Models, Inference & Algorithms (MIA)

Primer: Permutation Enhances the Rigor of Genomics Data Analysis

Jingyi Jessica Li

Meeting: mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu

November 5, 2025

Permutation Enhances the Rigor in Genomics Data Analysis

Jingyi Jessica Li
Professor and Program Head of Biostatistics
Fred Hutchinson Cancer Center

Affiliate Professor of Biostatistics
University of Washington

Junction of Statistics and Biology

Fred Hutch Cancer Center

Models, Inference & Algorithms (MIA)

Primer: Permutation Enhances the Rigor of Genomics Data Analysis

Jingyi Jessica Li

November 5, 2025

Permutation Enhances the Rigor in Genomics Data Analysis

Jingyi Jessica Li
Professor and Program Head of Biostatistics
Fred Hutchinson Cancer Center

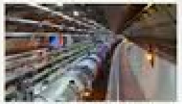
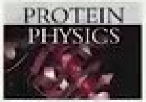
Affiliate Professor of Biostatistics
University of Washington

Junction of Statistics and Biology

Fred Hutch Cancer Center

Liberalism in scientific research

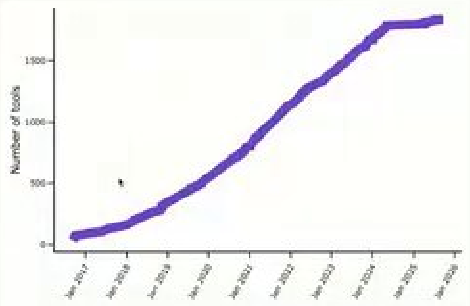
What is Physics?

	Object Fundamentalism Only general laws	Object Neutralism Not necessarily general laws	Object Liberalism Anything related to reality
Method Fundamentalism Only theories are Physics	 Only String Theory is Physics	 Condensed Matter Physics is also Physics	 Algebraic Topology is also Physics
Method Neutralism Experiments can also be Physics	 Accelerator Data Analysis is also Physics	 Quantum Optics is also Physics	 Life Science is also Physics
Method Liberalism Anything that involves experiment is Physics	 Divination is also Physics	 Long Material is also Physics	 **Neural Network is also Physics**

Translated from https://www.163.com/dy/article/JE0VILQC05118OGM.html (https://www.zhihu.com/pin/1513735925213917184) by Dr. Pan Liu

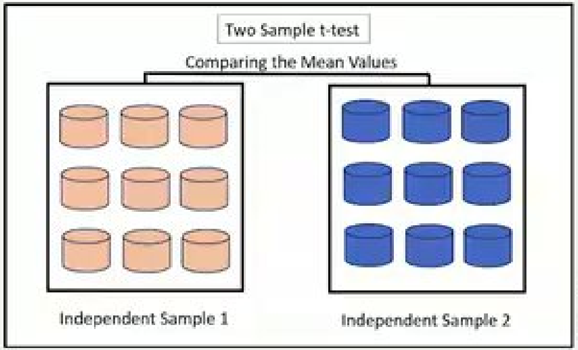
Genomics is a “liberal” discipline

1. Interdisciplinary nature
2. Data-driven focus
3. Rapid evolution of methods
4. Flexible analytical approaches



We currently track **1837** tools...

Statistics ensures rigor in data analysis



Guinness Brewery

Dublin, Ireland

William Sealy Gosset, who developed the "t-statistic" and published it under the pseudonym of "Student"

MCBRLab bilibili

Journal of the American Statistical Association ›
Volume 116, 2021 - Issue 536

Submit an article | Journal homepage

Enter keywords, authors, DO

Listen

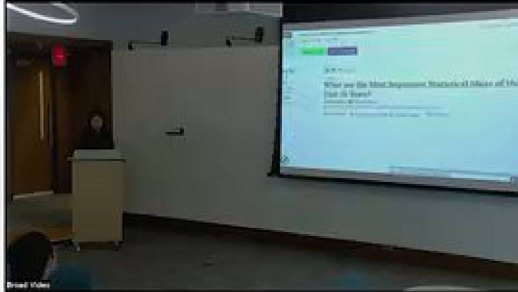
Reviews

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman & Aki Vehtari

Pages 2087-2097 | Received 30 Nov 2020, Accepted 23 May 2021, Published online: 08 Jul 2021

Cite this article https://doi.org/10.1080/01621459.2021.1938081

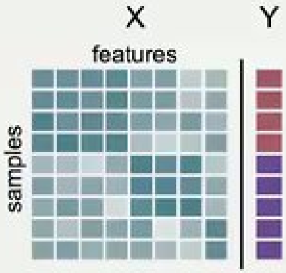


https://medium.com/bitgrit-data-science-publication/the-8-most-important-statistical-ideas-of-the-past-50-years-11220e46736f

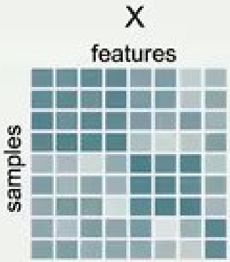
MCBRLab bilibili

How to permute data?

Supervised learning



Unsupervised learning



Teaser: bulk RNA-seq DE analysis

Genome Biology

Home About Articles Submission Guidelines Collections Submit manuscript

Short Report | Open access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li & Jingyi Jessica Li

Genome Biology **23**, Article number: 79 (2022) | Cite this article

57k Accesses | **230** Citations | **178** Altmetric | Metrics

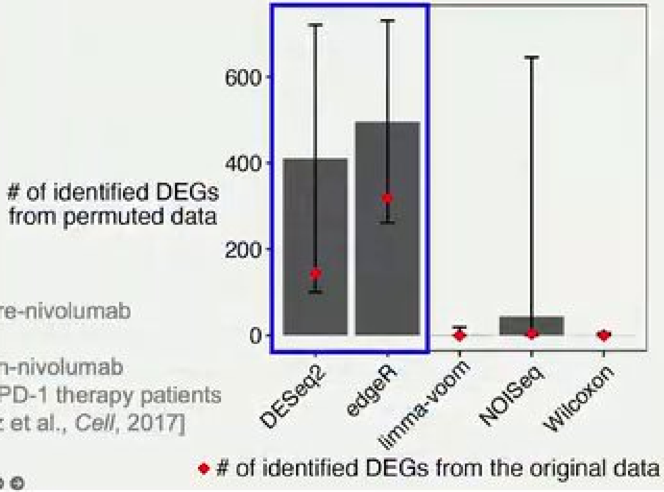
X/Twitter: @jsb_ucla

Yumei Li
(Wei Li Lab→
Soochow U)

Xinzhou Ge
(JSB→
OregonState)

Wei Li
(UC Irvine)

Teaser: bulk RNA-seq DE analysis



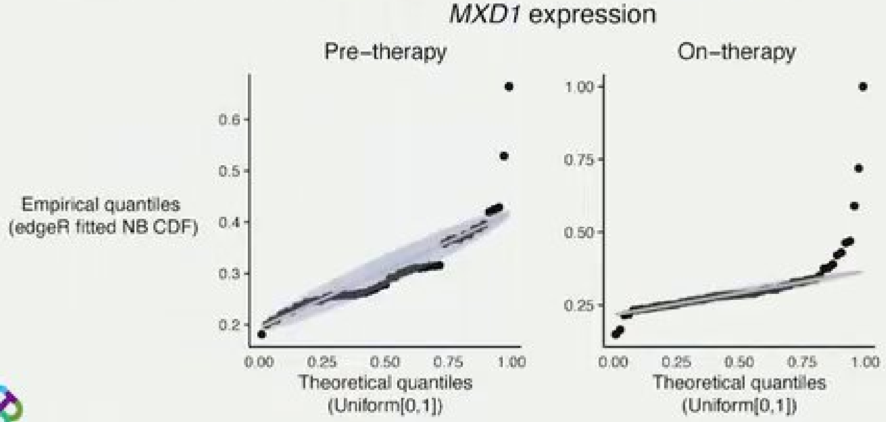
51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., *Cell*, 2017]

[Li*, Ge* et al., *Genome Biology*, 2022]

Teaser: bulk RNA-seq DE analysis

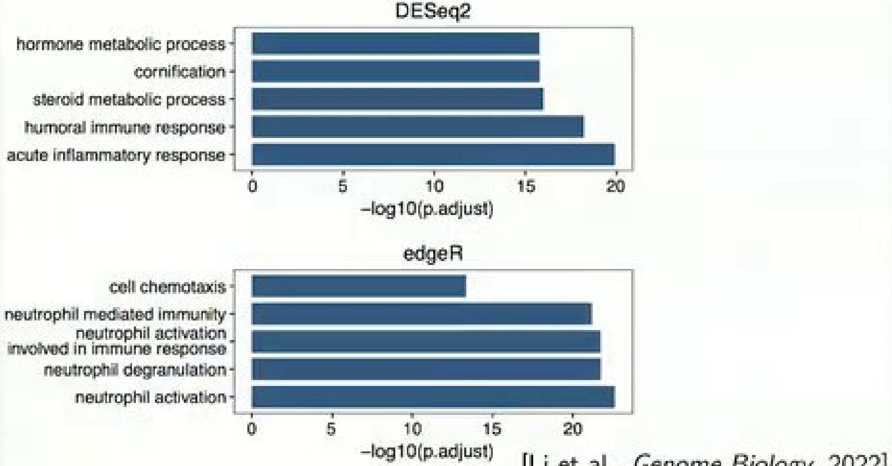
Q: Why are many genes identified as DE genes from permuted data?

A: The negative binomial assumption does not hold on this dataset.



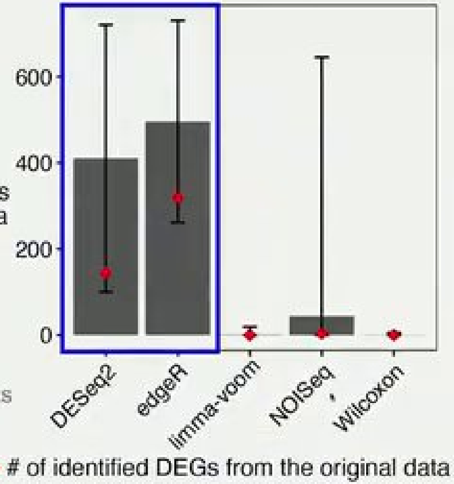
Teaser: bulk RNA-seq DE analysis

False discoveries may mislead scientific conclusions



[Li et al., *Genome Biology*, 2022]

Teaser: bulk RNA-seq DE analysis

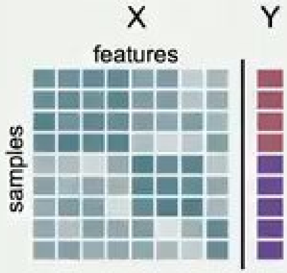


51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., *Cell*, 2017]

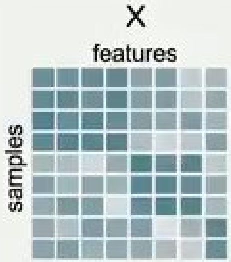
[Li*, Ge* et al.,
Genome Biology, 2022]

How to permute data?

Supervised learning

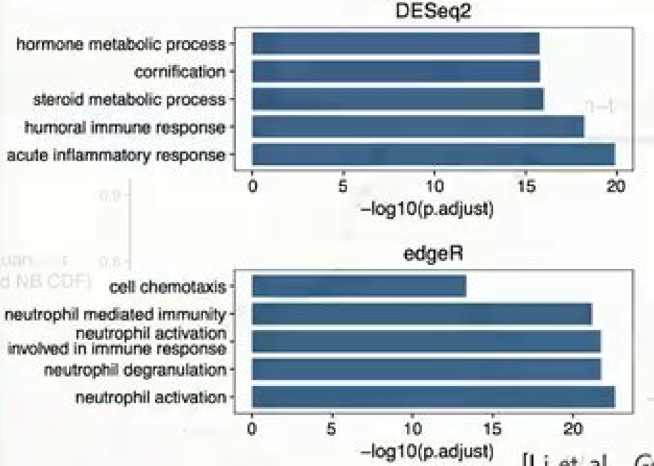


Unsupervised learning



Teaser: bulk RNA-seq DE analysis

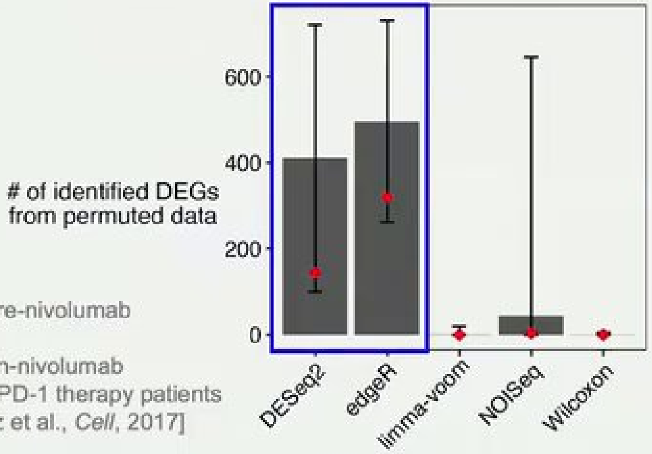
False discoveries may mislead scientific conclusions



[Li et al., *Genome Biology*, 2022]

Teaser: bulk RNA-seq DE analysis

Q: Why are many genes identified as DE genes from permuted data?



