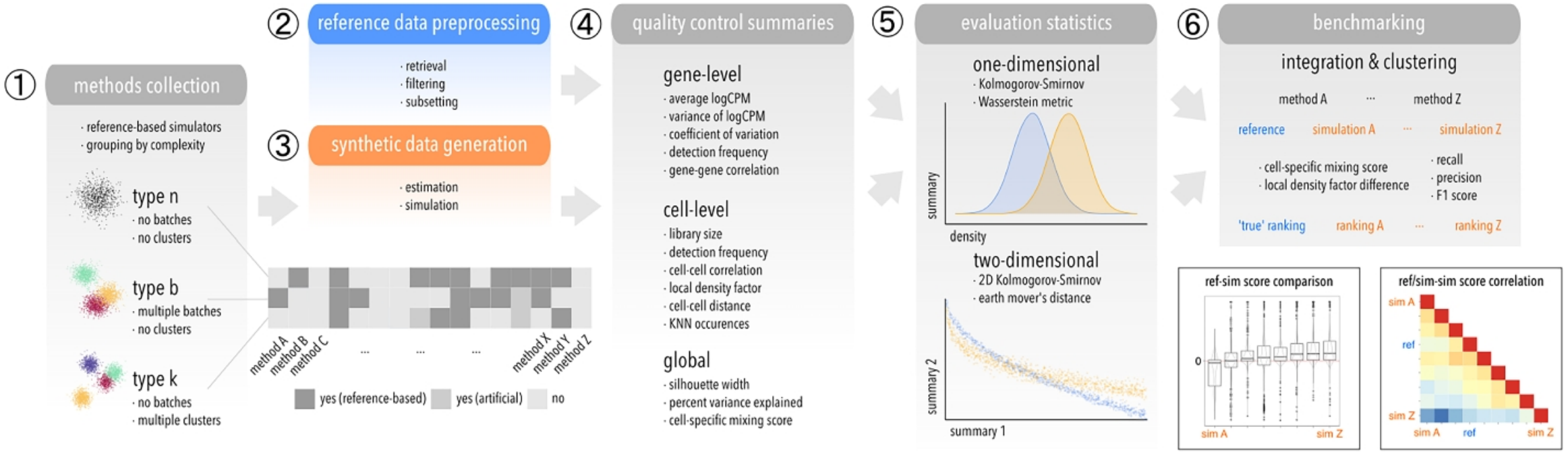
Benchmark design

- 16개의 시뮬레이션 방법 선택

1. 최소한의 수동적 수정(minor manual adjustment)을 통해 설치 및 실행 가능한 도구
2. 실제 참조 데이터로부터 매개변수를 추정하는 참조기반(referenced-base)시뮬레이션 방법

