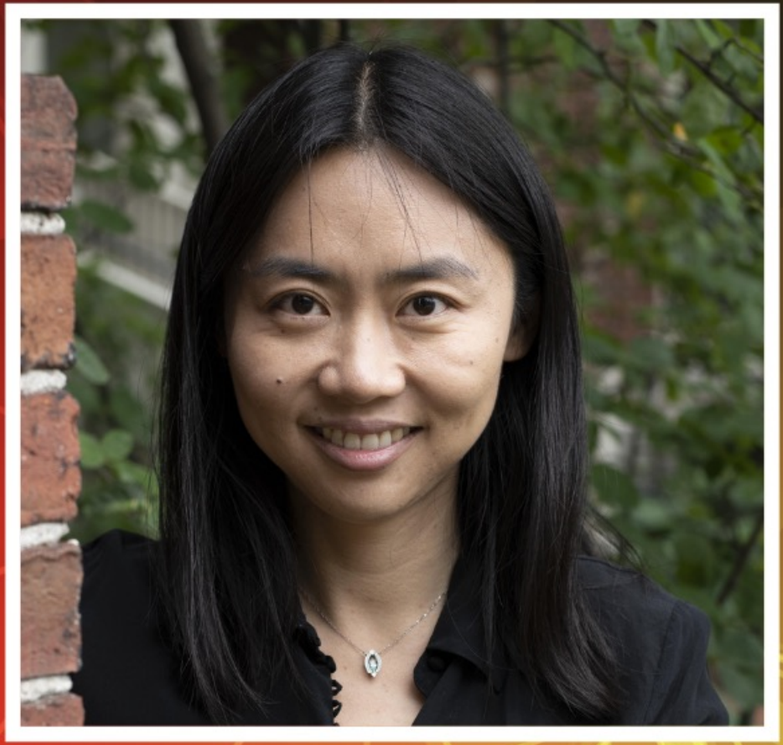


# Jingyi Jessica Li

*Professor of Statistics and Data Science*
*University of California, Los Angeles*
*Helen Putnam Fellow, 2022-23*
*Radcliffe Institute for Advanced Study at Harvard University*
*United States*

## ISCB OVERTON PRIZE KEYNOTE

## TUESDAY 8:45AM

# Using Synthetic Null Data to Enhance Statistical Rigor in Genomics

**Jingyi Jessica Li (李婧翌)**

Professor of Statistics and Data Science, Biostatistics, Computational Medicine, and Human Genetics

University of California, Los Angeles

Junction of Statistics and Biology

UCLA

# Using Synthetic Null Data to Enhance Statistical Rigor in Genomics

**Jingyi Jessica Li (李婧翌)**

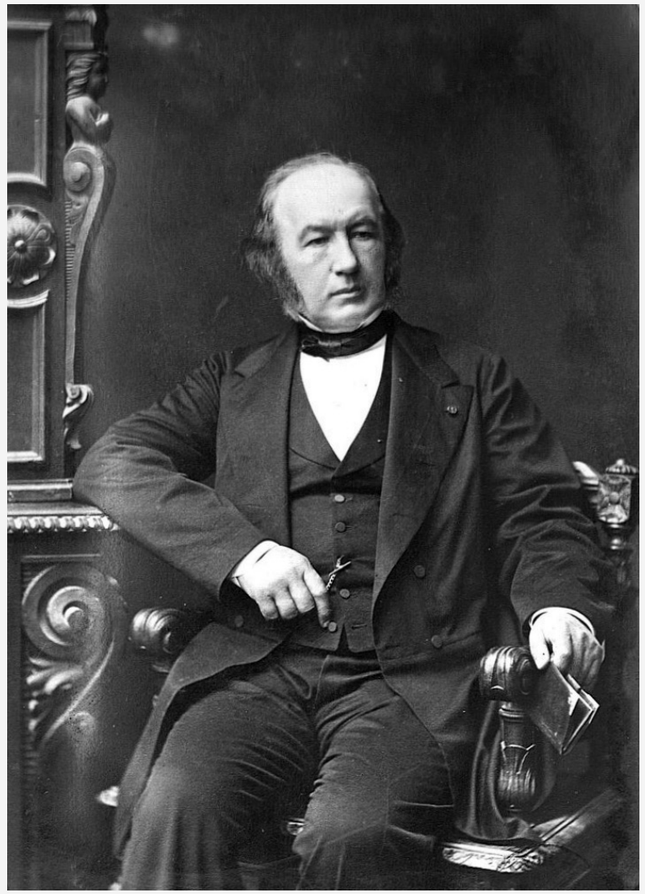
Professor of Statistics and Data Science, Biostatistics, Computational Medicine, and Human Genetics

University of California, Los Angeles

Junction of Statistics and Biology

UCLA

# Negative control in medicine

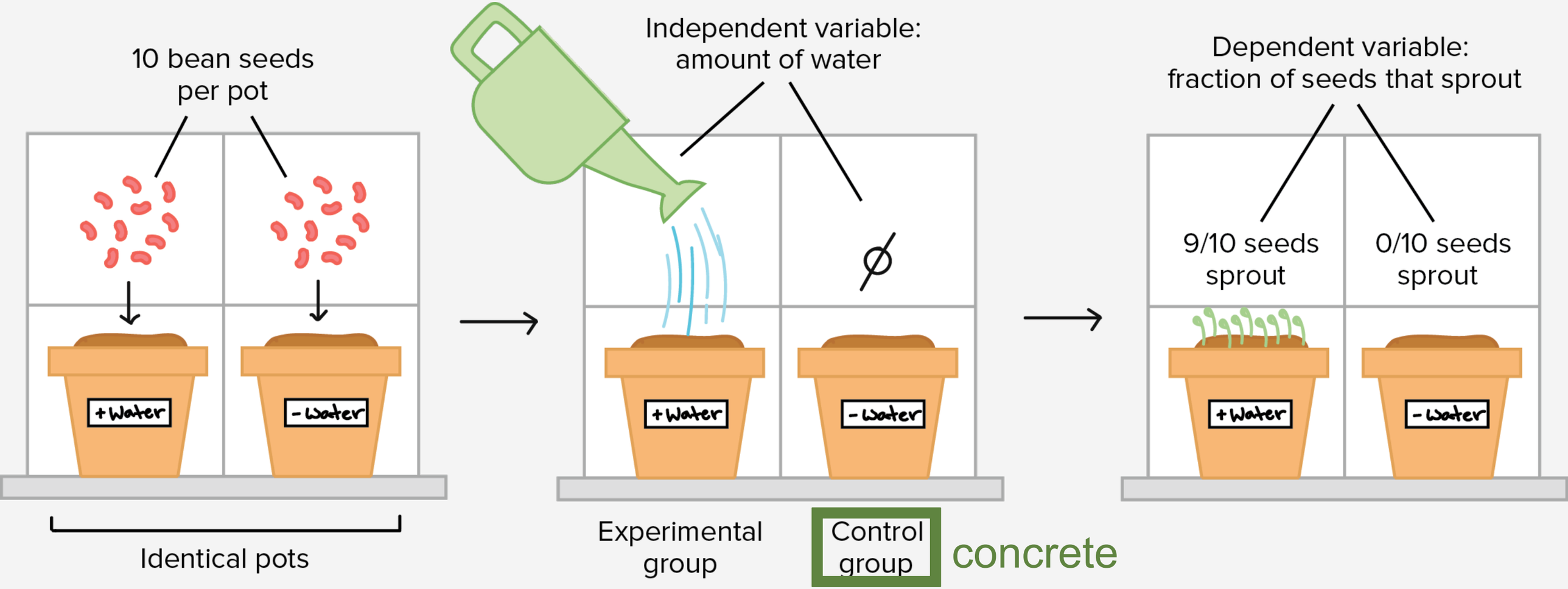


Claude Bernard
(1813 – 1878)

Université Claude Bernard Lyon 1

Source: https://fr.wikipedia.org/wiki/Claude_Bernard

# Negative control in biological experiments



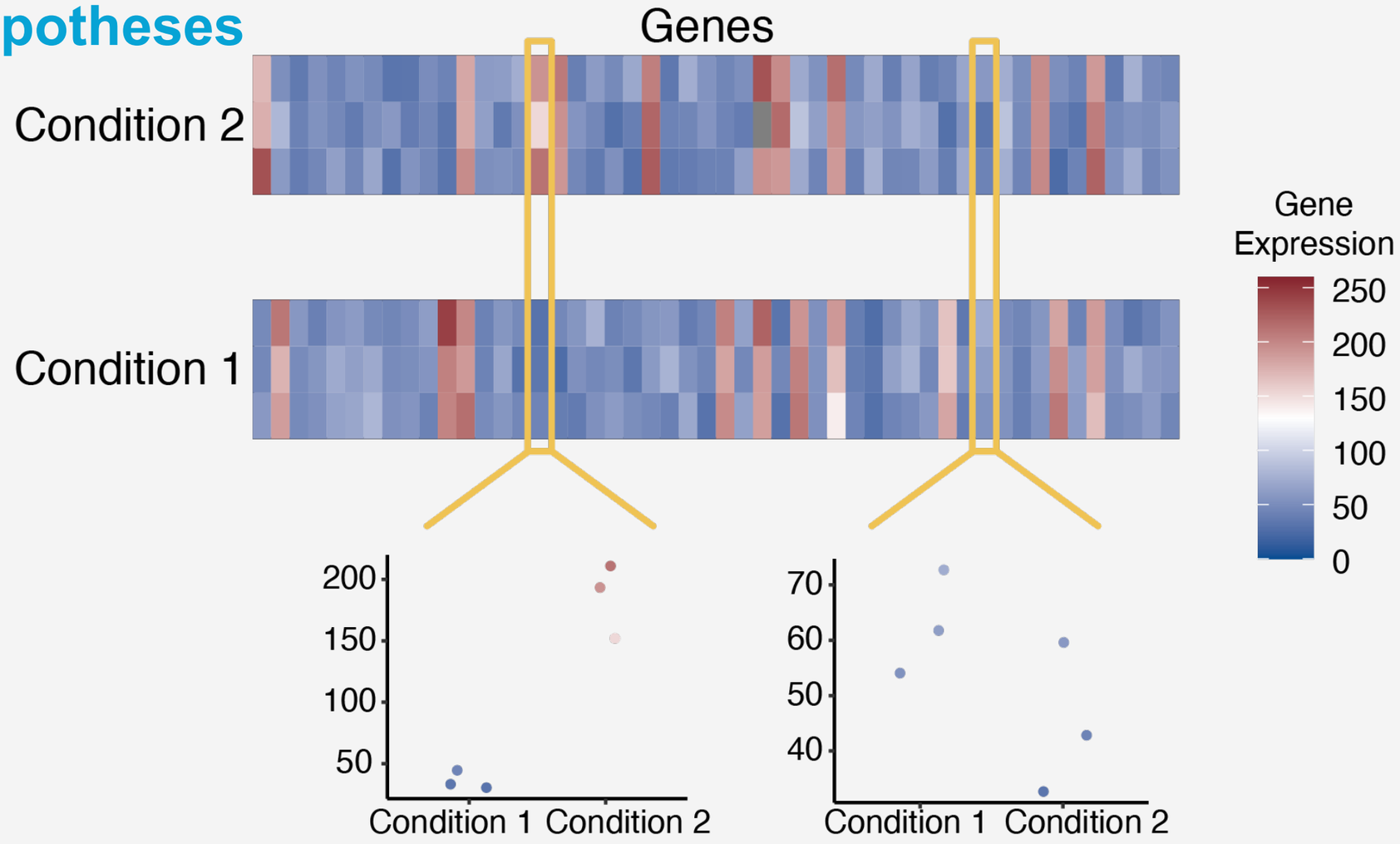

concrete

Source: https://www.khanacademy.org/science/biology/intro-to-biology/science-of-biology/a/experiments-and-observations

# “Negative control” in genomic data analysis

**Q: Where is the negative control?**

**A: Null hypotheses**




**One hypothesis test per gene: reject null hypothesis → DE gene**

# Null hypothesis in statistical hypothesis testing

abstract

A null hypothesis is a type of conjecture used in statistics that proposes that there is no difference between certain characteristics of a population or data-generating process.

Source: https://www.awesomefintech.com/term/null_hypothesis/

# Since null hypothesis is abstract, it is often misunderstood and misused

**Questions I will discuss in this talk**

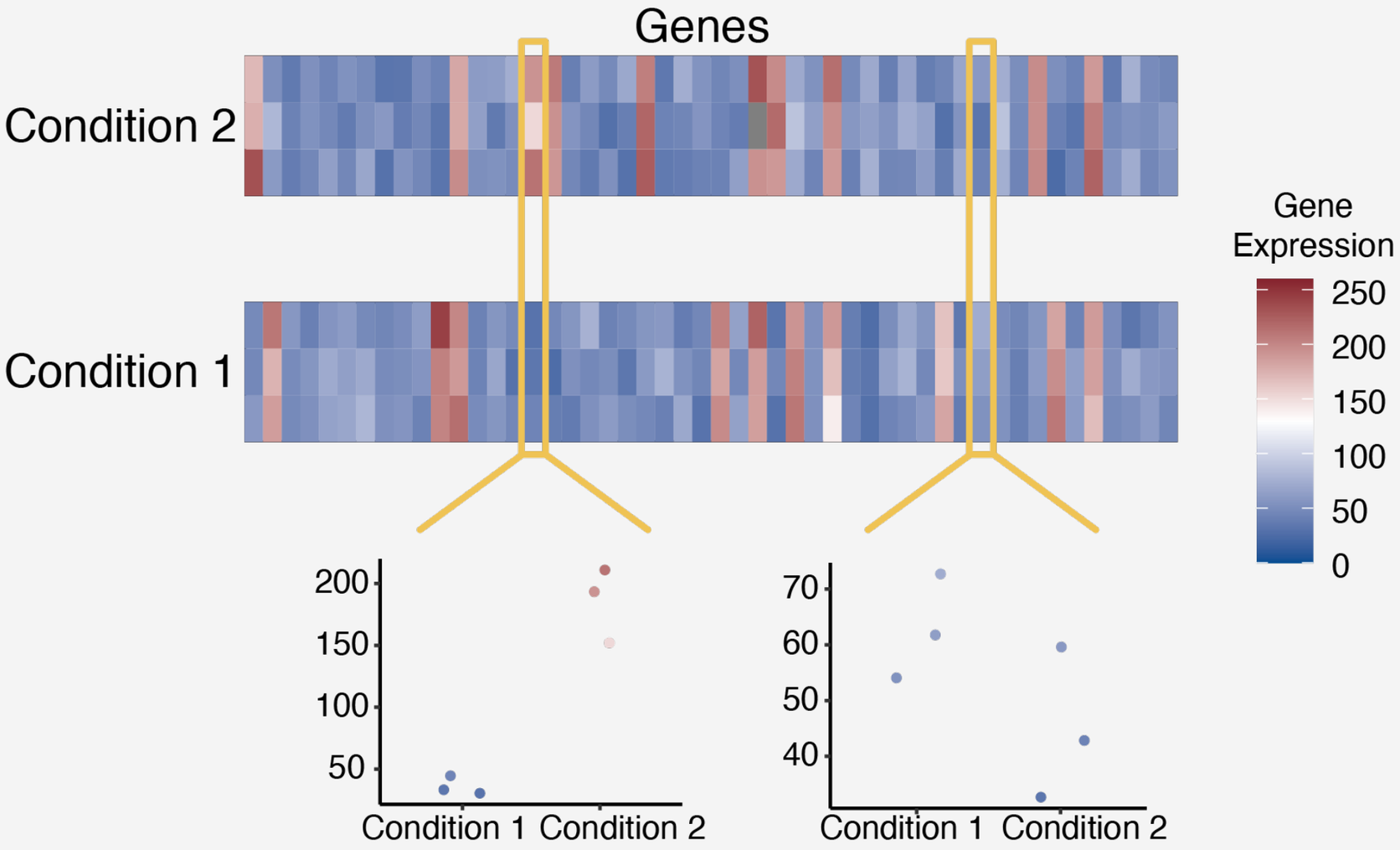
1. What is an appropriate null hypothesis?
   - Different null hypotheses → different discoveries/conclusions
2. How to make an abstract null hypothesis concrete?
   - Synthetic null data
3. How to use synthetic null data to reduce false discoveries?
   - Contrastive strategy

# Question 1

## What is an appropriate null hypothesis?

# Example 1: bulk RNA-seq DE analysis

## Q: What genes are differentially expressed (DE) between two conditions?




**One hypothesis test per gene: reject null hypothesis → DE gene**

# Example 1: bulk RNA-seq DE analysis

Popular methods (originally designed for **small** sample sizes):

- **edgeR** [Robinson et al., *Bioinformatics*, 2010]; cited > 31K times
- **DESeq2** [Love et al., *Genome Biol*, 2014]; cited > 52K times

# Example 1: bulk RNA-seq DE analysis

Popular methods (originally designed for **small** sample sizes):

- **edgeR** [Robinson et al., *Bioinformatics*, 2010]; cited > 31K times
- **DESeq2** [Love et al., *Genome Biol*, 2014]; cited > 52K times

