Analysis of single-cell RNA-seq data (II)

Hao Wu
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

Ziyi Li
Department of Biostatistics
The University of Texas MD
Anderson Cancer Center

ENAR 2021 short course
March 2021

Course outline

- 8-9:15: Intro and data preprocessing.
- 9:15-9:45: Lab: preprocessing and visualization.
- **10-11:15: Normalization, batch effect, imputation, DE, simulator.**
- 11:15-12: Lab: Normalization, batch effect, imputation, DE, simulator
- 12-1: lunch break
- 1-2: Clustering and pseudotime construction
- 2-2:30: Lab: Clustering and pseudotime construction
- 2:45–3:30: Supervised cell typing & related single cell data sources
- 3:30-4: Lab: supervised cell typing.
- 4:15-5: scRNA-seq in cancer

Outline for this session

- Statistical models for scRNA-seq data
- Data preprocessing
  - Normalization
  - Batch effect correction
  - Imputation
- Differential expression
- Data simulator
- Sample size calculator

Review: data model for bulk RNA-seq

- The most common model is the gene-wise gamma-Poisson (negative binomial) distribution:

$$Y_{ij} \mid \lambda_i \sim Poisson(\lambda_i), \lambda_i \sim Gamma(\alpha, \beta)$$

$$Y_{ij} \sim NB(\alpha, \beta)$$

- NB is over-dispersed Poisson:
  - Poisson: $var = \mu$
  - NB: $var = \mu + \mu^2\phi$
  - Dispersion parameter ϕ approximates the squared coefficient of variation: $\phi = \frac{var-\mu}{\mu^2} \approx \frac{var}{\mu^2}$

Data model for scRNA-seq

- The data distribution is more complex than bulk RNA-seq due to
  - Mixture of cell types
  - Drop out
- Often-used statistical models
  - Gene-wise: zero-inflated model, mixture model
  - Cell-wise: Dirichlet-multinomial.

Gene-wise modeling

- Many expressions follow multi-modal distribution.
- Most methods use mixture of distributions:
  - SCDE (Kharchenko et al., 2014): a mixture of NB and Poisson.
  - MAST (Finak et al., 2015): a generalized linear hurdle model.
  - SC2P (Wu et al. 2018): a mixture of zero-inflated Poisson (ZIP) and lognormal-Poisson.
- Recent discussions about the presence of zero-inflation: whether it's caused by UMI or droplet.
  - Cao et al. 2020 Nat Biotech; Svensson 2020 Nat Biotech.

Cell-wise modeling

- Counts for cells in one cell type follow Dirichlet-multinomial distribution
  - Counts for each cell follow a multinomial distribution
  - The MN means follow Dirichlet cross cells in the same cell type.
- For multiple cell types, the counts follow a mixture of Dirichlet-multinomial.
- Used more often in cell clustering methods (DIMM-SC, BAMM-SC).

Data normalization

- scRNA-seq is very noisy.
- Spike-in data is usually available.
  - Spike-ins from the external RNA Control Consortium (ERCC) panel contains 92 synthetic spikes based on bacterial genome with known expression level.
- UMI is helpful for removing amplification noise.
- A combination of spike-in and UMI can potentially be used for data normalization.
- Simple normalization (such as by sequencing depth) for bulk RNA-seq can be applied, e.g., TPM or FPKM.

Lun *et al. Genome Biology* (2016) 17:75
DOI 10.1186/s13059-016-0947-7

Genome Biology

METHOD Open Access

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

- Works for data without spike-in.
- The goal is to estimate a size factor for each cell.
- The idea is to normalize on summed expression values from pools of cells – it's more stable than using individual cell.
- Bioconductor package **scran.**