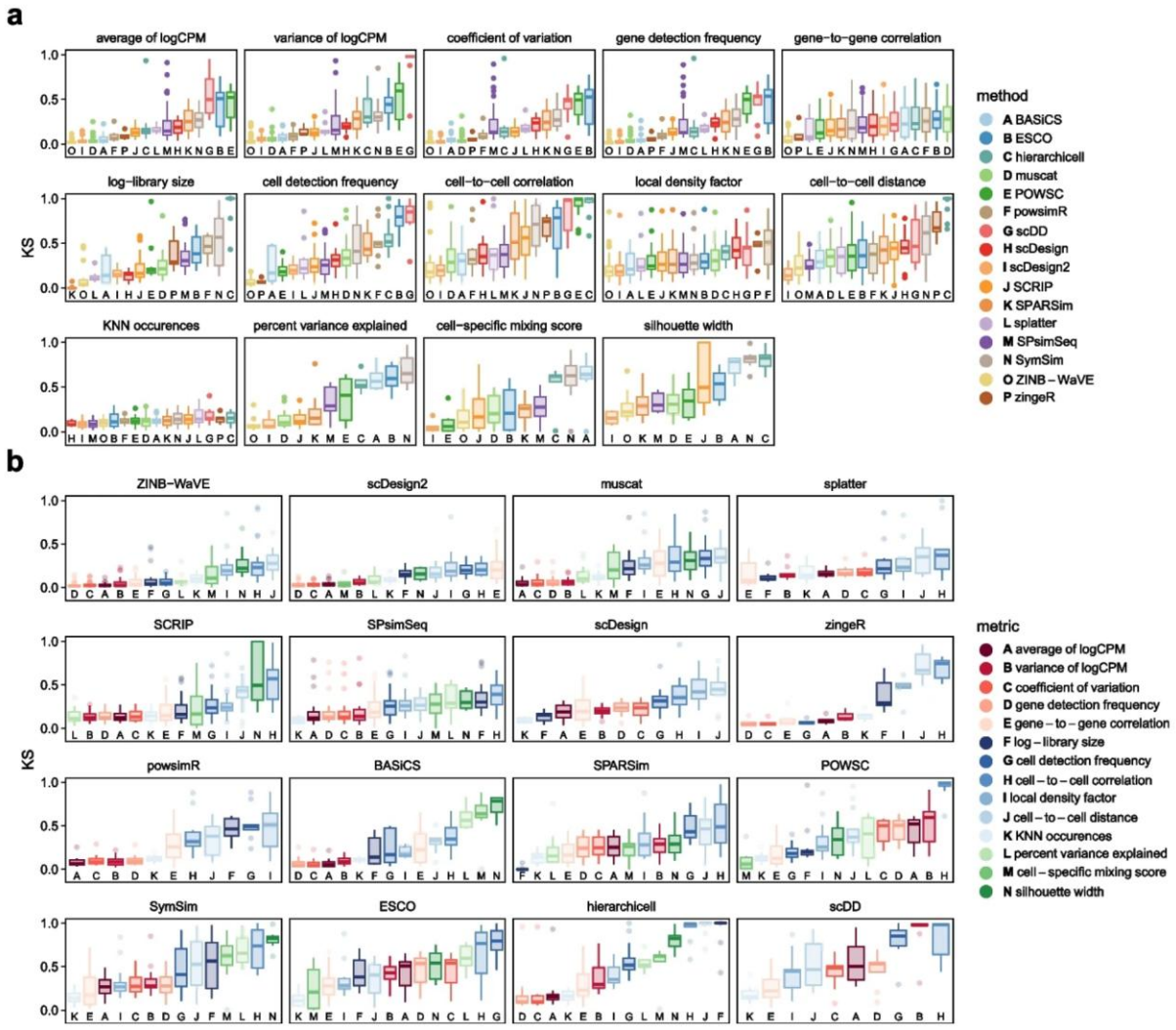
	Batches	Clusters	Type(s)	Cell #	Parallelization	Availability	Year	Model
BASiCS [37]	✓	✗	b	✗	✓✗	R/Bioc	2015	NB
ESCO [38]	✓	✓	n,b,k	✓	✓✓	R/GitHub	2020	Gamma-Poisson
hierarchicell [39]	✓	✗	n,b	✓	✗✗	R/GitHub	2021	NB
muscat [40]	✓	✓	n,b,k	(✓)†	✗✗	R/Bioc	2020	NB
POWSC [41]	✗	✓	n,k	(✓)†	✗✗	R/Bioc	2020	zero-inflated, log-normal Poisson mixture
powsimR [42]	✗	(✓)	n*	(✓)†	✓✓	R/GitHub	2017	NB
scDD [43]	✗	✗	n*	✓	✓✓	R/Bioc	2016	Bayesian NB mixture model
scDesign [44]	✗	(✓)	n	✓	○✓	R/GitHub	2019	Gamma-Normal mixture model
scDesign2 [45]	✗	✓	n,k	✓	✓✗	R/GitHub	2020	(zero-inflated) Poisson or NB + Gaussian copula for gene-gene correlations
SCRIP [46]	✓	✓	n,b,k	✓	✗✗	R/GitHub	2020	(Beta-)Gamma-Poisson
SPARSim [47]	✓	✗	n,b	(✓)‡	✗✗	R/GitLab	2020	Gamma-multivariate hypergeometric
splatter [15] (Splat model)	(✓)	(✓)	n	✓	✗✗	R/Bioc	2017	Gamma-Poisson
SPsimSeq [16]	✓	✗	n,b	✓	○✗	R/Bioc	2020	log-linear model-based density estimation + Gaussian copula for gene-gene correlations
SymSim [48]	✓	✗	n,b	✓	✗✗	R/GitHub	2019	kinetic model using MCMC
ZINB-WaVE [49]	✓	✓	n,b,k	✗	✗✗	R/Bioc	2018	zero-inflated NB
zingeR [50]	✗	✗	n	(✓)†‡	✗✗	R/GitHub	2017	zero-inflated NB