51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., *Cell*, 2017]

◆ # of identified DEGs from the original data

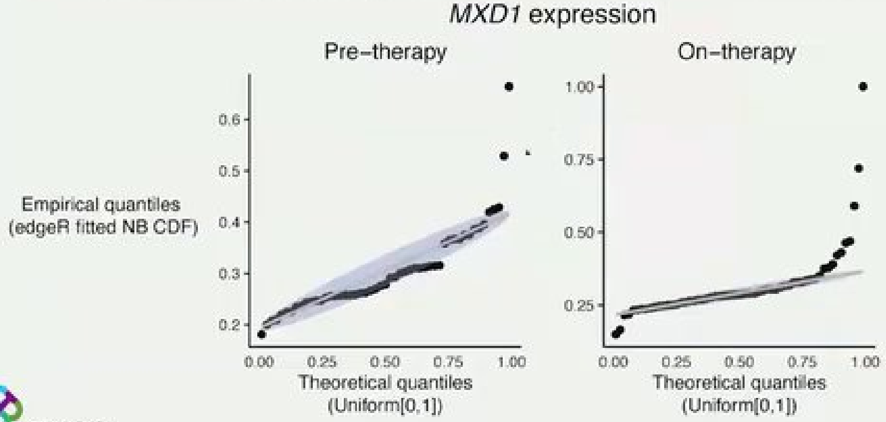
[Li*, Ge* et al., *Genome Biology*, 2022]

MCBRLab bilibili

Teaser: bulk RNA-seq DE analysis

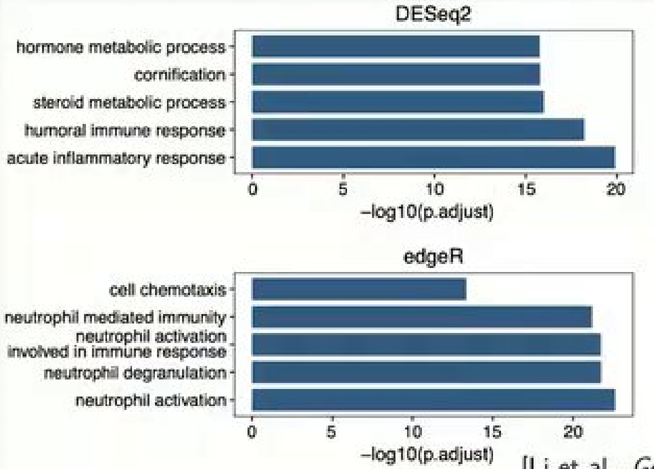
Q: Why are many genes identified as DE genes from permuted data?

A: The negative binomial assumption does not hold on this dataset.



Teaser: bulk RNA-seq DE analysis

False discoveries may mislead scientific conclusions

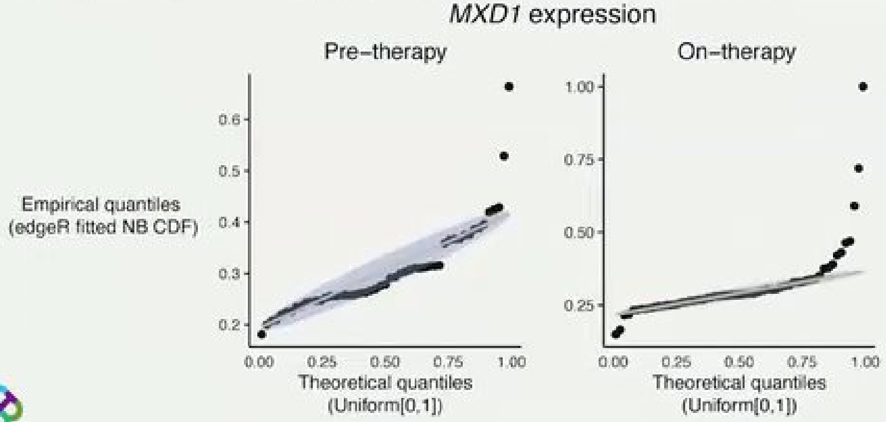


[Li et al., *Genome Biology*, 2022]

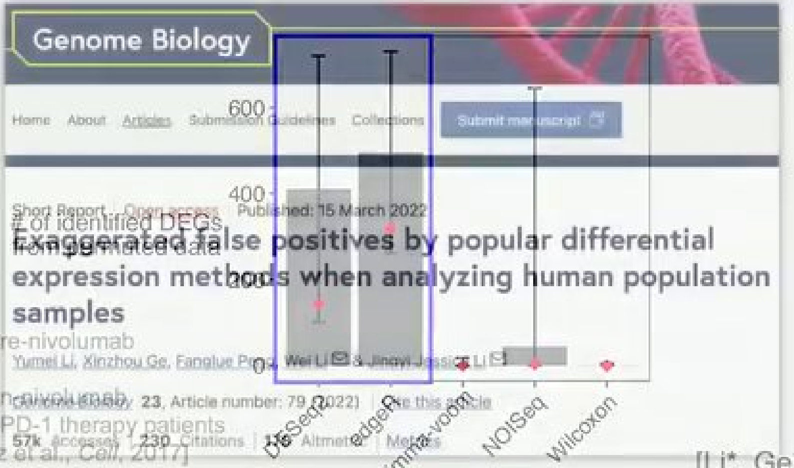
Teaser: bulk RNA-seq DE analysis

Q: Why are many genes identified as DE genes from permuted data?

A: The negative binomial assumption does not hold on this dataset.



Teaser: bulk RNA-seq DE analysis



51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., Cell, 2017]

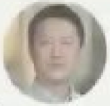
◆ # of identified DEGs from the original data

X/Twitter: @jsb_ucla

[Li*, Ge* et al., *Genome Biology*, 2022]

Yumei Li
(Wei Li Lab→
Soochow U)

Xinzhou Ge
(JSB→
OregonState)



Wei Li
(UC Irvine)

Teaser: bulk RNA-seq DE analysis

Genome Biology

Home About Articles Submission Guidelines Collections Submit manuscript

Short Report | Open access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li & Jingyi Jessica Li

Genome Biology 23, Article number: 79 (2022) | Cite this article

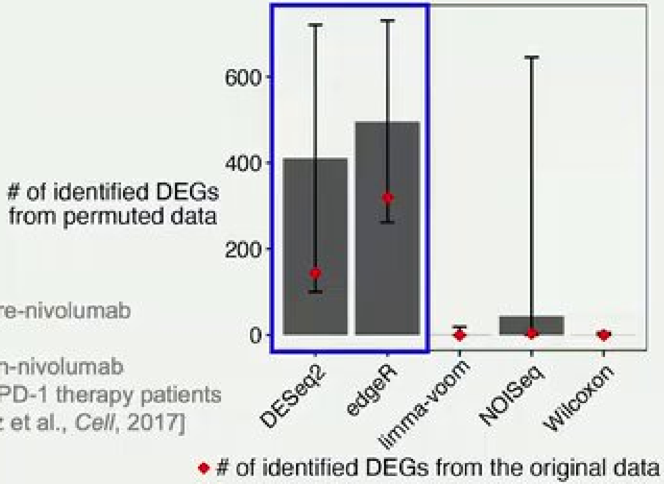
X/Twitter: @jsb_ucla

Yumei Li
(Wei Li Lab→ Soochow U)

Xinzhou Ge
(JSB→ OregonState)

Wei Li
(UC Irvine)

Teaser: bulk RNA-seq DE analysis

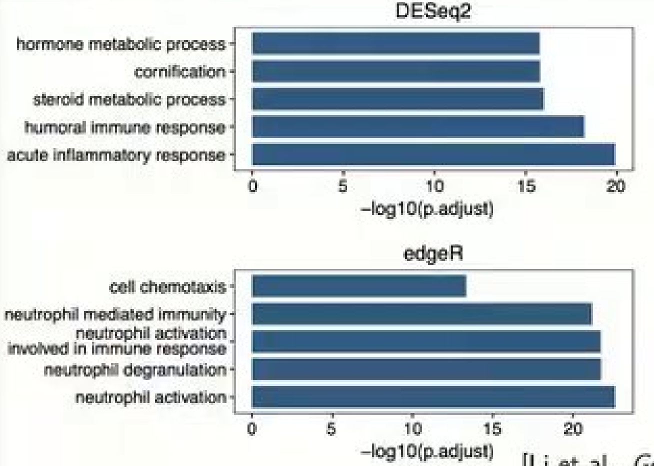


51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., *Cell*, 2017]

[Li*, Ge* et al.,
Genome Biology, 2022]

Teaser: bulk RNA-seq DE analysis

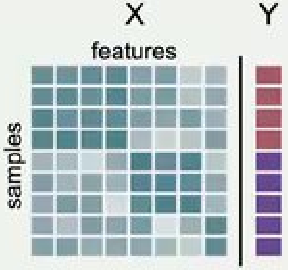
False discoveries may mislead scientific conclusions



[Li et al., *Genome Biology*, 2022]

How to permute data?

Supervised learning

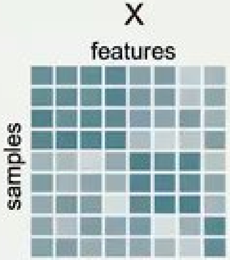


Bulk RNA-seq:

features = genes

Y = sample condition labels

Unsupervised learning



Single-cell RNA-seq:

samples = cells;

features = genes

Two examples where permutation helps

1. Single-cell data **visualization**

Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters

Lucy Xia, Christy Lee & Jingyi Jessica Li

Nature Communications **15**, Article number: 1753 (2024) | Cite this article

2. Aggregating single cells into **metacells**

mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu & Jingyi Jessica Li

Nature Communications **16**, Article number: 8602 (2025) | Cite this article

Example 1: dubious t-SNE/UMAP embeddings?

How to Use t-SNE Effectively

- **Hyperparameters** really matter
- **Distances between clusters** might not mean anything
- …

Source: https://distill.pub/2016/misread-tsne/

Example 1: dubious t-SNE/UMAP embeddings?

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › technology features › article

Technology Feature | Published: 24 May 2024

Seeing data as t-SNE and UMAP do

Vivien Marx

Nature Methods 21, 930–933 (2024) | Cite this article

Dimension reduction helps to visualize high-dimensional datasets. These tools should be used thoughtfully and with tuned parameters. Sometimes, these methods take a second thought.

Example 1: dubious t-SNE/UMAP embeddings?

Q: Is a cell's embedding dubious or trustworthy?

A: Examine the cell's neighbors before and after embedding

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature communications > articles > article

Article | Open access | Published: 26 February 2024

Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters

Lucy Xia, Christy Lee & Jingyi Jessica Li

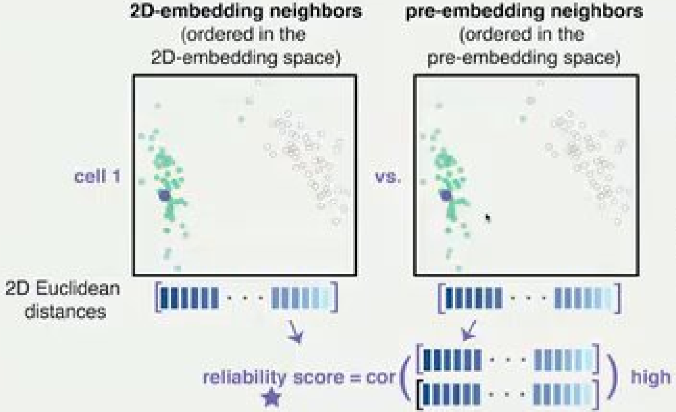
Nature Communications 15, Article number: 1753 (2024) | Cite this article


Lucy Xia (HKUST) Christy Lee (JSB)

Example 1: dubious t-SNE/UMAP embeddings?

scDEED intuition

A trustworthy cell embedding



pre-embedding neighbors

defined in the PC space

neighborhood size

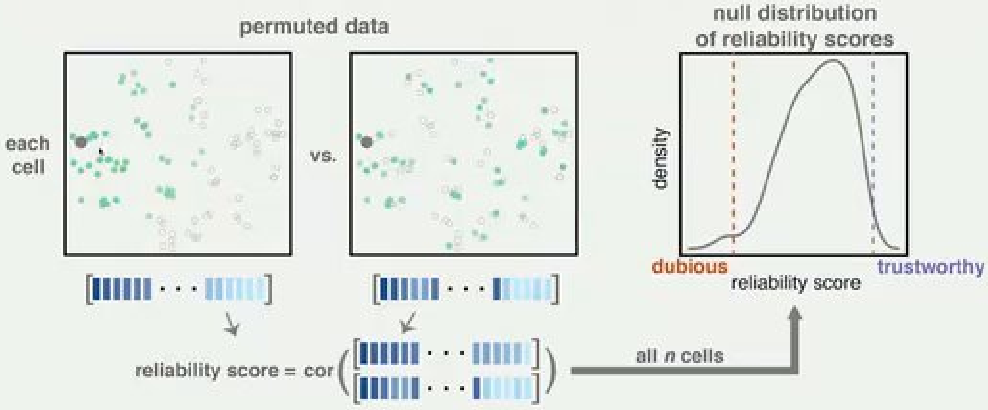
= half of all cells

"mid-range distances"

MCBRLab bilibili

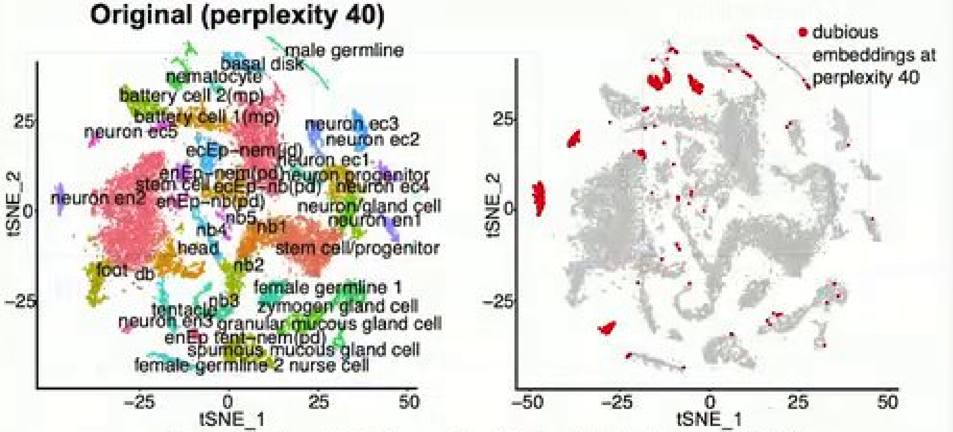
Example 1: dubious t-SNE/UMAP embeddings?