Both assume a **negative binomial (NB)** distribution per gene and condition

For each gene,

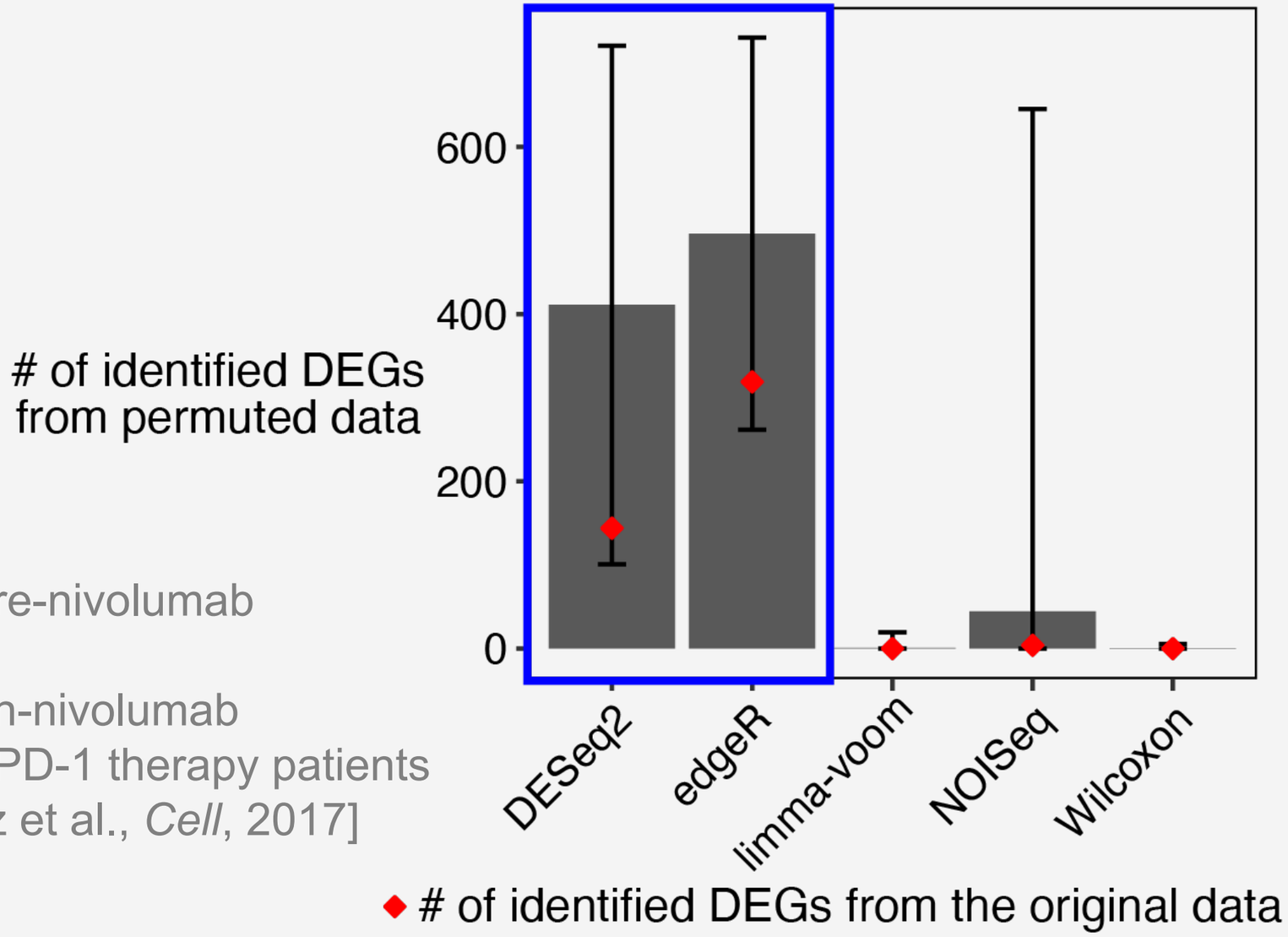
- Condition 1: $X_i \overset{\text{ind}}{\sim} \mathrm{NB}(\mu_1 s_i, \sigma_1),\ i = 1, \ldots, n$
- Condition 2: $Y_j \overset{\text{ind}}{\sim} \mathrm{NB}(\mu_2 s_j, \sigma_2),\ j = 1, \ldots, m$

**Null hypothesis** $H_0 : \mu_1 = \mu_2$

which is appropriate only if the NB assumption is reasonable

# Example 1: 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]

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

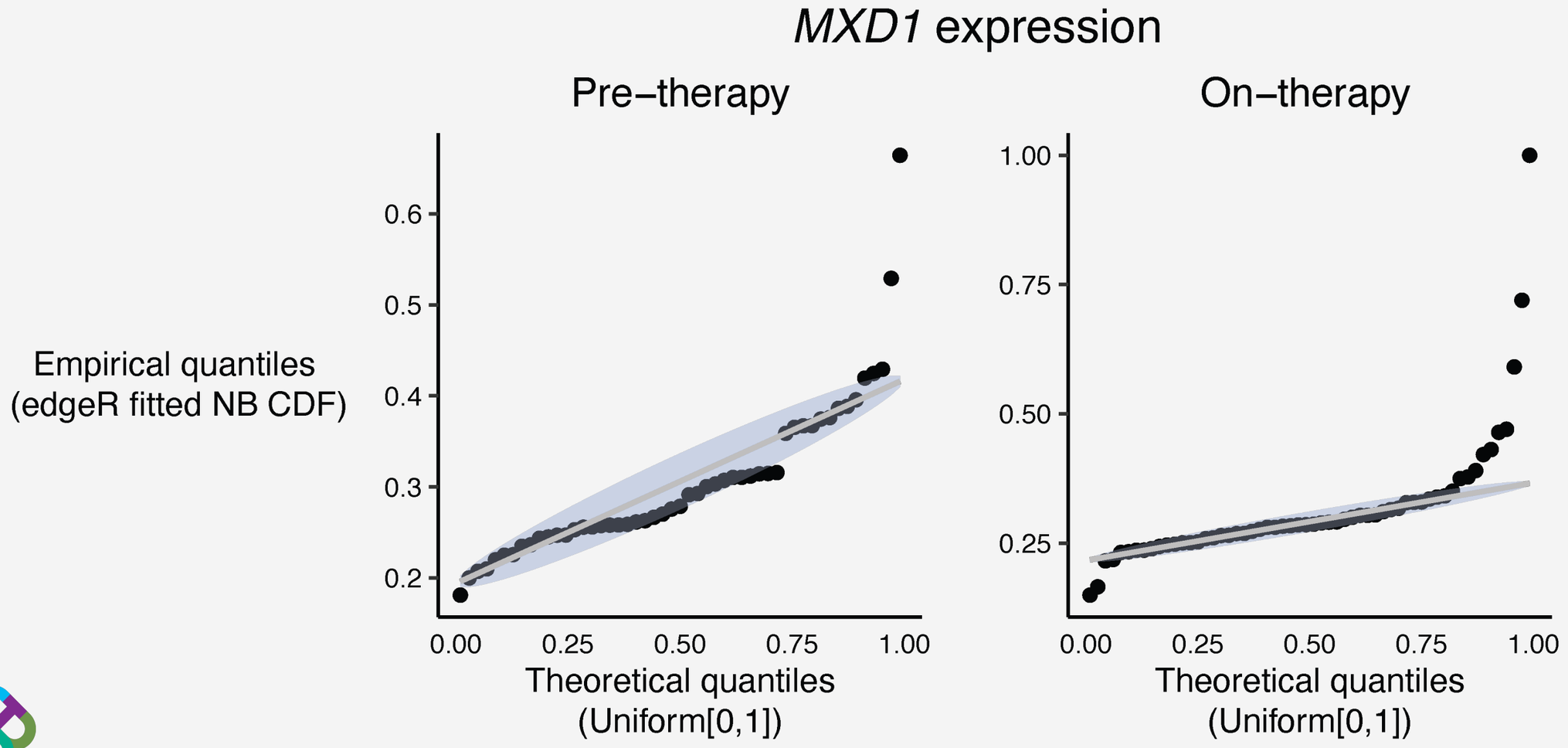
Yumei Li
(Wei Li Lab)

Xinzhou Ge
(JSB)

# Example 1: bulk RNA-seq DE analysis

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

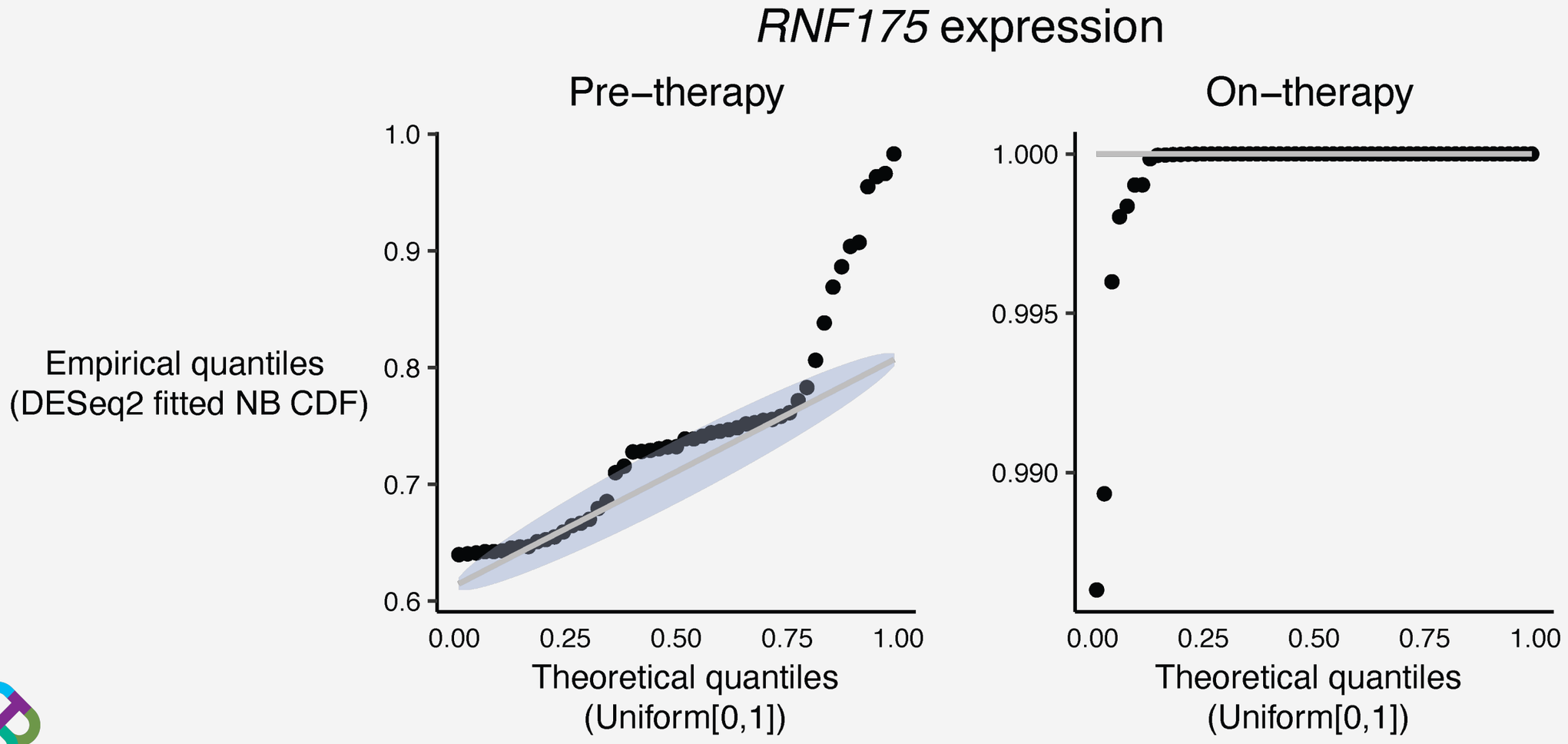
**A: The NB assumption does not hold on this dataset.**



# Example 1: bulk RNA-seq DE analysis

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

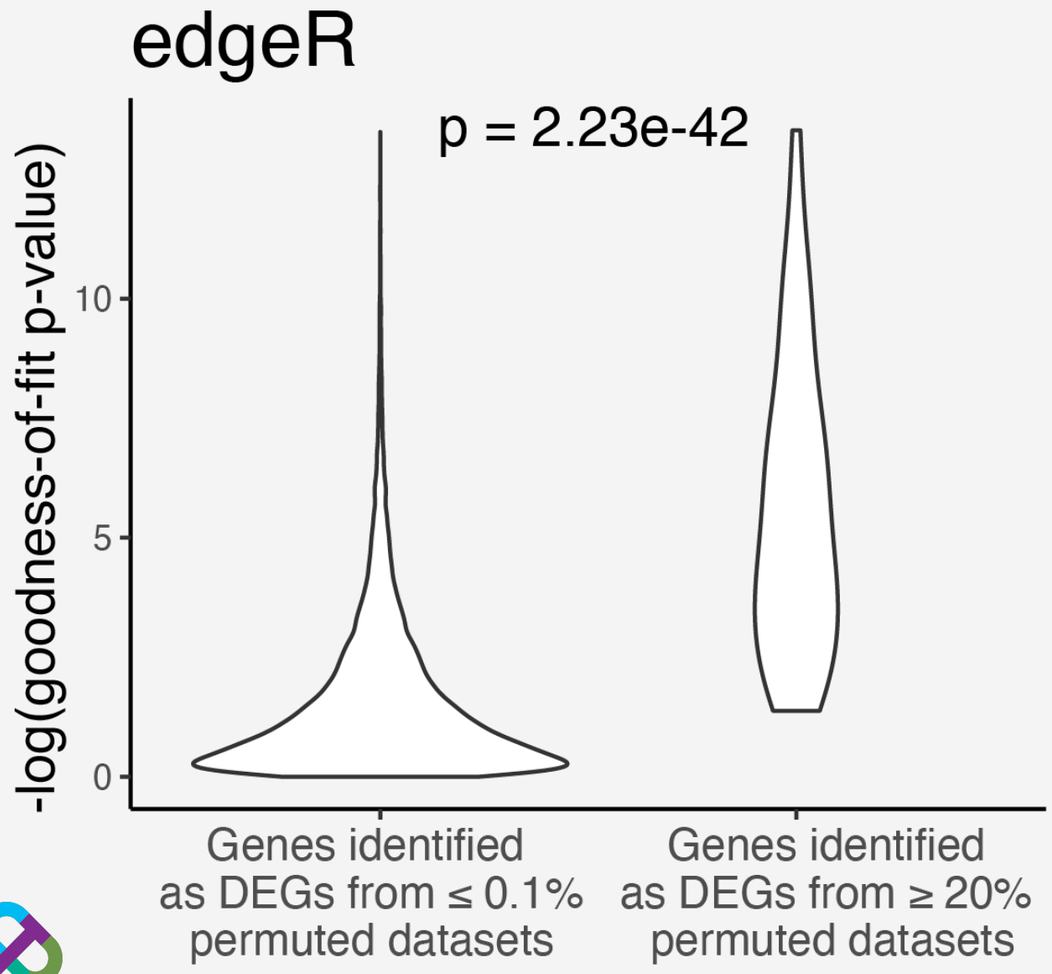
**A: The NB assumption does not hold on this dataset.**