Benchmark design



Results

Simulators vary in their ability to mimic scRNA-seq data characteristics



- splatter: 널리 사용되지만 대부분의 요약 지표에서 중간 수준의 성능

- 모델 비교

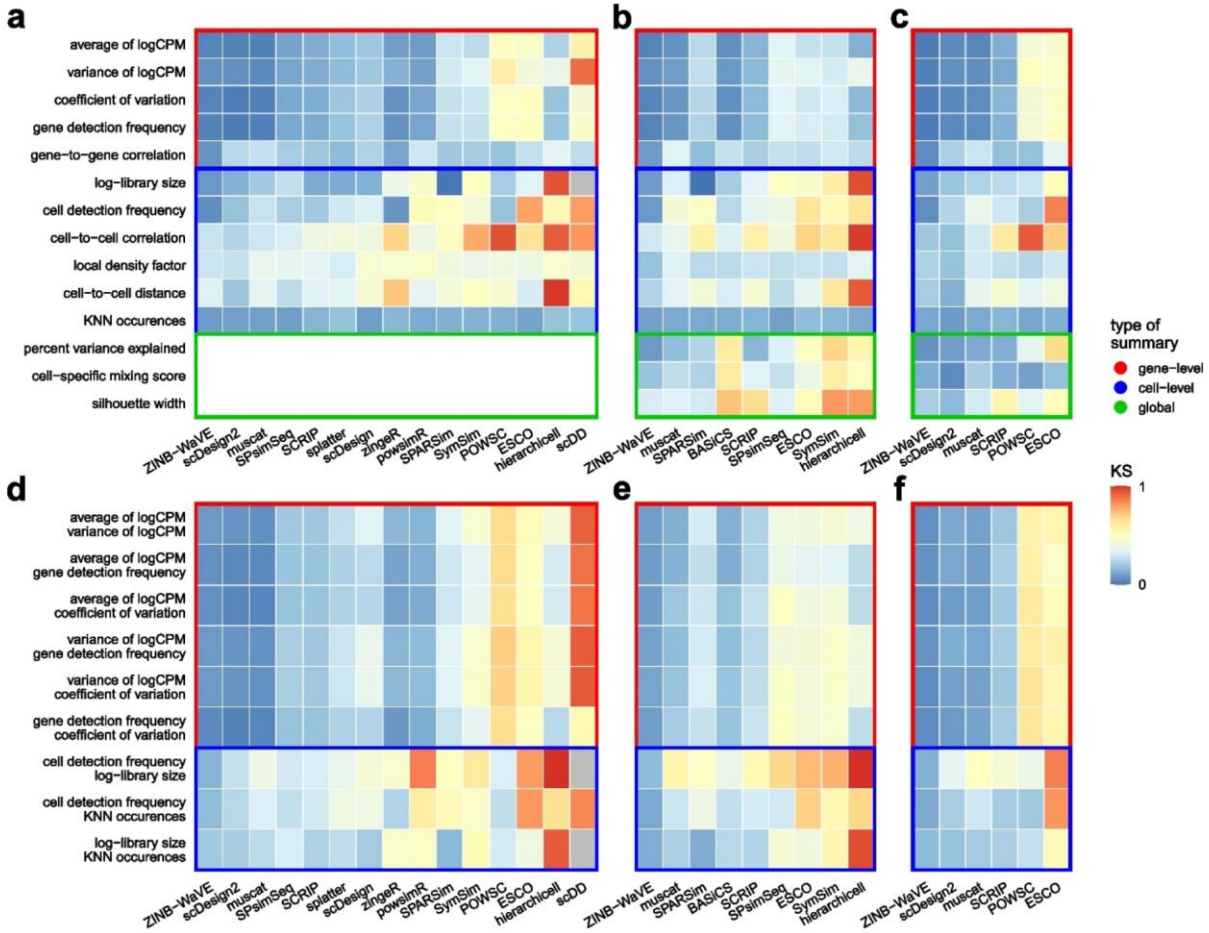
scDD와 hierarchicell, ESCO는 낮은 성능

ZINB-WaVE, scDesign2, muscat이 우수한 성능

- 다양한 지표와 데이터셋 전반에서 낮은 KS 통계량
- 특히 Cell-to-cell correlation 지표에서

실제와 시뮬레이션 데이터의 차이가 큼

Simulators vary in their ability to mimic scRNA-seq data characteristics



- 모델 비교

ZINB-WaVE, scDesign2, muscat, SPsimSeq: 유사하게 우수한 성능

POWSC, ESCO, hierarchicell, scDD: 다양한 지표에서 낮은 순위

- 지표 비교

LDF, cell-to-cell distance 및 correlation는 제대로 재현되지 않음.