Permuted cells are exchangeable → A cell's neighbors are random



Example 1: dubious t-SNE/UMAP embeddings?

scDEED detects dubious embeddings

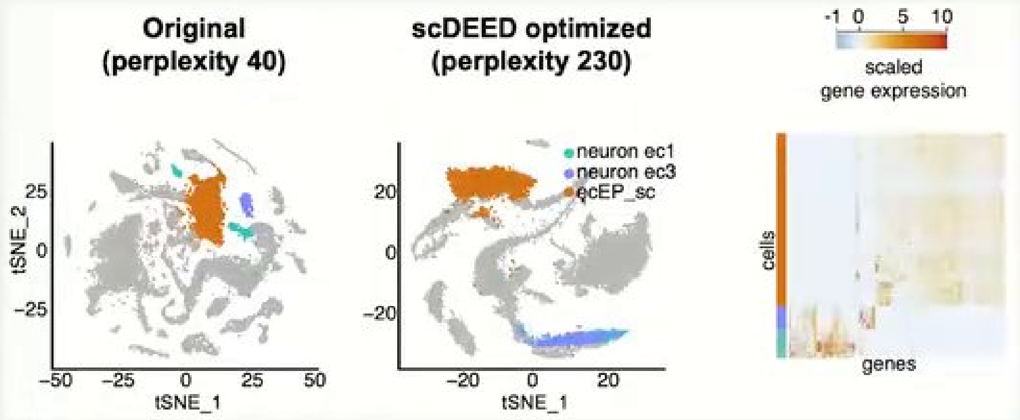


Hydra single-cell RNA-seq data [Siebert et al., *Science*, 2019]

MCBRLab bilibili

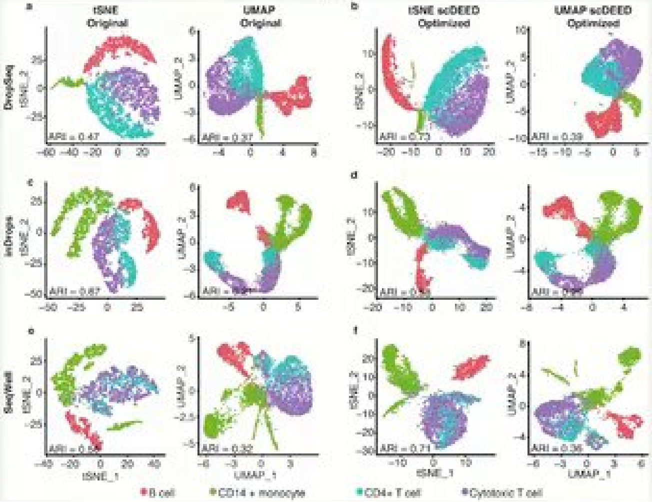
Example 1: dubious t-SNE/UMAP embeddings?

scDEED optimizes hyperparameters by minimizing dubious embeddings



Example 1: dubious t-SNE/UMAP embeddings?

scDEED enhances the consistency between t-SNE and UMAP



Two examples where permutation helps

1. Single-cell data **visualization**

Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters

Lucy Xia, Christy Lee & Jingyi Jessica Li

Nature Communications **15**, Article number: 1753 (2024) | Cite this article

2. Aggregating single cells into **metacells**

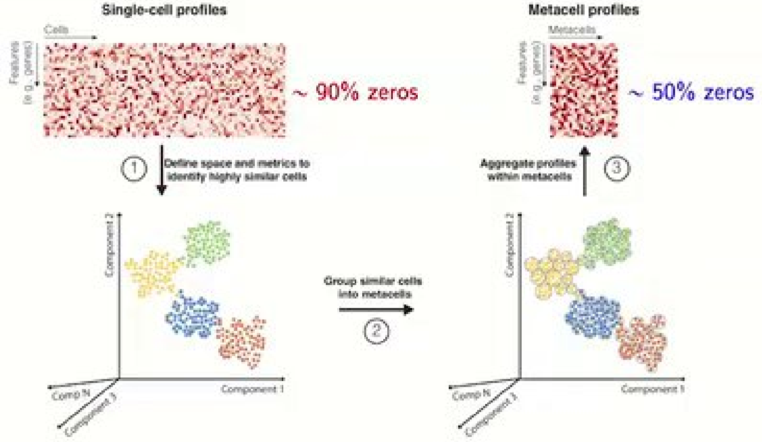
mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu & Jingyi Jessica Li

Nature Communications **16**, Article number: 8602 (2025) | Cite this article

Example 2: aggregating single cells into metacells

Metacell: a heuristic solution to the sparsity issue in single-cell data

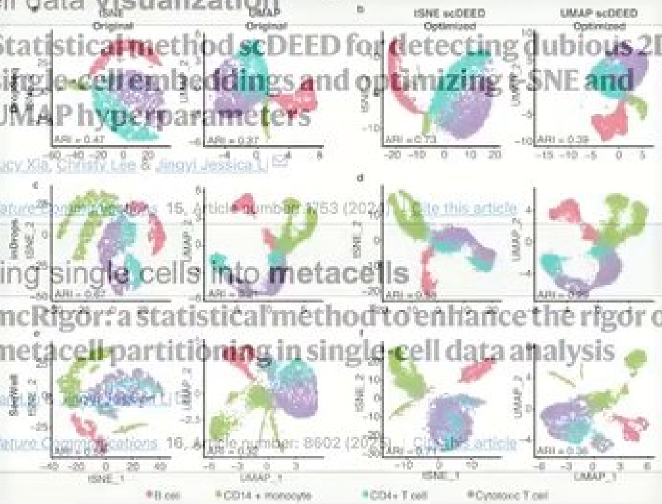


Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

MCBRLab bilibili

Example 1: dubious t-SNE/UMAP embeddings?

scDEED enhances the consistency between t-SNE and UMAP



1. Single-cell data visualization

Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters

Lucy Xia, Christy Lee & Jingyi Jessica Li

Nature Communications 15, Article number: 1753 (2024) | Cite this article

2. Aggregating single cells into metacells

mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

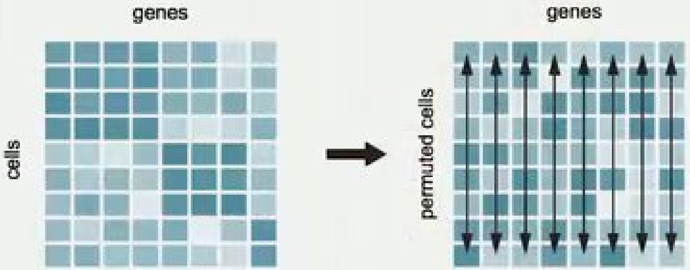
Pan Liu & Jingyi Jessica Li

Nature Communications 16, Article number: 8602 (2025) | Cite this article

Example 1: dubious t-SNE/UMAP embeddings?

Q: What is preserved by within-gene permutation?

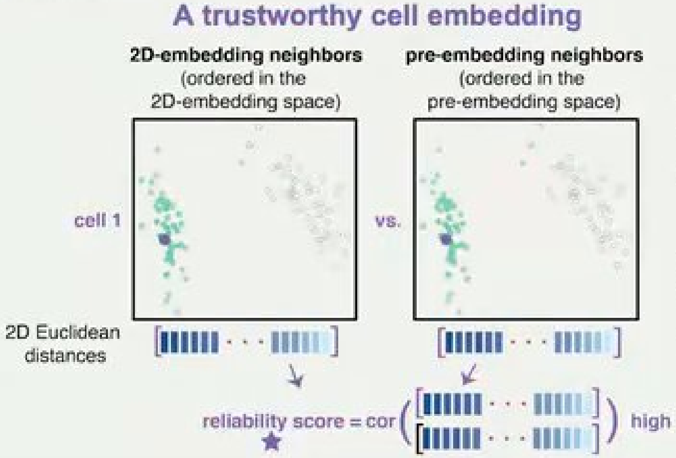
A: Every gene's distribution.



within-gene permutation

Example 1: dubious t-SNE/UMAP embeddings?

scDEED intuition



pre-embedding neighbors

defined in the PC space

neighborhood size

= half of all cells

"mid-range distances"

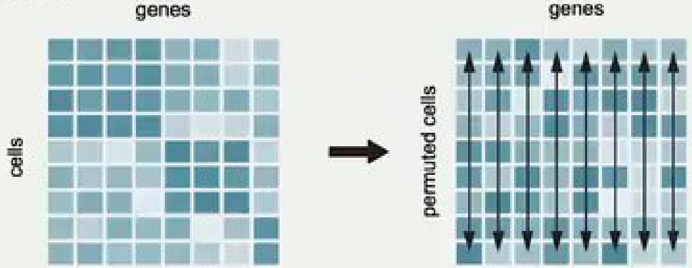
Example 1: dubious t-SNE/UMAP embeddings?

Q: What is the null hypothesis?

A: A cell's neighbors are random after embedding.

Q: How to obtain such a case?

A: Permutation.



within-gene permutation

MCBRLab bilibili

Two examples where permutation helps

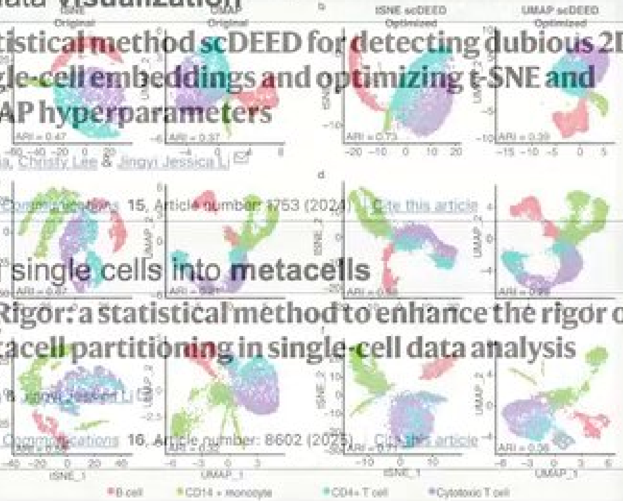
scDEED enhances the consistency between t-SNE and UMAP

1. Single-cell data **visualization**

Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters

Lucy Xia, Christy Lee & Jingyi Jessica Li

Nature Communications 15, Article number: 1753 (2024) | Cite this article



2. Aggregating single cells into **metacells**

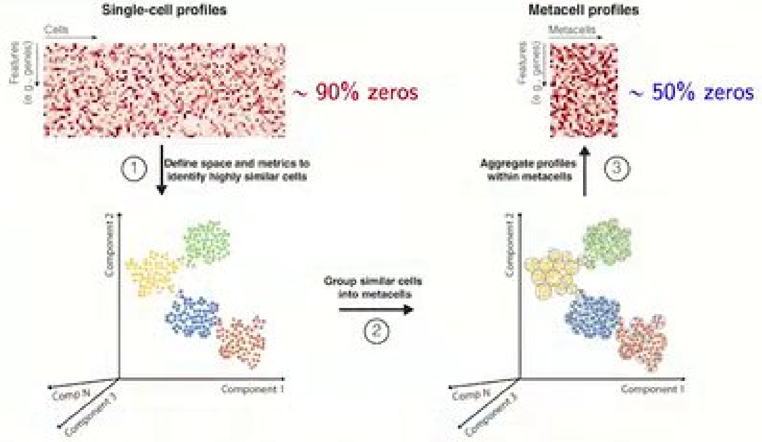
mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu & Jingyi Jessica Li

Nature Communications 16, Article number: 8602 (2025) | Cite this article

Example 2: aggregating single cells into metacells

Metacell: a heuristic solution to the sparsity issue in single-cell data



Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

Example 2: aggregating single cells into metacells

A statistical definition of "metacell"

"A homogeneous collection of single-cell profiles that could have been resampled from the same original cell."

$\Longrightarrow$ Variation within a metacell is attributed exclusively to measurement error

Two-layer observation model: Cell (observation) $i = 1, \ldots, n$; Feature $j = 1, \ldots, p$

Expression model: $\boldsymbol{\lambda}_i \sim \mathcal{F}(\cdot|\mathbf{x}_i)$

Measurement model: $y_{ij} \sim \mathcal{G}(y_{i+}\lambda_{ij})$

$\Downarrow$

Statistical definition: A **metacell** is a group of single cells that share the same $\boldsymbol{\lambda}$

$\Downarrow$

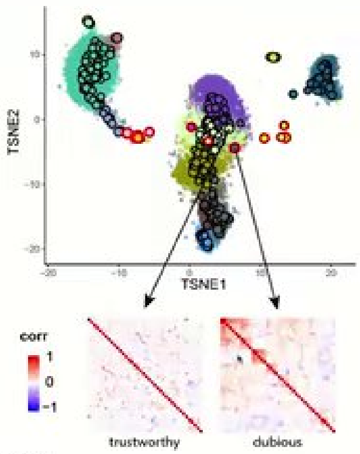
Satisfying this definition?

Yes: **trustworthy metacells** — No: **dubious metacells**

$\boxed{\text{A statistical problem}}$

Example 2: aggregating single cells into metacells

Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

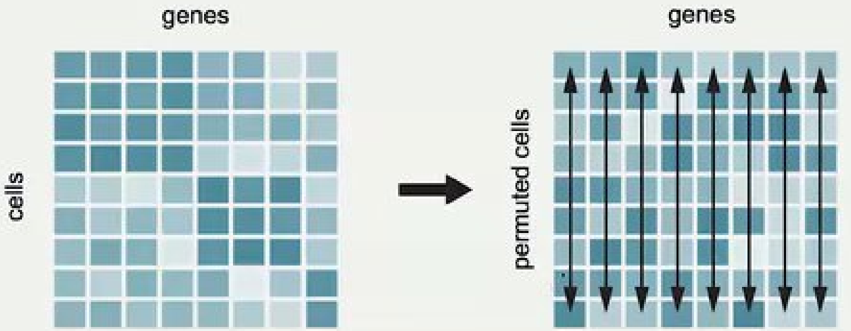
$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

in a data-specific way

Pan Liu
(JSB)

Example 2: aggregating single cells into metacells

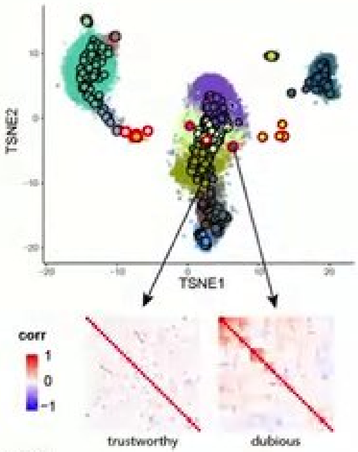
Q: Is within-gene permutation enough?



within-gene permutation

Example 2: aggregating single cells into metacells

Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

in a data-specific way

Intuition:

- Within a trustworthy metacell, features are approximately uncorrelated

Pan Liu
(JSB)

Example 2: aggregating single cells into metacells

A statistical definition of "metacell"

"A homogeneous collection of single-cell profiles that could have been resampled from the same original cell."

$\Longrightarrow$ Variation within a metacell is attributed exclusively to measurement error

Two-layer observation model: Cell (observation) $i = 1, \ldots, n$
Feature $j = 1, \ldots, p$

Expression model: $\boldsymbol{\lambda}_i \sim \mathcal{F}(\cdot|\mathbf{x}_i)$

Measurement model: $y_{ij} \sim \mathcal{G}(y_{i+}\lambda_{ij})$

$\Downarrow$

Statistical definition: A **metacell** is a group of single cells that share the same $\boldsymbol{\lambda}$

$\Downarrow$

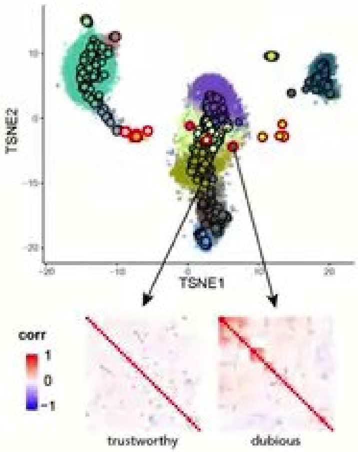
Satisfying this definition?

