# KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

# Robust scRNA-seq Data Analysis: An Automated Pipeline from Dimensionality Reduction to Clustering

김현 _ IBS

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

⚠

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

# KSBi-BIML 2026

## Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

**한국생명정보학회장 류 성 호**

## 강의개요

# Robust scRNA-seq Data Analysis: An Automated Pipeline from Dimensionality Reduction to Clustering

단일 세포 전사체 데이터(scRNA-seq)는 고차원적 특성과 함께 극심한 기술적 노이즈를 내재하고 있다. 이러한 노이즈를 제어하고 유의미한 생물학적 신호를 추출하기 위해 차원 축소(Dimensionality reduction) 알고리즘이 필수적으로 사용된다. 그러나 기존 알고리즘은 축소할 차원의 수를 사용자가 직접 결정하게 함으로써 사용자 편향을 야기한다. 이러한 편향은 분석가의 주관에 따라 분석 결과를 크게 달라지게 만드는 원인이 된다.

뿐만 아니라, 세포 유형 식별을 위해 이어지는 클러스터링 단계에서도 몇 개의 그룹으로 세포를 나눌지 결정하는 파라미터 설정을 사용자에게 의존하므로, 분석 전반에 걸쳐 사용자 편향이 반복적으로 개입된다. 이러한 한계들은 결국 연구 결과의 신뢰성 및 재현성 저하로 직결된다.

본 강의에서는 이러한 문제를 해결하기 위한 수학적 방법론과 자동화 알고리즘을 다룬다. 구체적으로, 데이터 내 신호 왜곡을 효과적으로 제거하는 최적의 데이터 전처리부터, 랜덤행렬이론(Random Matrix Theory, RMT)에 기반하여 최적 차원을 결정하는 자동 차원 축소, 그리고 안정성 지표를 활용한 강건한 클러스터링 분석까지의 전 과정을 심도 있게 학습한다. 본 강의를 통해 수강생들은 복잡한 파라미터 튜닝 없이도 데이터 본연의 신호에 근거한 견고하고 신뢰할 수 있는 scRNA-seq 분석 결과를 도출하는 핵심 역량을 갖춘다.

* 교육생준비물:

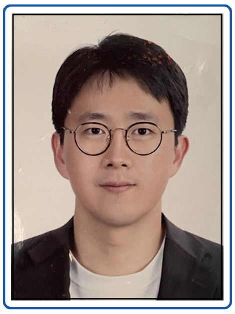
노트북 (메모리 16GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 중급

* 강의: 김현 박사 (기초과학연구원 의생명수학그룹)

# Curriculum Vitae

## Speaker Name: Hyun Kim, Ph.D.



### ▶ Personal Info

| | |
|---|---|
| Name | Hyun Kim |
| Title | Senior researcher |
| Affiliation | IBS Biomedical Mathematics Group |

### ▶ Contact Information

| | |
|---|---|
| Address | 55, Expo-ro, Yuseong-gu, Daejeon, Republic of Korea |
| Email | kimman3803@gmail.com |

**Research Interest**

Big Data Analysis, Single cell sequencing data analysis, Random matrix theory, Dimensionality reduction, Nonlinear dynamics, Biophysics, Neural circuits, Neuronal dynamics, Suprachiasmatic nucleus (SCN) network, biological clock, and oscillatory network.

**Educational Experience**

| | |
|---|---|
| 2013 | B.S. in Physics, Korea University, Republic of Korea |
| 2021 | Ph.D. Physics (Nonlinear dynamics and Biophysics), Korea University, Republic of Korea |

**Professional Experience**

| | |
|---|---|
| 2021-2024 | Post-doc research fellow, Biomedical Mathematics Group, IBS, Republic of Korea |
| 2024- | Senior Researcher, Biomedical Mathematics Group, IBS, Republic of Korea |

**Selected Publications (3 maximum)**

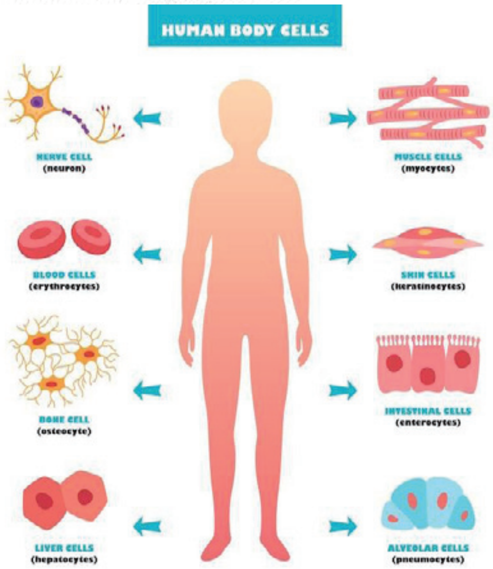
1. Kim, H., Park, I, Park, J. E., Kim, Jong K., Seo, M., & Kim, Jae K. scICE: Enhancing clustering reliability and efficiency of scRNA-seq data with multi-cluster label consistency evaluation. Nat Commun, 16, 6031. (2025)
2. Kim, H., Chang, W., Chae, S.J. et al. scLENS: data-driven signal detection for unbiased scRNA-seq data analysis. Nat Commun 15, 3575 (2024).
3. Kim H., Min C, Jeong B, Lee KJ. Deciphering clock cell network morphology within the biological master clock, suprachiasmatic nucleus: From the perspective of circadian wave dynamics. PLoS Comput Biol. 2022 18(6): e1010213.

# KSBi-BIML 2026

**Robust scRNA-seq Data Analysis:** An Automated Pipeline from Dimensionality Reduction to Clustering

IBS 의생명수학그룹
김현

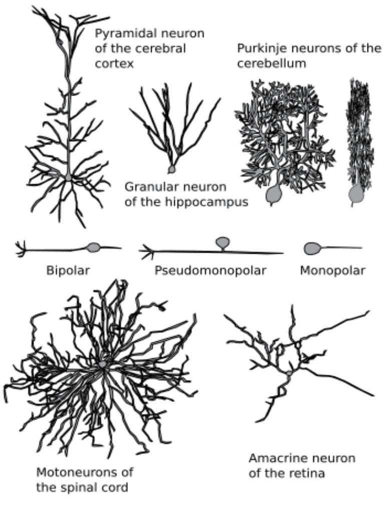
## Comprehending the human body necessitates understanding its 200+ diverse cell types.


HUMAN BODY CELLS
NERVE CELL (neuron)
MUSCLE CELLS (myocytes)
BLOOD CELLS (erythrocytes)
SKIN CELLS (keratinocytes)
BONE CELL (osteocyte)
INTESTINAL CELLS (enterocytes)
LIVER CELLS (hepatocytes)
ALVEOLAR CELLS (pneumocytes)

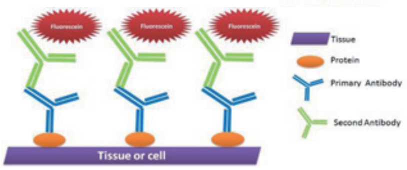
## The evolution of cell type identification

### Immunotechniques - Molecular Markers

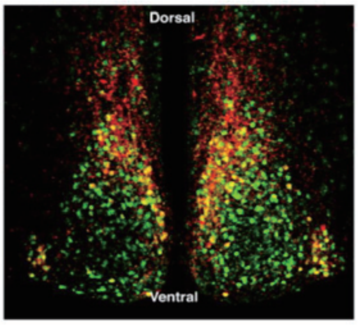
### Microscopy - Cell morphology


Pyramidal neuron of the cerebral cortex
Purkinje neurons of the cerebellum
Granular neuron of the hippocampus
Bipolar
Pseudomonopolar
Monopolar
Motoneurons of the spinal cord
Amacrine neuron of the retina