특히 global summaries에서는 PVE와 silhouette width도

type b와 k에서 낮은 재현성을 보임.

- 런타임 비교

BASiCS이 가장 느리며

ESCO, hierarchicell, muscat, POWSC, splatter이 가장 빠른 그룹

Batch simulators yield over-optimistic but faithful integration method performance

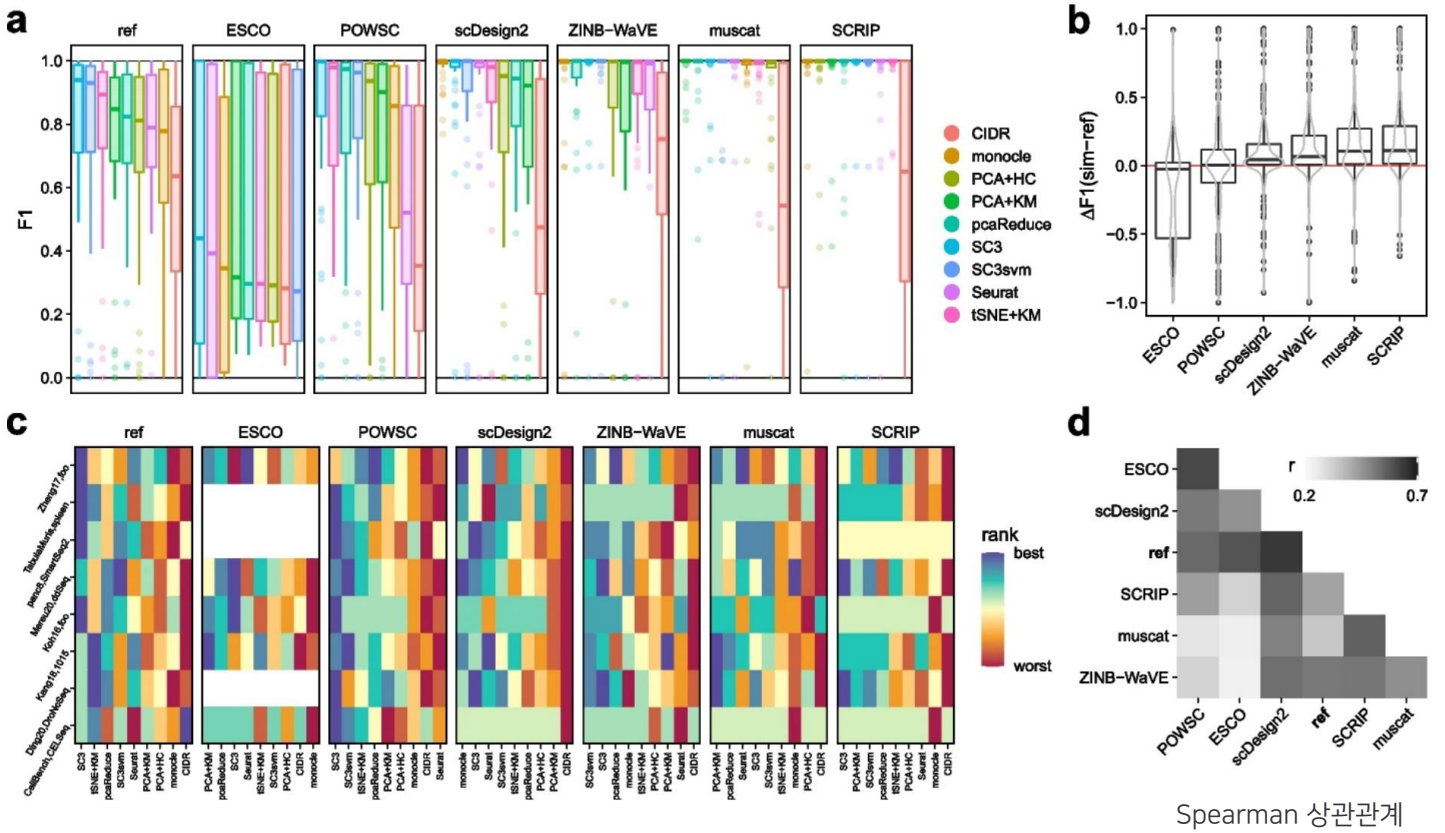
- 실제 데이터와 시뮬레이션에서 배치 보정 성능 순위가 일치하는지 확인하기 위해서 성능 비교
- 8개의 type b 실제 데이터 사용, 6가지 단일세포 RNA 시퀀싱 매치 보정 방법 성능 비교