Yes: **trustworthy metacells** No: **dubious metacells**

$\boxed{\text{A statistical problem}}$

Example 2: aggregating single cells into metacells

Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

in a data-specific way

Intuition:

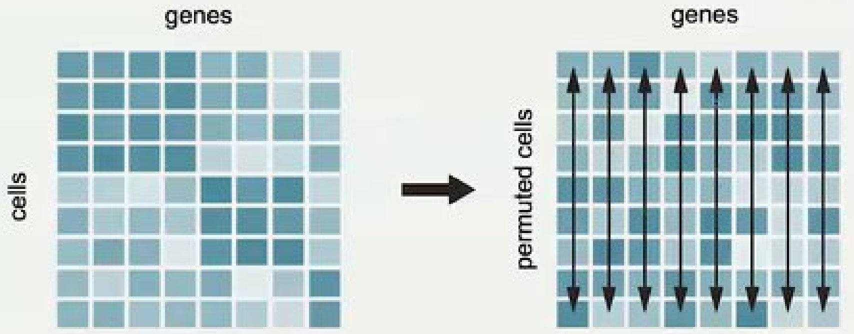
- Within a trustworthy metacell, features are approximately uncorrelated

Pan Liu
(JSB)

Example 2: aggregating single cells into metacells

Q: Is within-gene permutation enough?

A: Genes become uncorrelated, but cell library sizes are gone.



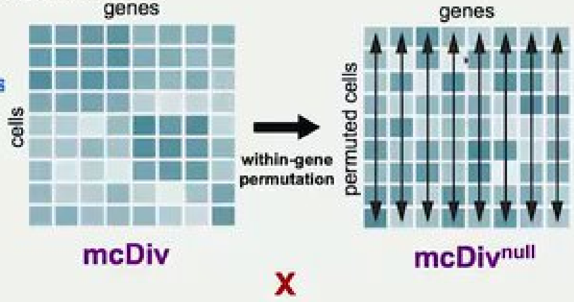
within-gene permutation

Example 2: aggregating single cells into metacells

Double permutation

Within-gene permutation:

- preserves genes marginal distributions
- removes gene correlations
- removes cell library sizes

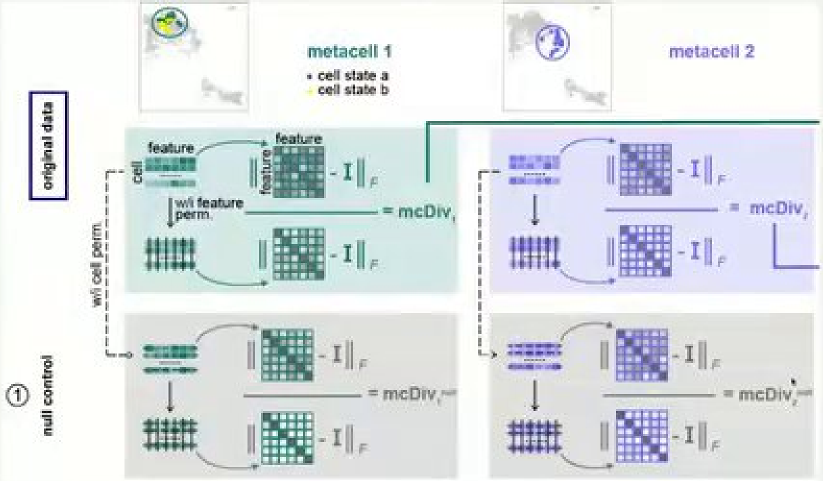


Pan Liu
(JSB)

MCBRLab bilibili

Example 2: aggregating single cells into metacells

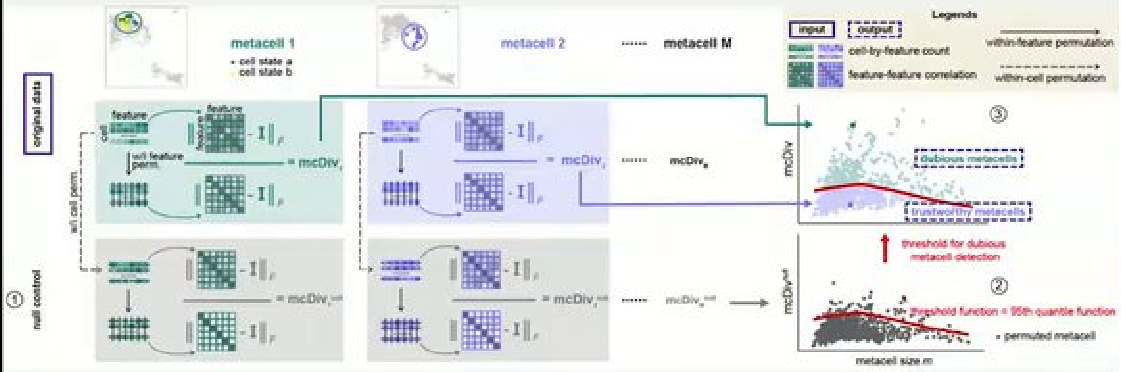
mcRigor function 1: detecting dubious metacells



MCBRLab bilibili

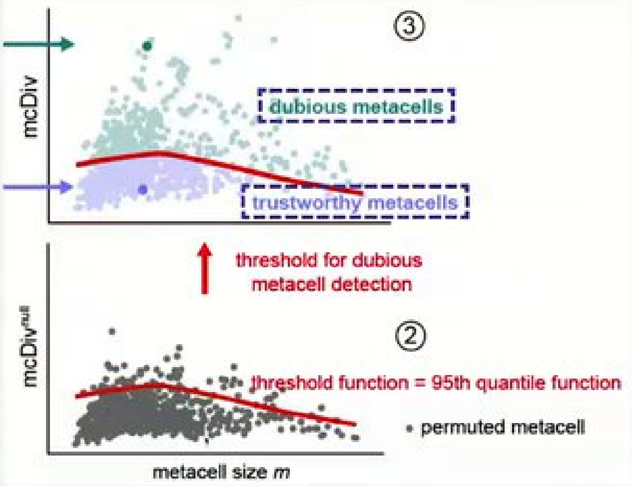
Example 2: aggregating single cells into metacells

mcRigor function 1: detecting dubious metacells

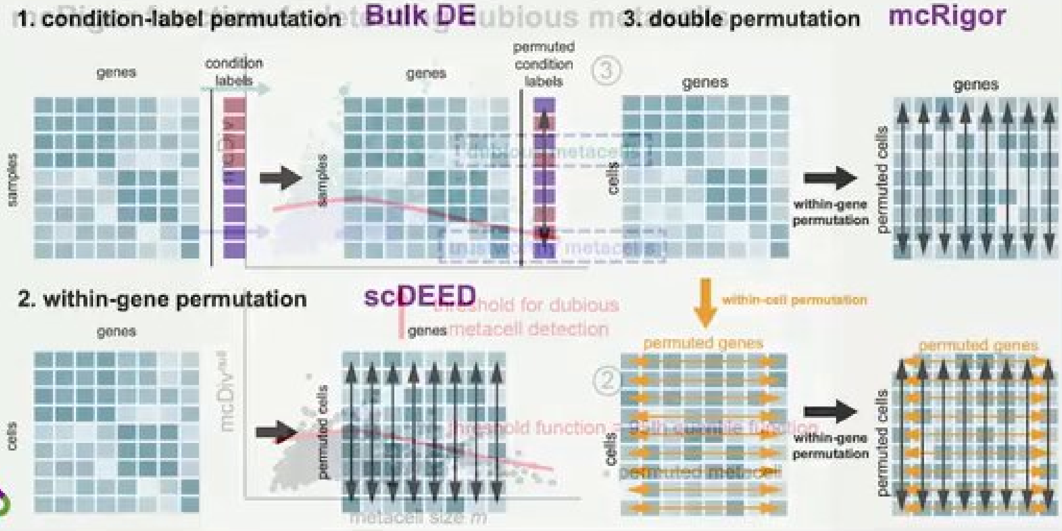


Example 2: aggregating single cells into metacells

mcRigor function 1: detecting dubious metacells



Summary


1. condition-label permutation
Bulk DE
3. double permutation
mcRigor
genes
condition labels
samples
permuted condition labels
cells
within-gene permutation
permuted cells
2. within-gene permutation
scDEED
within-cell permutation
permuted genes

Acknowledgements

Bulk DE analysis — **scDEED** — **mcRigor**

Yumei Li (Wei Li Lab→ Soochow U)

Xinzhou Ge (JSB→ OregonState)

Wei Li (UC Irvine)

Lucy Xia (HKUST)

Christy Lee

Pan Liu

NIH NIGMS · NSF · Alfred P. Sloan Foundation · Johnson & Johnson WiSTEM²D · W.M. Keck Foundation · Chan Zuckerberg Initiative

More generally,

Synthetic null data

ClusterDE: a post-clustering DE method for single-cell and spatial transcriptomics data

Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses

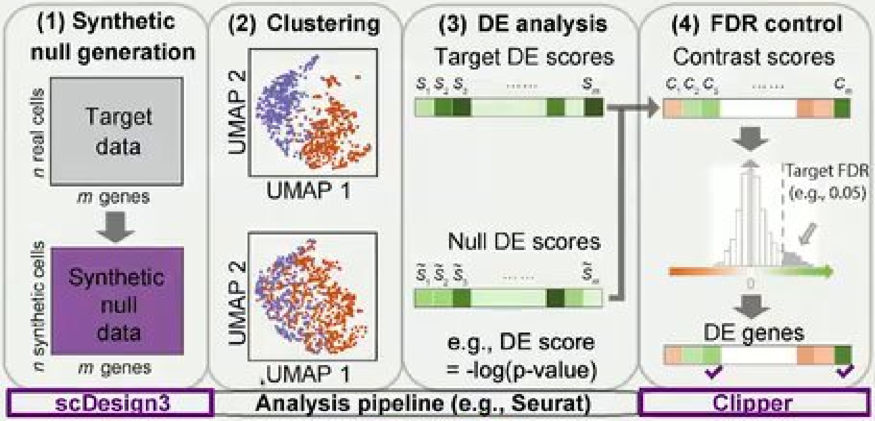
Dongyuan Song, Siqi Chen, Christy Lee, Kexin Li, Xinzhou Ge, Jingyi Jessica Li

doi: https://doi.org/10.1101/2023.07.21.550107

Dongyuan Song Kexin Li Christy Lee Siqi Chen

Single-cell post-clustering DE analysis

Q: How to control false discoveries using synthetic null data?



ClusterDE

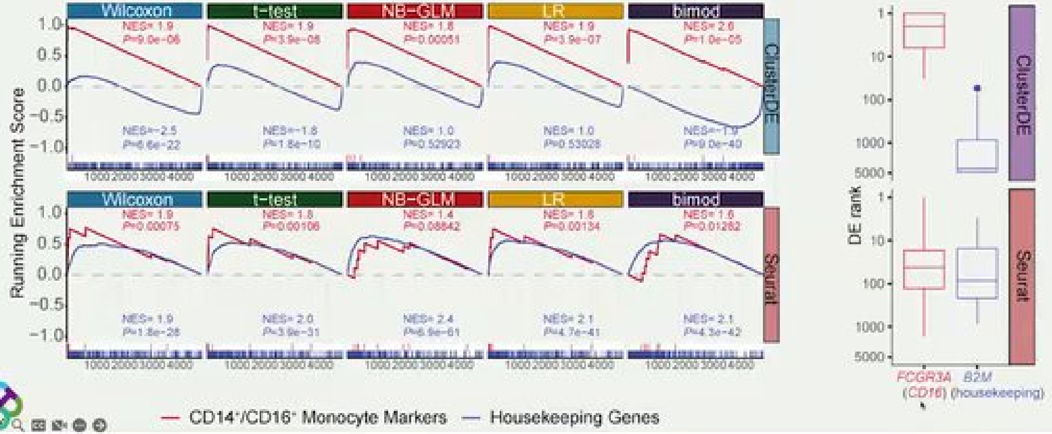
Dongyuan Song

Kexin Li
(JSB→Microsoft)

Single-cell post-clustering DE analysis

Expectation 1: Cell-type marker genes should be found as top DE genes.

Expectation 2: Housekeeping genes should NOT be found as top DE genes.

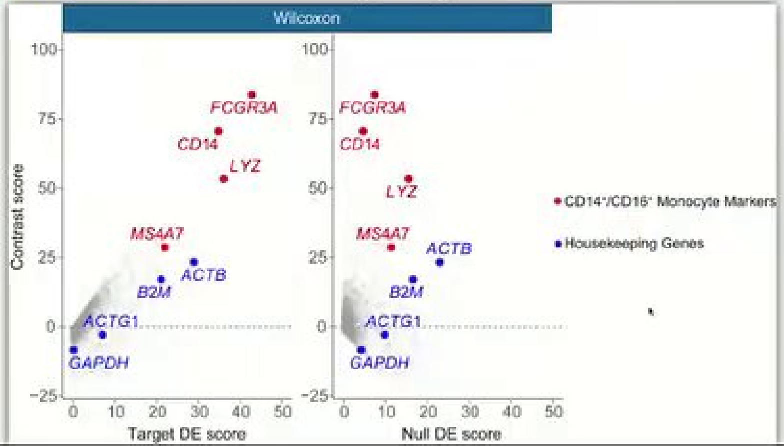


MCBRLab bilibili

Single-cell post-clustering DE analysis

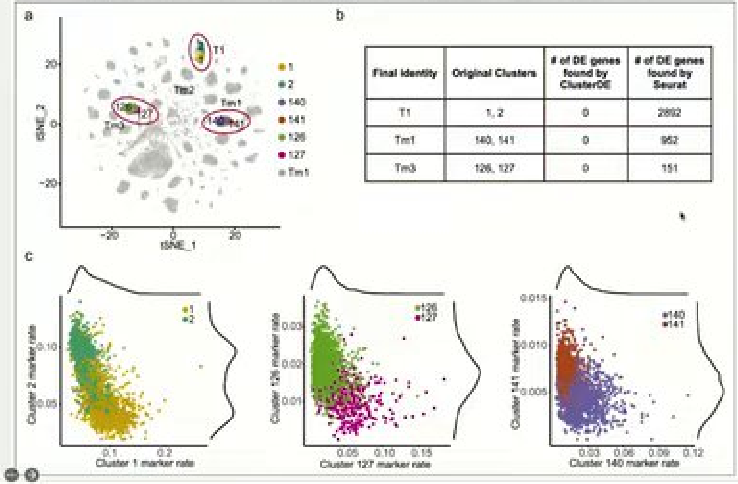
Q: Why does ClusterDE NOT identify housekeeping genes as top DE genes?