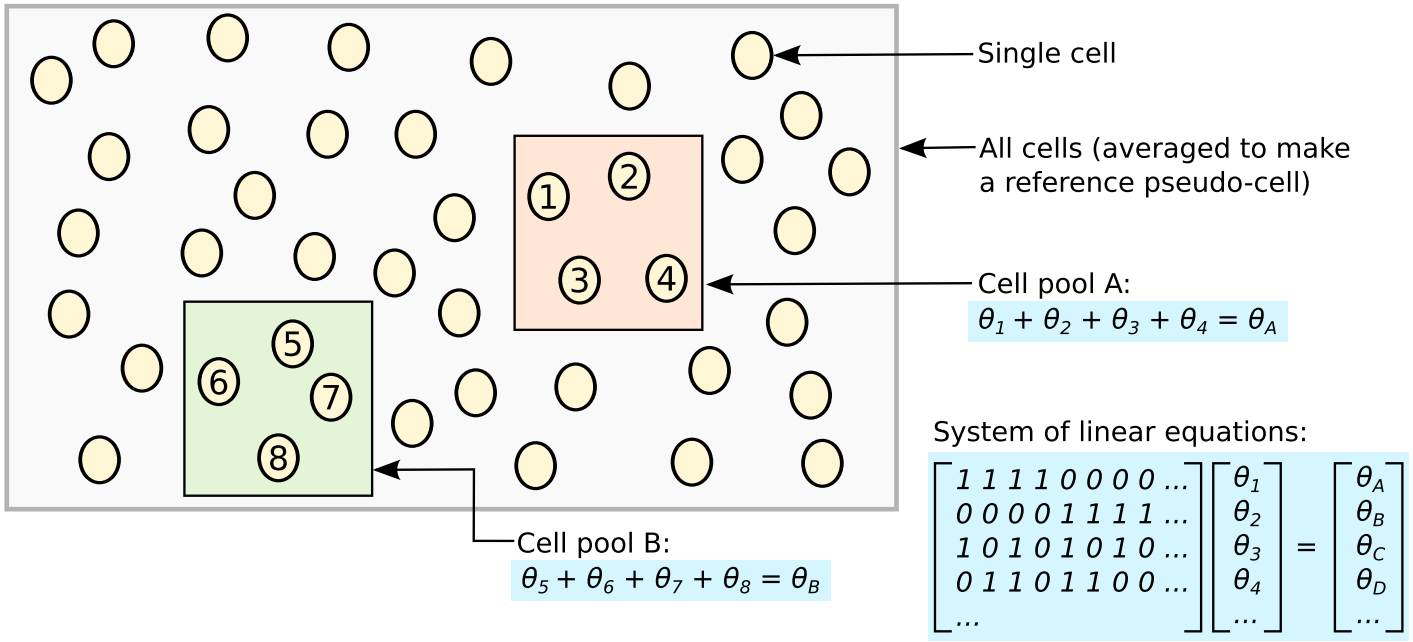


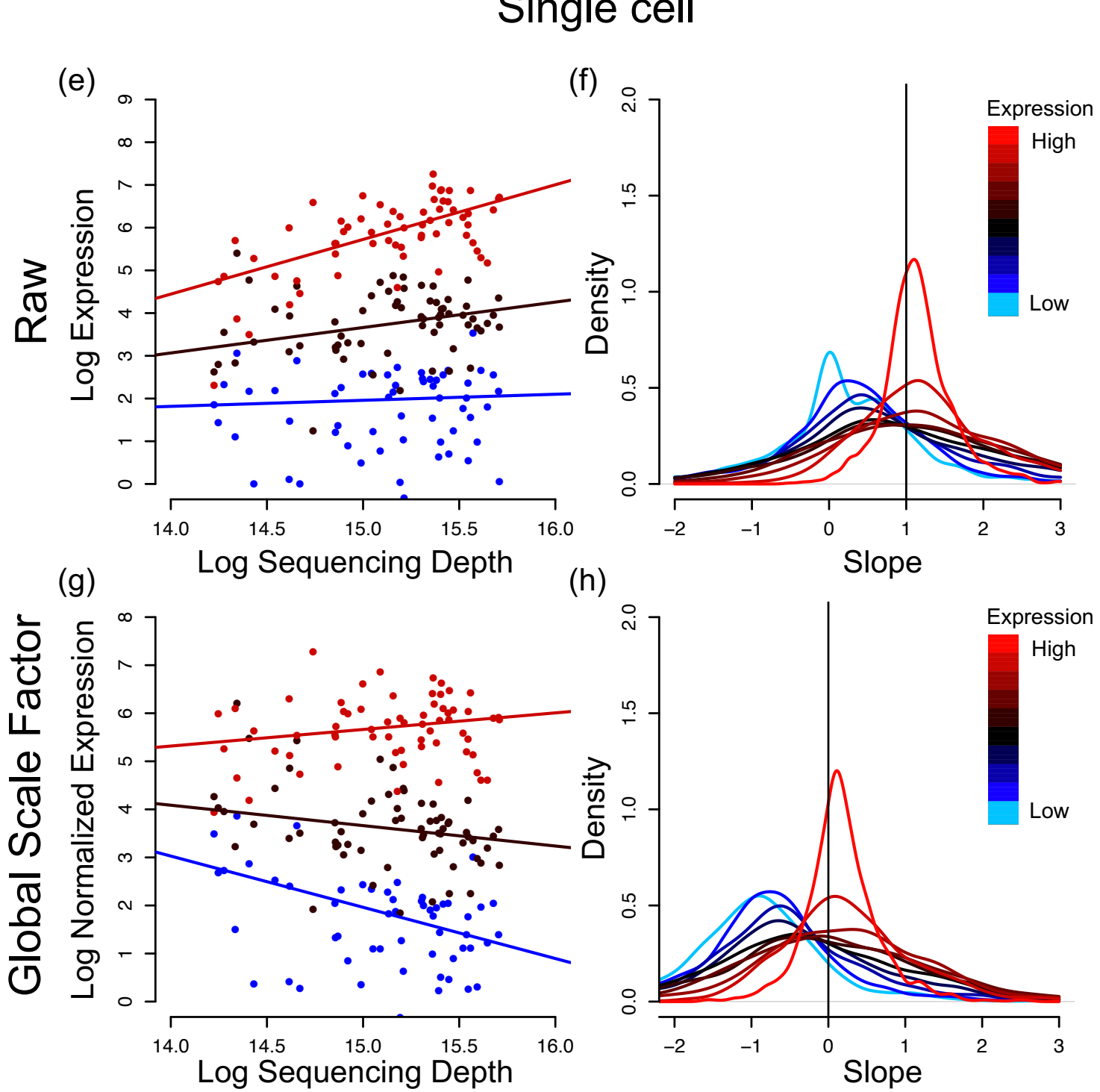
Fig. 3 Schematic of the deconvolution method. All cells in the data set are averaged to make a reference pseudo-cell. Expression values for cells in pool A are summed together and normalized against the reference to yield a pool-based size factor θ_A. This is equal to the sum of the cell-based factors θ_j for cells $j = 1$–4 and can be used to formulate a linear equation. (For simplicity, the t_j term is assumed to be unity here.) Repeating this for multiple pools (e.g., pool B) leads to the construction of a linear system that can be solved to estimate θ_j for each cell j

SCnorm: robust normalization of single-cell RNA-seq data

584 | VOL.14 NO.6 | JUNE 2017 | NATURE METHODS

Rhonda Bacher[1,5], Li-Fang Chu[2,5], Ning Leng[2], Audrey P Gasch[3], James A Thomson[2], Ron M Stewart[2], Michael Newton[1,4] & Christina Kendziorski[4]

- Basic idea: one normalization factor per cell doesn't fit all genes.
- Relationships of read counts and sequencing depths vary and depend on the expression levels.


Single cell
(e)
(f)
Raw
Log Expression
Log Sequencing Depth
Density
Slope
Expression
High
Low
(g)
(h)
Global Scale Factor
Log Normalized Expression
Log Sequencing Depth
Density
Slope
Expression
High
Low

SCnorm Solution

- Uses quantile regression to estimate the dependence of read counts on sequencing depth for every gene.
- Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group.
- Bioconductor package **SCnorm**.

Batch effect

- Batch effect in scRNA-seq can be severe.
- Can be difficult to randomize the design, i.e., batch is confounded with individual, so it causes trouble for analyzing data from multiple individuals (more on this later).
- Bulk data method such as Combat doesn’t work well.