- Batch Correction Score(BCS) = CMS* (Cell-specific Mixing Score) + LDF* (Local Density Factor)

- LDF*는 실제 데이터와 시뮬레이션 데이터 간 꽤 일치 했지만, CMS*는 대부분의 방법에서 일치도가 낮음

- 가장 높은 유사성: SPsimSeq, ZINB-WaVE, SPARsim, SCRIP이고,
- 가장 낮은 유사성: muscat과 Symsim

Cluster simulatiors affect the performance of clustering methods



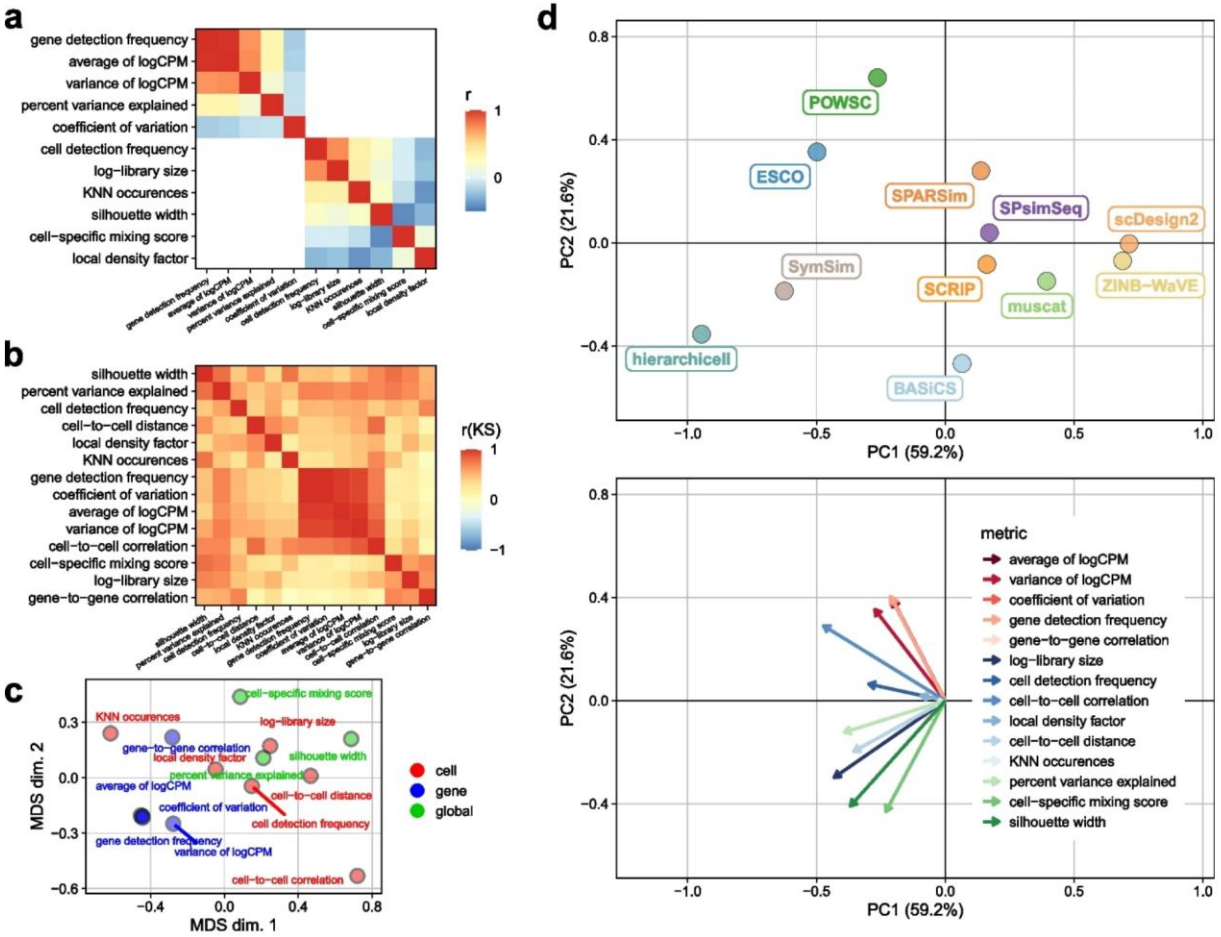
- type k 기반으로 클러스터링 방법 9개 비교
- 8개의 type k 실제 데이터를 사용해서 9가지 단일세포 클러스터링 알고리즘 평가