A: ClusterDE uses contrast scores (= target DE score – null DE score).



Single-cell post-clustering DE analysis

ClusterDE guides the merging of spurious clusters.



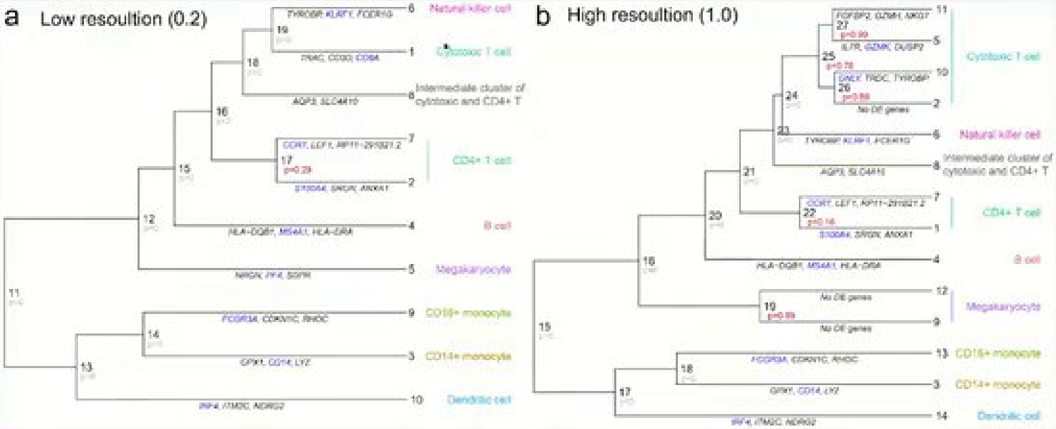
Final identity	Original Clusters	# of DE genes found by ClusterDE	# of DE genes found by Seurat
T1	1, 2	0	2892
Tm1	140, 141	0	952
Tm3	126, 127	0	151

Drosophila visual system developmental atlas

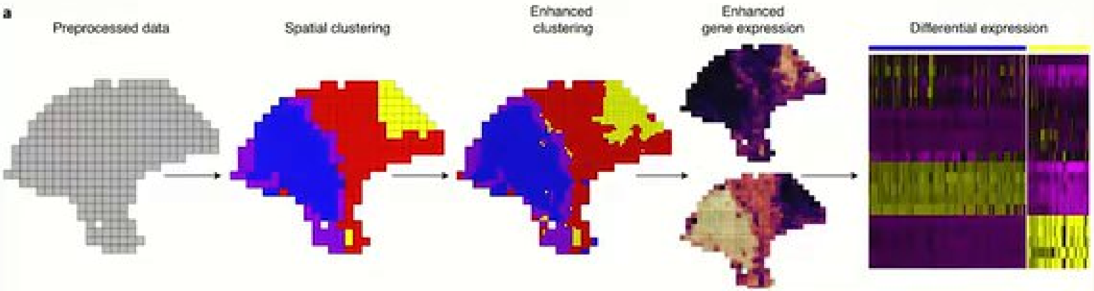
Ozel et al. *Nature* (2021)

Single-cell post-clustering DE analysis

ClusterDE identifies cell-type markers in a cell cluster hierarchy.

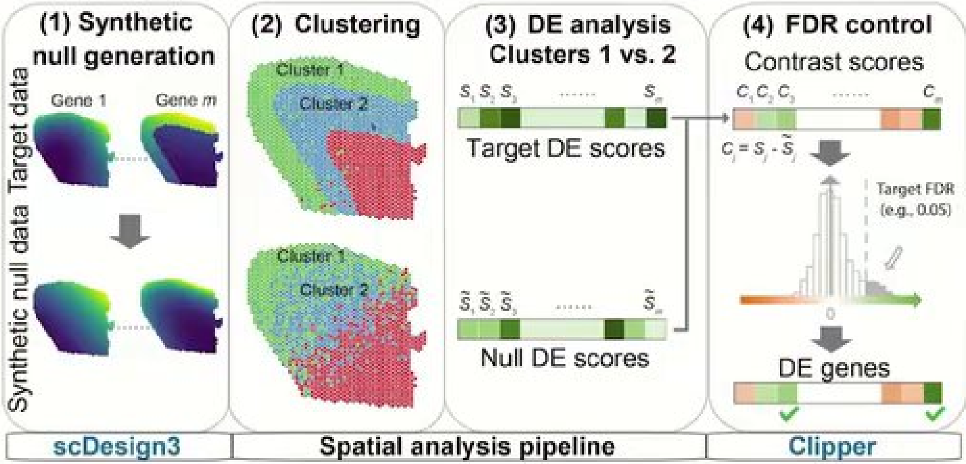


Spatial post-clustering DE analysis

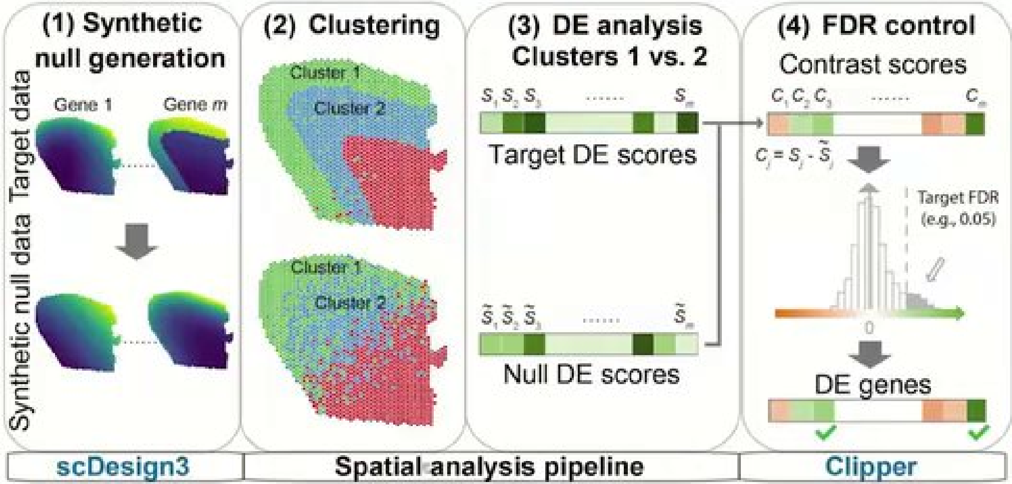


Source: https://www.nature.com/articles/s41587-021-00935-2

Spatial post-clustering DE analysis

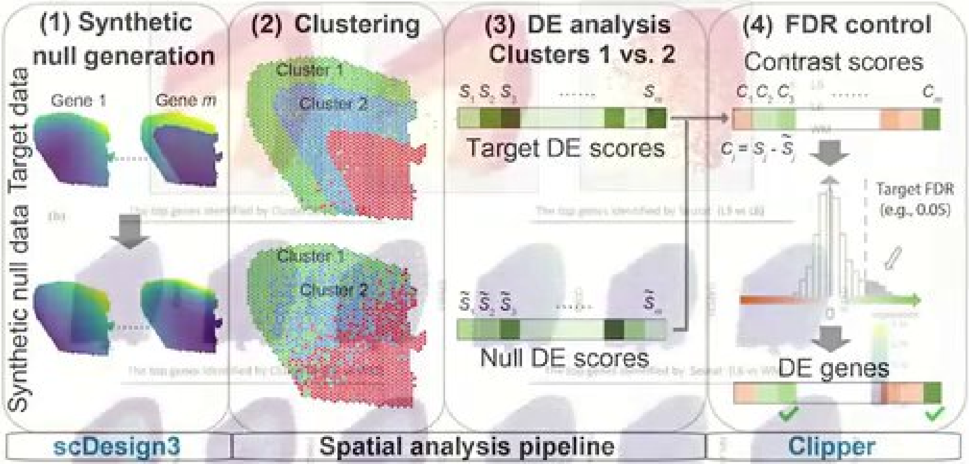


Spatial post-clustering DE analysis

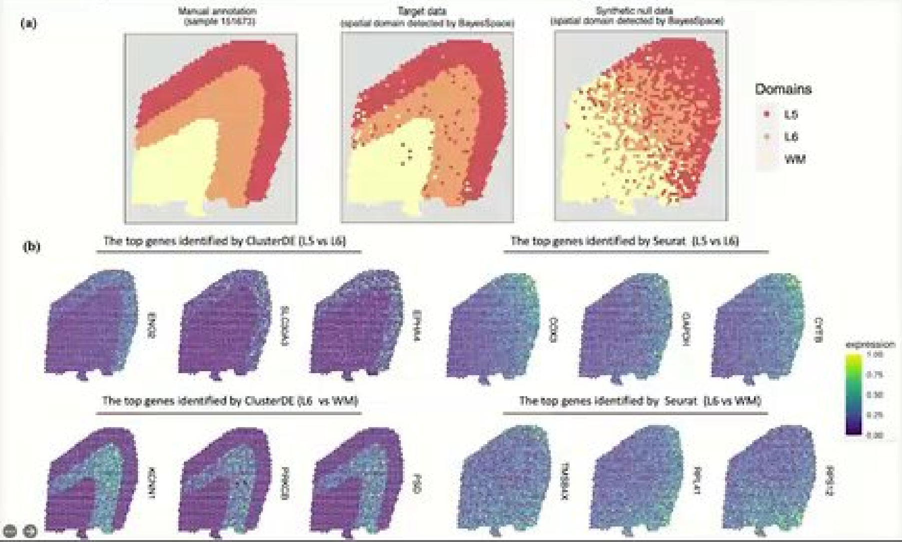


MCBRLab bilibili

Spatial post-clustering DE analysis



Spatial post-clustering DE analysis



ClusterDE: a synthetic control method for removing clustering-induced bias

Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses

Dongyuan Song, Siqi Chen, Christy Lee, Kexin Li, Xinzhou Ge, Jingyi Jessica Li

doi: https://doi.org/10.1101/2023.07.21.550107

Dongyuan Song Kexin Li Christy Lee Siqi Chen

ClusterDE: a synthetic control method for removing clustering-induced bias

More generally: Nullstrap

A general framework for statistical inference that

- Constructs **synthetic null data** using a model trained under the null
- Applies the same selection and inference procedure to both the observed and synthetic null datasets
- Calibrates the results from the two datasets to control the **FDR**
- Avoids data alteration and preserves **power**

Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses

Dongyuan Song, Siqi Chen, Christy Lee, Kexin Li, Xinzhou Ge, Jingyi Jessica Li

doi: https://doi.org/10.1101/2023.07.21.550107

arXiv > stat > arXiv:2501.05012

Statistics > Methodology

[Submitted on 9 Jan 2025 (v1), last revised 15 Jul 2025 (this version, v2)]

Nullstrap: A Simple, High-Power, and Fast Framework for FDR Control in Variable Selection for Diverse High-Dimensional Models

Changhu Wang, Ziheng Zhang, Jingyi Jessica Li

Dongyuan Song Kexin Li Christy Lee Siqi Chen

More Generally: Nullstrap

- A general framework for statistical inference that:
- Constructs **synthetic null data** using a model trained under the null
- Applies **the same selection and inference procedure** to both the observed and synthetic null datasets
- Calibrates the results from the two datasets to control the **FDR**
- Avoids data alteration and preserves **power**

arXiv > stat > arXiv:2501.05012

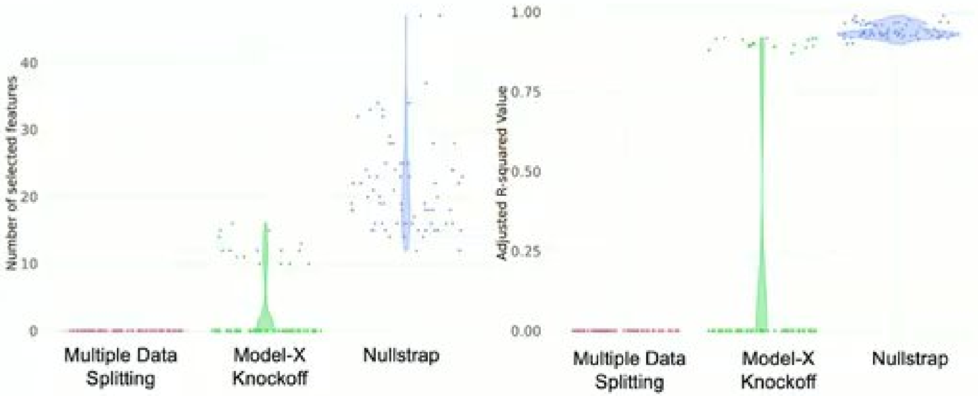
Statistics > Methodology

(Submitted on 9 Jan 2025 (v1), last revised 15 Jul 2025 (this version, v2))

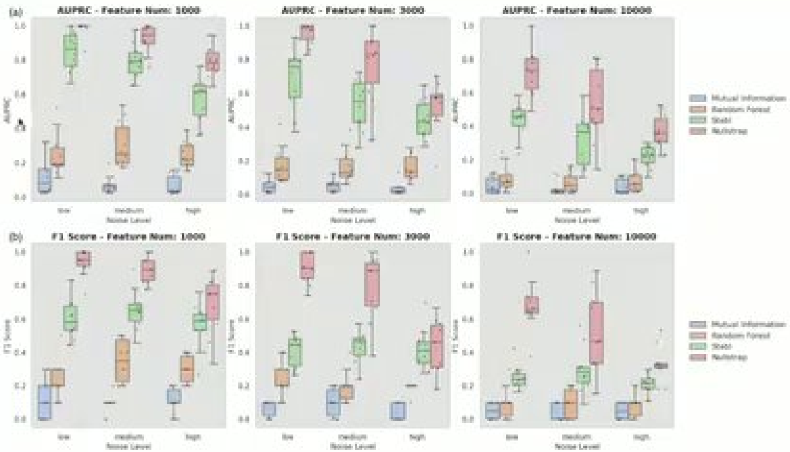
Nullstrap: A Simple, High-Power, and Fast Framework for FDR Control in Variable Selection for Diverse High-Dimensional Models