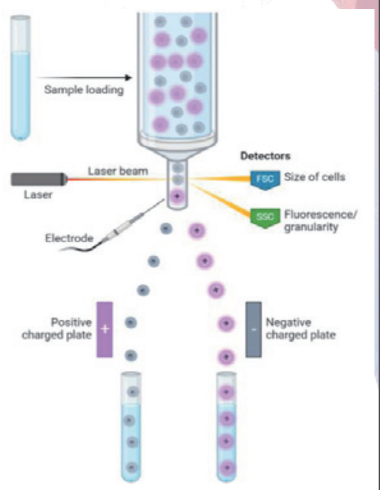
Immunofluorescence



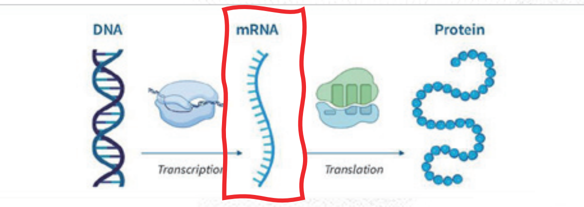
(a)

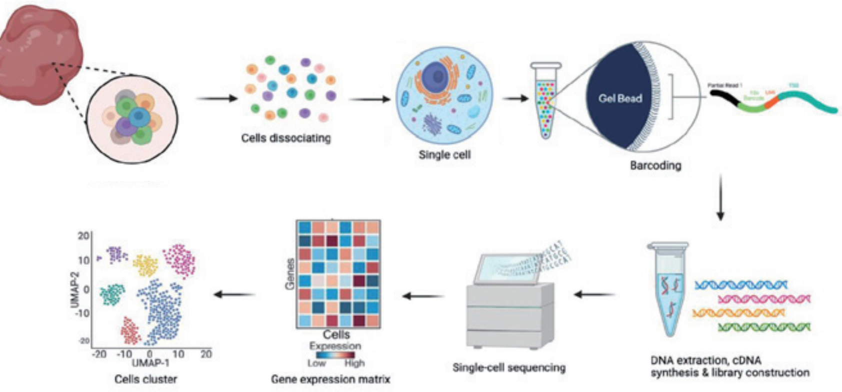


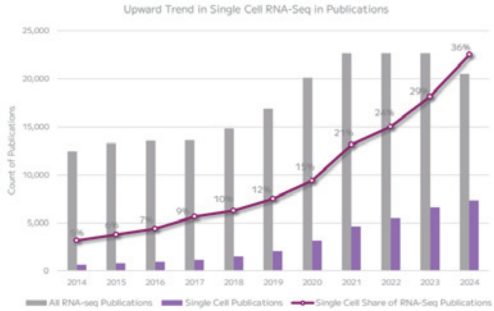
Flow cytometry



## The latest method for classifying cell types involves utilizing the similarity of transcriptomic profiles assessed through scRNA-seq.



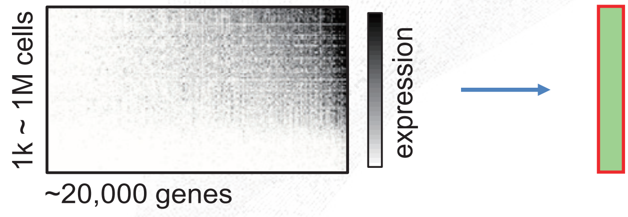




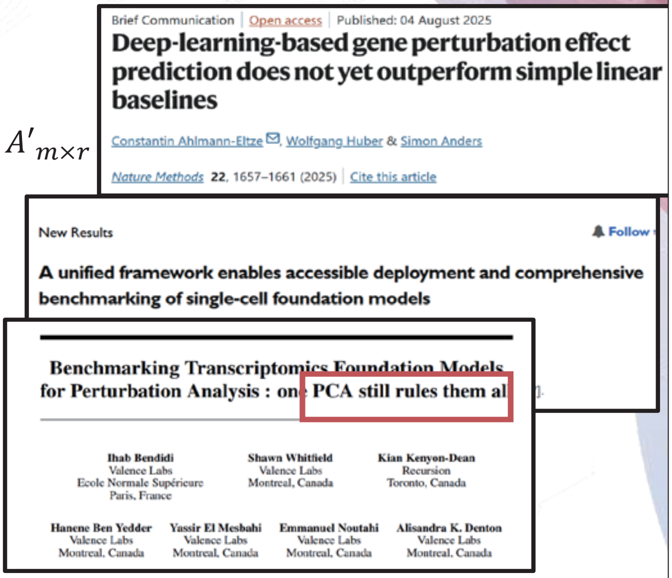
Ref: https://www.parsebiosciences.com

## Principal Component Analysis (PCA) is mostly used for dimensionality reduction in scRNA-seq data.

Noisy and High-Dimensional data



$A'_{m\times r}$



Brief Communication | Open access | Published: 04 August 2025

**Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines**

Constantin Ahlmann-Eltze, Wolfgang Huber & Simon Anders

*Nature Methods* 22, 1657–1661 (2025) | Cite this article

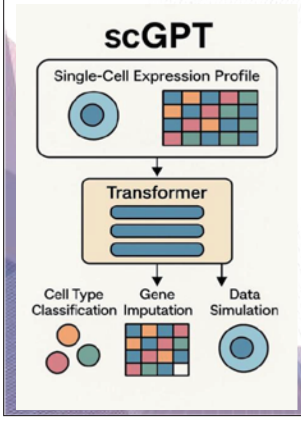
New Results

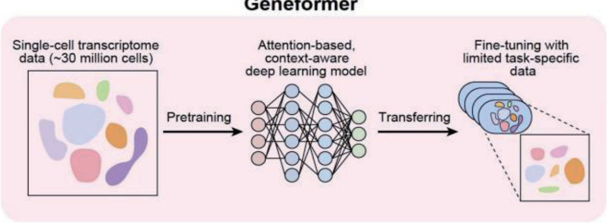
Follow

**A unified framework enables accessible deployment and comprehensive benchmarking of single-cell foundation models**

**Benchmarking Transcriptomics Foundation Models for Perturbation Analysis : on PCA still rules them al**

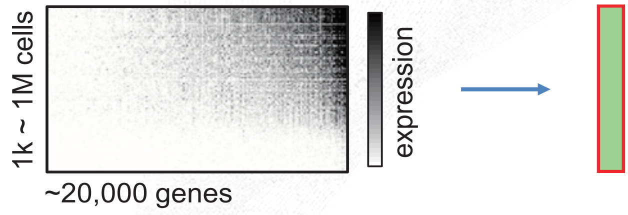
Ihab Bendidi, Valence Labs, Ecole Normale Supérieure, Paris, France
Shawn Whitfield, Valence Labs, Montreal, Canada
Kian Kenyon-Dean, Recursion, Toronto, Canada
Hanene Ben Yedder, Valence Labs, Montreal, Canada
Yassir El Mesbahi, Valence Labs, Montreal, Canada
Emmanuel Noutahi, Valence Labs, Montreal, Canada
Alisandra K. Denton, Valence Labs, Montreal, Canada



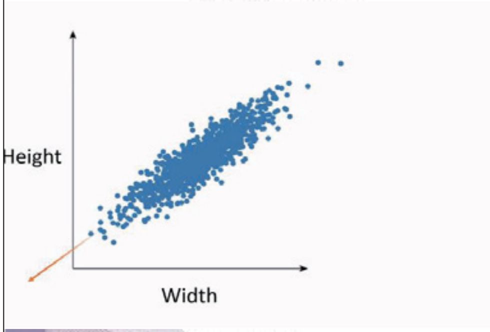

Geneformer
Single-cell transcriptome data (~30 million cells)
Pretraining
Attention-based, context-aware deep learning model
Transferring
Fine-tuning with limited task-specific data

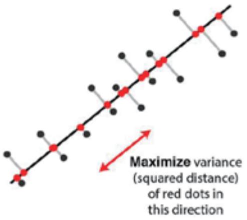
## Principal Component Analysis (PCA) is mostly used for dimensionality reduction in scRNA-seq data.