Batch effect correction methods

- The cells are from different cell types, which complicates the problem.
- Most methods are developed for cell clustering, i.e., jointly perform batch correction and cell clustering.
- The goal is to minimize the impact of batch effects on cell clustering.

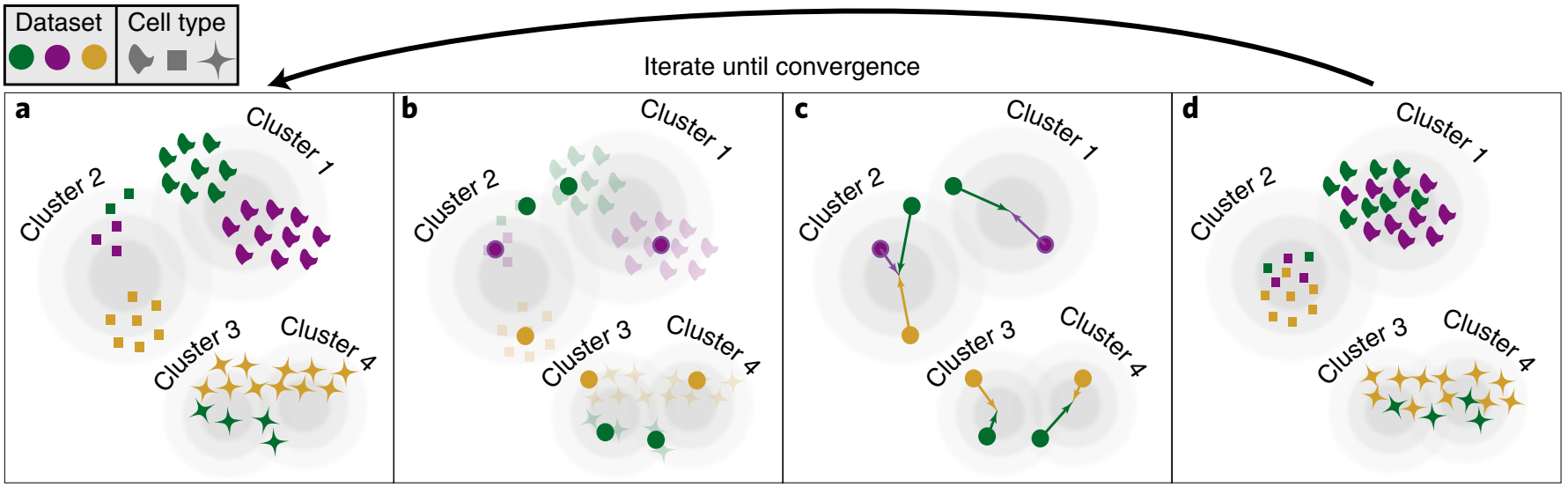
nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0619-0

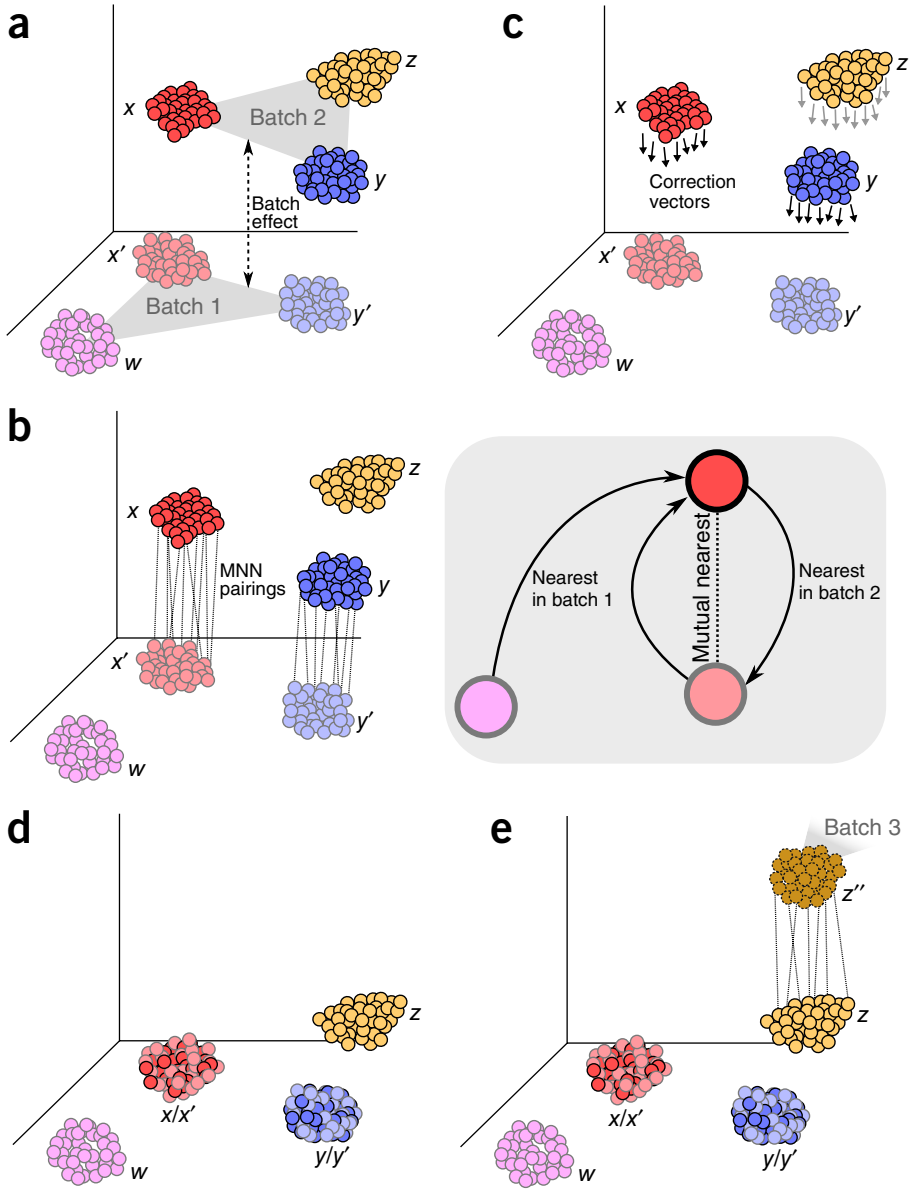
Fast, sensitive and accurate integration of single-cell data with Harmony