Changhu Wang, Ziheng Zhang, Jingyi Jessica Li

Real data application: the triple-omic dataset ($n = 150$, $p = 6331$)



An independent benchmark: performance comparison result



https://doi.org/10.1101/2025.09.09.675248

An independent benchmark: runtime comparison result

Method	$p = 1000$	$p = 3000$	$p = 10\,000$
Mutual Information	1.96 ± 0.02	5.87 ± 0.05	19.52 ± 0.19
Random Forest	41.03 ± 0.77	122.16 ± 1.90	406.27 ± 7.78
Stabl	92.60 ± 2.13	348.66 ± 29.11	1644.52 ± 151.70
Nullstrap	1.17 ± 0.16	3.08 ± 0.40	8.50 ± 0.75

https://doi.org/10.1101/2025.09.09.675248

Summary: Toward Principled Inference After Unsupervised Learning

- **Inference Target**: *selective (conditional)* or *unconditional* inference?
- **Challenge**: Selective inference is method-specific and requires
  – Complex derivations (feasible for few algorithms)
  – Or strong assumptions (e.g., data splitting),
  making it difficult to apply broadly in unsupervised learning.
- **Nullstrap**:
  A *numerical analog* of the likelihood ratio test, akin to bootstrap.
  – Avoids analytical derivation
  – Uses synthetic null data to calibrate inference
- **Outlook**: More theoretical development and applications

Acknowledgements

ClusterDE


Dongyuan Song (Uconn)


Kexin Li


Christy Lee


Siqi Chen

ClusterDE Theory


Changhu Wang


Xinzhou Ge (Oregon State)

Nullstrap


Changhu Wang


Ziheng Zhang

NIH NIGMS · NSF · Alfred P. Sloan Foundation · Johnson & Johnson WiSTEM2D · W.M. Keck Foundation · Chan Zuckerberg Initiative

Models, Inference & Algorithms (MIA)

Meeting: mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu

November 5, 2025

mcRigor: a statistical method to enhance the rigor of metacell (mc) partitioning in single-cell data analysis

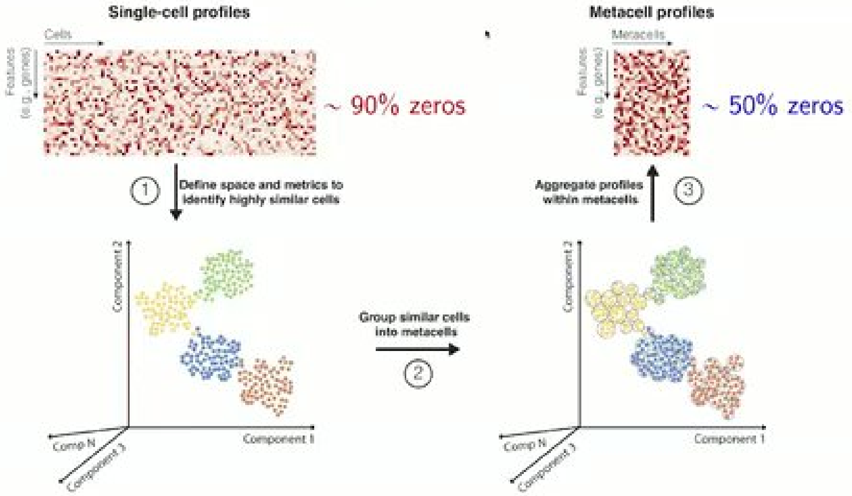
Pan Liu

Postdoctoral researcher

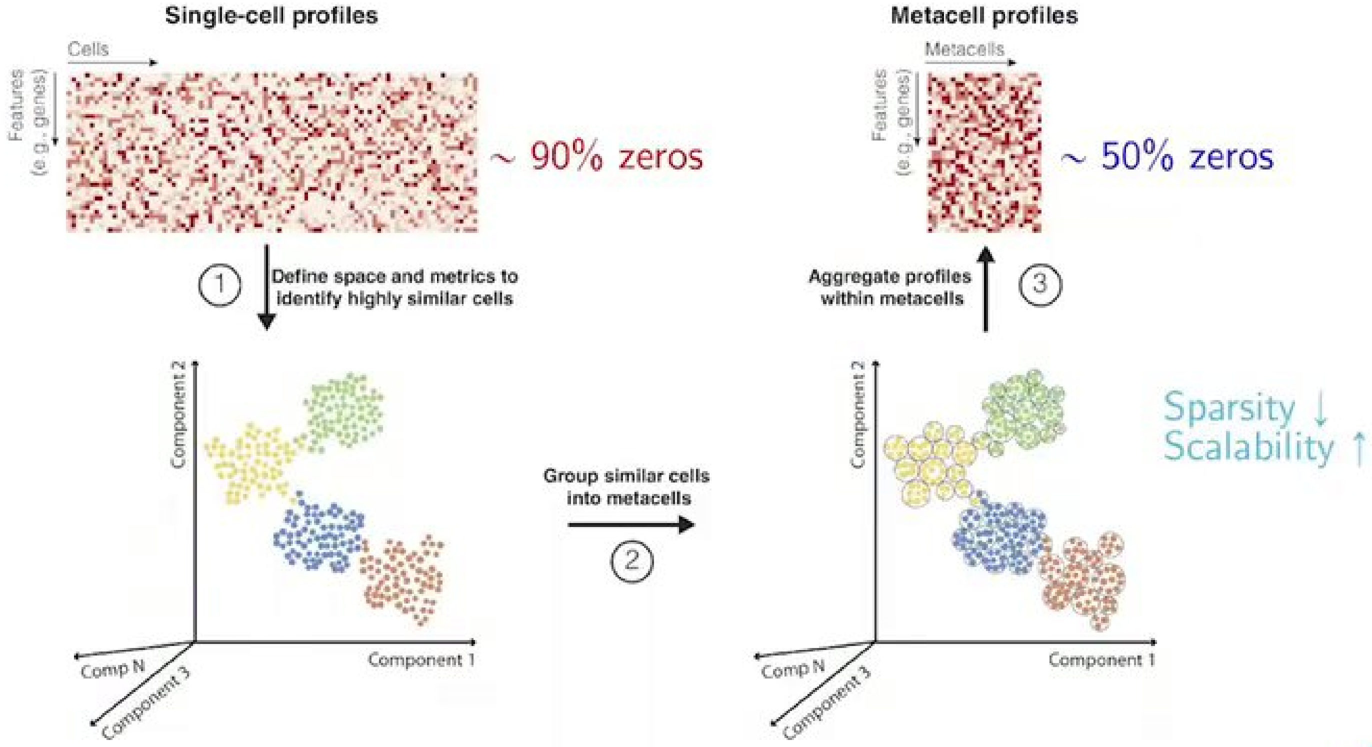
Biostatistics Program, Fred Hutchinson Cancer Center
Department of Statistics and Data Science, University of California, Los Angeles

with Prof. Jingyi Jessica Li

Metacell: a heuristic solution to the sparsity issue in single-cell data



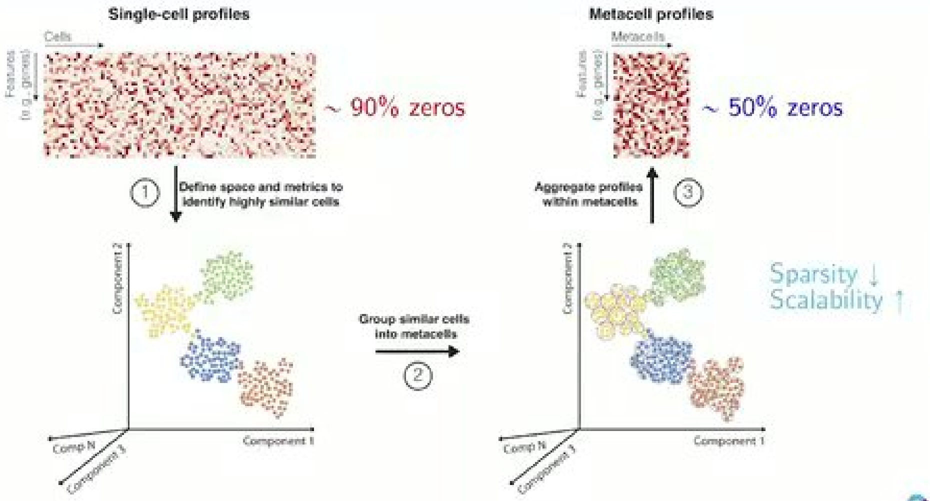
Metacell: a heuristic solution to the sparsity issue in single-cell data



Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

MCBRLab bilibili

Metacell: a heuristic solution to the sparsity issue in single-cell data



Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

MCBRLab bilibili

Metacell methods and applications

Metacell Methods

MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions

MetaCell

Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis

MetaCell-2

Metacells untangle large and complex single-cell transcriptome networks

SuperCell

SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data

SEACells

MetaQ: fast, scalable and accurate metacell inference via single-cell quantization

MetaQ

Metacell Applications

Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis

Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells

NASH limits anti-tumour surveillance in immunotherapy-treated HCC

Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer

Questions we wish to answer

Q: How to define a "metacell"?

Q: How to detect dubious metacells?

Q: How to optimize metacell partitioning?

Article | Open access | Published: 29 September 2025

mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu & Jingyi Jessica Li

Nature Communications **16**, Article number: 8602 (2025) | Cite this article

How to define a "metacell"?

The first publication that proposed the "metacell" concept

Method | Open access | Published: 11 October 2019

MetaCell: analysis of single-cell RNA-seq data using *K*-nn graph partitions

Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz & Amos Tanay

Genome Biology **20**, Article number: 206 (2019) | Cite this article

40k Accesses | **163** Citations | **46** Altmetric | Metrics

"A homogeneous collection of single-cell profiles that could have been resampled from the same original cell."

How to define "measurement error" in single-cell sequencing data?

Perspective | Published: 24 May 2021

Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis

Abhishek Sarkar & Matthew Stephens

Nature Genetics 53, 770–777 (2021) | Cite this article

Expression model: distribution of true expression levels

Measurement model: distribution of observed counts | true expression levels

How to define "metacell" in a statistical way?

"A homogeneous collection of single-cell profiles that could have been resampled from the same original cell."

$\Longrightarrow$ Variation within a metacell is attributed exclusively to measurement error

Two-layer observation model:

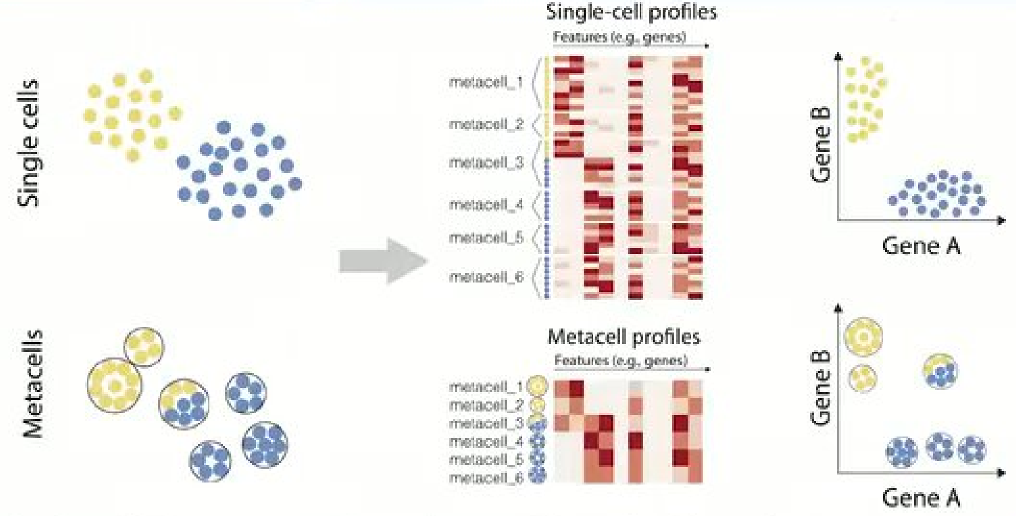
$$\textbf{Expression model:} \quad \boldsymbol{\lambda}_i \sim \mathcal{F}(\cdot|\mathbf{x}_i)$$

$$\textbf{Measurement model:} \quad y_{ij}|\boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i\lambda_{ij})$$

Cell (observation) $i = 1, \ldots, n$, Feature $j = 1, \ldots, p$

- λ_{ij}: the relative expression level of feature j in cell i; $\boldsymbol{\lambda}_i = [\lambda_{i1}, \ldots, \lambda_{ip}]^\top$
- $\mathbf{x}_i \in \mathbb{R}^q$: the covariates of cell i, e.g., cell types, pseudotimes in lineage trajectories
- $\mathcal{F}$: the p-dim expression distribution that captures meaningful biological information
- y_{ij}: the observed count of feature j in cell i; $y_{i+} = \sum_{j=1}^{p} y_{ij}$; $c_i = \mathrm{E}[y_{i+}|\boldsymbol{\lambda}_i]$