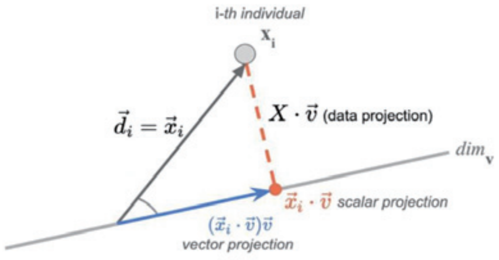
Noisy and High-Dimensional data



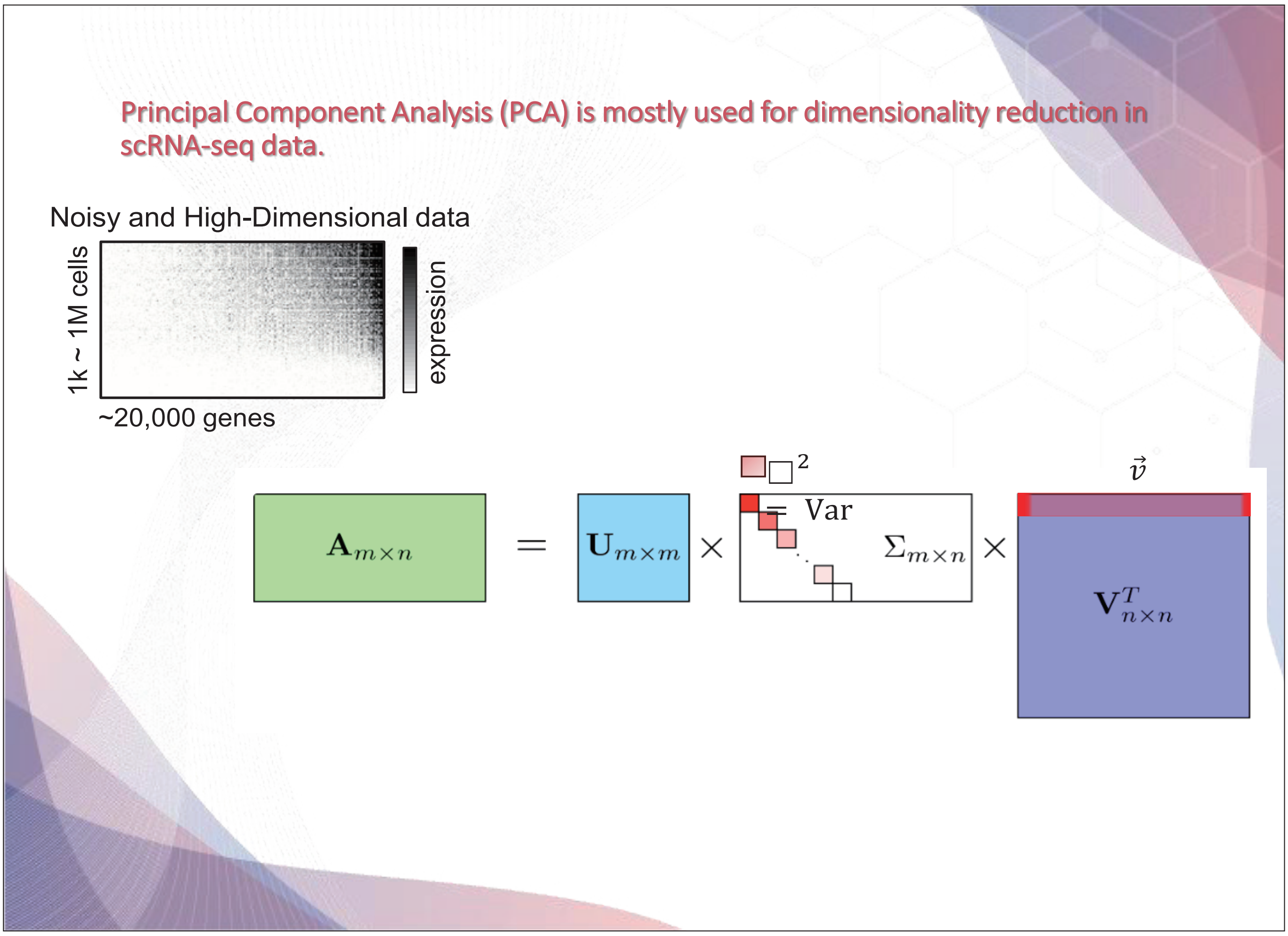
$A'_{m\times r}$




Maximize variance (squared distance) of red dots in this direction

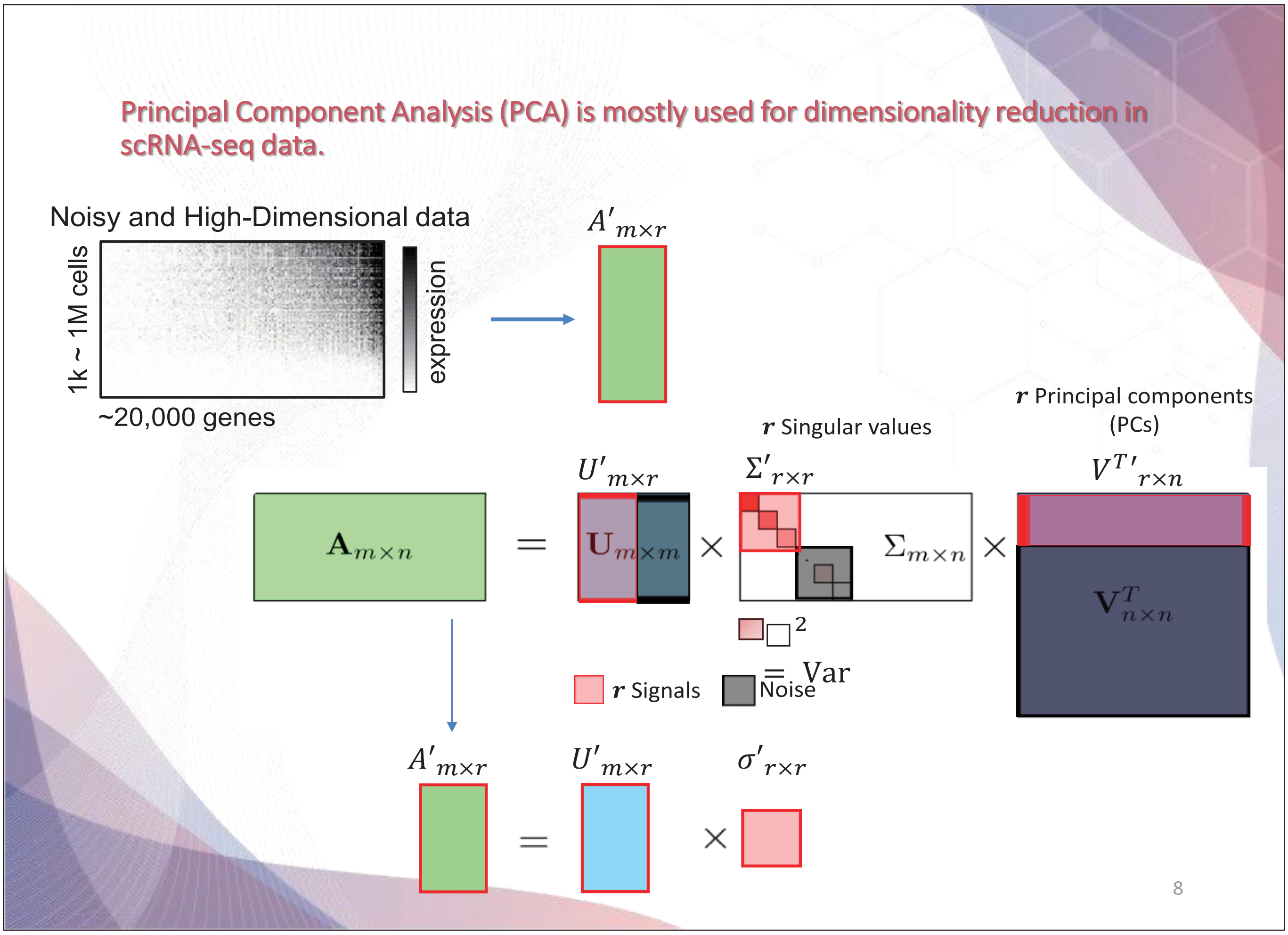

i-th individual
$x_i$
$\vec{d}_i = \vec{x}_i$
$X \cdot \vec{v}$ (data projection)
$dim_v$
$\vec{x}_i \cdot \vec{v}$ scalar projection
$(\vec{x}_i \cdot \vec{v})\vec{v}$
vector projection