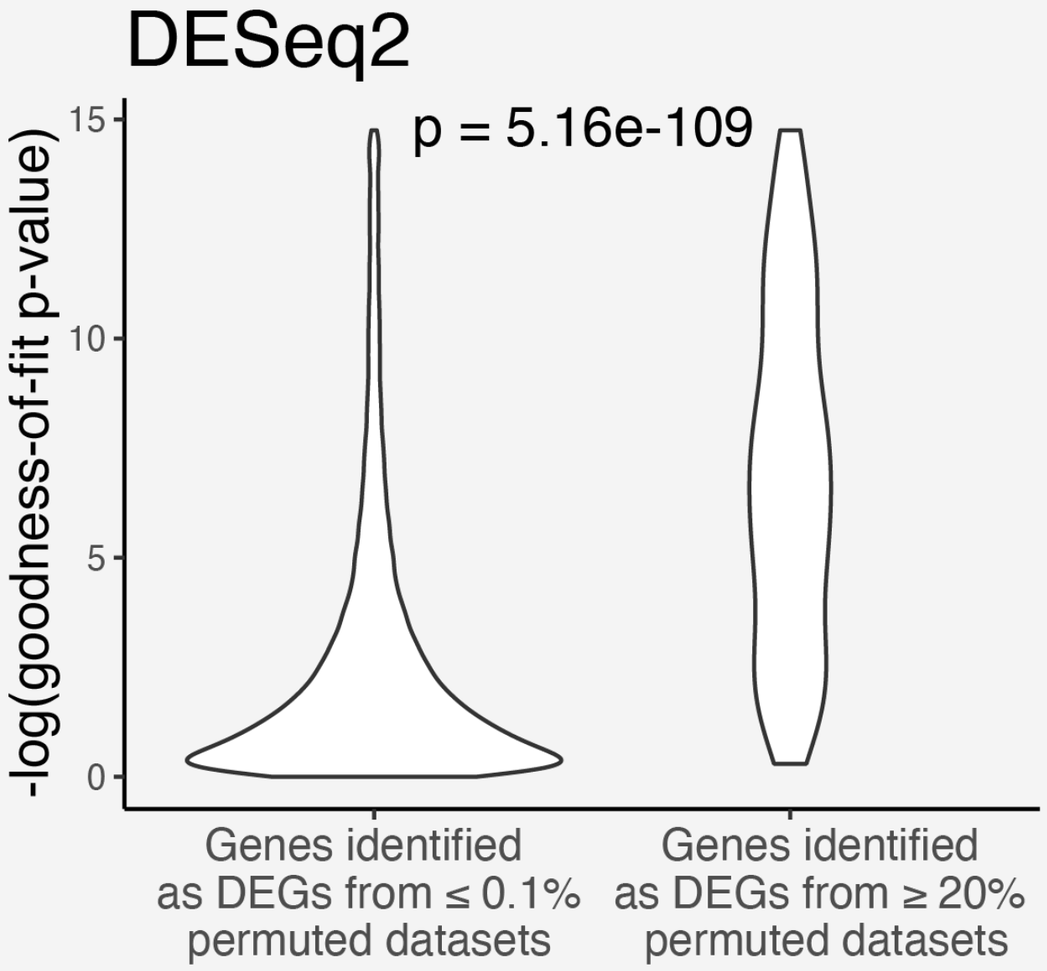


# Example 1: bulk RNA-seq DE analysis

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

**A: The NB assumption does not hold on this dataset.**

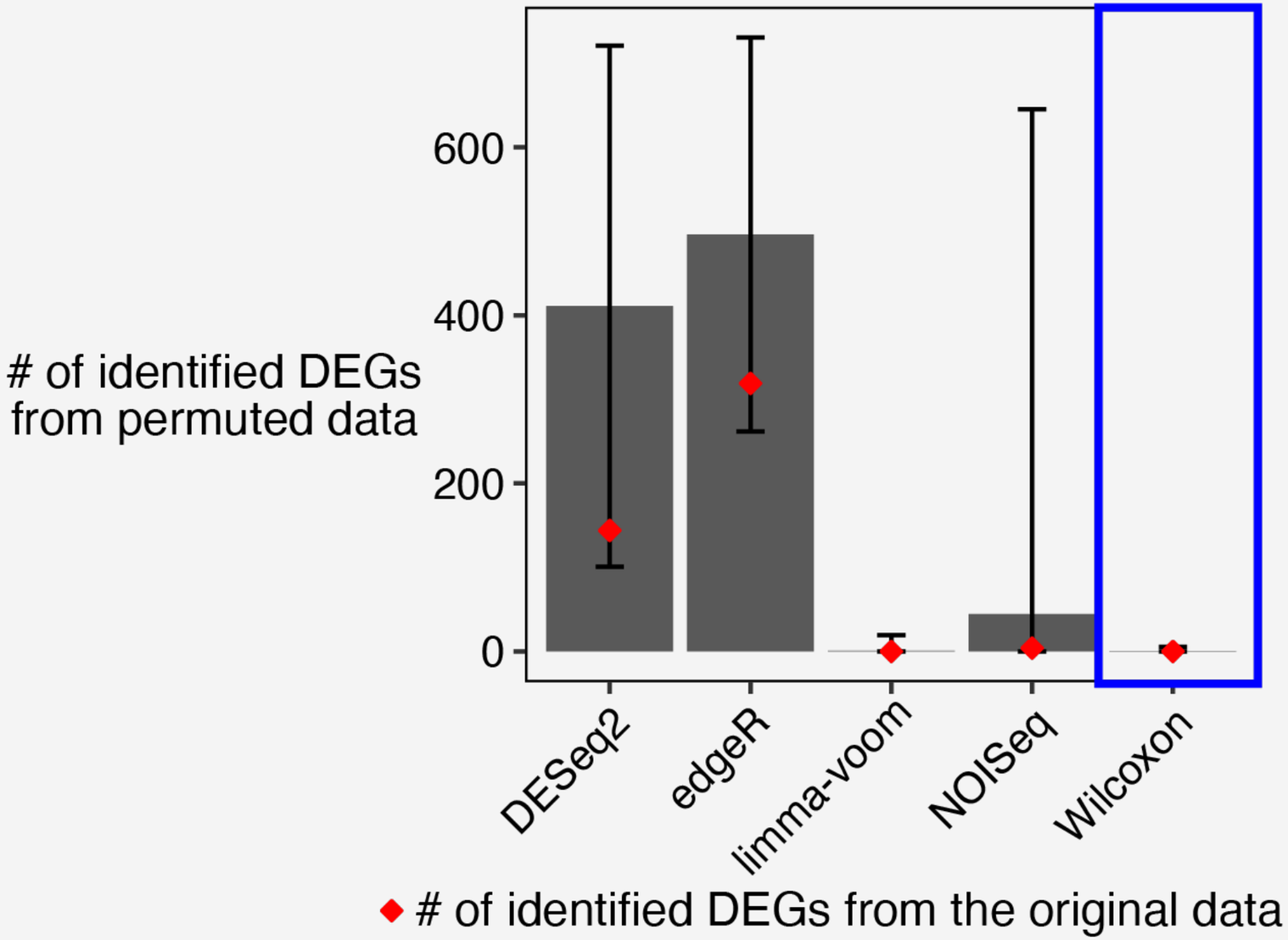



Chenxin Jiang
(CUHK→JSB)

# Example 1: bulk RNA-seq DE analysis

## Q: Why does Wilcoxon NOT identify DE genes from permuted data?



# Example 1: bulk RNA-seq DE analysis

**Q: Why does Wilcoxon NOT identify DE genes from permuted data?**

**A: It has a different null hypothesis.**

For each gene, the normalized counts

Condition 1: $\widetilde{X}_i,\ i = 1, \ldots, n$

Condition 2: $\widetilde{Y}_j,\ j = 1, \ldots, m$

**Null hypothesis (approximate, ignoring ties):**

$$\boldsymbol{H_0} : \mathbb{P}(\widetilde{X}_i > \widetilde{Y}_j) = 0.5,\ \text{for all } i, j$$

which does NOT have the **NB assumption**

# Example 1: bulk RNA-seq DE analysis

Genome Biology

Home About Articles Submission Guidelines 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



Yumei Li
(Wei Li Lab)

Xinzhou Ge
(JSB)

Prof. Wei Li
(UC Irvine)

**Twitter: @jsb_ucla**

**Which null hypothesis is more appropriate?**

**Too abstract a question?**

**Intuition: DE genes found from permuted data are not trustworthy**

**Then what does permutation provide?**

**Synthetic null (*in silico* negative control)**

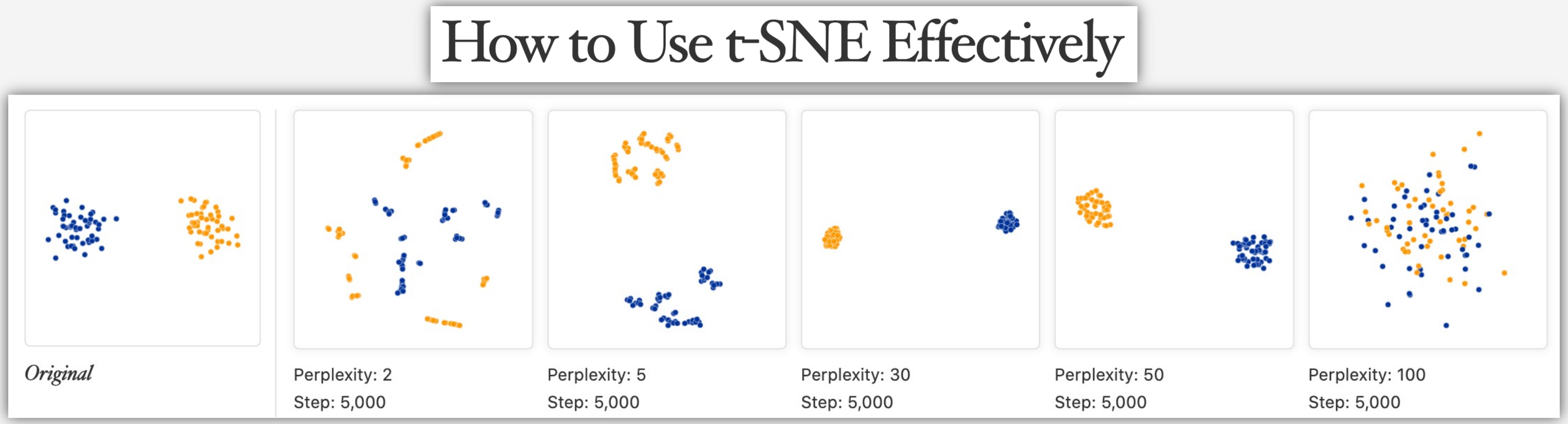
# Question 2:

## How to make an abstract null hypothesis concrete?

## Synthetic null data

# Another example where permutation helps

# Example 2: dubious t-SNE/UMAP embeddings?




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

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

# Example 2: dubious t-SNE/UMAP embeddings?

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

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

**scDEED: a statistical method for detecting dubious 2D single-cell embeddings**

Lucy Xia, Christy Lee, Jingyi Jessica Li

**doi:** https://doi.org/10.1101/2023.04.21.537839

Under revision at *Nat Comms*

Tuesday 12:10 pm in Salle Rhone 1
(BioVis COSI)

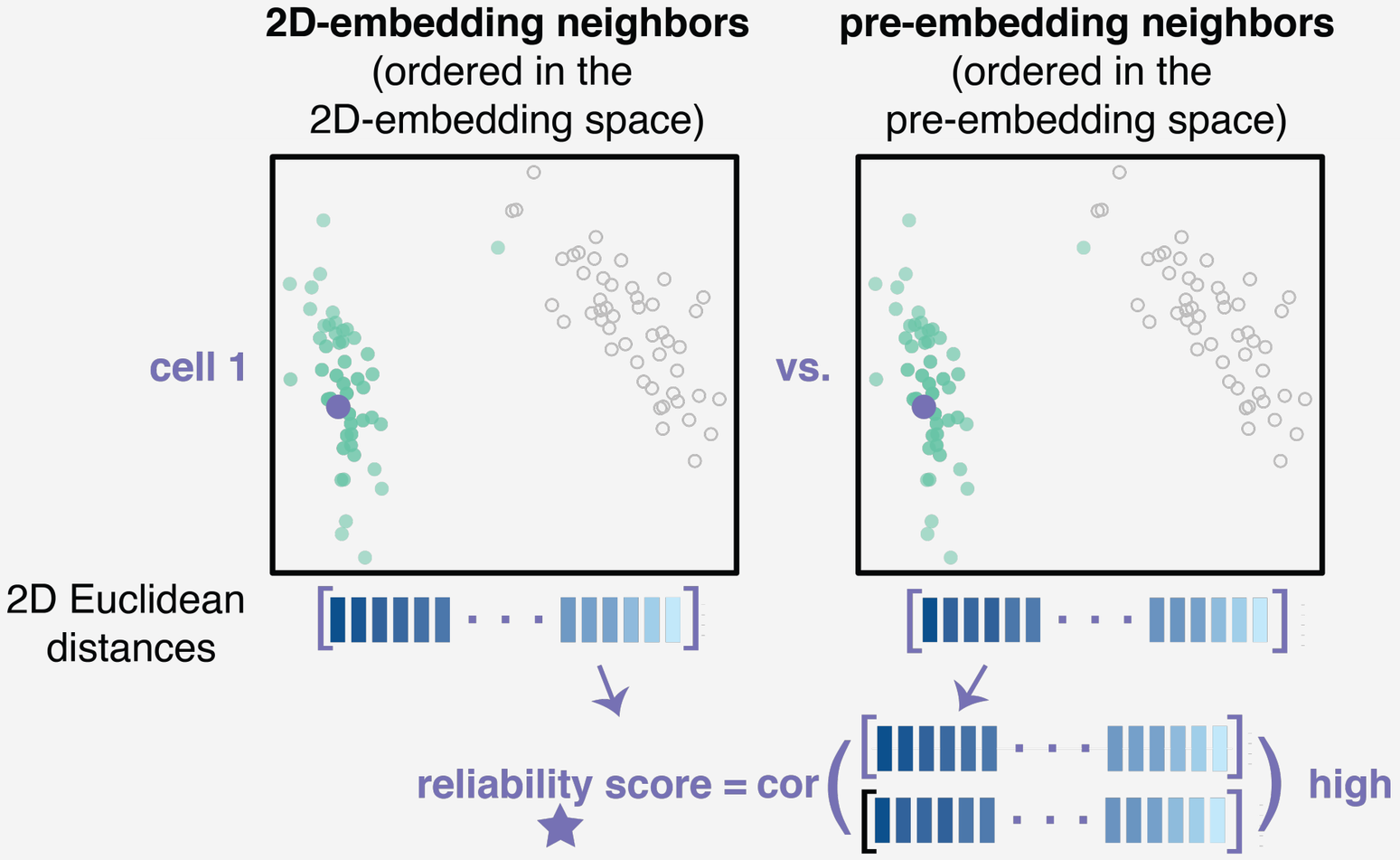
Lucy Xia
(HKUST)

Christy Lee
(JSB)