Ilya Korsunsky[1,2,3,4], Nghia Millard[1,2,3,4], Jean Fan[5], Kamil Slowikowski[1,2,3,4], Fan Zhang[1,2,3,4], Kevin Wei[2], Yuriy Baglaenko[1,2,3,4], Michael Brenner[2], Po-ru Loh[1,3,4] and Soumya Raychaudhuri[1,2,3,4,6*]



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors

Laleh Haghverdi[1,2], Aaron T L Lun[3], Michael D Morgan[4] & John C Marioni[1,3,4]



A comprehensive comparison paper

Tran *et al. Genome Biology* (2020) 21:12
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

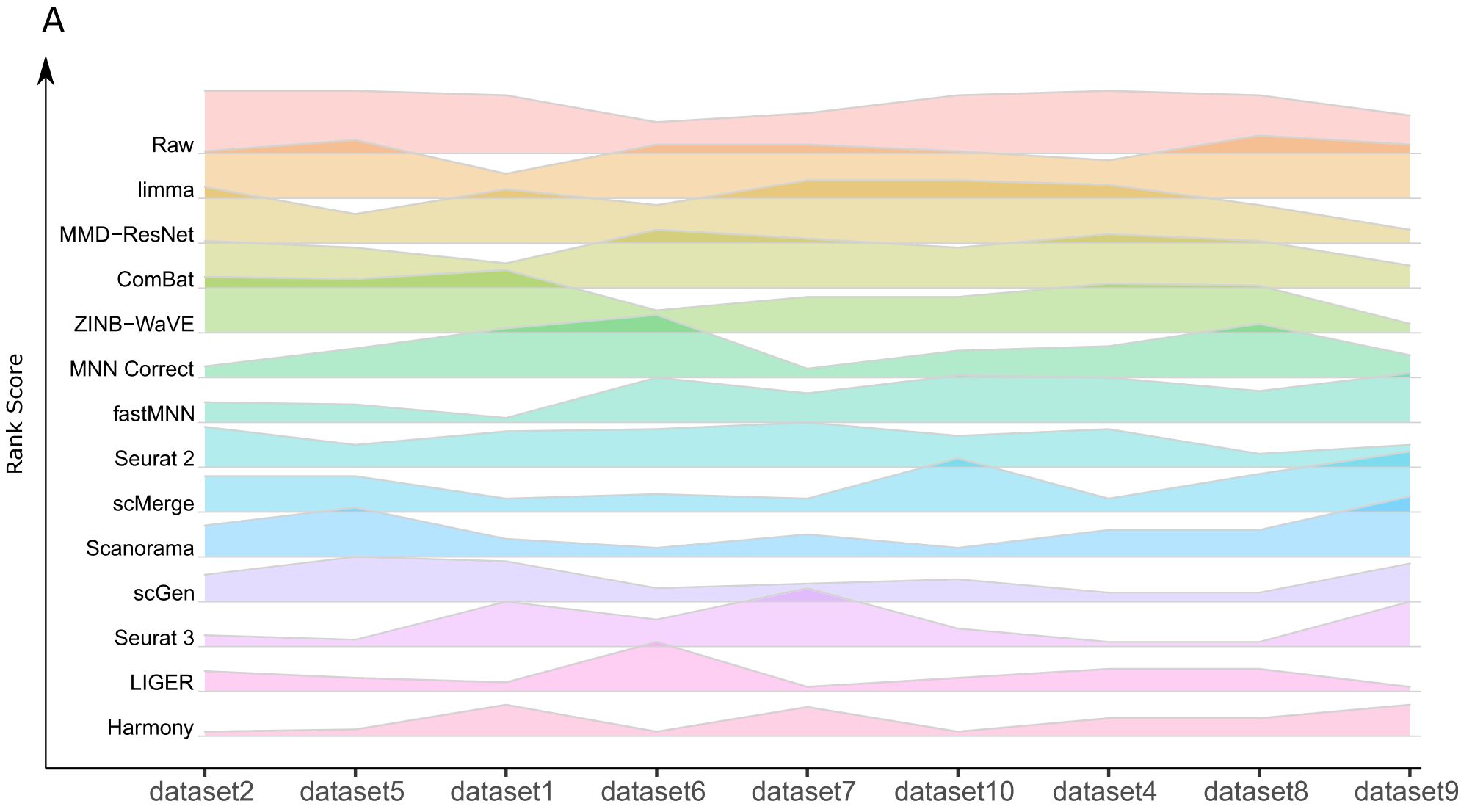
RESEARCH Open Access

A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Hoa Thi Nhu Tran[†], Kok Siong Ang[†], Marion Chevrier[†], Xiaomeng Zhang[†], Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen[*]