$$\max_{\|\vec{v}\|=1} \|X\vec{v}\|^2 = \mathrm{Var}$$



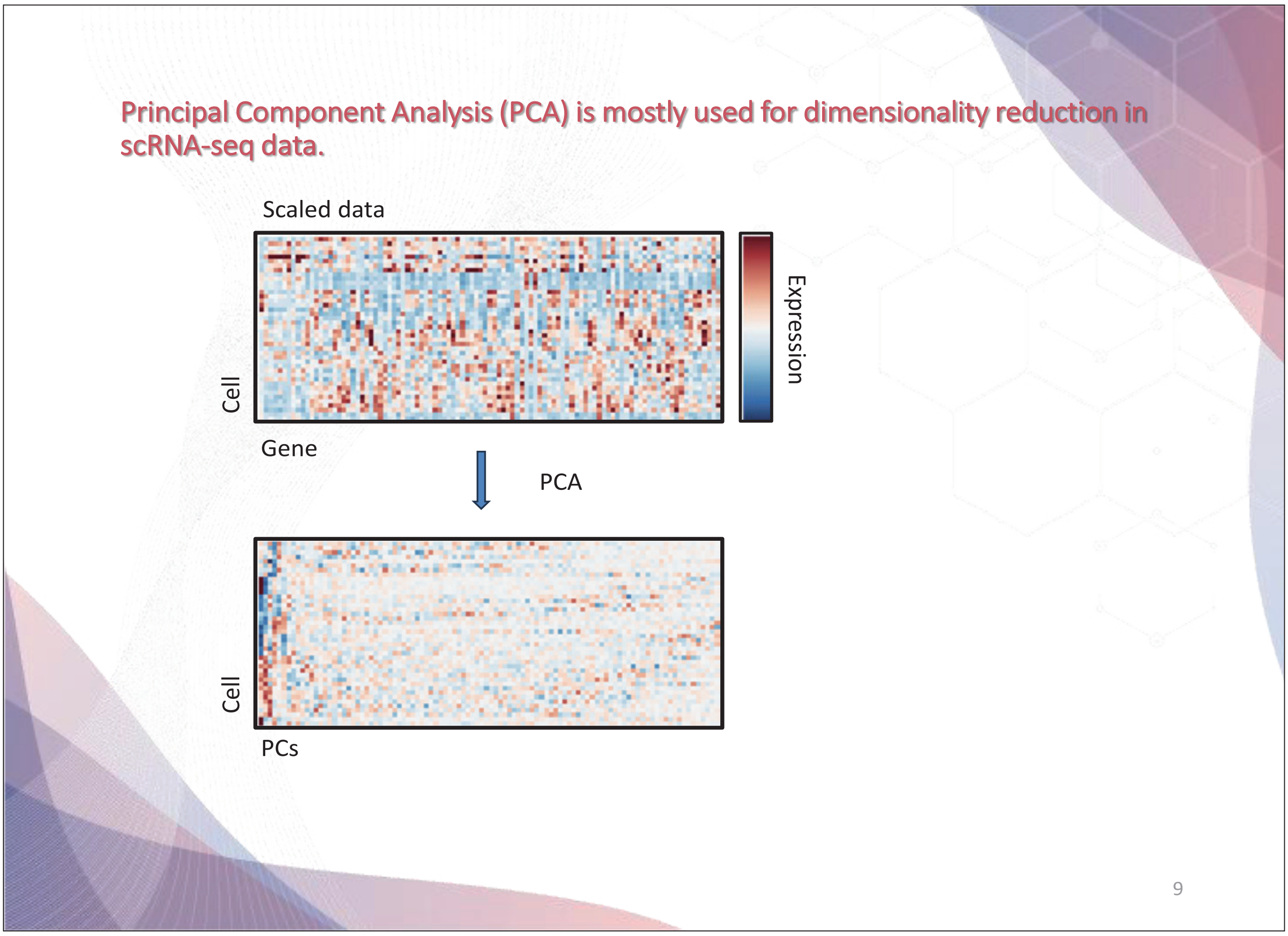
Principal Component Analysis (PCA) is mostly used for dimensionality reduction in scRNA-seq data.

Noisy and High-Dimensional data
1k ~ 1M cells
~20,000 genes
expression
$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m} \times \Sigma_{m\times n} \times \mathbf{V}^T_{n\times n}$
$\square^2 = \mathrm{Var}$
$\vec{v}$

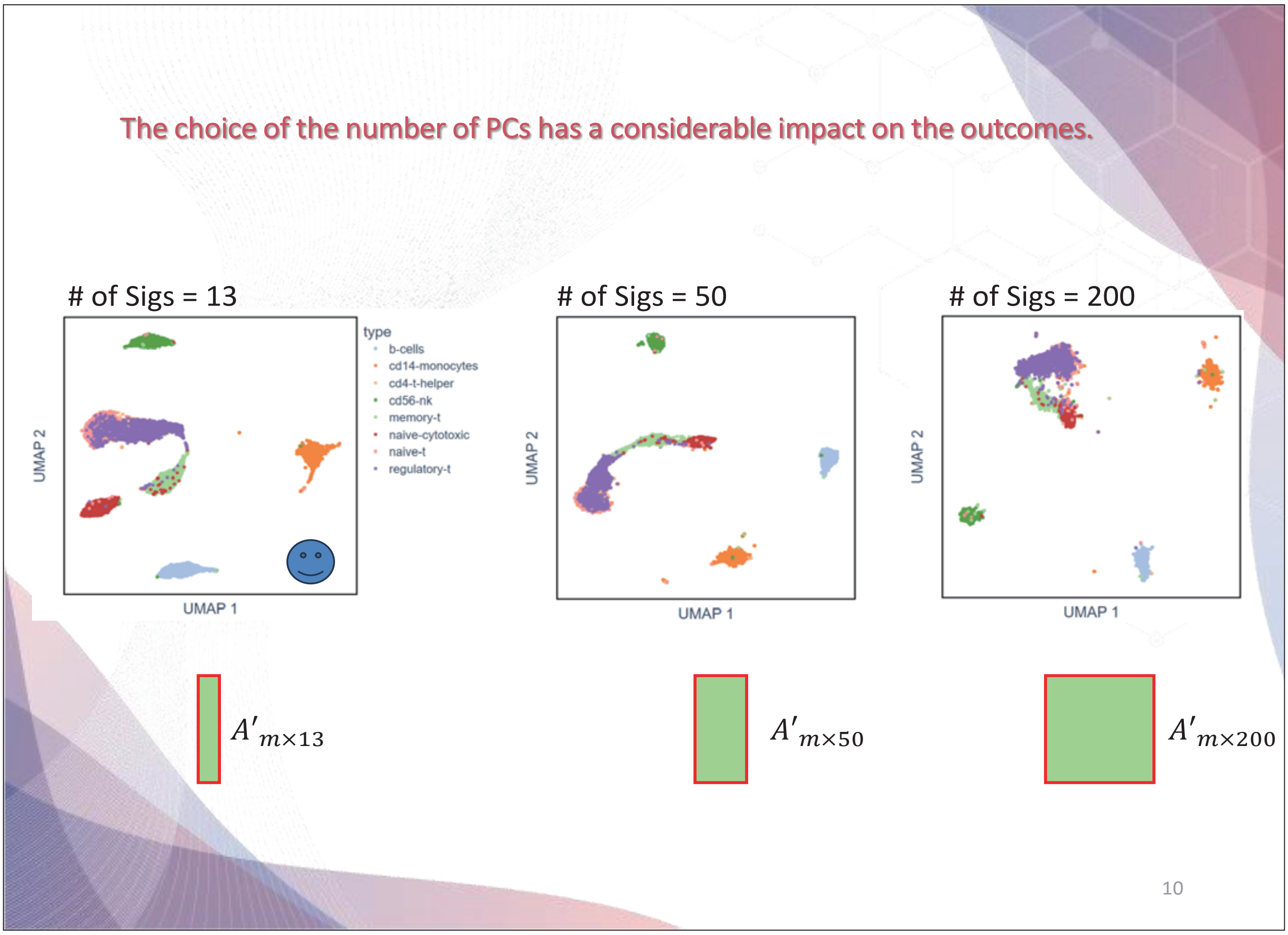


Principal Component Analysis (PCA) is mostly used for dimensionality reduction in scRNA-seq data.