# Example 2: dubious t-SNE/UMAP embeddings?

## scDEED intuition

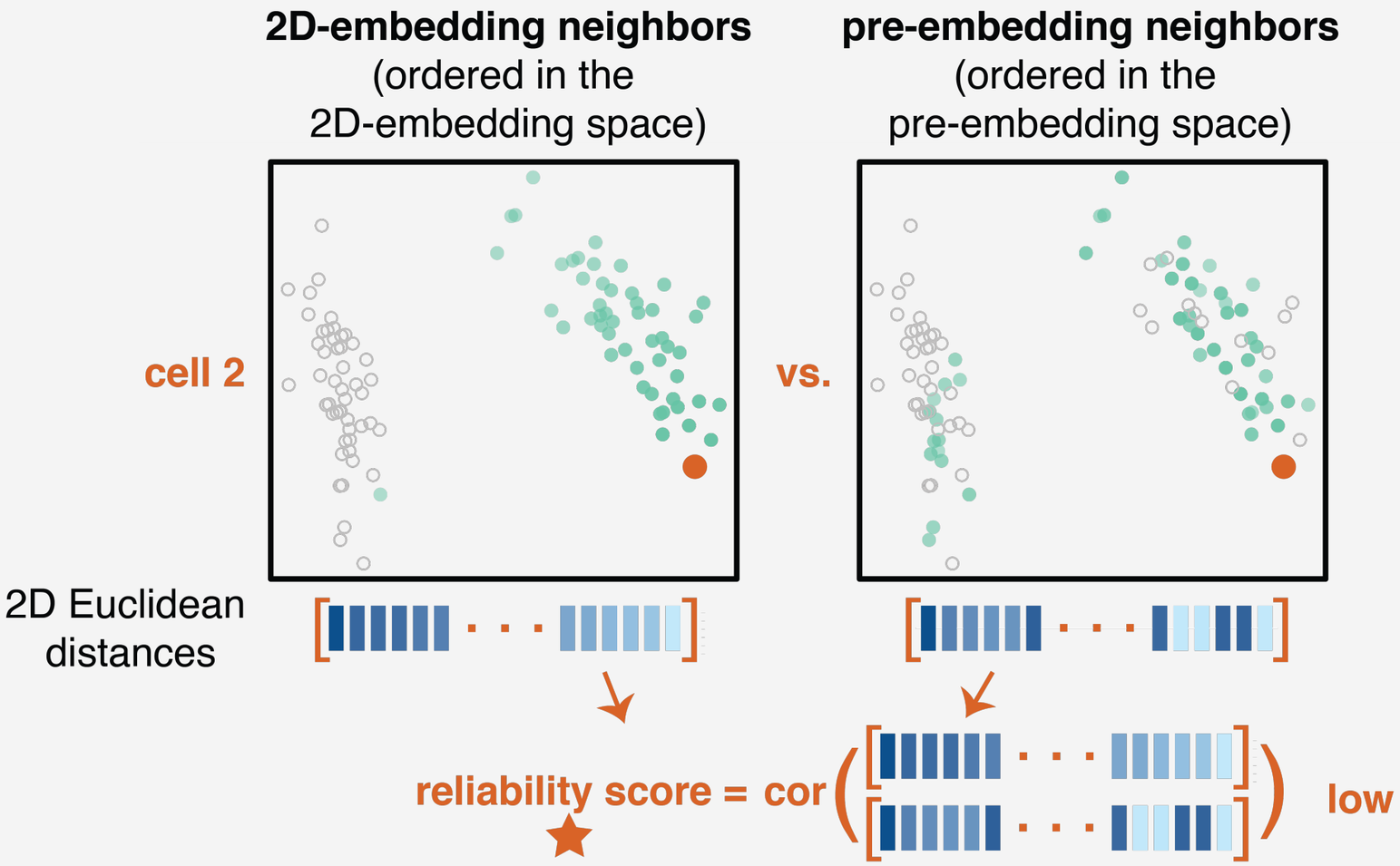
### A trustworthy cell embedding



# Example 2: dubious t-SNE/UMAP embeddings?

## scDEED intuition

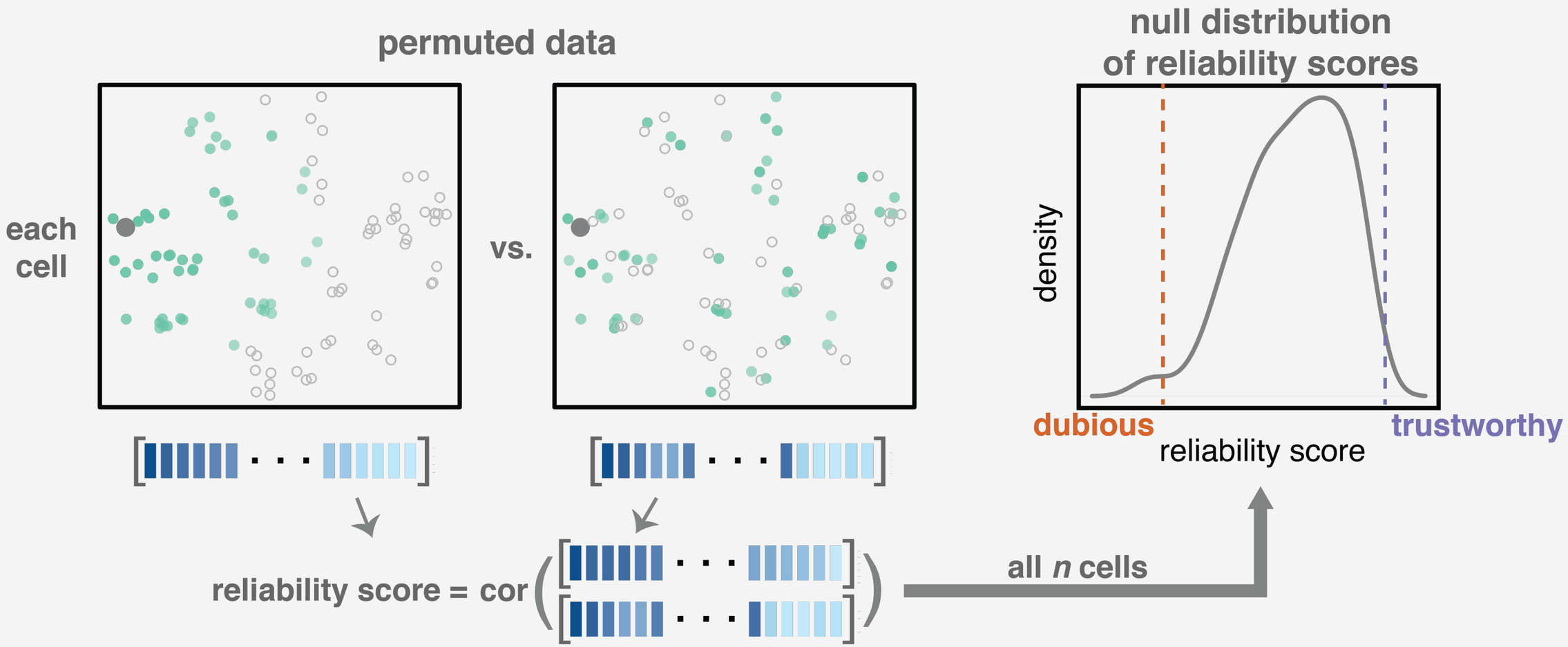
### A dubious cell embedding



# Example 2: dubious t-SNE/UMAP embeddings?

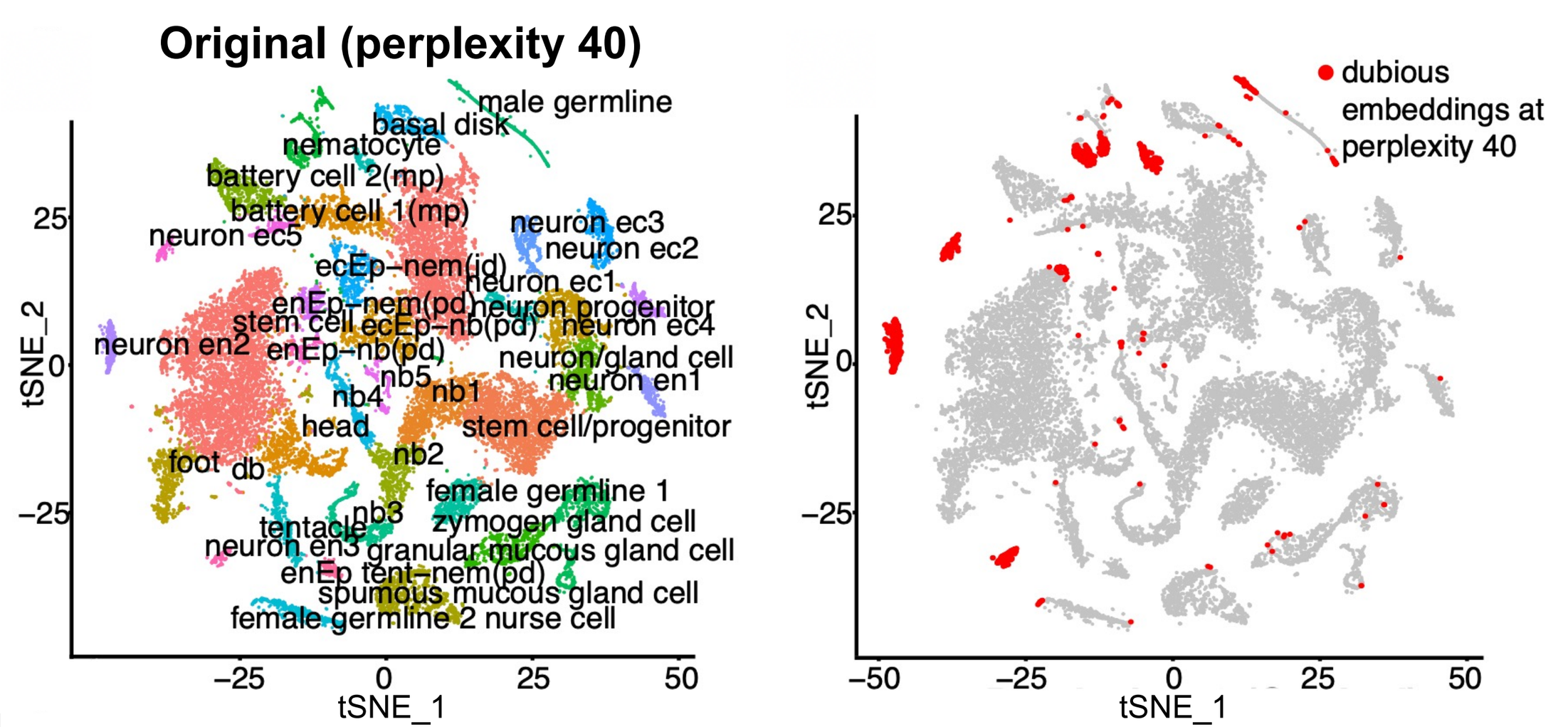
**Q: What is the null hypothesis?**

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



# Example 2: dubious t-SNE/UMAP embeddings?

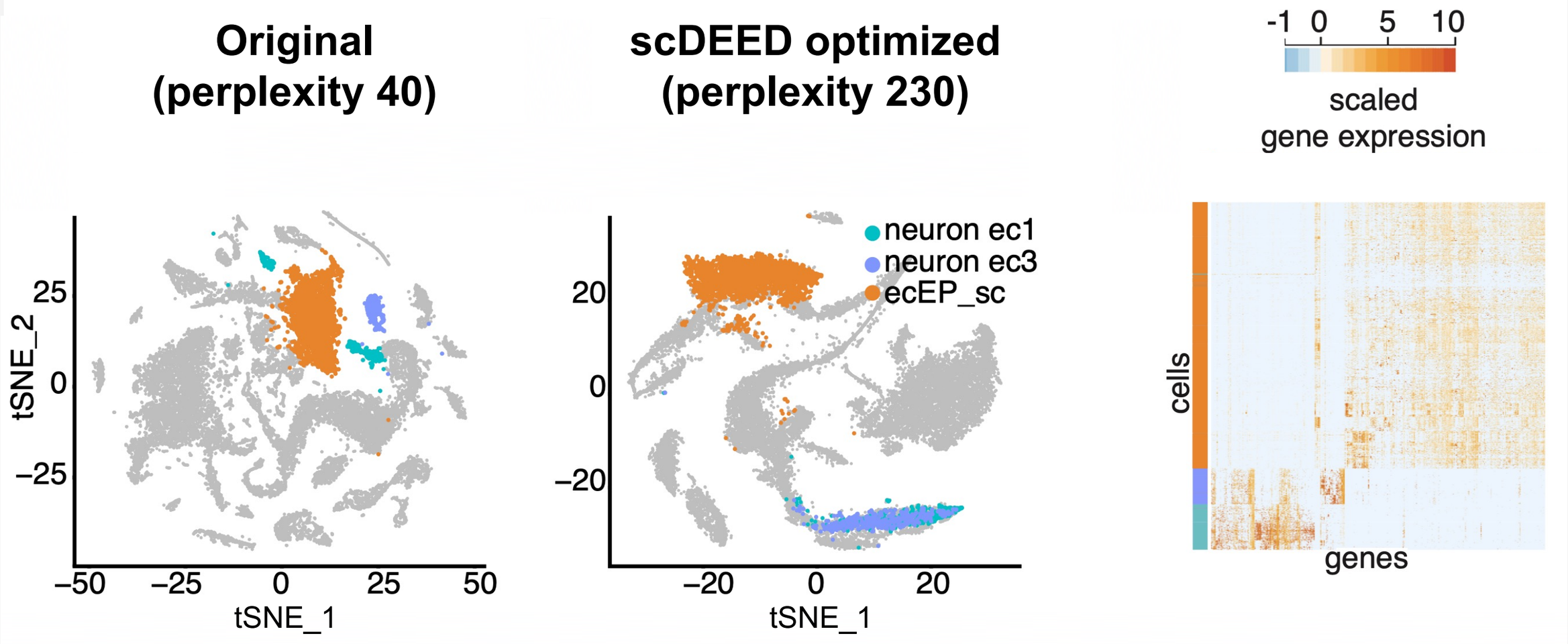
## scDEED detects dubious embeddings




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

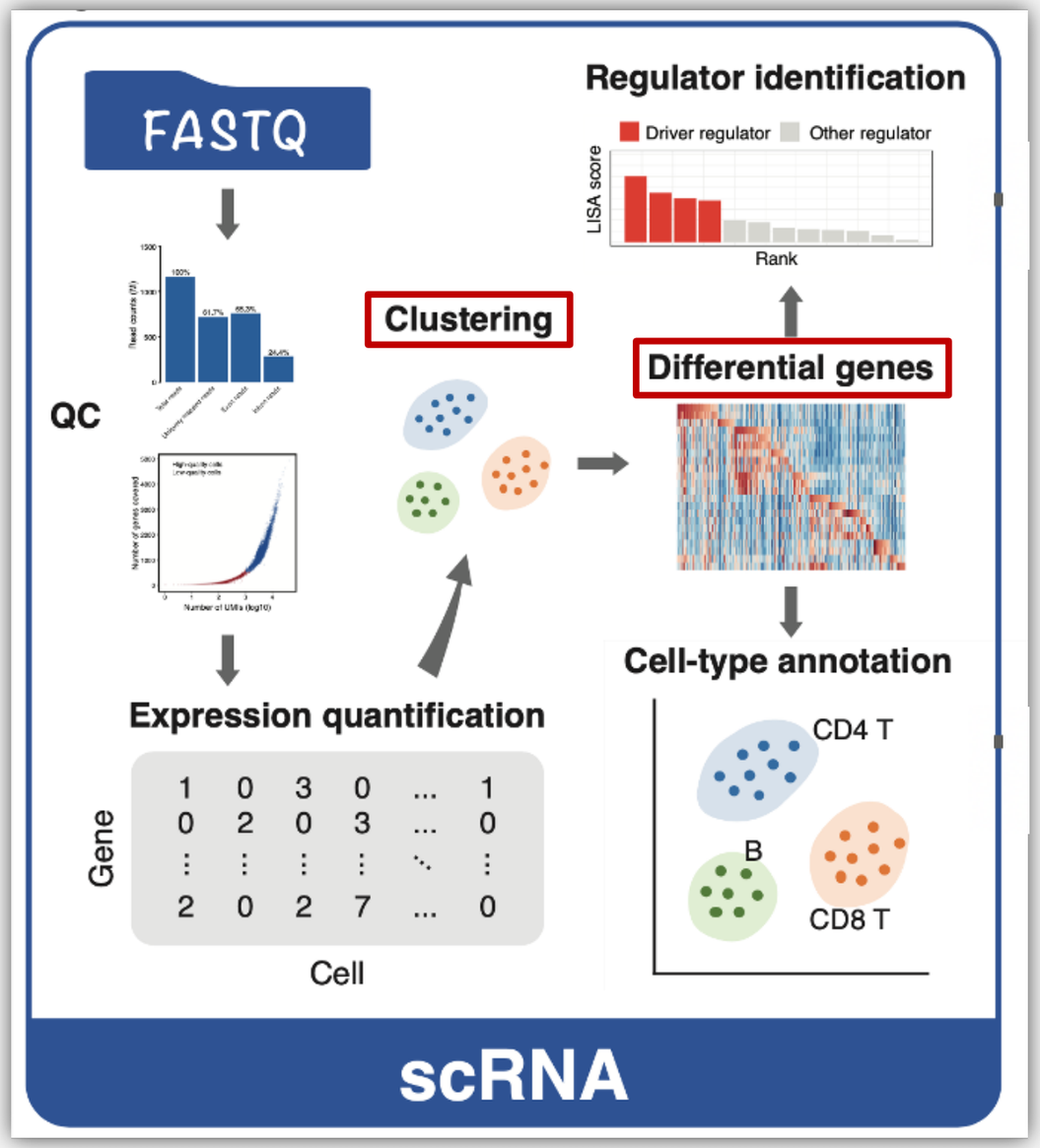
# Example 2: dubious t-SNE/UMAP embeddings?

**scDEED** optimizes hyperparameters by minimizing # of dubious embeddings



# Synthetic null generation beyond permutation

# Example 3: single-cell post-clustering DE analysis

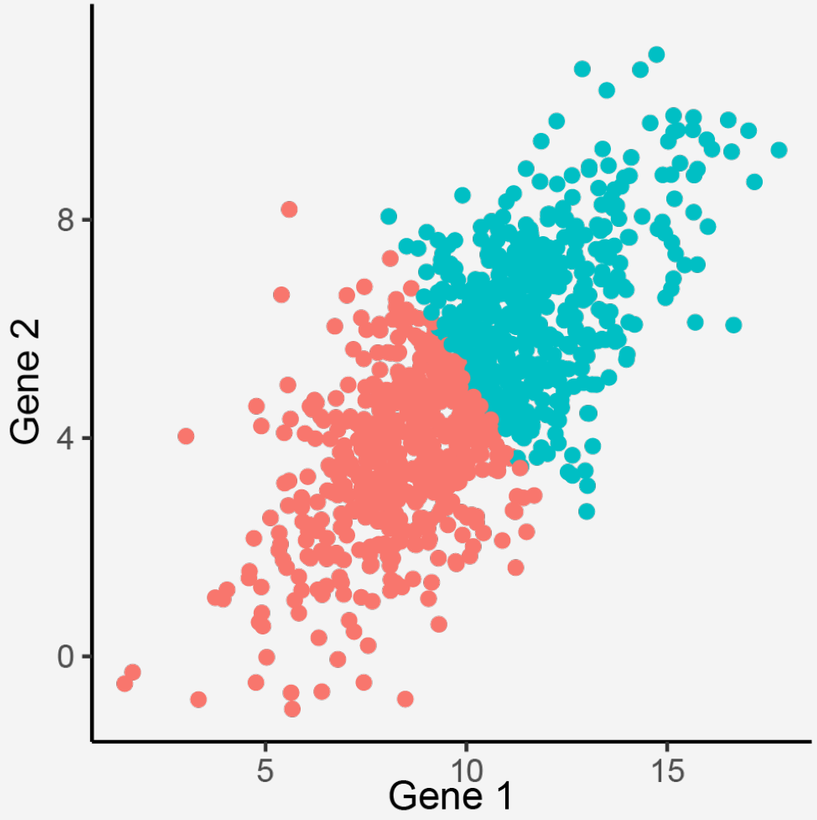


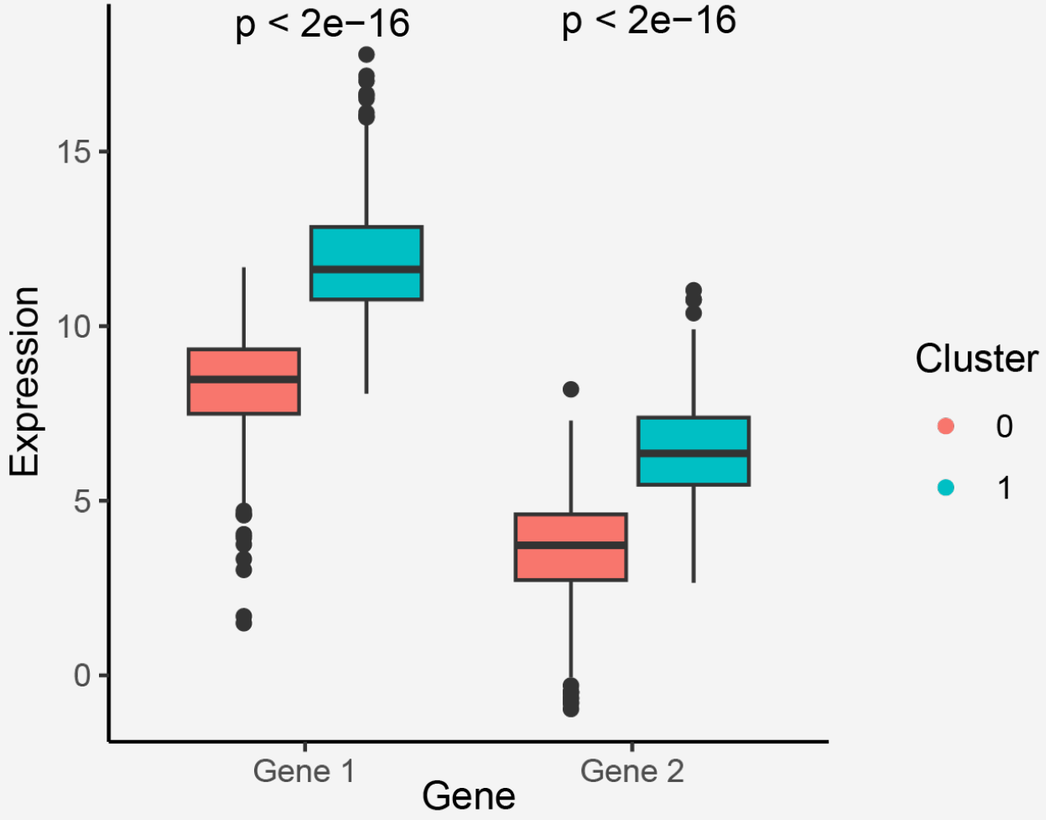

Source: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02116-x

# Example 3: single-cell post-clustering DE analysis

**Double dipping**: same data used twice

1. Clustering: define cell clusters based on gene expression
2. DE: test if every gene has the same mean expression between cell clusters