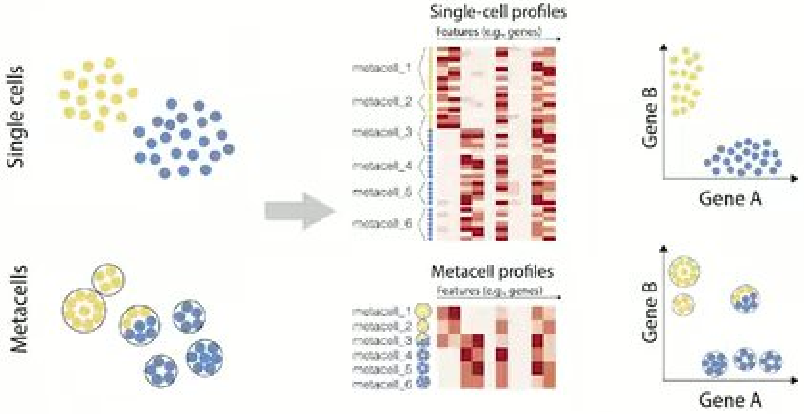
Dubious metacells can bias analysis



Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

MCBRLab bilibili

Dubious metacells introduce bias and artifacts

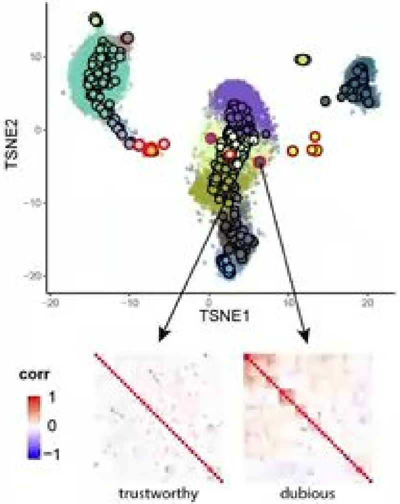


no correlation at single cell resolution

artifact correlation introduced by dubious metacells

Bilous, Mariia, et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

in a data-specific way

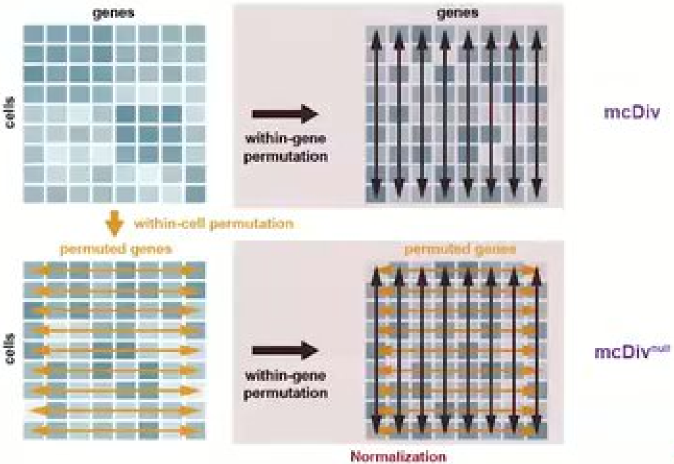
Double permutation for null construction

Within-gene permutation:

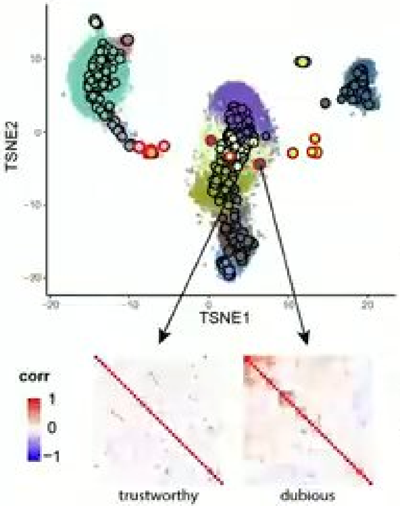
- removes genes correlations
- removes cell library sizes
- preserves genes marginal distributions

Within-cell permutation:

- preserves cell library sizes
- removes genes correlations
- removes genes marginal distributions



Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

Measurement model:

$$y_{ij} \mid \boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_{ij}), \text{ where } c_i = \text{E}[y_{i+} \mid \boldsymbol{\lambda}_i]$$

$$\Rightarrow (y_{i1}, \ldots, y_{ip}) \mid \boldsymbol{\lambda}_i, y_{i+} \overset{\text{ind}}{\sim} \text{Mult}(y_{i+}, \lambda_{i1}, \ldots, \lambda_{ip})$$

Strategy:

- Per-metacell statistic (mcDiv)
- Null: Cell-library-size-preserved permutation

MCBRLab bilibili

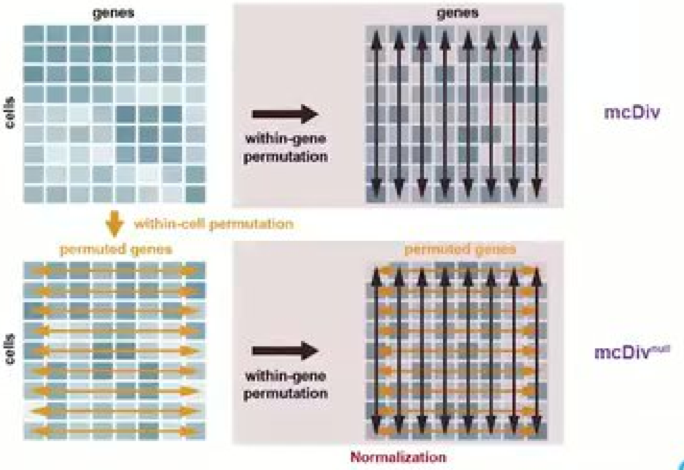
Double permutation for null construction

Within-gene permutation:

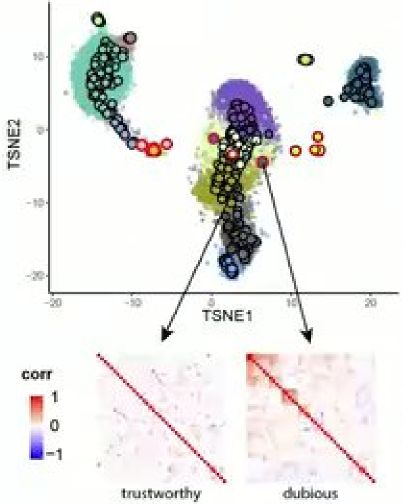
- removes genes correlations
- removes cell library sizes
- preserves genes marginal distributions

Within-cell permutation:

- preserves cell library sizes
- removes genes correlations
- removes genes marginal distributions



Our proposal: mcRigor



Goals: a statistical criterion to

- Identify dubious metacells consisting of single cells from different cell states
- Nominate the top-performing metacell method and optimize its hyperparameter

$$\text{granularity level } \gamma = \frac{\#\text{single cells}}{\#\text{metacells}}$$

Measurement model:

$$y_{ij} \mid \boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_{ij}), \text{ where } c_i = \mathrm{E}[y_{i+} \mid \boldsymbol{\lambda}_i]$$

$$\Rightarrow (y_{i1}, \ldots, y_{ip}) \mid \boldsymbol{\lambda}_i, y_{i+} \overset{\text{ind}}{\sim} \text{Mult}(y_{i+}, \lambda_{i1}, \ldots, \lambda_{ip})$$

Strategy:

- Per-metacell statistic (mcDiv)
- Null: Cell-library-size-preserved permutation

How to define “metacell” in a statistical way?

“A homogeneous collection of single-cell profiles that could have been resampled from the same original cell.”

$\Longrightarrow$ Variation within a metacell is attributed exclusively to measurement error

Two-layer observation model:

$$\textbf{Expression model:} \quad \boldsymbol{\lambda}_i \sim \mathcal{F}(\cdot|\mathbf{x}_i)$$

$$\textbf{Measurement model:} \quad y_{ij} \,|\, \boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_{ij})$$

$$\Downarrow$$

Statistical definition: A **metacell** is a group of single cells that share the same $\boldsymbol{\lambda}$

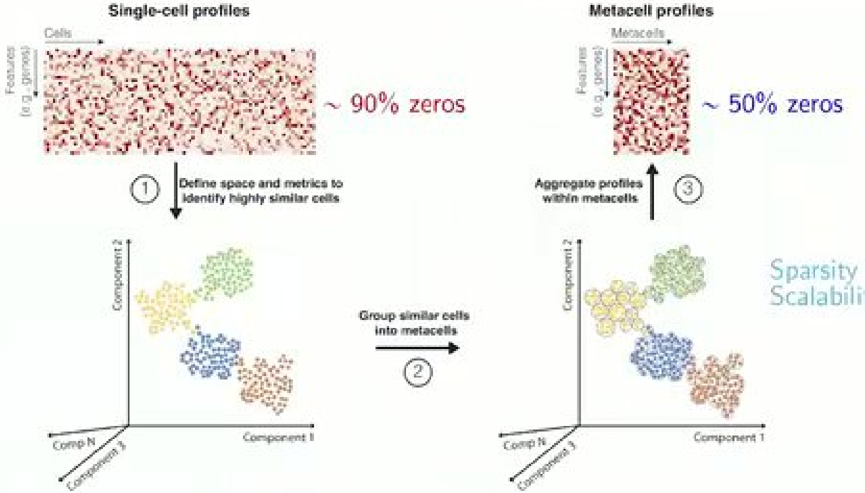
$$\Downarrow$$

Satisfying this definition?

Yes: **trustworthy metacells** No: **dubious metacells**

$\boxed{\text{A statistical problem}}$

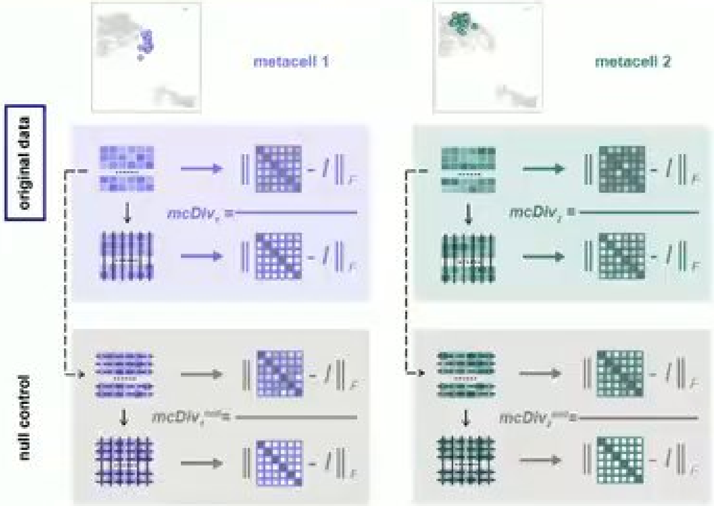
Metacell: a heuristic solution to the sparsity issue in single-cell data



Sparsity ↓
Scalability ↑

Bilous, M., et al. "Building and analyzing metacells in single-cell genomics data." Molecular Systems Biology (2024): 1-23.

mcRigor method—mcDiv and mcDiv^{null}



MCBRLab bilibili

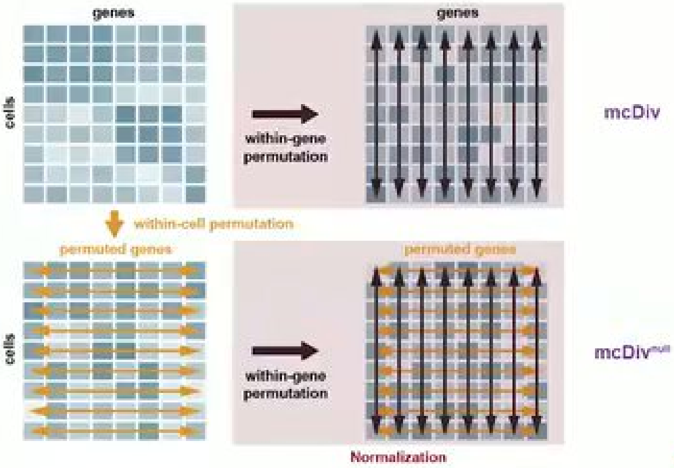
Double permutation for null construction

Within-gene permutation:

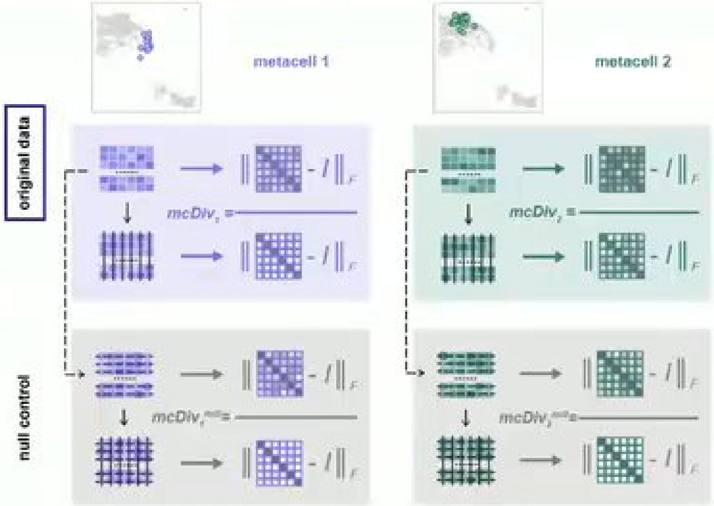
- removes genes correlations
- removes cell library sizes
- preserves genes marginal distributions

Within-cell permutation:

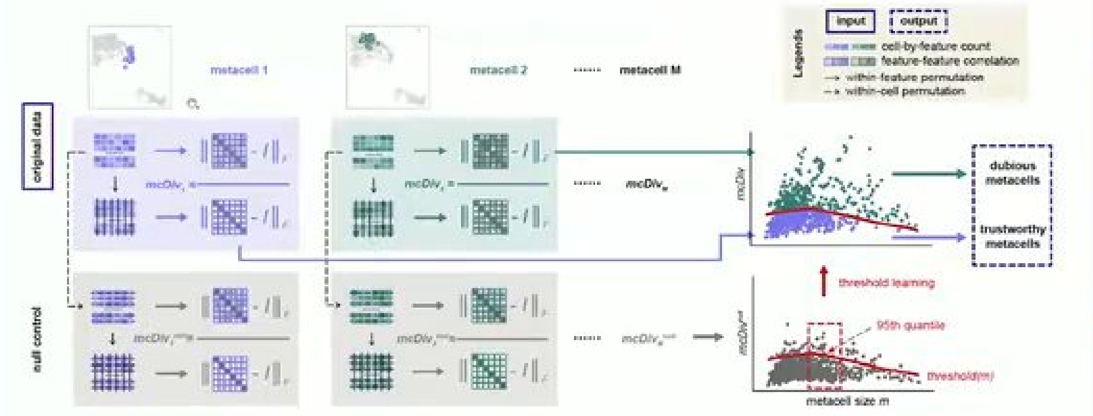
- preserves cell library sizes
- removes genes correlations
- removes genes marginal distributions



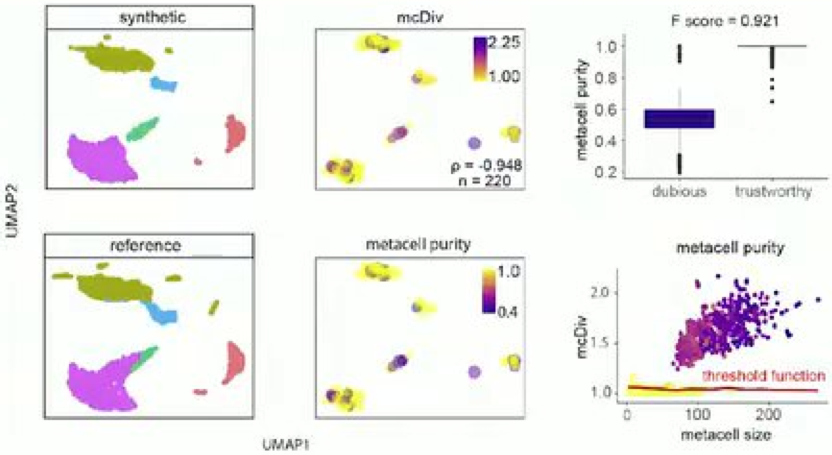
mcRigor method—mcDiv and mcDiv^{null}



mcRigor method



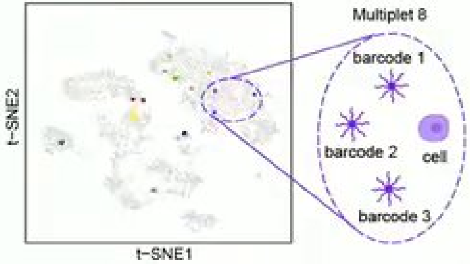
mcRigor effectively detects dubious metacells with high accuracy

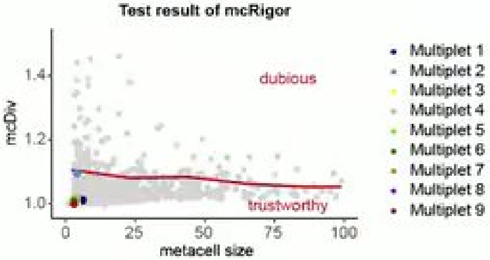


Simulator: *Song, D., et al. "scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics." Nat Biotech (2024)*
MetaCell: *Baran, Y., et al. "MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions." Genome Biology (2019)*

Test of mcRigor on barcode multiplets

Barcode multiplet: a set of cell-like observations in which each observation is assigned a unique cell barcode but actually originates from the same physical cell.