Noisy and High-Dimensional data
1k ~ 1M cells
~20,000 genes
expression
$A'_{m\times r}$
$\boldsymbol{r}$ Singular values
$\boldsymbol{r}$ Principal components (PCs)
$U'_{m\times r}$
$\Sigma'_{r\times r}$
$V^{T\prime}_{r\times n}$
$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m} \times \Sigma_{m\times n} \times \mathbf{V}^T_{n\times n}$
$\square^2 = \mathrm{Var}$
$\boldsymbol{r}$ Signals
Noise
$A'_{m\times r} = U'_{m\times r} \times \sigma'_{r\times r}$
8

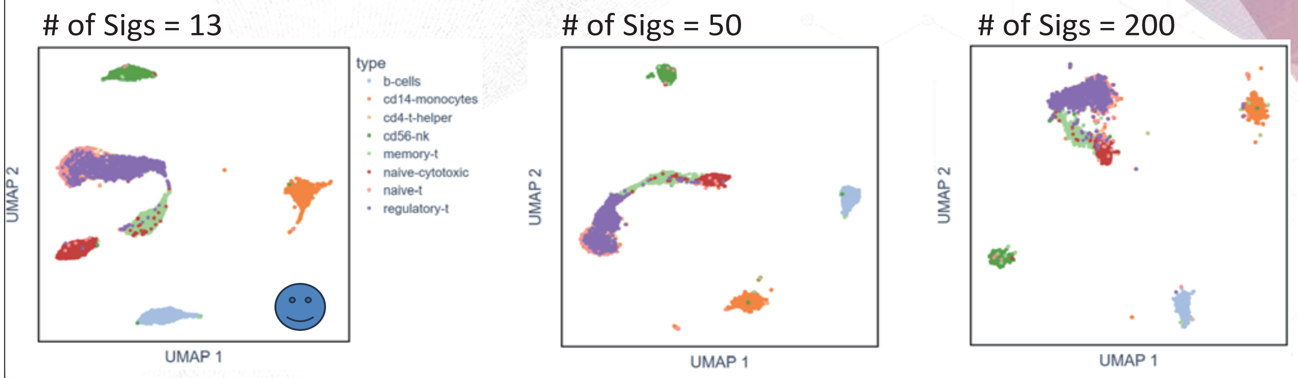


Principal Component Analysis (PCA) is mostly used for dimensionality reduction in scRNA-seq data.



The choice of the number of PCs has a considerable impact on the outcomes.

## Lack of consensus in selecting number of signals

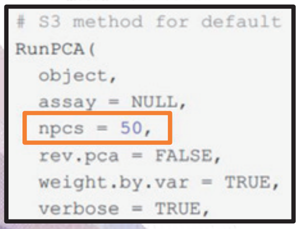


# of Sigs = 13

# of Sigs = 50

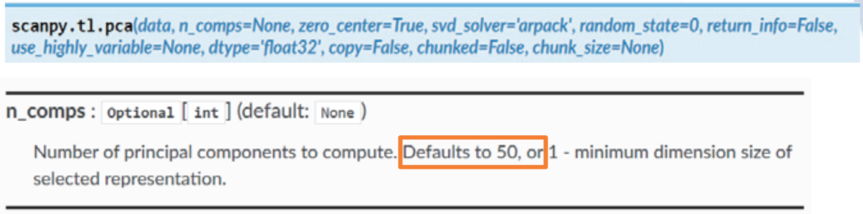
# of Sigs = 200

Seurat



```
# S3 method for default
RunPCA(
  object,
  assay = NULL,
  npcs = 50,
  rev.pca = FALSE,
  weight.by.var = TRUE,
  verbose = TRUE,
```

Scanpy



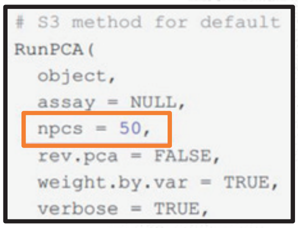
`scanpy.tl.pca(data, n_comps=None, zero_center=True, svd_solver='arpack', random_state=0, return_info=False, use_highly_variable=None, dtype='float32', copy=False, chunked=False, chunk_size=None)`

**n_comps** : `Optional` [ `int` ] (default: `None` )

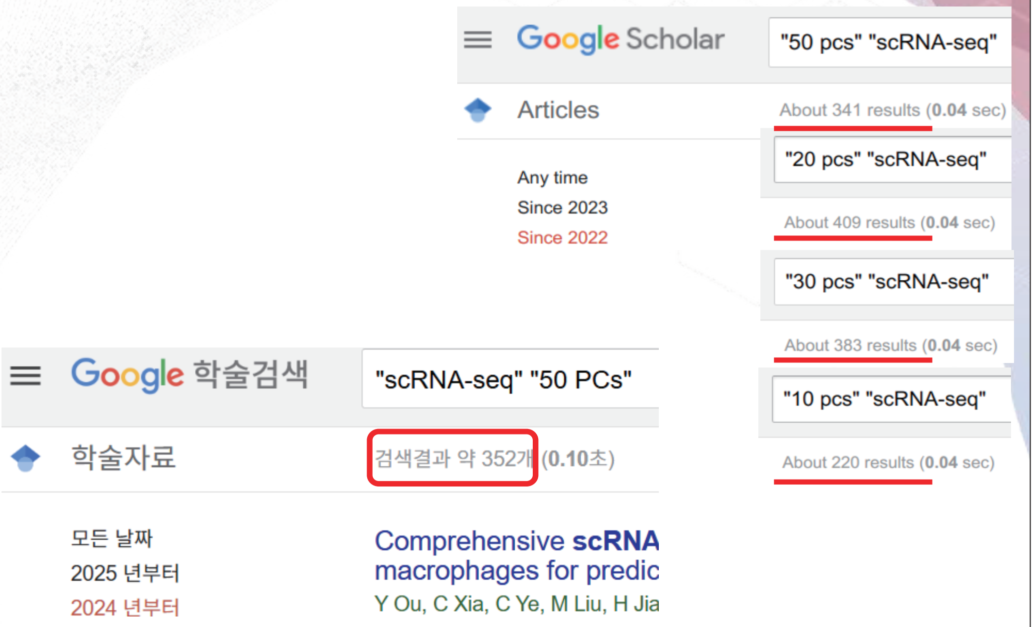
Number of principal components to compute. Defaults to 50, or 1 - minimum dimension size of selected representation.