# Example 3: single-cell post-clustering DE analysis

**Q: Why inflated false discoveries?**

**A: Two different null hypotheses**

Expression level of *m* genes: $Y_1, \ldots, Y_m$

Cell type (latent): $Z \in \{0, 1\}$

Cell cluster (based on $Y_1, \ldots, Y_m$): $\widehat{Z} \in \{0, 1\}$

**The ideal null hypothesis** $\boldsymbol{H_0} : \mathbb{E}[Y_j \mid Z = 0] = \mathbb{E}[Y_j \mid Z = 1]$

**The post-clustering double-dipping (DD) null hypothesis**

$$\boldsymbol{H_0}^{\mathrm{DD}} : \mathbb{E}[Y_j \mid \widehat{Z} = 0] = \mathbb{E}[Y_j \mid \widehat{Z} = 1]$$

$\boldsymbol{H_0}^{\mathrm{DD}}$ **does not hold but** $\boldsymbol{H_0}$ **holds** $\rightarrow$ **false-positive cell-type marker gene**

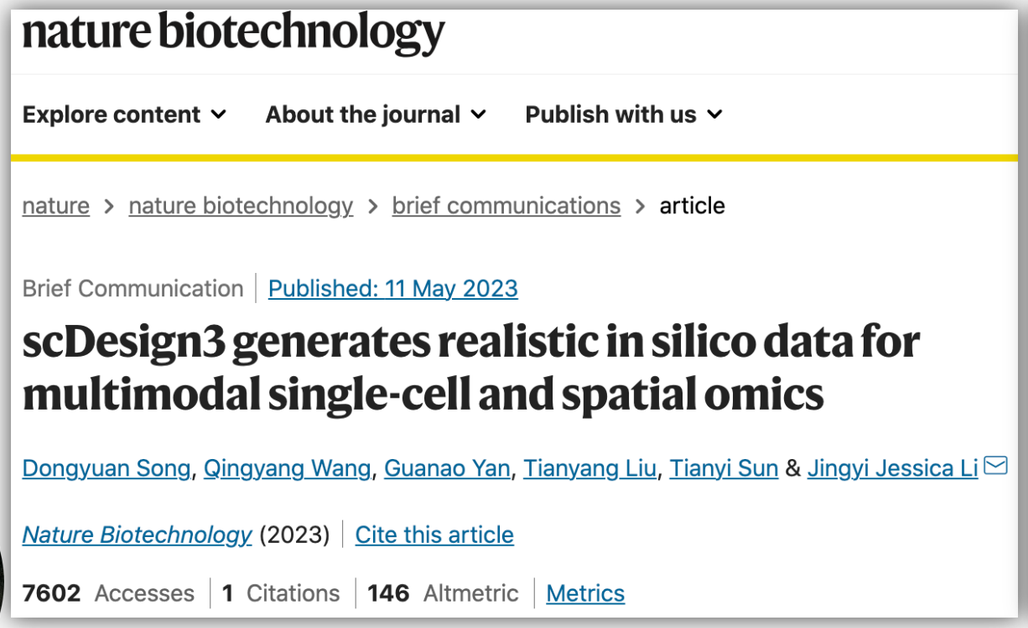
# Example 3: single-cell post-clustering DE analysis

Q: What is a meaningful "negative control" for cell type discovery?

A: All cells in one "hypothetical" cell type

→ all genes satisfy the ideal null hypotheses

Q: A model for synthetic null generation? A: scDesign2/3


nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature biotechnology > brief communications > article

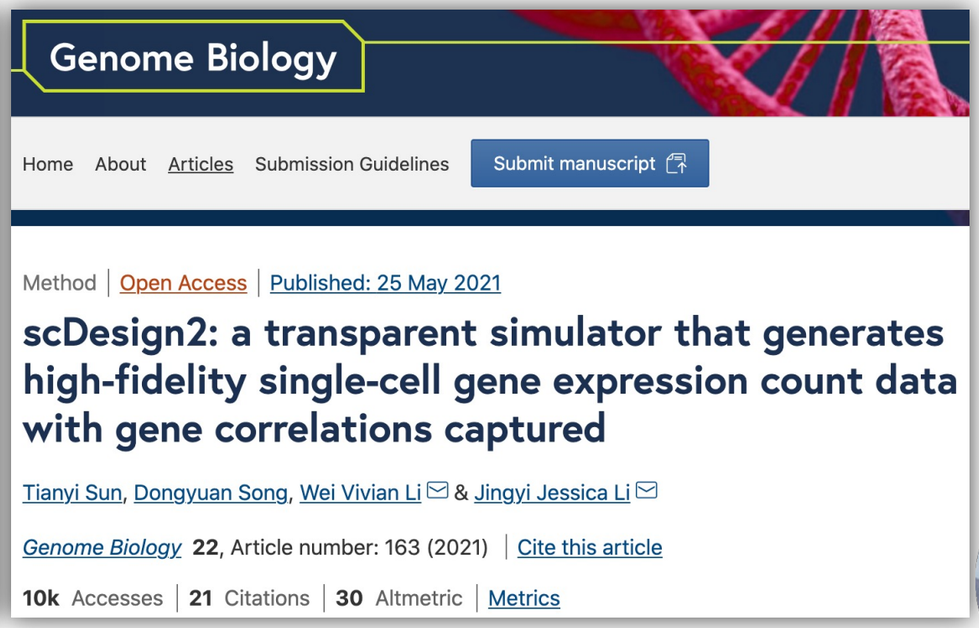
Brief Communication | Published: 11 May 2023

scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun & Jingyi Jessica Li

*Nature Biotechnology* (2023) | Cite this article

7602 Accesses | 1 Citations | 146 Altmetric | Metrics


Genome Biology

Home About Articles Submission Guidelines Submit manuscript

Method | Open Access | Published: 25 May 2021

scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

Tianyi Sun, Dongyuan Song, Wei Vivian Li & Jingyi Jessica Li

*Genome Biology* 22, Article number: 163 (2021) | Cite this article

10k Accesses | 21 Citations | 30 Altmetric | Metrics

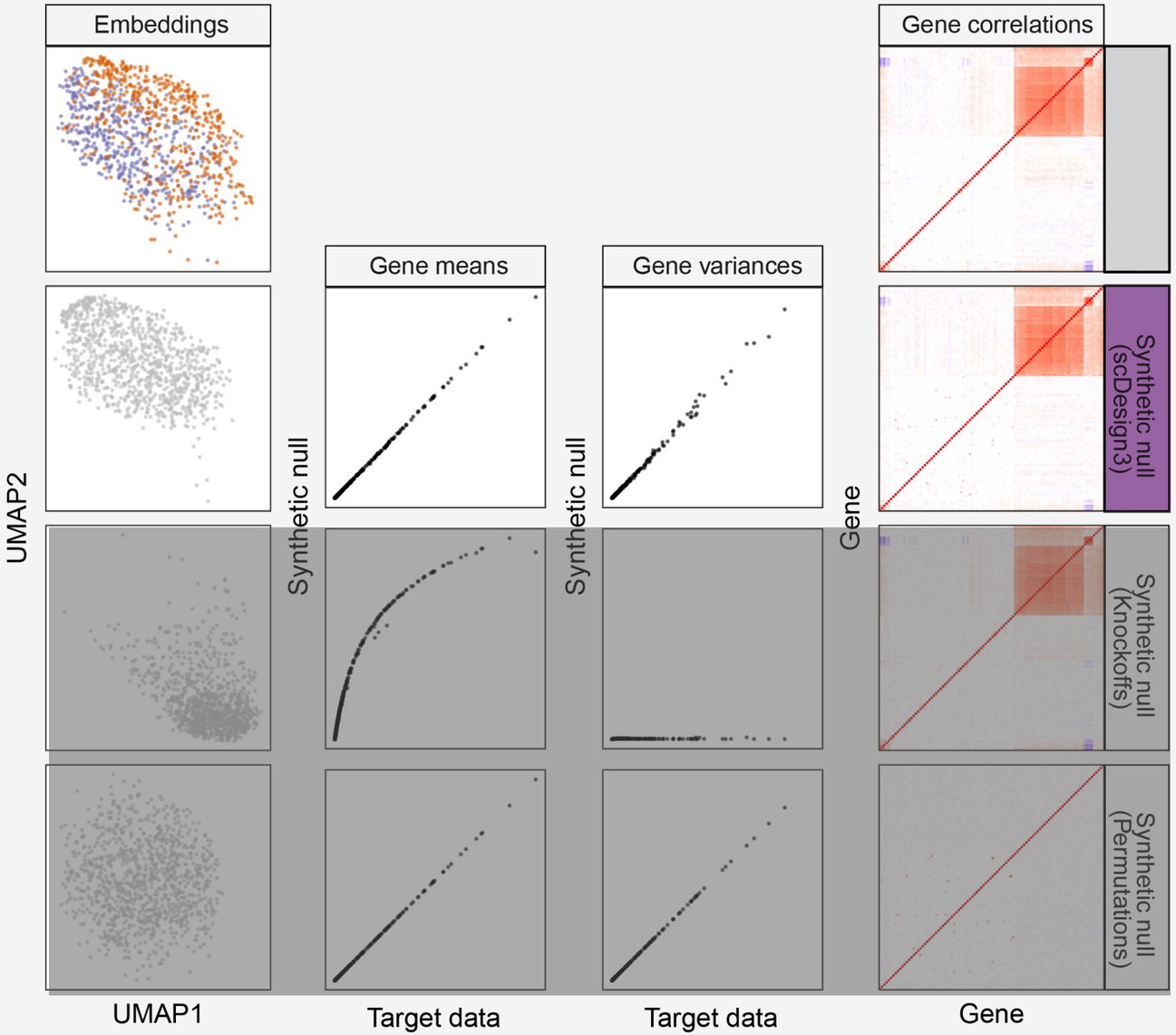
Dongyuan Song
(JSB)

Tuesday 5:20 pm in Salle Rhone 2
(RegSys COSI)

Tianyi Sun
(JSB)