Table 1 Description of the 14 batch-effect correction methods

Tools	Programming language	Batch-effect-corrected output	Methods	Reference package version
Seurat 2 (CCA, MultiCCA)	R	Normalized canonical components (CCs)	Canonical correlation analysis and dynamic time warping	Butler et al. [4], Seurat package version 2.3.4
Seurat 3 (Integration)	R	Normalized gene expression matrix	Canonical correlation analysis and mutual nearest neighbors-anchors	Stuart et al. [12], Seurat package version 3.0.1
Harmony	R	Normalized feature reduction vectors (Harmony)	Iterative clustering in dimensionally reduced space	Korsunsky et al. [13], Harmony version 0.99.9
MNN Correct	R	Normalized gene expression matrix	Mutual nearest neighbor in gene expression space	Haghverdi et al. [5], Scran package version 1.12.0
fastMNN	R	Normalized principal components	Mutual nearest neighbor in dimensionally reduced space	Haghverdi et al. [5], Lun ATL [7], Scran package version 1.12.0
ComBat	R	Normalized gene expression matrix	Adjusts for known batches using an empirical Bayesian framework	Johnson et al. [1]
limma	R	Normalized gene expression matrix	Linear model/empirical Bayes model	Smyth et al. [2], limma version 3.38.3
scGen	Python	Normalized gene expression matrix	Variational auto-encoders neural network model and latent space	Lotfollahi et al. [16], 2019, scGen version 1.0.0
Scanorama	Python/R	Normalized gene expression matrix	Mutual nearest neighbor and panoramic stitching	Hie et al. [9], Scanorama version 1.4.
MND-ResNet	Python	Normalized principal components	Residual neural network for calibration	Shaham et al. [15] updated code to Python 3
ZINB-WaVE	R	Normalized feature reduction vectors (ZINB-WaVE)/normalized gene expression matrix	Zero-inflated negative binomial model, extension of RUV model	Risso et al. [6], ZINB-WaVE version 1.6.0
scMerge	R	Normalized gene expression matrix	Stably expressed genes (scSEGs) and RUVIII model	Lin et al. [18], scMerge version 1.1.3
LIGER	R	Normalized feature reduction vectors (LIGER)	Integrative non-negative matrix factorization (iNMF) and joint clustering + quantile alignment	Welch et al. [14], liger version 1.0
BBKNN	Python/R	Connectivity graph and normalized dimension reduction vectors (UMAP)	Batch balanced *k*-nearest neighbors	Polański et al. [10], bioRxiv. BBKNN version 1.3.2


A
Rank Score
Raw
limma
MMD−ResNet
ComBat
ZINB−WaVE
MNN Correct
fastMNN
Seurat 2
scMerge
Scanorama
scGen
Seurat 3
LIGER
Harmony
dataset2
dataset5
dataset1
dataset6
dataset7
dataset10
dataset4
dataset8
dataset9

Other interesting methods

ARTICLE

https://doi.org/10.1038/s41467-020-16905-2 OPEN

Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction

Fangda Song[1], Ga Ming Angus Chan[1] & Yingying Wei[1✉]

ARTICLE

https://doi.org/10.1038/s41467-020-15851-3 OPEN

Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis

Xiangjie Li[1,2,3], Kui Wang[1,4], Yafei Lyu[1], Huize Pan[5], Jingxiao Zhang[2], Dwight Stambolian[6], Katalin Susztak[7], Muredach P. Reilly[5], Gang Hu[1,8✉] & Mingyao Li[1✉]

Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis

Jian Hu[1], Xiangjie Li[2], Gang Hu[3], Yafei Lyu[1], Katalin Susztak[4] and Mingyao Li[1✉]

Data imputation

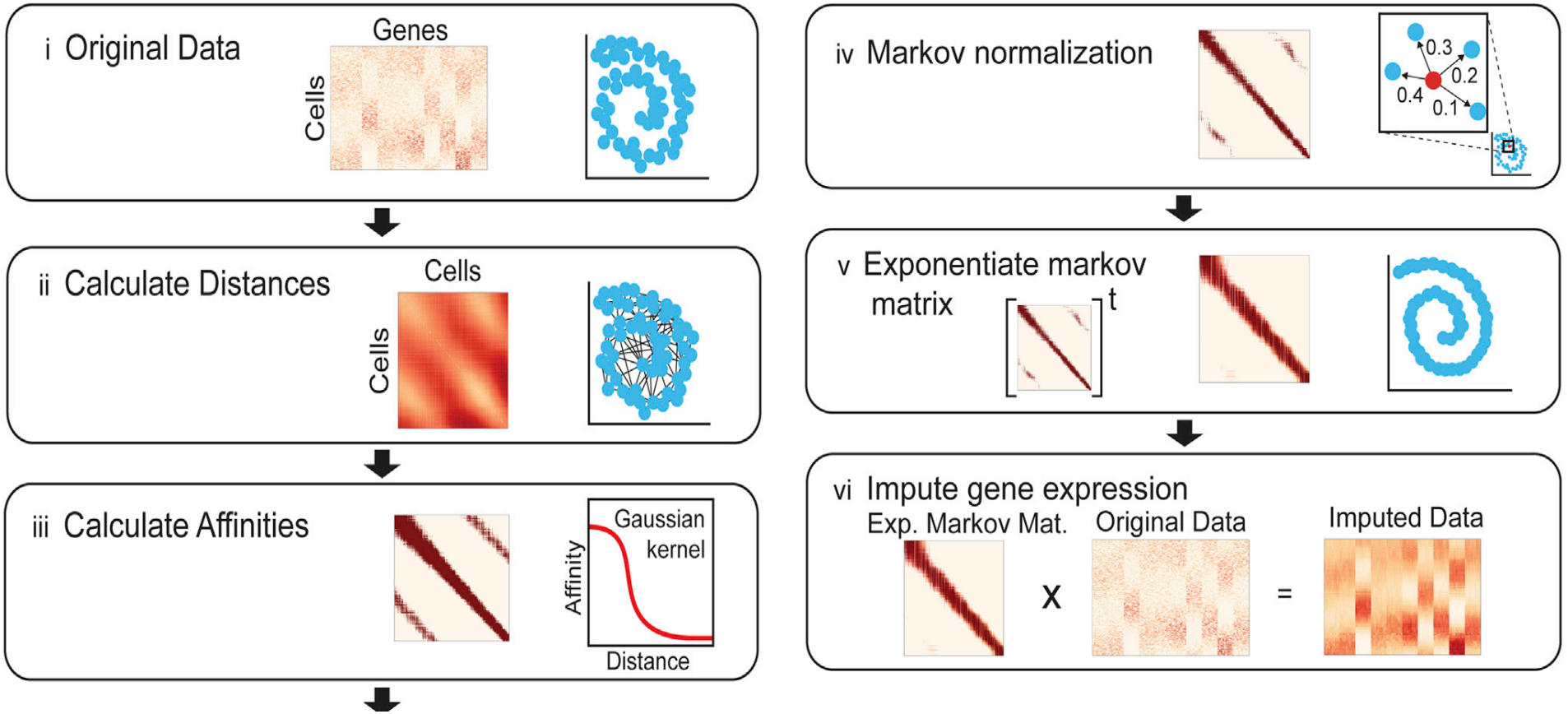
- scRNA-seq has lots of missing data (dropout).
- Imputing the missing data help the downstream analyses.
- There are a number of methods:
  - SAVER (Huang et al. 2018 Nat. Methods)
  - ScImpute (Li et al. 2018 Nat. Comm.)
  - MAGIC (van Dijk et al. 2018 Cell)
  - SCRABBLE (Peng et al. 2019 GB)

General strategy for imputation

- The problem is similar to a “recommendation system”.
  - First compute the similarities among genes and cells.
  - To impute one element, borrow information from similar gene/cell.

Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

David van Dijk,[1] Roshan Sharma,[1,2] Juozas Nainys,[1,3] Kristina Yim,[4] Pooja Kathail,[1,5] Ambrose J. Carr,[1,5] Cassandra Burdziak,[1] Kevin R. Moon,[4,6] Christine L. Chaffer,[7] Diwakar Pattabiraman,[8] Brian Bierie,[8] Linas Mazutis,[1] Guy Wolf,[6] Smita Krishnaswamy,[4,6,9,*] and Dana Pe'er[1,9,10,*]



ARTICLE

DOI: 10.1038/s41467-018-03405-7 OPEN

An accurate and robust imputation method scImpute for single-cell RNA-seq data

Wei Vivian Li [1] & Jingyi Jessica Li [1,2]

- scImpute: base on Gamma-Normal mixture model to estimate and impute dropout values.
- Steps:
  - Learn each gene's dropout probability in each cell
  - Impute dropout values of genes in a cell by borrowing information of the same gene in other "similar" cells, which are selected based on genes not severely affected by dropout events.

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0033-z

SAVER: gene expression recovery for single-cell RNA sequencing

Mo Huang[1], Jingshu Wang[1], Eduardo Torre[2,3], Hannah Dueck[4], Sydney Shaffer[3], Roberto Bonasio[5], John I. Murray[4], Arjun Raj[3,4], Mingyao Li[6] and Nancy R. Zhang[1]*

- Use gene-to-gene relationships to recover the true expression levels.
- Assume gamma-Poisson (NB) for counts
- Estimate the gamma prior parameters in an empirical Bayes-like approach with a Poisson LASSO regression, using the expression of other genes as predictor.
- The posterior mean are the imputed expression.

SAVER model

- Assume: $Y_{gc} \sim Poisson(s_c \lambda_{gc}), \lambda_{gc} \sim Gamma(\alpha_{gc}, \beta_{gc})$
- Obtain: $\lambda_{gc} | Y_{gc}, \hat{\alpha}_{gc}, \hat{\beta}_{gc} \sim Gamma(Y_{gc} + \hat{\alpha}_{gc}, s_c + \hat{\beta}_{gc})$
  - $\mu = \alpha / \beta$, $v = \alpha / \beta^2$
  - Penalized Poisson LASSO regression: $\log E(\frac{Y_{gc}}{s_c} | Y_{g'c})$
    $= \log \mu_{gc} = \gamma_{g0} + \sum_{g' \neq g} \gamma_{gg'} \log[\frac{Y_{g'c}+1}{s_c}]$
  - $\hat{v}_{gc}$ is obtained with MLE of marginal likelihood of Y under certain assumptions
- Estimate: $\hat{\lambda}_{gc} = \frac{Y_{gc}+\hat{\alpha}_{gc}}{s_c+\hat{\beta}_{gc}} = \frac{s_c}{s_c+\hat{\beta}_{gc}} \frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c+\hat{\beta}_{gc}} \hat{\mu}_{gc}$

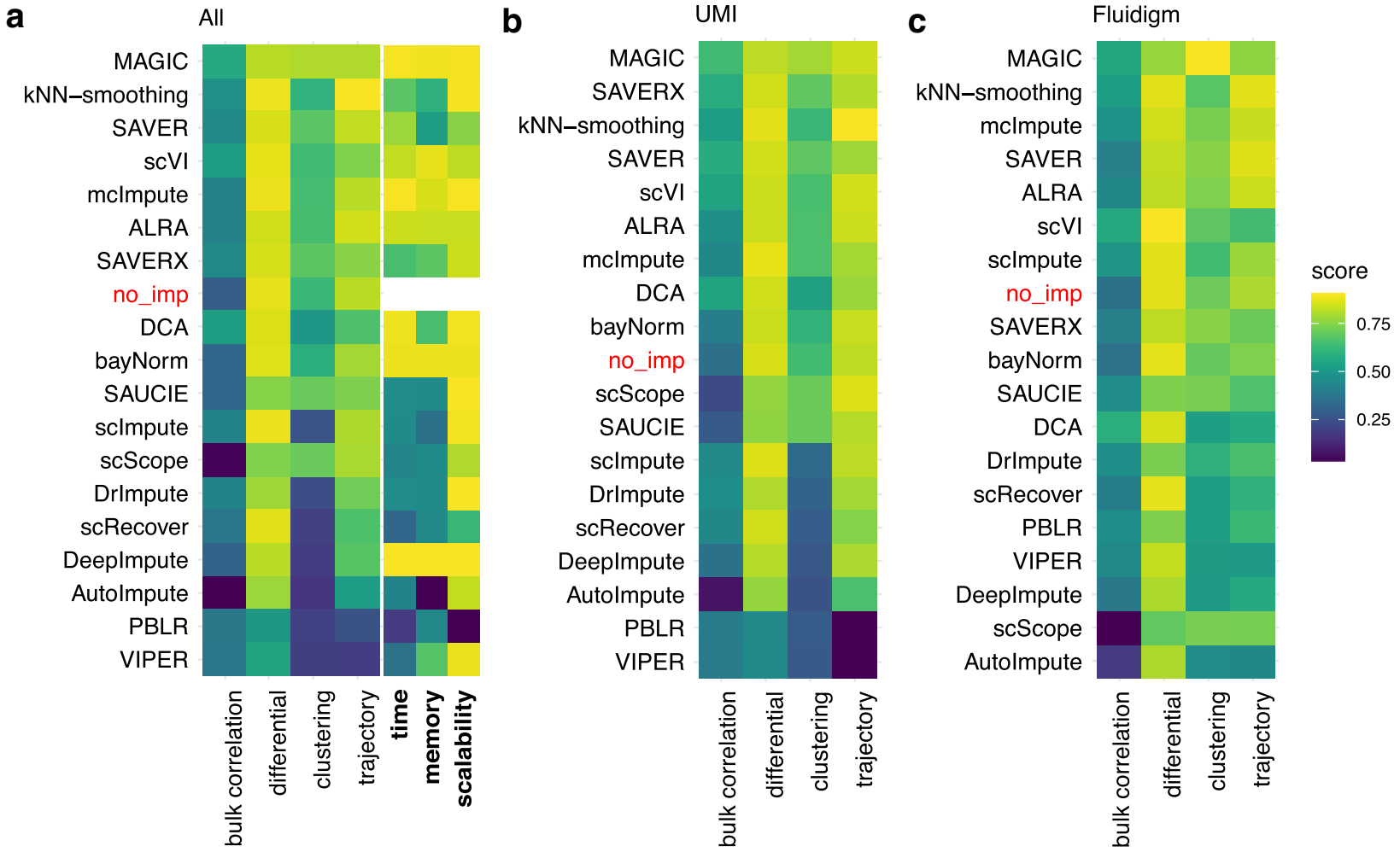
Hou *et al. Genome Biology* (2020) 21:218
https://doi.org/10.1186/s13059-020-02132-x