- 정답 라벨과 생성한 클러스터를 Hungarian algorithm을 통해 매칭, F1 score 계산

- 시뮬레이션 데이터에서 F1이 실제보다 항상 높게 나옴
⟶ 과대평가(over-optimistic)

- scDesign2, POWSC는 실제와 순위 유사,
muscat, SCRIP는 불일치 심함

Meta-analysis of summaries



- 시뮬레이션 데이터를 평가할 때 사용하는 summary의 중복성과 핵심 summary 판별

- MDS(Multidimensional Scaling) 분석을 통한 summary 간 거리 시각화

- PCA(principal component analysis) 분석을 통해 시뮬레이터와 데이터셋의 조합을 summary 통계량의 조합으로 표현

⇒ 중복이 많은 gene summary 등은 줄일 수 있으되, global summary는 다양하게 보는 것이 좋음

- summary의 필요성은 데이터 구조에 따라 달라질 수 있음

Discussion

- 시뮬레이터 평가로 gene/cell/global summary 기준으로 비교, 전체 성능 순위화
  - ZINB-WaVE는 거의 모든 평가에서 좋은 성능
- 시뮬레이터와 벤치마킹의 신뢰성 문제
  - 벤치마크 평가를 위해 ground trurh 필요,
  실험적으로 정답을 확보하기 어려운 경우를 위해 유연하고 신뢰성 있는 시뮬레이션 프레임워크 개발 필수적
- 시뮬레이터를 사용한 분석의 한계
  - 대부분 시뮬레이터는 단일 그룹만 생성 가능
  - 실제 데이터와 비교 시 클러스터링 결과 왜곡 존재
  - 클러스터링의 경우 시뮬레이션 기반 평가에서는 과대평가
  - 품질 평가 도구(scater, countsimQC) 활용의 부족
  - 성능 외 고려할 요소
    - splatter는 사용하기 쉽고 문서화가 잘 되어있으나, 다른 시뮬레이터는 성능은 좋더라도 사용성이 낮고 파라미터 해석이 어려움
- 목적별로 중요한 summary 지표를 정의하고, 평가 리포트를 표준화할 필요성
- trajectory 기반 시뮬레이터도 존재하기에 trajectory 관련 신뢰성 확보 필요(해당 연구는 type n, b, k로 시뮬레이터들을 분류)
- 향후 유연하고 해석 가능한 시뮬레이터 필요

Conclusions

- 현재 시뮬레이터들의 한계
  - 표현할 수 있는 데이터 복잡성 수준에 제한이 있음
  - 유전자 발현 차이를 만들기 위해 사용자의 인위적 입력에 의존함
  - 시뮬레이터가 얼마나 현실적인 데이터 특성을 반영할 수 있느냐에 따라,
  다른 분석 도구들을 평가하는 데 적합한 정도도 달라짐

- 시뮬레이션 기반 벤치마크 연구 결과는 사용하는 시뮬레이터에 따라 달라지며
시뮬레이터의 성능이 높다고 해서 integration이나 클러스터링 방법의 평가 결과의 신뢰도가 높지 않음

- 어떤 품질 지표(summary)를 어떻게 선택하느냐가 평가 결과에 영향을 미침
  - scRNA-seq 데이터 구조를 충실하게 반영할 수 있는 summary의 종류, 개수, 중요도를 정의하는 연구 필요