# Example 3: single-cell post-clustering DE analysis

**scDesign3 preserves per-gene mean, variance, and gene-gene correlations.**

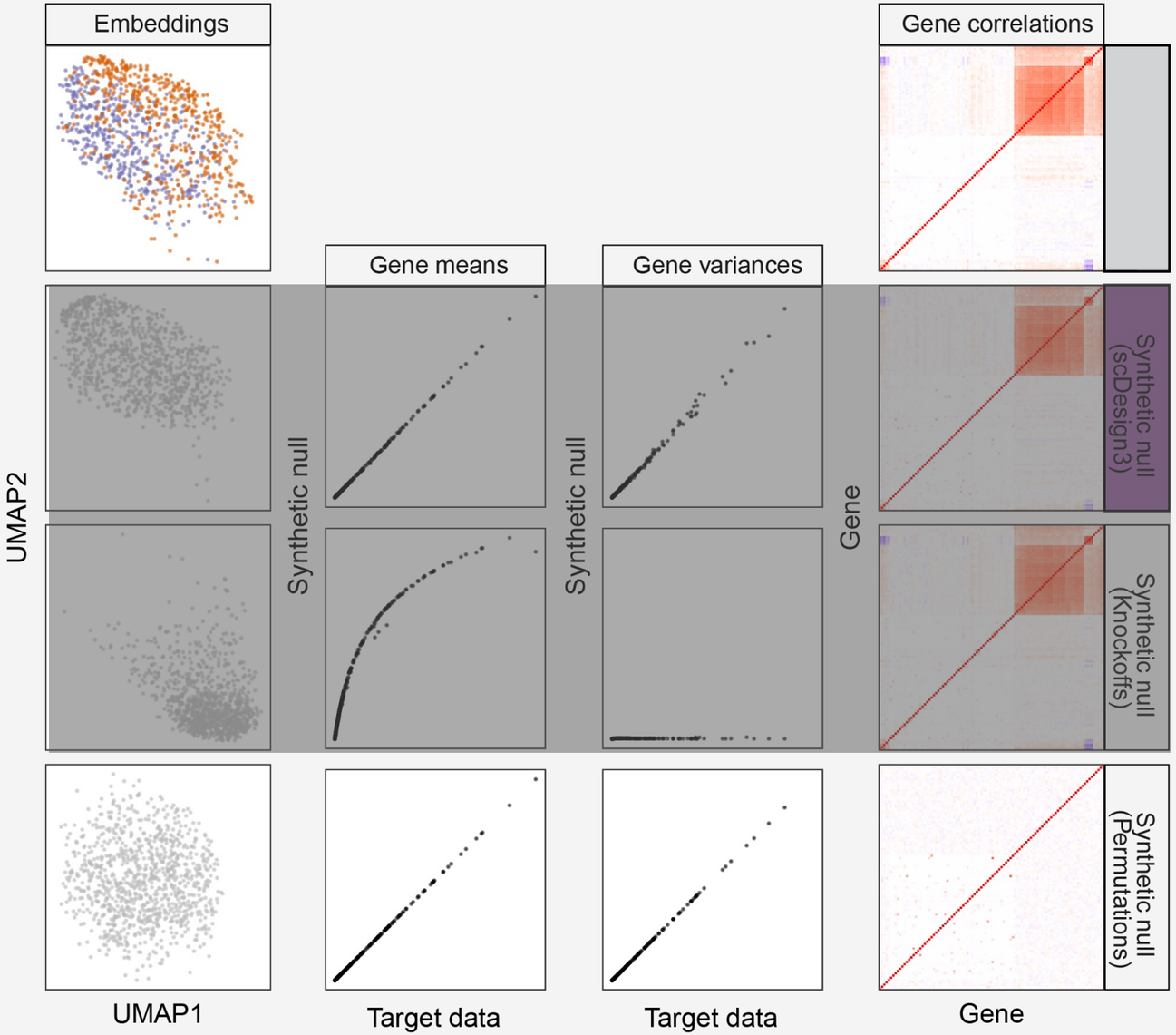


# Example 3: single-cell post-clustering DE analysis

## scDesign3 preserves per-gene mean, variance, and gene-gene correlations.

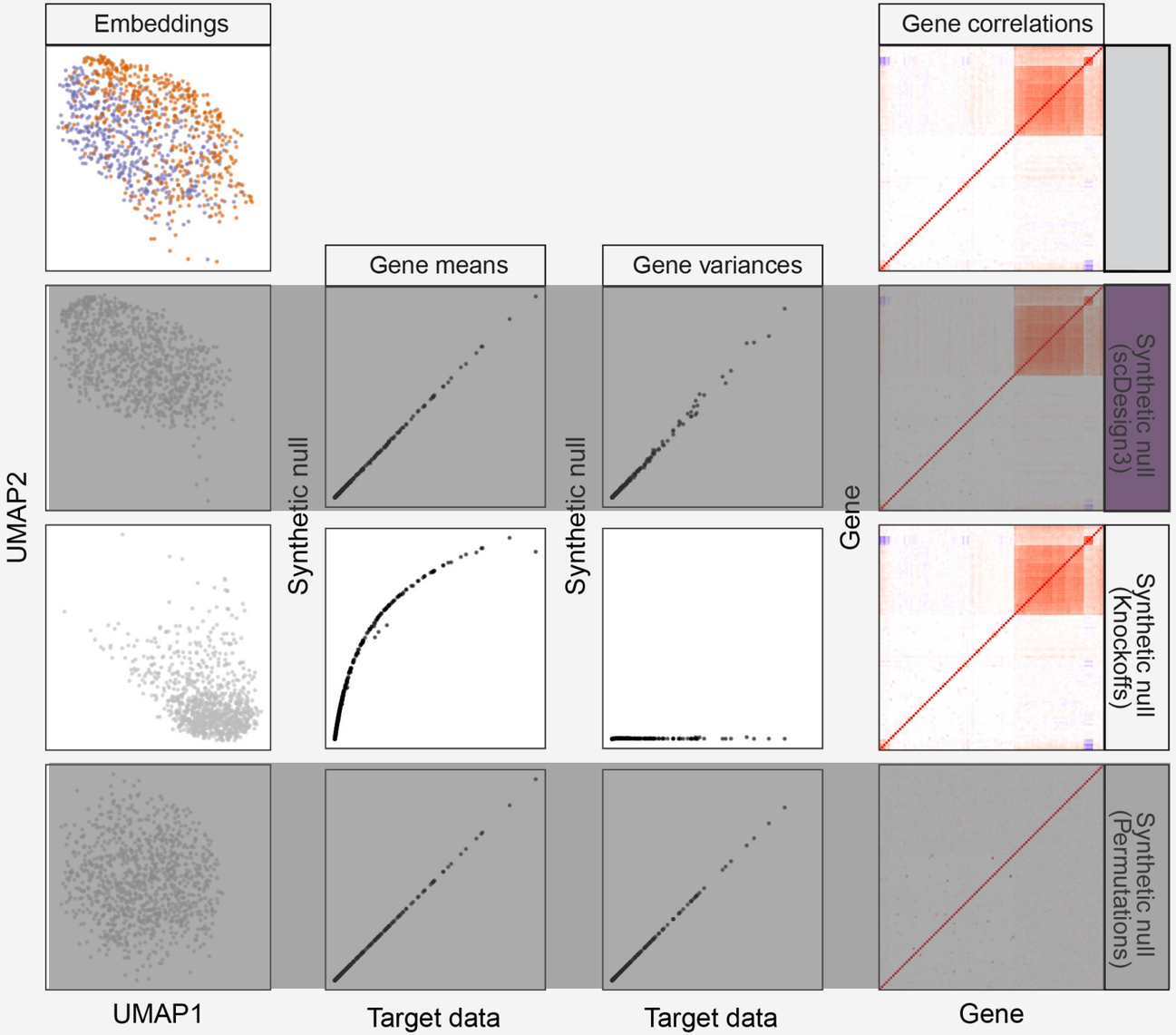
**Q: Why not permutation?**

**A: Gene-gene correlations are crucial for clustering.**



# Example 3: single-cell post-clustering DE analysis

**scDesign3 preserves per-gene mean, variance, and gene-gene correlations.**

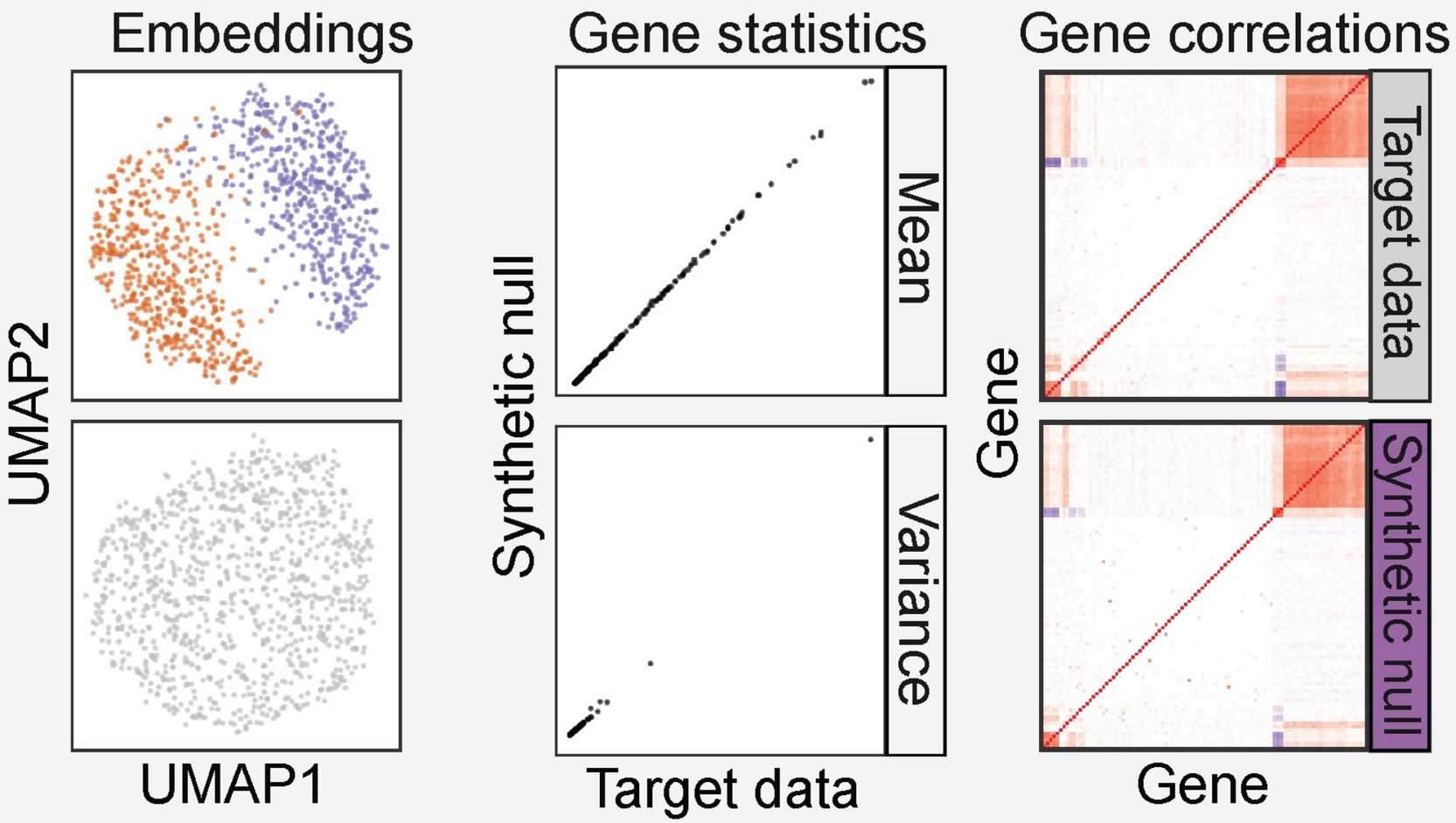



**Q: Why not knockoffs? [Barber and Candes, *Ann Stat*, 2015]**

**A: There is no outcome variable; not a supervised learning setting.**

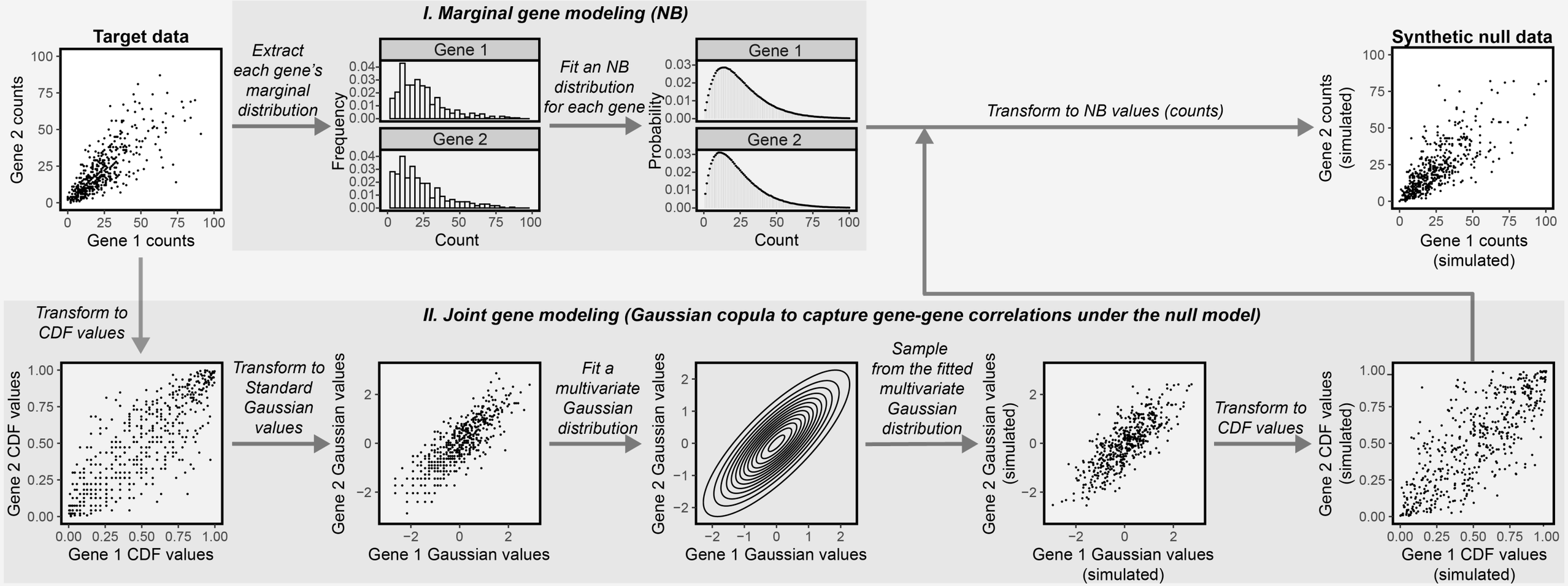
# Example 3: single-cell post-clustering DE analysis

**scDesign3 preserves per-gene mean, variance, and gene-gene correlations.**



# Example 3: single-cell post-clustering DE analysis

## scDesign3 synthetic null generation (marginal NB + Gaussian copula)

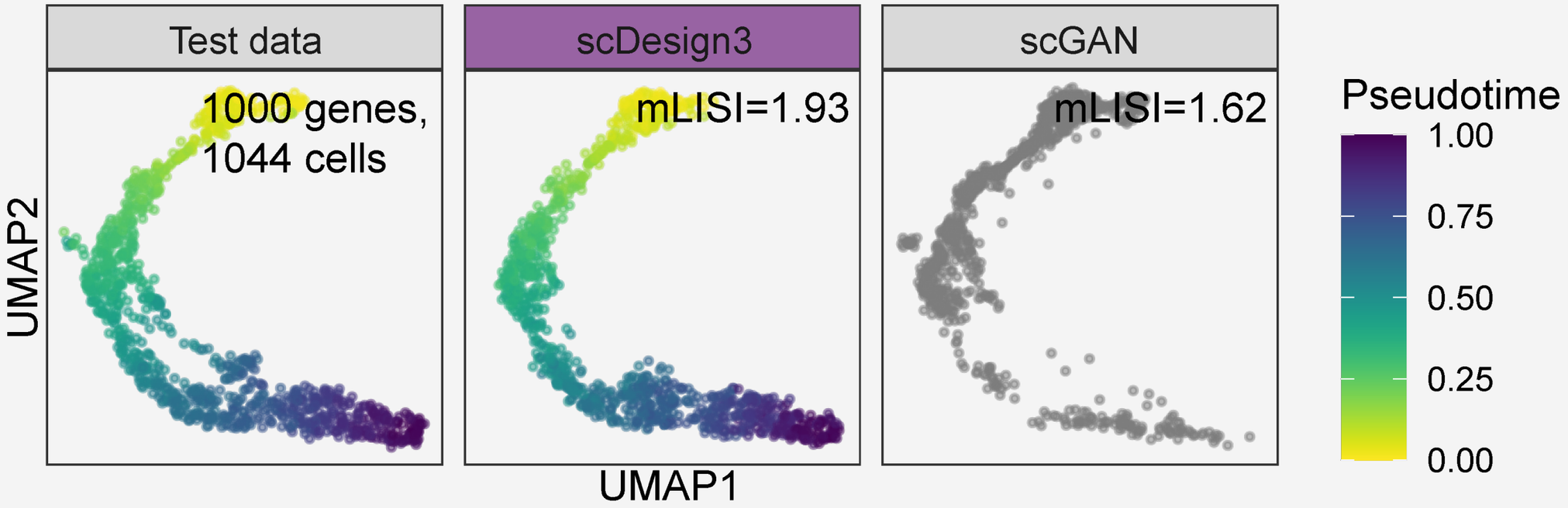


# Example 3: single-cell post-clustering DE analysis

**Q: Why NOT use deep learning (e.g., GAN) to generate synthetic data?**

**A: Unclear how to generate synthetic null data by modifying parameters.**

**scDesign3 vs. scGAN**




2 mins w/ 1 core; ~ 24 hrs

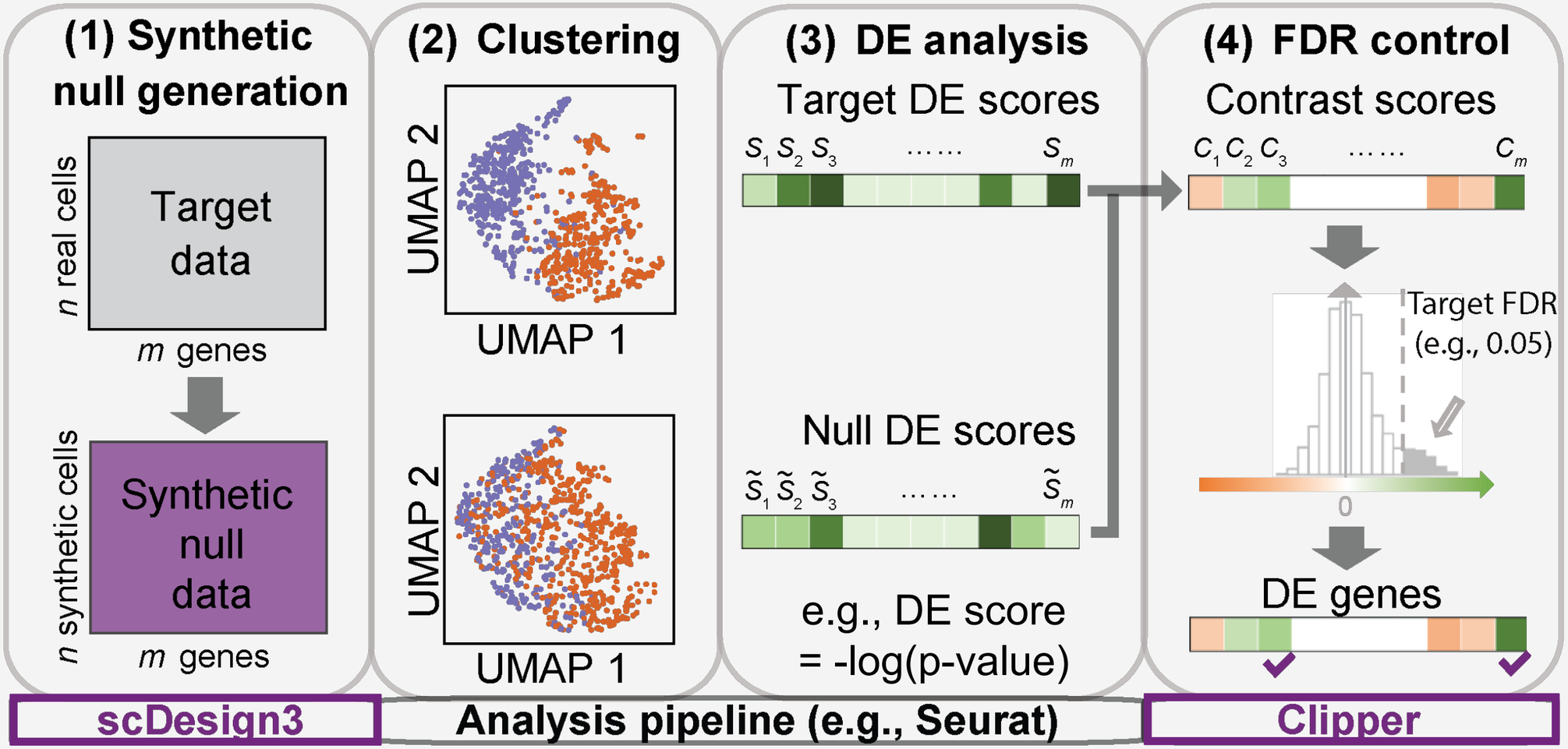
# Question 3:

# How to use synthetic null data to reduce false discoveries?

## Contrastive strategy

# Example 3: single-cell post-clustering DE analysis

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




Dongyuan Song
(JSB)

**ClusterDE**

Kexin Li
(JSB)

# Example 3: single-cell post-clustering DE analysis

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

**A: Clipper — a contrastive strategy for p-value-free FDR control**

Genome Biology

Home About Articles Submission Guidelines Submit manuscript

Method | Open Access | Published: 11 October 2021

## Clipper: *p*-value-free FDR control on high-throughput data from two conditions

Xinzhou Ge, Yiling Elaine Chen, Dongyuan Song, MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Ning Wang, Wei Li, Leo D. Wang & Jingyi Jessica Li