Genome Biology

RESEARCH Open Access

A systematic evaluation of single-cell RNA-sequencing imputation methods

Wenpin Hou, Zhicheng Ji, Hongkai Ji* and Stephanie C. Hicks*



Differential expression (DE)

- DE analysis is the most important task for bulk expression data (microarray or RNA-seq).
- Popular tools:
  - Microarray: limma
  - Bulk RNA-seq: DESeq2, edgeR
- Important method:
  - Variance shrinkage

DE in scRNA-seq

- Considering cell types:
  - Compare cross cell types: identify cell type specific genes.
  - Compare the same cell type cross biological conditions.
  - Need cell clustering first.
- Method consideration:
  - Traditional methods test mean changes
  - The consideration and modeling of “drop-out” is important in scRNA-seq data.

DE methods

- SCDE (Kharchenko et al. 2014 Nat. Methods)
- MAST (Finik et al. 2015 GB)
- SC2P (Wu et al. 2018 Bioinformatics)
- Seurat and monocle provide DE functions.
- Bulk methods (DESeq, edgeR) are sometimes used.
- A comparison paper: Soneson and Robinson (2018) Nat. Methods

Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

- MAST: "Model-based Analysis of Single- cell Transcriptomics."
- Bioconductor package **MAST**.

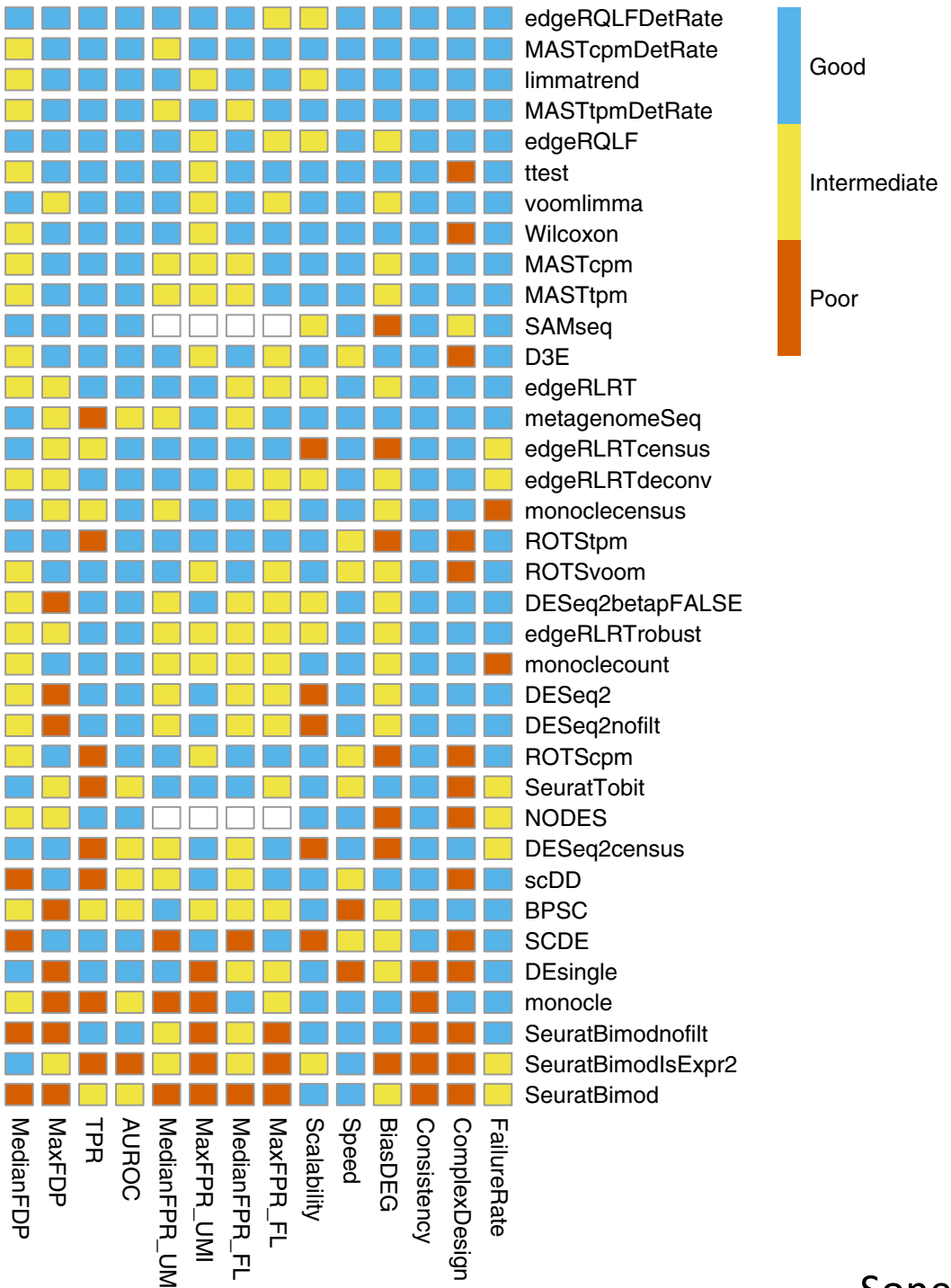
MAST for DE

- Main ideas:
  - Use log2(TPM+1) as input data
  - Both dropout probability and expression level depends on experimental conditions.

$$logit\left(Pr(Z_{ig}=1)\right)=\mathrm{X_i}\,\beta_{\mathrm{g}}^{\mathrm{D}}$$

$$\Pr\left(Y_{ig}=y|Z_{ig}=\ 1\right)=\ \mathrm{N}\left(\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{C}},\ \sigma_g^2\right)$$

  - Model fitting with some regularization.
  - DE is based on chi-square or Wald test.



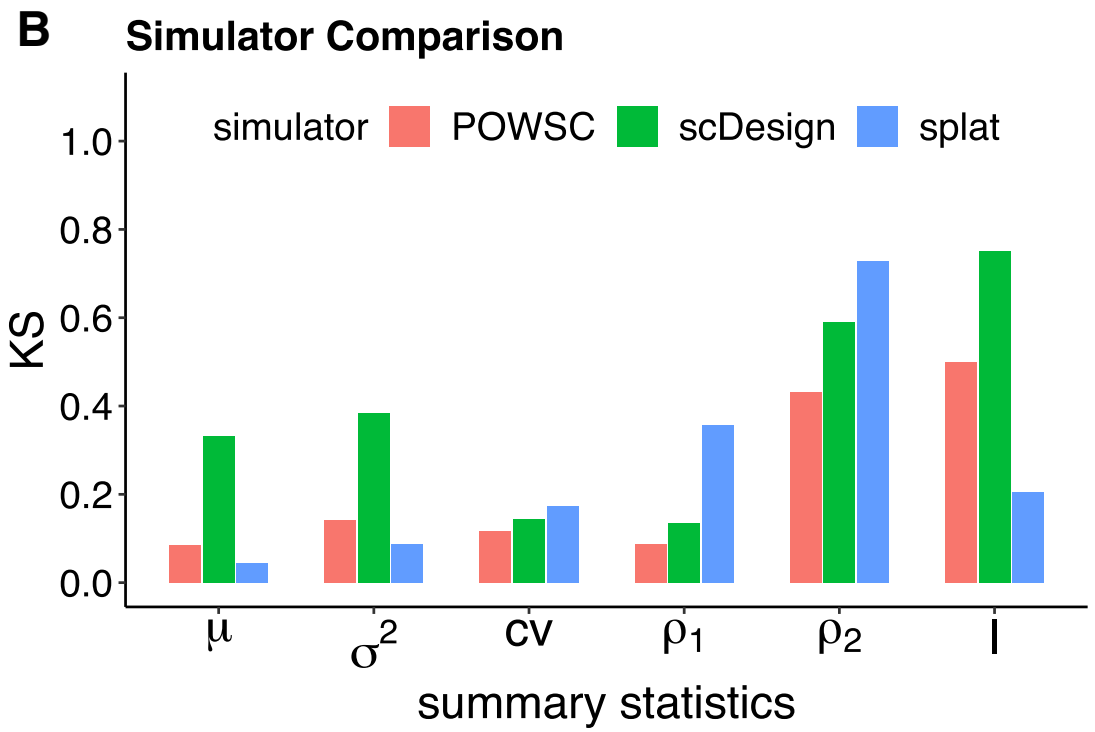
Soneson and Robinson (2018) Nat. Methods

scRNA-seq Data simulators