Reference datasets

Dataset	Description	Preprocessing	Batches	Clusters	Features	Observations	Source
CellBench	three human lung adenocarcinoma cell lines (HCC827, H1975, H2228) mixed in equal proportions and sequenced across three different platforms (CEL-Seq2, Drop-Seq, Chromium)	–	3	3	13575	1401	GSE118767
Gierahn17	human HEK293 (embryonic kidney cells) cell line sequenced with Seq-Well	–	–	–	24187	1453	GSE92495
Ding20	two mouse cortex snRNA-seq experiments (Cortex1 and Cortex2), each comprising 4 technologies (10x Chromium, DroNc-seq, sci-RNA-seq, Smart-Seq2)	retaining only first experiment (Cortex1) and cells that received a type annotation	4	8	28692	4523	SCP425
Kang18	droplet-based scRNA-seq data of PBMCs from eight patients, each measured before and after 6h treatment with IFN-β	retaining untreated samples only, removing multiplets and cells that did not receive a type annotation	8	8	17198	12315	GSE96583
Koh16	in vitro cultured H7 human embryonic stem cells (WiCell) and H7-derived downstream early mesoderm progenitors	–	–	9	60483	498	GSE85066
MCA20	Mouse Cell Atlas (MCA) dataset of Microwell-seq data from >28 tissues (2-4 replicates each) and cultures	retaining only features that are shared across all replicates (of a given tissue), and observations for which metadata was available	1-4	170	>10,000	>1,200,00	GSE108097
Mereu20	PBMC data from 13 platforms (Chromium, Chromium(sn), in-Drop, C1HT-small and -medium, CEL-Seq2, ddSEQ, Drop-Seq, ICELL8, MARS-Seq, Quartz-Seq2, mcSCRB-Seq, and Smart-Seq2)	–	13	9	23381	20237	GSE133549
Oetjen18	Droplet-based scRNA-seq of bone marrow mononuclear cells from 20 healthy donors of different sex and age (25 samples in total)	removal of replicated samples (Ck, C1, C2, Sk1, Sk2, S1, S2)	18	–	33694	72241	GSE120221
panc8	eight human pancreatic islet cell datasets from five technologies (CEL-Seq, CEL-Seq2, inDrop (four replicates), Fluidigm C1, SMART-Seq2)	retaining the inDrop (technical) replicate with the highest number of cells	5	13	23600	10963	GSE81076, GSE85241, GSE86469, E-MTAB-5061
TabulaMuris	droplet-based scRNA-seq data from Mus musculus (8 male and female mice) across 20 organs and tissues	–	10	13	23341	17404	GSE109774
Tung17	triplicated Fluidigm's C1 data of induced pluripotent stem cell (iPSC) lines of three individuals (9 samples in total)	–	–	3	20327	864	GSE77288
Zheng17	droplet-based scRNA-seq data of PBMCs from a single healthy individual	T cell subpopulations merged into CD4+ and CD8+	–	9	32738	68579	10x Genomics

Reference datasets

Dataset	Subset(s)	Type	Batch(es)	Cluster(s)
CellBench	✗	b,k	3	3
	H2228	b	3	H2228
	celseq	k	sc_celseq	3
Ding20	✗	b,k	4	8
	10x.InhibNeuron	n	10x Chromium	Inhibitory neuron
	ExcitNeuron	b	4	Excitatory neuron
	DroNcSeq	k	DroNc-seq	5
Gierahn17	✓	n	0	0
Kang18	✗	b,k	8	8
	1015	k	1015	6
	B	n	1015	B cells
	NK	n	1015	NK cells
Koh16	✓	k	0	7
MCA20	✗	b,k	13	9
	gland.AT2	b	4	T cell_Cd8b1 high
	lung.AT2	b	4	AT2 Cell
Mereu20	✗	b,k	13	9
	CD4T	b	13	CD4 T cells
	ddSeq	k	ddSeq	9
Oetjen18	✓	b	18	0
	R	n	R	0
panc8	✗	b,k	5	9
	inDrop1.beta	n	indrop1	beta
	inDrop.ductal	b	indrop1-4	ductal
	SmartSeq2	k	smartseq2	7
TabulaMuris	✗	b,k	10	31
	limb.MSCs	n	Limb_Muscle	mesenchymal stem cell
	spleen	k	Spleen	4
Tung17	✓	b	3	0
	NA19101	n	NA19101	0
Zheng17	✓	k	0	7
	HSCs	n	0	HSCs CD34+
	Monocytes	n	0	Monocytes CD14+