*Genome Biology* **22**, Article number: 288 (2021) | Cite this article



Xinzhou Ge
(JSB)

# Example 3: single-cell post-clustering DE analysis

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

**A: Clipper — a contrastive strategy for p-value-free FDR control**

- **NO requirement of**
  - high-resolution p-values
  - parametric distributions
  - large sample sizes
- **Foundation: knockoffs**
- **Two components**
  - **contrast scores**
  - **cutoff**

**Goal**: marginal screening for **interesting** features

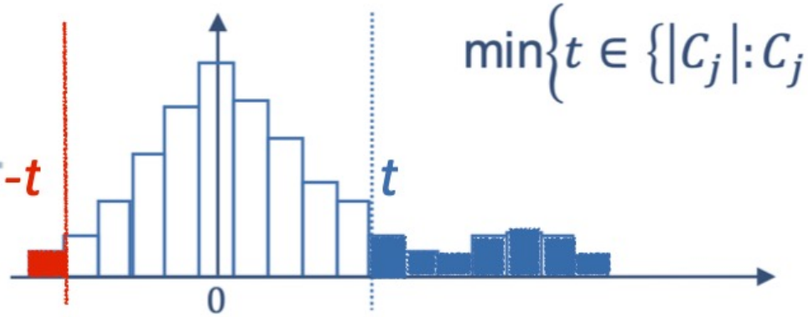
$m$ features **$m$ must be large** FDR threshold $q$

Contrast scores

$C_1$

$\vdots$

$C_m$




Distribution of $C_1, \ldots, C_m$

Contrast score cutoff

$$\min\left\{t \in \{|C_j| : C_j \neq 0\} : \frac{1+\#\{j : C_j \leq -t\}}{\#\{j : C_j \geq t\} \vee 1} \leq q\right\}$$

Knockoffs
[Barber and Candes, *Ann Stat*, 2015]

# Example 3: single-cell post-clustering DE analysis

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

A: **Clipper** — a contrastive strategy for p-value-free FDR control

**Clipper core idea**: **contrast score** of feature $j = 1, \ldots, m$:

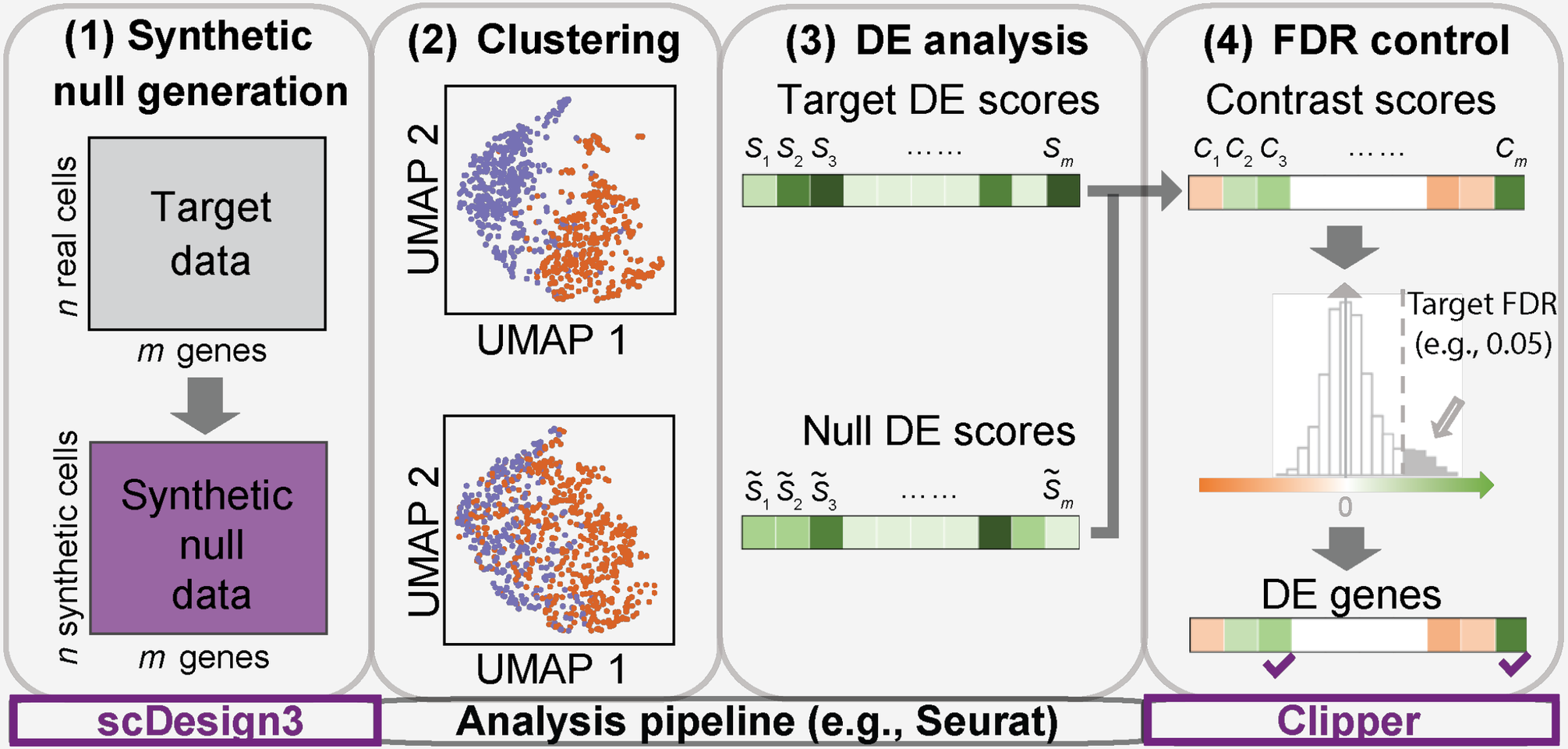
$$C_j := t(\textbf{target data}) - t(\textbf{synthetic null data}),$$

where $t(\cdot)$ can be a **complex pipeline** (e.g., clustering + DE)

| | **target data** | **synthetic null data** (*in silico* negative control) |
|---|---|---|
| **ClusterDE** | real cells | **scDesign3** synthetic cells from one "hypothetical" type |

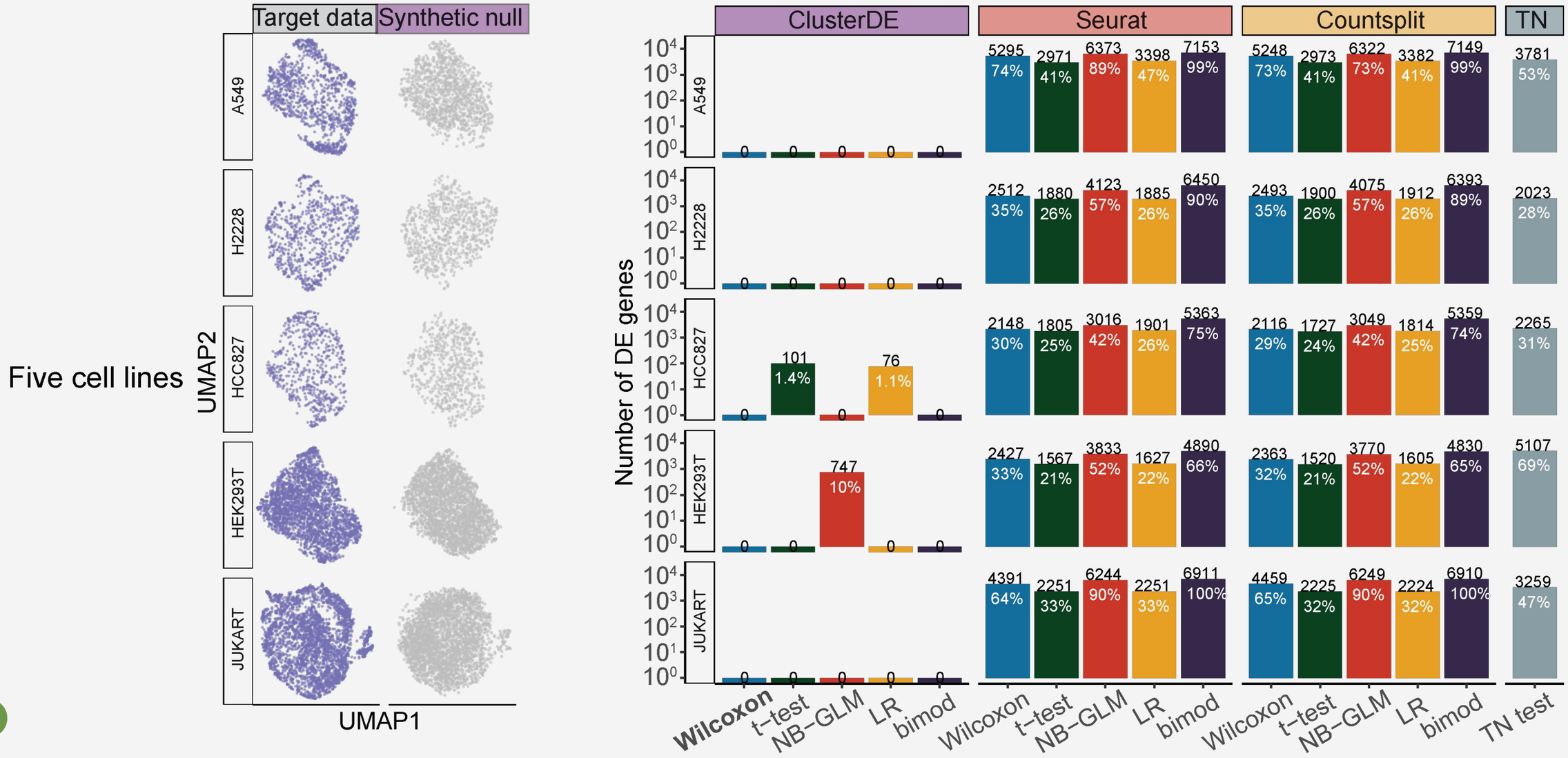
# Example 3: single-cell post-clustering DE analysis

**ClusterDE**: a post-clustering DE method robust to double dipping



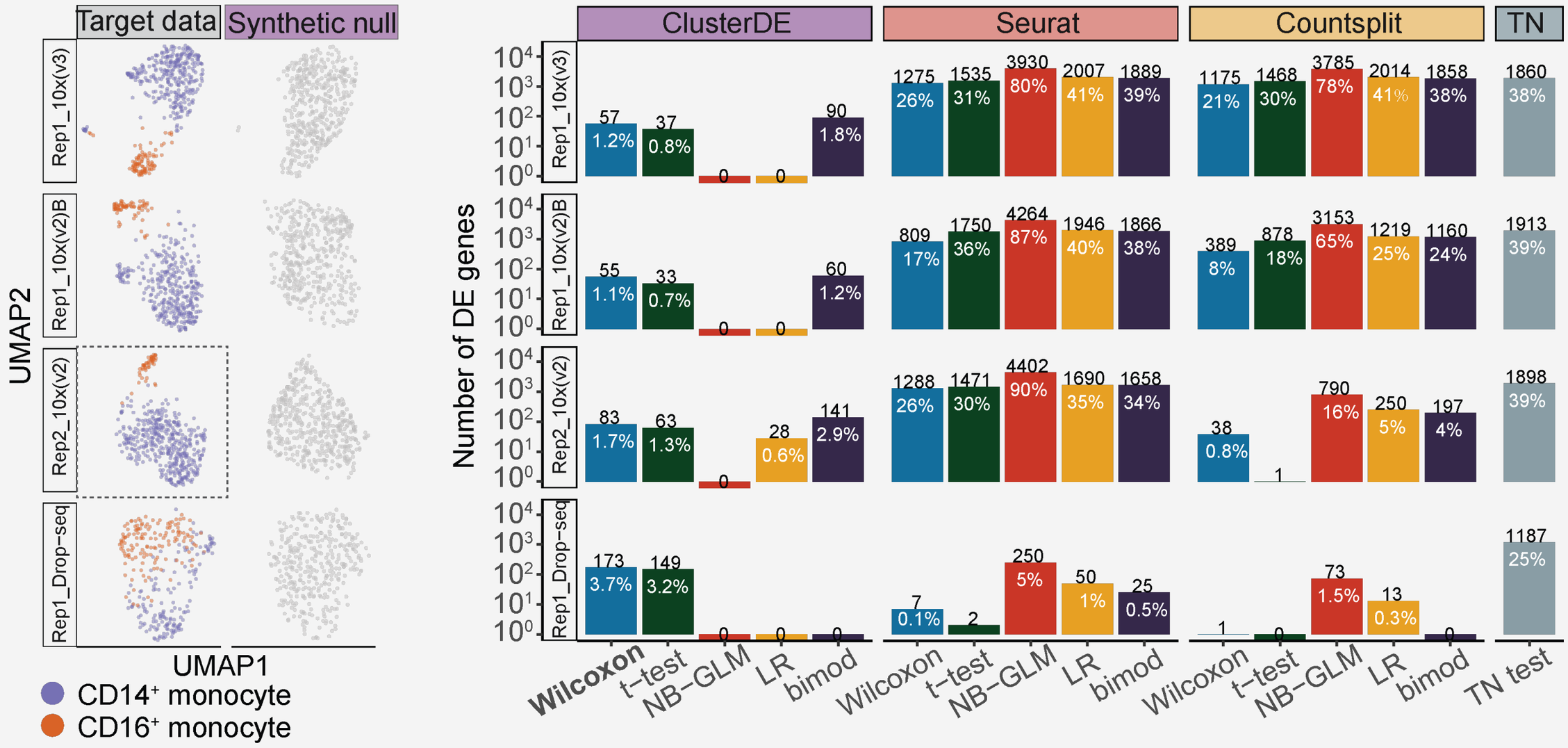
# Example 3: single-cell post-clustering DE analysis

## Expectation 1: No cell-type marker genes should be found from a cell line.



# Example 3: single-cell post-clustering DE analysis

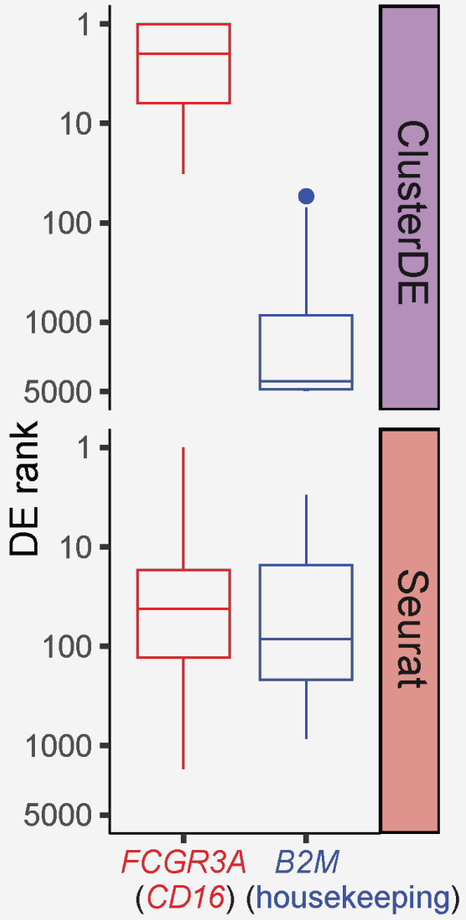
## Expectation 2: Cell-type marker genes should be found as top DE genes.