Method	Model Assumption	Reference
powsimR	NB (ZI-NB optional) on counts for both RNA-seq and scRNA-seq	Vieth et al., 2017, Bioinfo
scDesign	Gamma-normal mixture on log counts (same as scImpute)	Li et al., 2019, Bioinfo
Splatter (splat)	Gamma-Poisson hierarchical model on normalized counts (fitdistrplus)	Zappia et al., 2017, GB
POWSC	ZIP-LNP mixture on log normalized counts	Su et al., 2020, Bioinfo
scDBM	Deep generative models (Boltzmann Machines to the NB distribution)	Treppner et al., 2020, Preprint
scPOWER	NB on gene-level counts (DESeq)	Schmid et al., 2020, Preprint

Comparisons

- Based on real data.
- Estimate model parameters.
- Simulate the data.
- Compare between simulated and real expression matrices.
  - 4 **gene-wise** parameters
  - 2 **cell-wise** parameters.



Sample size calculation

- Power evaluation and sample size calculation is an important consideration at the experimental design stage, and required by almost all grant applications.
- Determining scRNA-seq sample size (# of cells) is difficult, i.e., no closed form solution.
- A few methods available under the context of DE, rely on some simulation procedures.
- Report number of cells required in order to achieve certain power for DE detection.

Power evaluator (in the context of DE analysis)

Method	Approach	Notes
powsimR	Use a series of established tools: edgeR, limma, and DESeq2. MAST, BPSC, and scDD	Report stratified power by mean expression levels
scDesign	Top 1000 genes ranked by effect score as reference true DE genes	Precision, recall, F1 score, and etc.
POWSC	Use MAST or SC2P to report two forms of DE genes	Stratified, marginal, and overall power evaluation
scPOWER	Use Mkmisc package for DE genes and use F test for eQTLs	Overall power by considering both power from DE genes and eQTLs