## User subjectivity in determining the number of PCs can compromise the reliability of outcomes

Seurat



```
# S3 method for default
RunPCA(
  object,
  assay = NULL,
  npcs = 50,
  rev.pca = FALSE,
  weight.by.var = TRUE,
  verbose = TRUE,
```



"50 pcs" "scRNA-seq" — About 341 results (0.04 sec)

"20 pcs" "scRNA-seq" — About 409 results (0.04 sec)

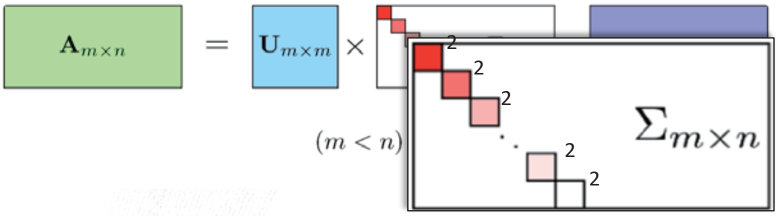
"30 pcs" "scRNA-seq" — About 383 results (0.04 sec)

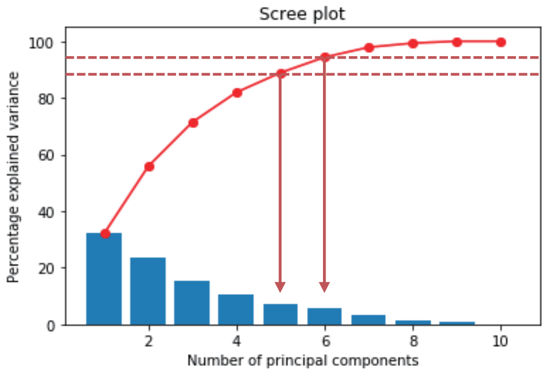
"10 pcs" "scRNA-seq" — About 220 results (0.04 sec)

"scRNA-seq" "50 PCs" — 검색결과 약 352개 (0.10초)

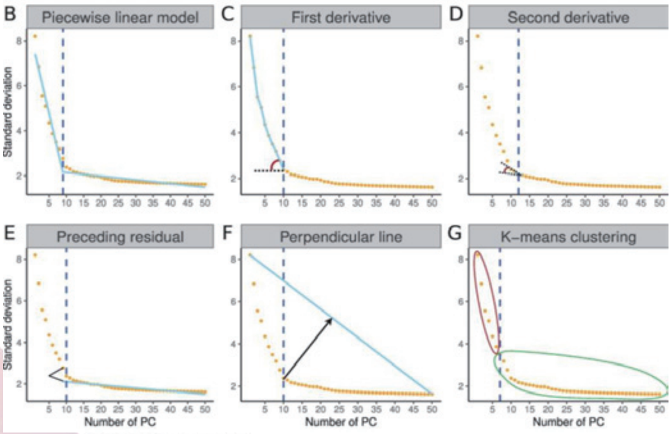
Comprehensive scRNA
macrophages for predic
Y Ou, C Xia, C Ye, M Liu, H Jia

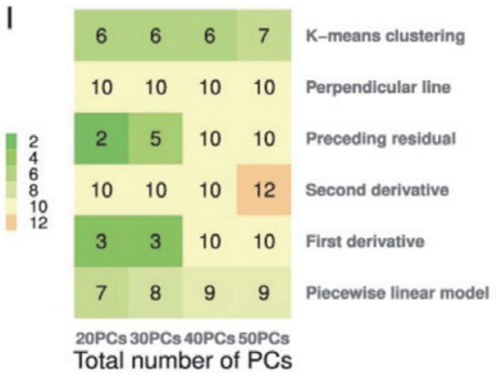
## Reducing subjectivity in determining the number of signals: Elbow method and Variance explained criterion.





## Reducing subjectivity in determining the number of signals: Elbow method and Variance explained criterion.





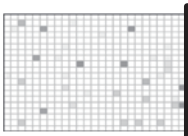
| 20PCs | 30PCs | 40PCs | 50PCs | |
|---|---|---|---|---|
| 6 | 6 | 6 | 7 | K−means clustering |
| 10 | 10 | 10 | 10 | Perpendicular line |
| 2 | 5 | 10 | 10 | Preceding residual |
| 10 | 10 | 10 | 12 | Second derivative |
| 3 | 3 | 10 | 10 | First derivative |
| 7 | 8 | 9 | 9 | Piecewise linear model |

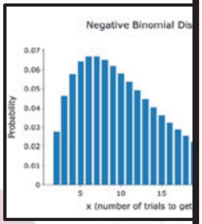
Total number of PCs

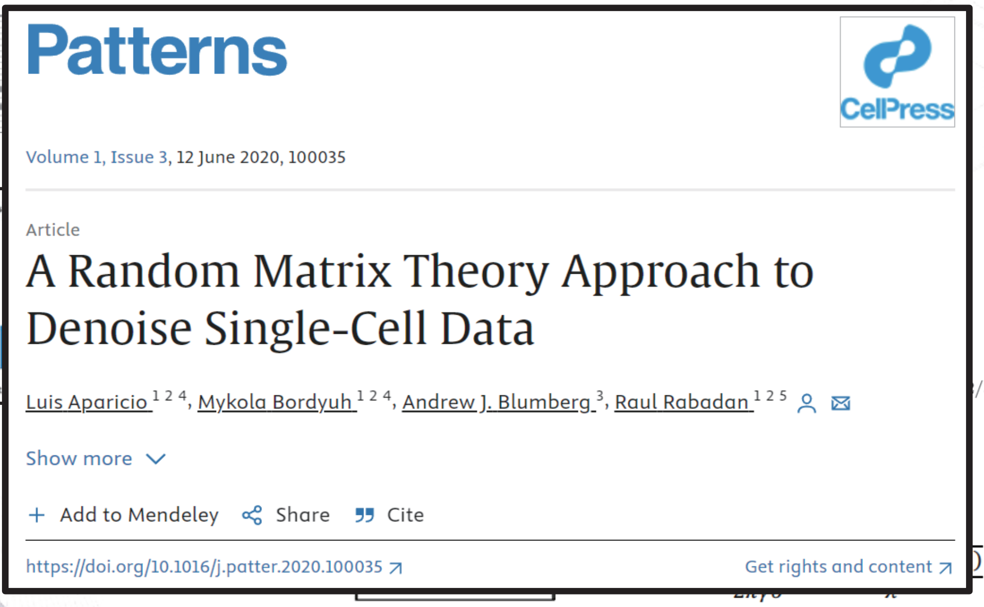
These methods still rely partly on user subjectivity!

## Random matrix theory (RMT) provides an objective threshold

Random matrix







**Patterns**

CellPress

Volume 1, Issue 3, 12 June 2020, 100035

Article

A Random Matrix Theory Approach to Denoise Single-Cell Data

Luis Aparicio [1 2 4], Mykola Bordyuh [1 2 4], Andrew J. Blumberg [3], Raul Rabadan [1 2 5]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.patter.2020.100035