Quality control summaries

Summary	Description/Interpretation	Formula/Implementation
mean of logCPM	expression mean	$\mu = \frac{1}{C}\sum_{c=1}^{C}\mathbf{Y}_{gc}$
variance of logCPM	expression variance	$\sigma = \frac{1}{C-1}\sum_{c=1}^{C}(\mathbf{Y}_{gc}-\mu)^2$
coefficient of variation	expression variability relative its mean	$\sqrt{\sigma}/\mu$
gene detection frequency	fraction of cells with non-zero count (for a given gene)	$\frac{1}{C}\sum_{c=1}^{C}\mathbb{1}(\mathbf{X}_{gc}\neq 0)$
gene-to-gene-correlation	expression association between pairs of genes	$\frac{\mathrm{cov}(\mathbf{Y}_g,\mathbf{Y}_{g'})}{\sigma_g\sigma_{g'}}$
log-library size	log1p-transformed total counts	$\log(1+\sum_{g=1}^{G}\mathbf{X}_{gc})$
cell detection frequency	fraction of detected genes (for a given cell)	$\frac{1}{G}\sum_{g=1}^{G}\mathbb{1}(\mathbf{X}_{gc}\neq 0)$
cell-to-cell-correlation	expression association between pairs of cells	$\mathrm{cov}(\mathbf{Y}_c,\mathbf{Y}_{c'})/(\sigma_c\cdot\sigma_{c'})$
local density factor	relative measure of a cell's local density compared to those within its neighbourhood (in PCA space)	custom wrapper of functions from the CellMixS package with PCs of $\mathbf{Z}$ as input
cell-to-cell distance	expression (dis)similarity between pairs of cells	Euclidean distance in PCA space of $\mathbf{Z}$
KNN occurences	number of times a cell is a k-nearest neighbor (KNN)	RANN's nn2 function on PCs of $\mathbf{Z}$ with k set to 5% of cells
percent variance explained	fraction of expression variance accounted for by batch/cluster	variancePartition's fitExtractVarPartModel function with $\mathbf{Z}$ as input
silhouette width	similarity of a cell to its own group (batch/cluster) compared to others	cluster's silhouette function on Euclidean distances in PCA space of $\mathbf{Z}$
cell-specific mixing score	probability of being in an equally 'mixed' (same batch/cluster) neighborhood (in PCA space)	CellMixS's cms function with PCs of $\mathbf{Z}$ as input

Evaluation statistics

- 각 summary 지표에 대해 참조 데이터와 시뮬레이션 데이터 간의 차이 평가 방법
  - Kolmogorov-Smornov(KS) 통계량: `stats` 패키지의 `ks.test( )` 함수 사용
  - Wasserstein Distance(Earth Mover's Distance (EMD)): `waddR`의 `Wasserstein_metirc( )` 함수 사용

- 2차원 지표 간의 결합 분포를 비교
  - 2차원 KS 통계량: `MASS` 패키지의 `kde2d( )` 함수 사용
  - 2차원 EMD: `emdist` 패키지의 `emd2d( )` 함수 사용
  - 2차원 비교는 유전자 수준과 세포 수준 summary 조합 중 의미 있는 것만 사용하며,
    Global summary와 상관계수 기반 summary는 제외

🔶 Integration evaluation

- 통합 분석 방법
    - ComBat
    - Harmony
    - fastMNN, mnnCorrect
    - limma
    - Seurat

- 방법의 성능 평가
    - CMS(Cell-specific Mixing Score): `CellMix` 패키지의 `cms( )` 함수 사용
        - CMS ⟶ CMS*: 0.5를 빼서 평균이 0이 되도록 조정
    - LDF(Local Density Factor)Difference: `ldfDiff( )` 함수 사용
        - LDF ⟶ LDF*: 평균 0, 값의 범위 0 ~ 1로 스케일 조정

⟶ 서로 다른 세포들이 서로 다른 배치 간에 얼마나 잘 섞였는지 평가하는 지표.
값이 0에 가까울수록 잘 섞인 상태

Clustering evaluation

- 클러스터링 방법(additional file1의 sec 5.2 참고)
    - CIDR
    - 계층적 클러스터링(HC)
    - 주성분 분석(PCA)을 이용한 k-평균 클러스터링(KM)
    - pcaReduce
    - SC3
    - Seurat
    - TSCAN
    - t-SNE 기반의 KM 클러스터링

- 적용 가능한 경우 클러스터의 개수는 실제 클러스터 수와 일치하도록 설정
- 방법의 성능 평가
    - Hungarian 알고리즘: 실제 클러스터 라벨과 예측된 클러스터 라벨 매칭
    - 클러스터 단위의 정밀도(precision), 재현율(recall), F1 score 계산