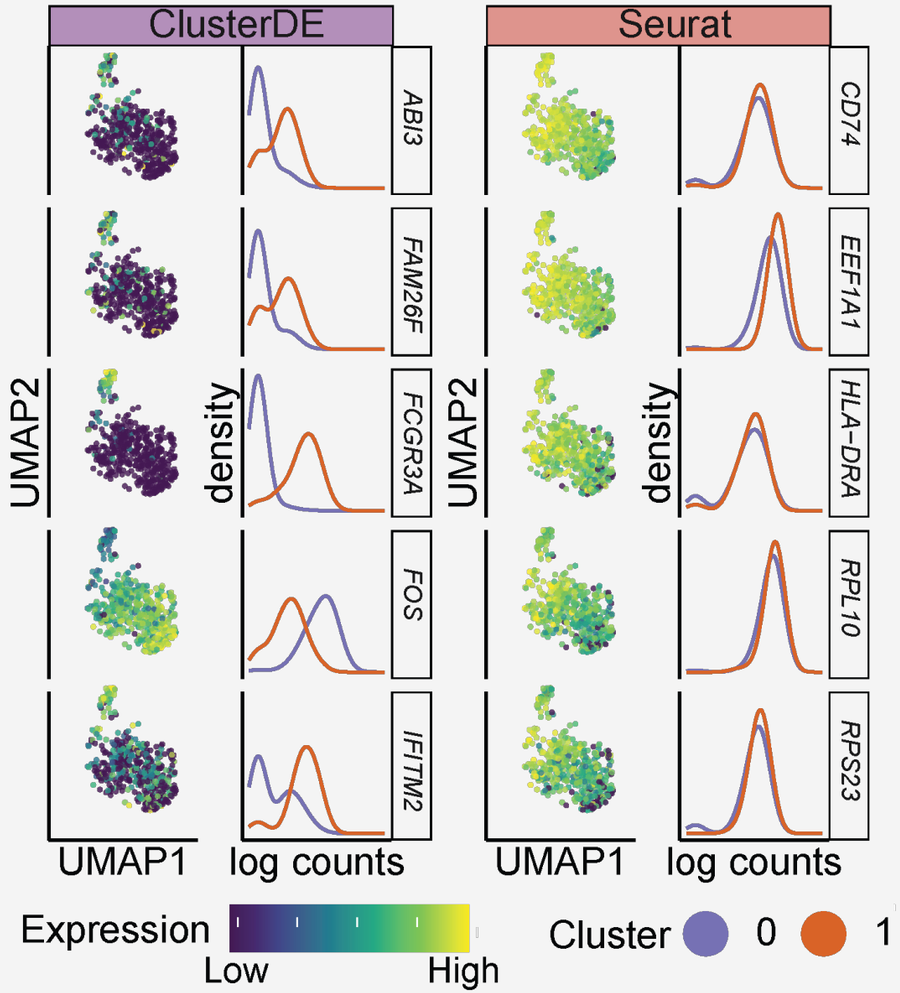


# Example 3: single-cell post-clustering DE analysis

**Expectation 2: Cell-type marker genes should be found as top DE genes.**

**Expectation 3: Housekeeping genes should NOT be found as top DE genes.**

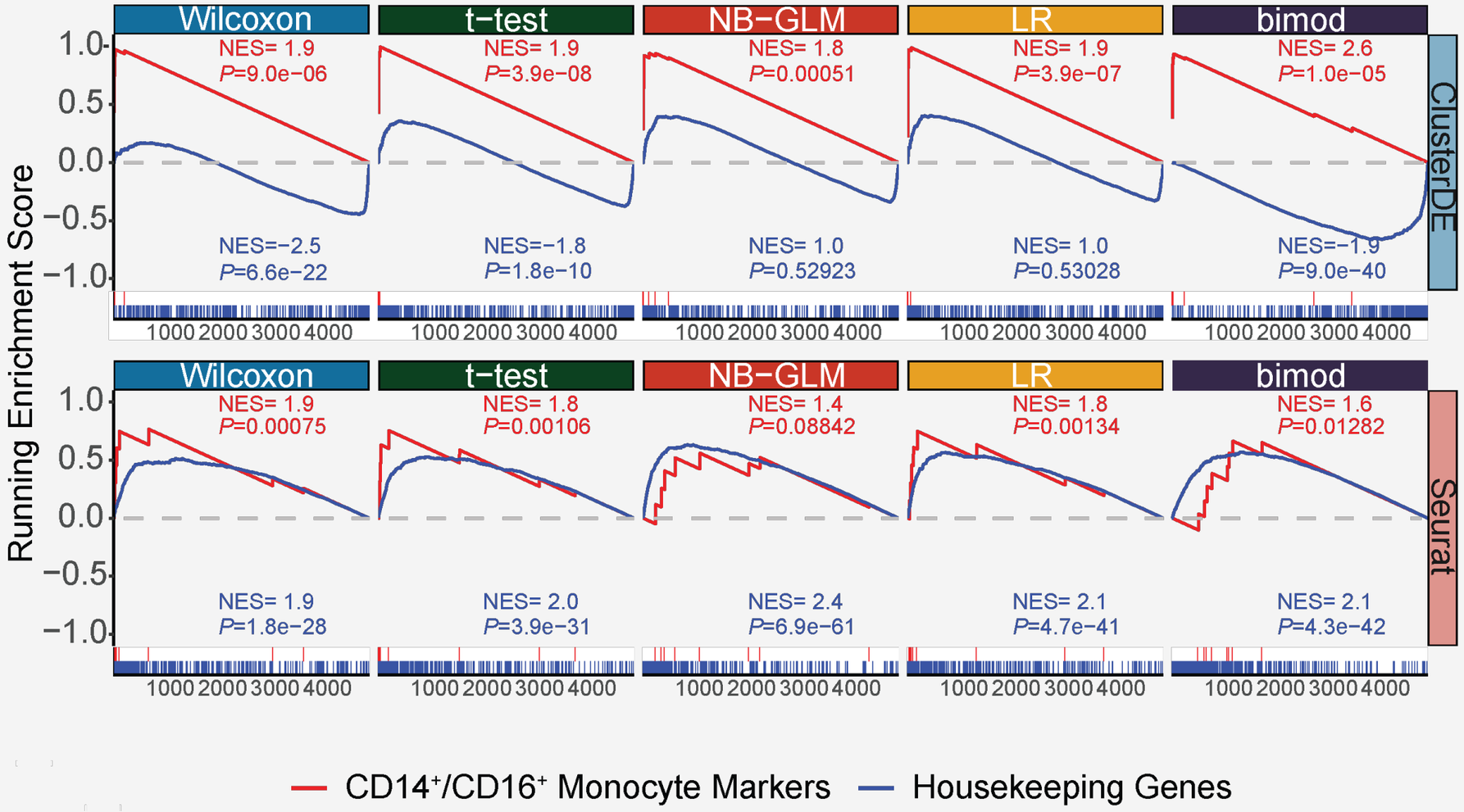


# Example 3: single-cell post-clustering DE analysis

**Expectation 2: Cell-type marker genes should be found as top DE genes.**

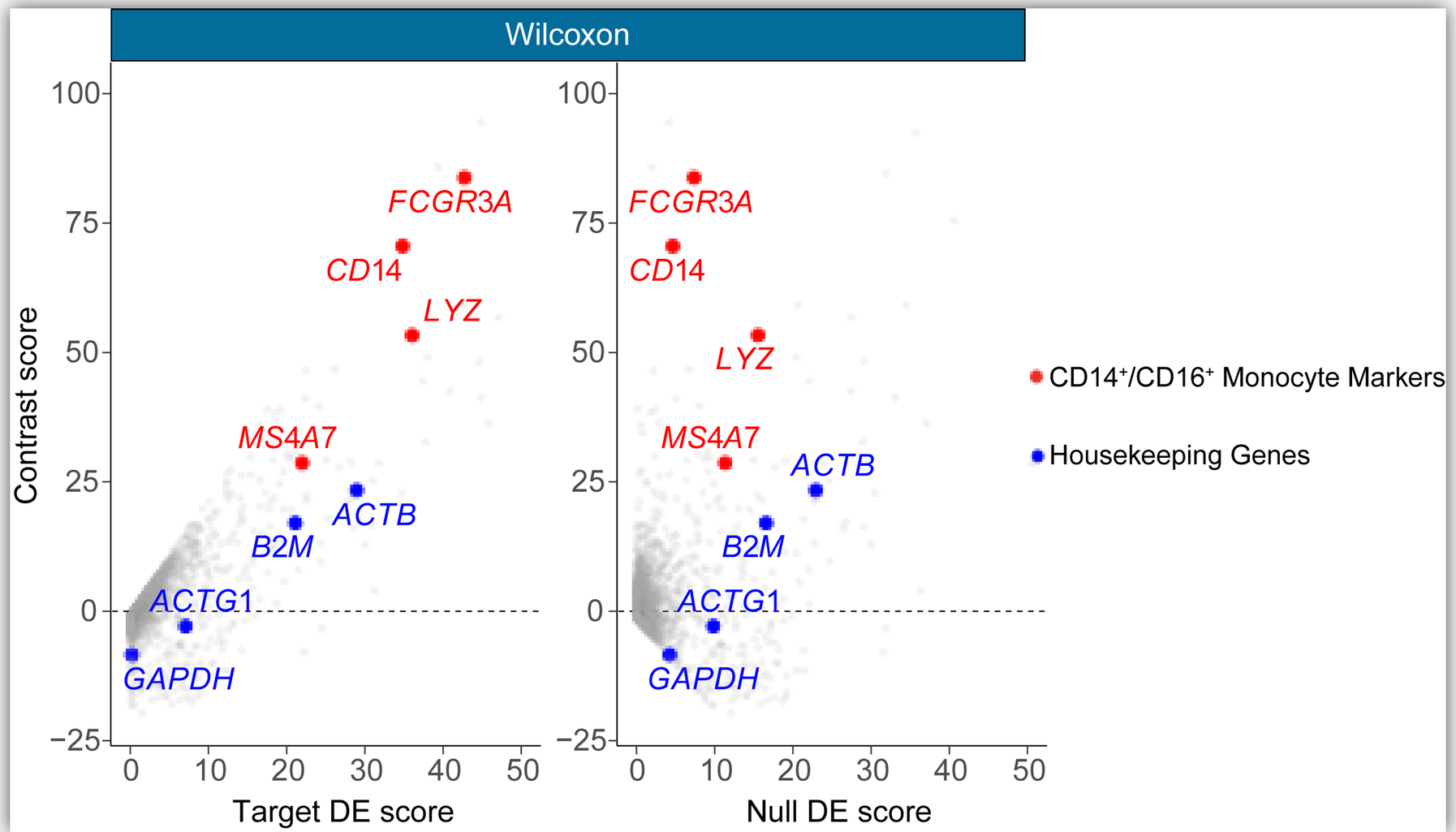
**Expectation 3: Housekeeping genes should NOT be found as top DE genes.**



# Example 3: single-cell post-clustering DE analysis

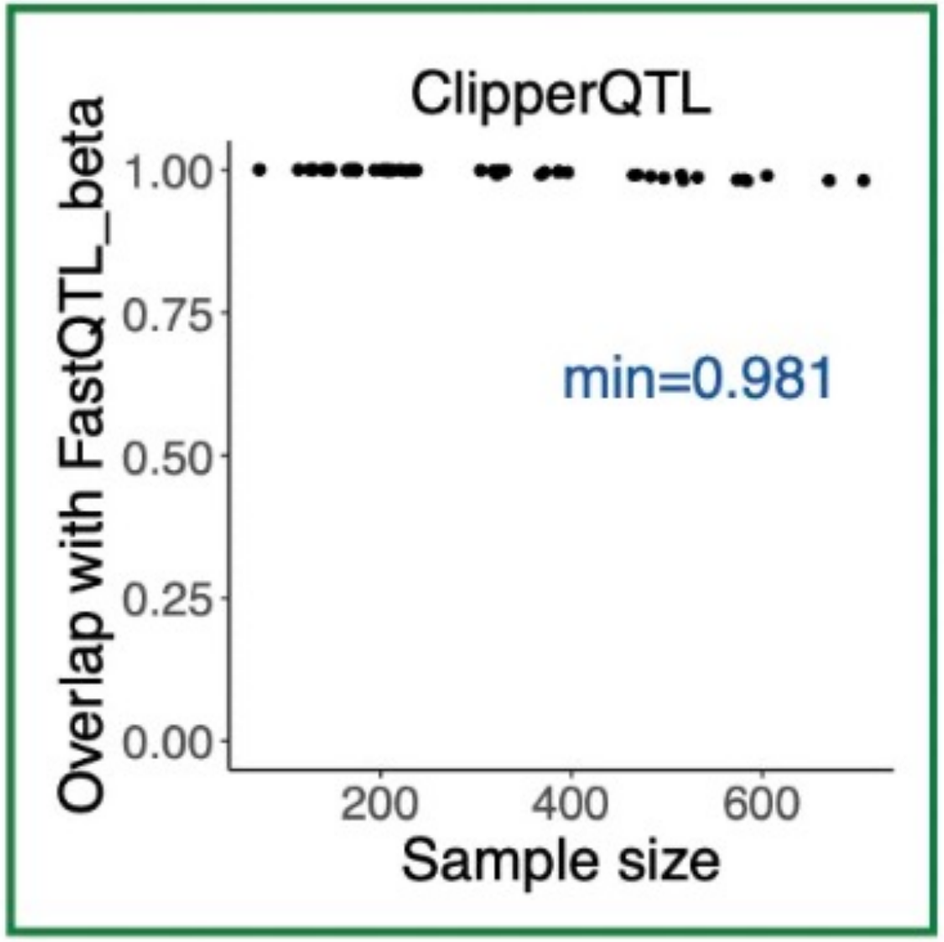
**Q: Why does ClusterDE NOT identify housekeeping genes as top DE genes?**

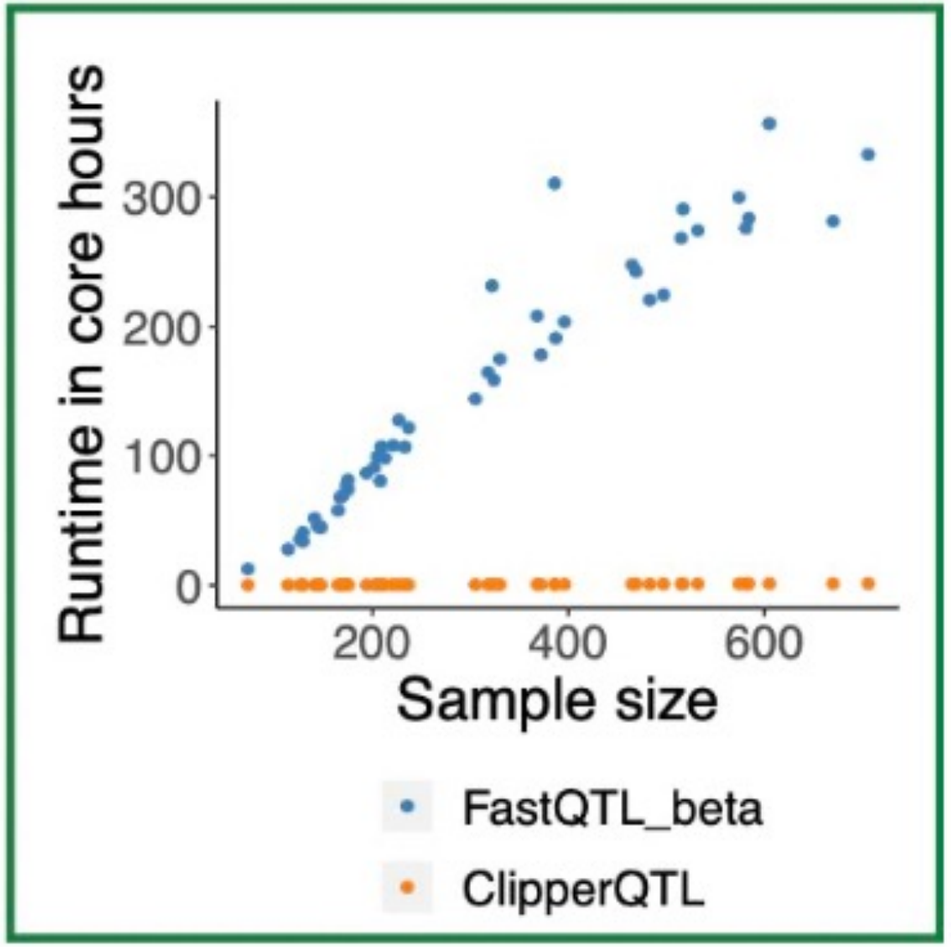
**A: ClusterDE uses contrast scores (= target DE score – null DE score).**



# Contrastive strategy is computationally efficient

**ClipperQTL (contrastive strategy) vs. FastQTL (p-value-based strategy)**




Heather Zhou
(JSB)

**Ongoing work**

# Summary

1. **What is an appropriate null hypothesis?**
   - Different null hypotheses → different discoveries/conclusions

   Example 1: bulk RNA-seq DE analysis: NB vs. Wilcoxon? Permutation

2. **How to make an abstract null hypothesis concrete?**
   - Synthetic null

   Example 2: dubious t-SNE/UMAP embeddings? Permutation → scDEED

   Example 3: single-cell post-clustering DE analysis: scDesign3

3. **How to use synthetic null data to reduce false discoveries?**
   - Contrastive strategy (Clipper) vs. p-value-based strategy: ClipperQTL

   Example 3: single-cell post-clustering DE analysis:

   **ClusterDE**: scDesign3 → clustering + DE → Clipper

# Take-home message 1

**Synthetic null data can make an abstract null hypothesis concrete and enable contrastive data analysis**

Synthetic null data generation is real-data-specific and problem-specific

*"Teaching someone to fish is better than giving them a fish"* — Chinese proverb

## Take-home message/question 2

**Less is more (?)**

- *Occam's razor: the principle of parsimony*
- Fewer but more reliable discoveries → science

# Acknowledgements


**Ph.D. advisors @ Berkeley**

– Peter J. Bickel
– Haiyan Huang

**Collaborators**

– Wei Li & Yumei Li @ UCI
– Lucy Xia @ HKUST
– Mark D. Biggin @ LBNL
– Xin Tong @ USC

**Nominators**

– Wei Li @ UCI
– Shirley Liu @ GV20
– Chongzhi Zang @ UVA

**Trainees @ UCLA**

– Xinzhou Ge (bulk DE; Clipper) will join Oregon State University
– Christy Lee (scDEED)
– Dongyuan Song (ClusterDE; scDesign3) will be on the job market
– Tianyi Sun (scDesign2)
– Kexin Li (ClusterDE)
– Heather Zhou (ClipperQTL)

Former trainees

– Wei Vivian Li @ UC Riverside
– Nan Miles Li @ Loyola Univ Chicago