Get rights and content

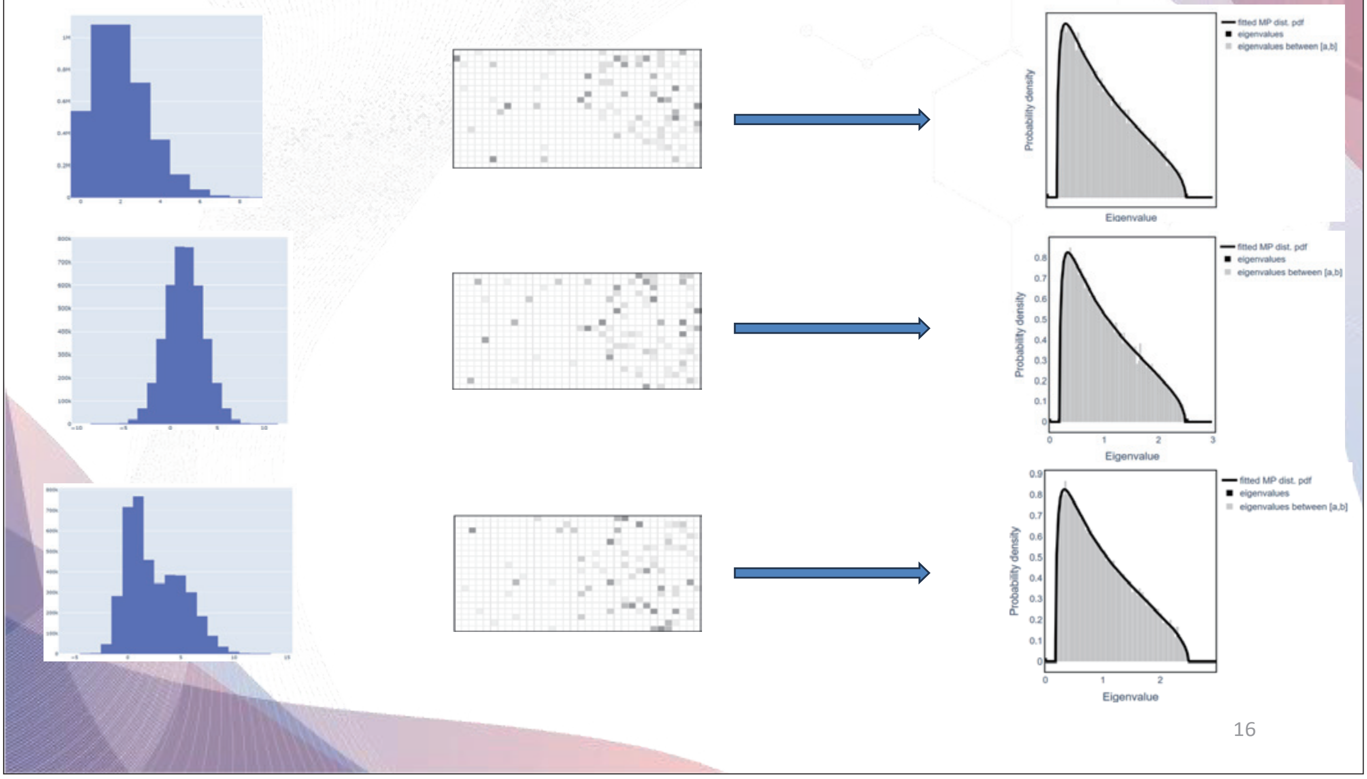
$^{/3}(t - a_m))$

$\mathbb{I}_{[a_-,a_+]}$

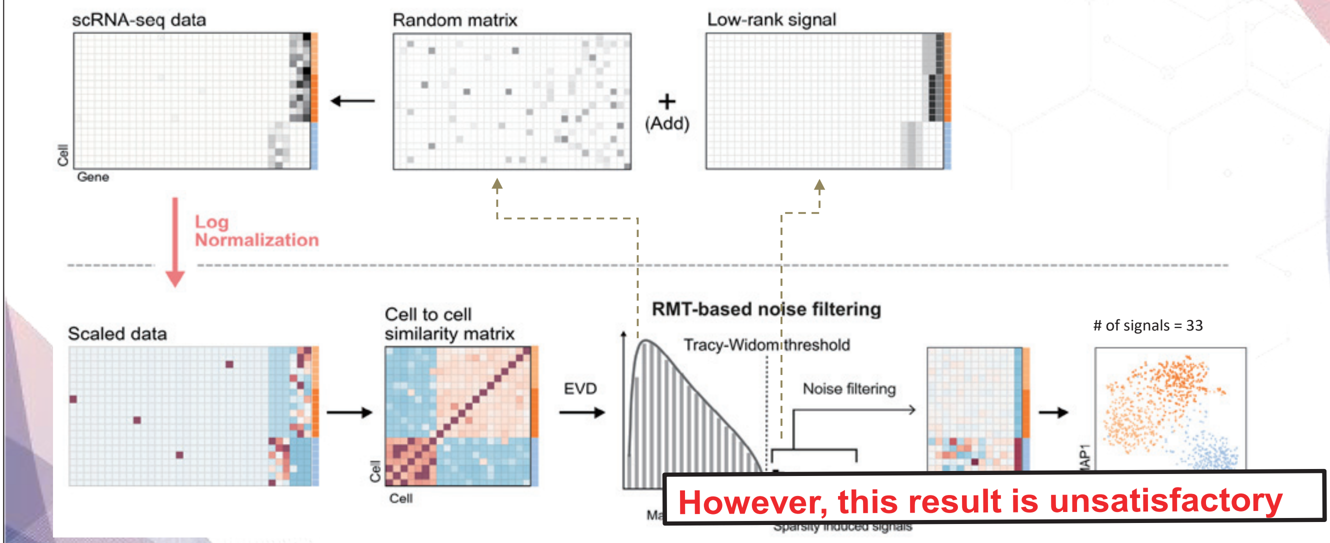
Eigenvalue

where $a_{\pm} = \sigma^2(1 \pm \sqrt{\gamma})^2, \quad \gamma = m/n$

## Random matrix theory's universality allows for noise analysis without the need for specific models.



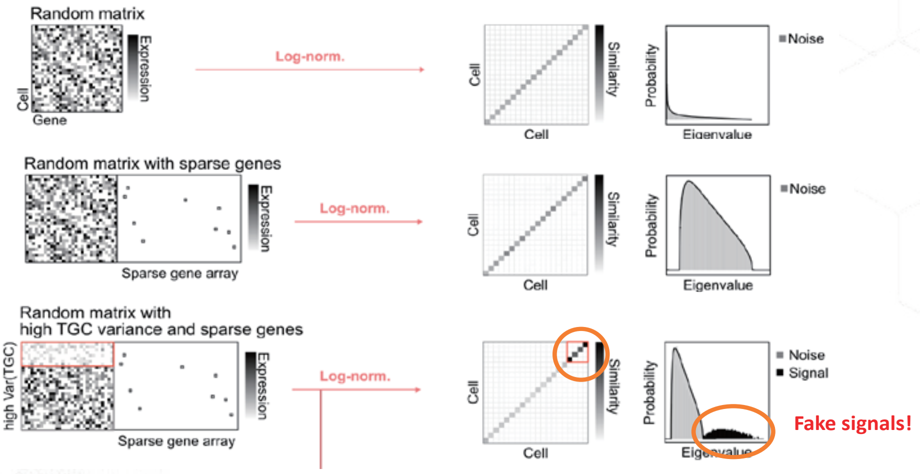
## Random matrix theory (RMT) provides a objective threshold



However, this result is unsatisfactory

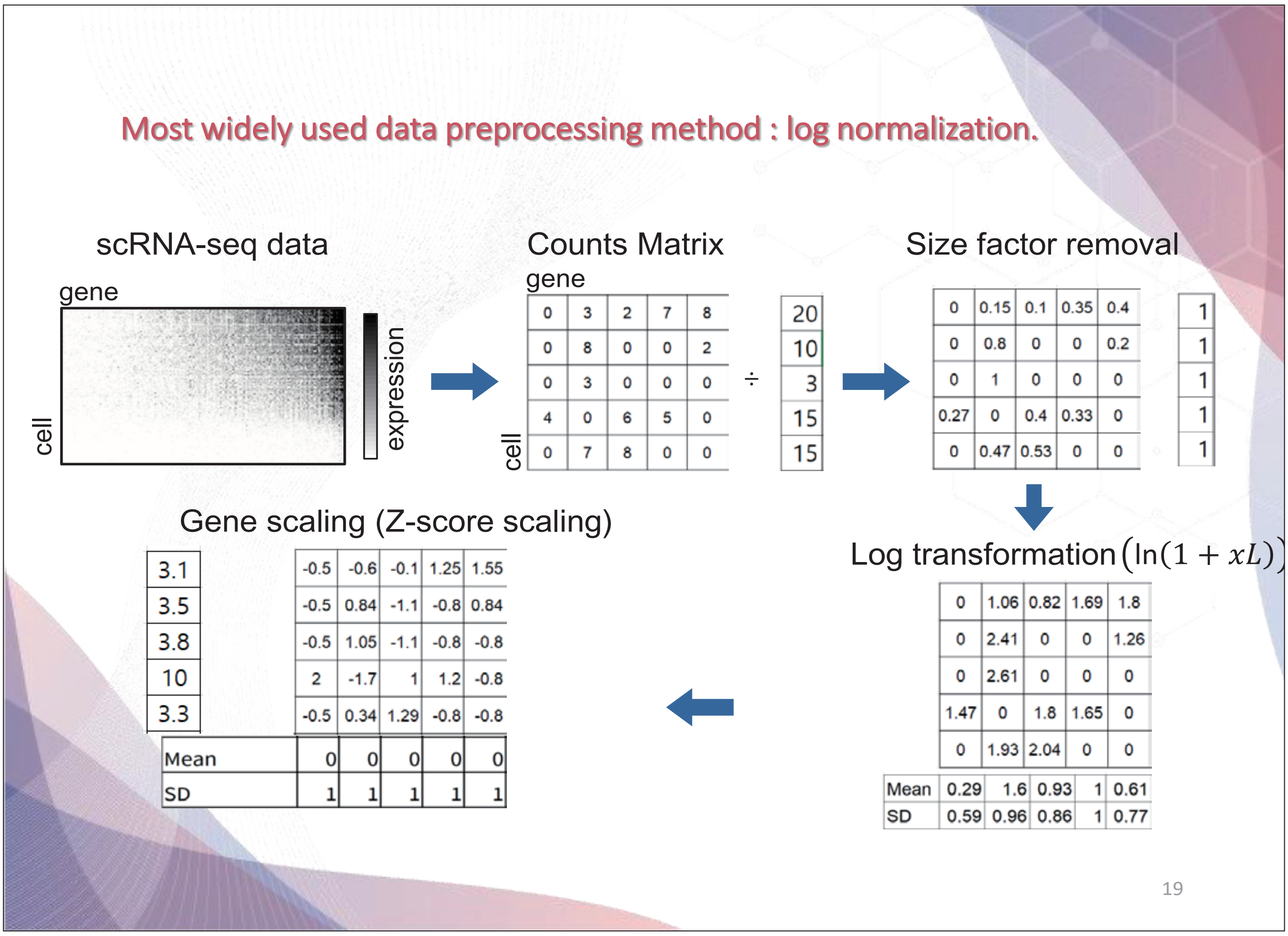
Luis Aparicio et al. 2020, Patterns; Kim et al. 2024. Nat. Commun.