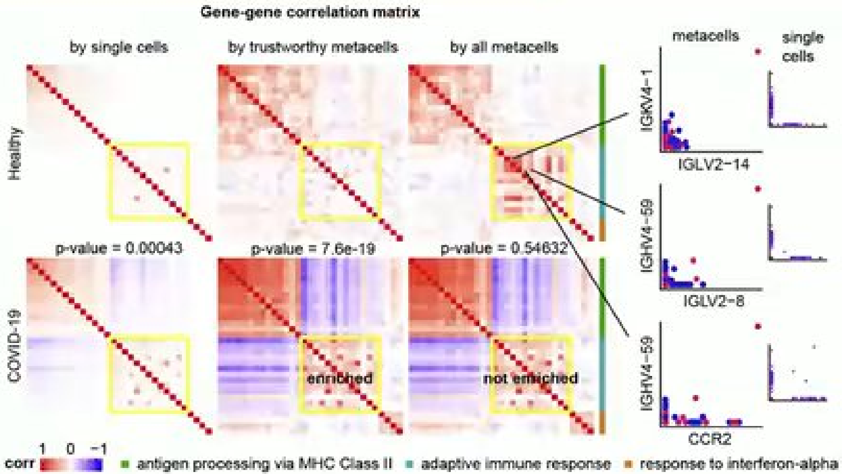




Lareau, C. A, et al. "Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility." Nat Biotech (2019)
SuperCell: Bilous, M., et al. "Metacells untangle large and complex single-cell transcriptome networks." BMC Bioinformatics (2022)

MCBRLab bilibili

mcRigor enhances gene co-expression analysis

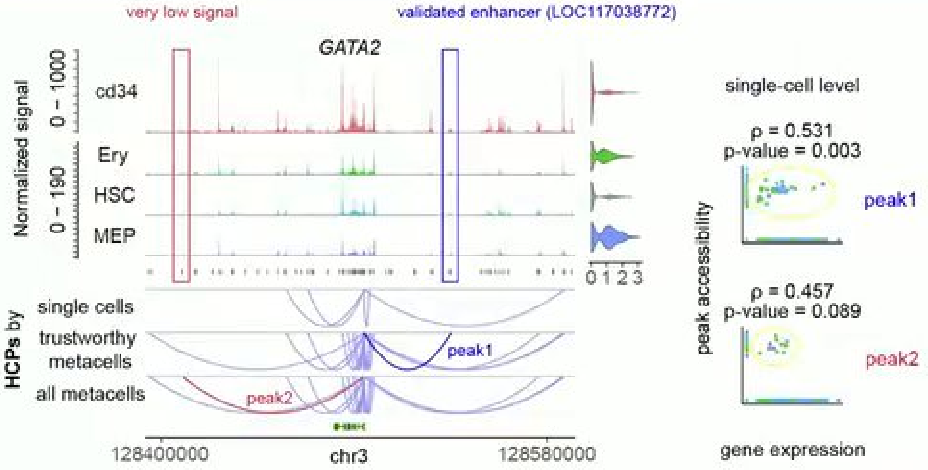


SuperCell: *Bilous, M., et al. "Metacells untangle large and complex single-cell transcriptome networks." BMC Bioinformatics (2022)*

Data: *Wilk, A. J., et al. "A single-cell atlas of the peripheral immune response in patients with severe covid-19." Nat Med (2020)*

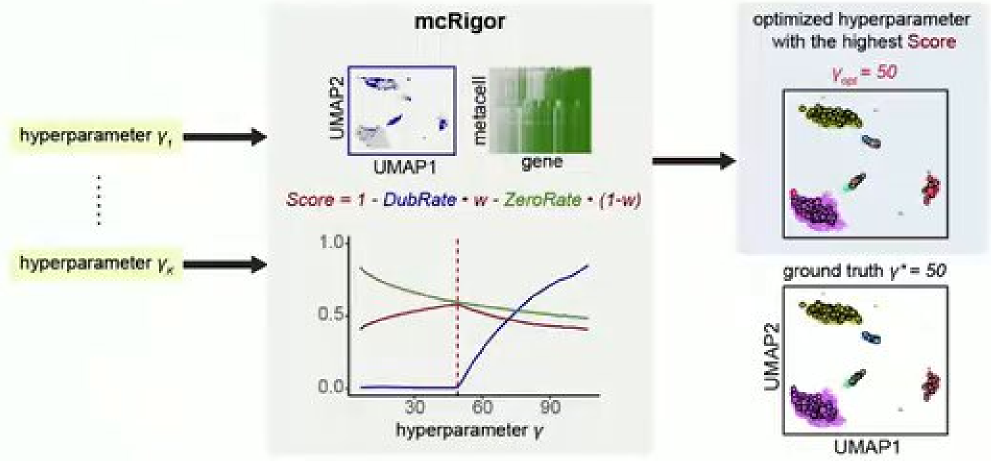
MCBRLab bilibili

mcRigor improves the reliability of gene regulatory inference

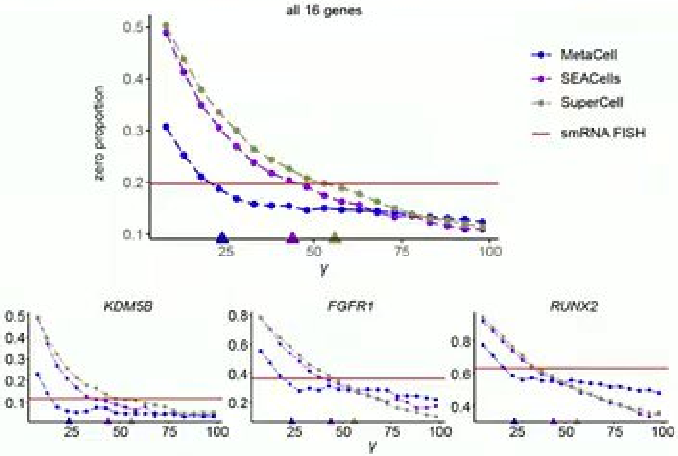


SEACells: *Persad, S., et al. "SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data." Nat Biotech (2023).*

mcRigor optimizes hyperparameter γ by balancing sparsity and dubiousness



mcRigor helps distinguish biological zeros from technical zeros

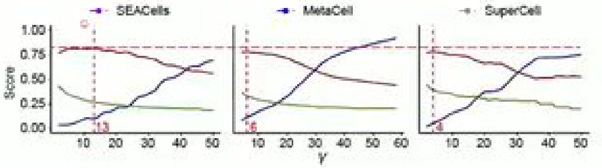


Data: *Torre, Eduardo, et al. "Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH." Cell systems (2018).*

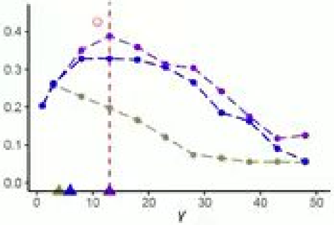
mcRigor selects optimal metacell method and γ for DEG detection

Optimal method-hyparameter configuration:

SEACells with γ = 13



F score
for DE genes derived from different metacell partitions (q-val<0.05)
v.s.
DE genes derived from bulk RNA data (q-val<0.05)



Data: *Chu, L.-F. et al. "Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm." Genome Biol. (2016).*

Q: How could dubious metacells be handled appropriately?

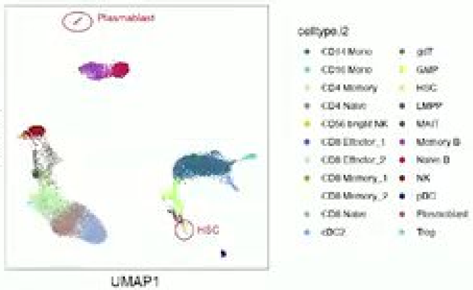
mcRigor two-step: an extension of mcRigor

Step 1: A method–hyperparameter configuration (i.e., a metacell partitioning method with a granularity level γ_1) is either specified by the user or selected by mcRigor. This configuration is then applied to partition single cells into metacells. If mcRigor detects dubious metacells within the partition, it is re-applied to the same partition using a lower divergence score threshold (as below) to label more metacells as dubious.

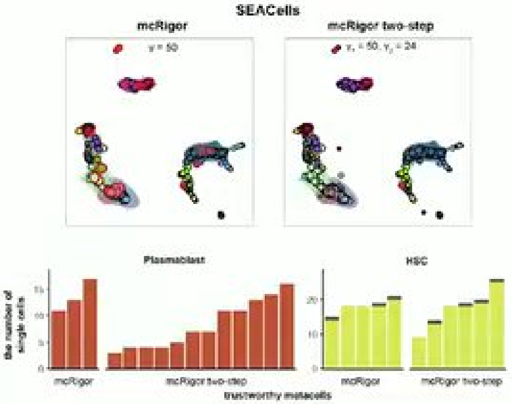
$$\theta(m_k) = q_{0.85}\left(\left\{\mathrm{mcDiv}_{k'}^{\mathrm{null}} : m_{k'} \in [m_k - h, m_k + h],\ k' = 1, \ldots, M\right\}\right)$$

Step 2: The selected metacell partitioning method is re-applied to the subset of single cells that belong to the metacells now marked as dubious. This yields a refined metacell partition under a new granularity level $\gamma_2 < \gamma_1$, which can be selected by mcRigor from the candidate set of granularity levels $2, \ldots, \gamma_1 - 1$.

mcRigor two-step effectively resolves rare cell types



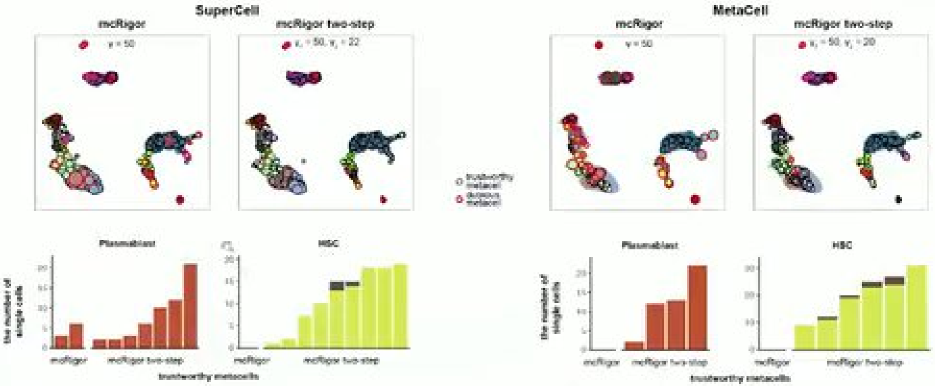
13408 total cells
106 Plasmablast cells (0.8%), 144 HSC cells (1.1%)



SEACells: *Persad, S., et al. "SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data." Nat Biotech (2023).*
Data: *Stuart, T., et al. "Comprehensive integration of single-cell data." Cell (2019).*

MCBRLab bilibili

mcRigor two-step effectively resolves rare cell types



MetaCell *Baran, Y., et al. "MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions." Genome Biology (2019).*
SuperCell: *Bilous, M., et al. "Metacells untangle large and complex single-cell transcriptome networks." BMC Bioinformatics (2022).*
Data: *Stuart, T., et al. "Comprehensive integration of single-cell data." Cell (2019).*

Conclusion

A: We give a statisical definition of "metacell" based on the two-layer model.

A: We detect dubious metacells using per-metacell mcDiv statistics and null contructed through double permutation.

A: We optimize metacell partitioning by balancing sparsity and dubiousness.

Article | Open access | Published: 29 September 2025

mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu & Jingyi Jessica Li

Nature Communications **16**, Article number: 8602 (2025) | Cite this article

R package and tutorial on Github

mcRigor 1.0 Reference Articles ▾

Functionality 1: detect dubious metacells for a given metacell partition
Functionality 2: optimize metacell partitioning
Implementing metacell partitioning methods
Extension: mcRigor two-step

mcRigor

The R package **mcRigor** is a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis. It can be used as an add-on for any existing metacell partitioning methods for obtaining more reliable metacells. mcRigor has two main functionalities: 1) detecting dubious metacells, which are composed of heterogeneous single cells, for a given metacell partition, and 2) optimizing the hyperparameter of a metacell partitioning method. The core of mcRigor is a feature-correlation-based statistic that measures the heterogeneity of a metacell, with its null distribution derived from a double permutation scheme. The following figure illustrates the schematics of mcRigor for dubious metacell detection (a) and hyperparameter optimization (b).

a

License
MIT + file LICENSE

Citation
Citing mcRigor

Developers
Pan Liu
Maintainer

R package: https://github.com/JSB-UCLA/mcRigor **Tutorial:** https://jsb-ucla.github.io/mcRigor

Acknowledgements

Thank Prof. Jingyi Jessica Li and
all members of the Junction of Statistics and Biology (JSB) lab!!!

Grants:

NIH NIGMS

CHAN ZUCKERBERG INITIATIVE

BROAD INSTITUTE