## High variability among samples and the sparseness of data lead to false signals following log normalization.



Fake signals!

Kim et al. 2024. Nat. Commun.



## Most widely used data preprocessing method : log normalization.

scRNA-seq data

Counts Matrix

gene (columns), cell (rows)

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| 0 | 3 | 2 | 7 | 8 | ÷ | 20 |
| 0 | 8 | 0 | 0 | 2 |   | 10 |
| 0 | 3 | 0 | 0 | 0 |   | 3 |
| 4 | 0 | 6 | 5 | 0 |   | 15 |
| 0 | 7 | 8 | 0 | 0 |   | 15 |

Size factor removal

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 0 | 0.15 | 0.1 | 0.35 | 0.4 | 1 |
| 0 | 0.8 | 0 | 0 | 0.2 | 1 |
| 0 | 1 | 0 | 0 | 0 | 1 |
| 0.27 | 0 | 0.4 | 0.33 | 0 | 1 |
| 0 | 0.47 | 0.53 | 0 | 0 | 1 |

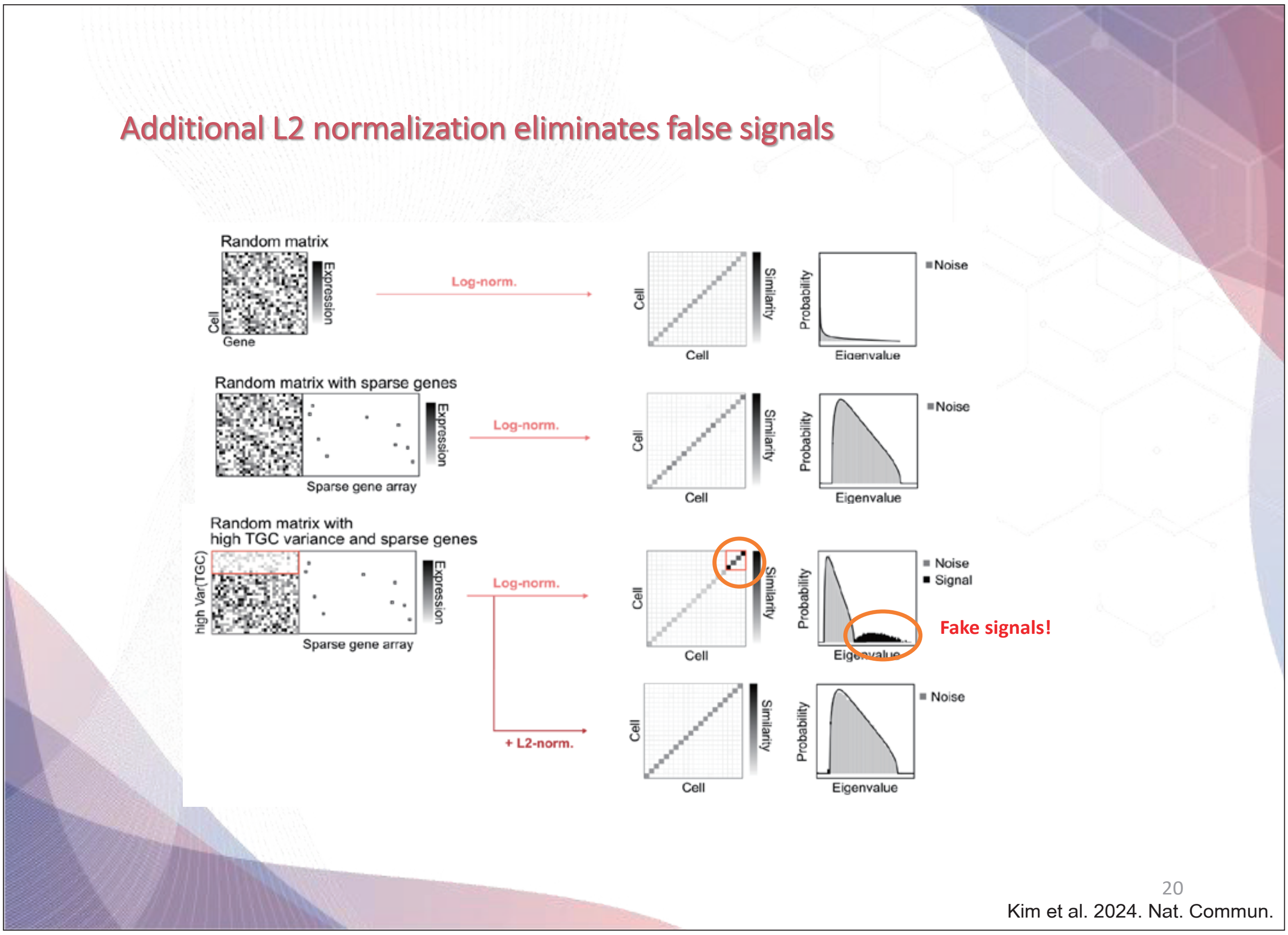
Log transformation $(\ln(1+xL))$

|   |   |   |   |   |   |
|---|---|---|---|---|---|
|   | 0 | 1.06 | 0.82 | 1.69 | 1.8 |
|   | 0 | 2.41 | 0 | 0 | 1.26 |
|   | 0 | 2.61 | 0 | 0 | 0 |
|   | 1.47 | 0 | 1.8 | 1.65 | 0 |
|   | 0 | 1.93 | 2.04 | 0 | 0 |
| Mean | 0.29 | 1.6 | 0.93 | 1 | 0.61 |
| SD | 0.59 | 0.96 | 0.86 | 1 | 0.77 |

Gene scaling (Z-score scaling)

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 3.1 | -0.5 | -0.6 | -0.1 | 1.25 | 1.55 |
| 3.5 | -0.5 | 0.84 | -1.1 | -0.8 | 0.84 |
| 3.8 | -0.5 | 1.05 | -1.1 | -0.8 | -0.8 |
| 10 | 2 | -1.7 | 1 | 1.2 | -0.8 |
| 3.3 | -0.5 | 0.34 | 1.29 | -0.8 | -0.8 |
| Mean | 0 | 0 | 0 | 0 | 0 |
| SD | 1 | 1 | 1 | 1 | 1 |

19



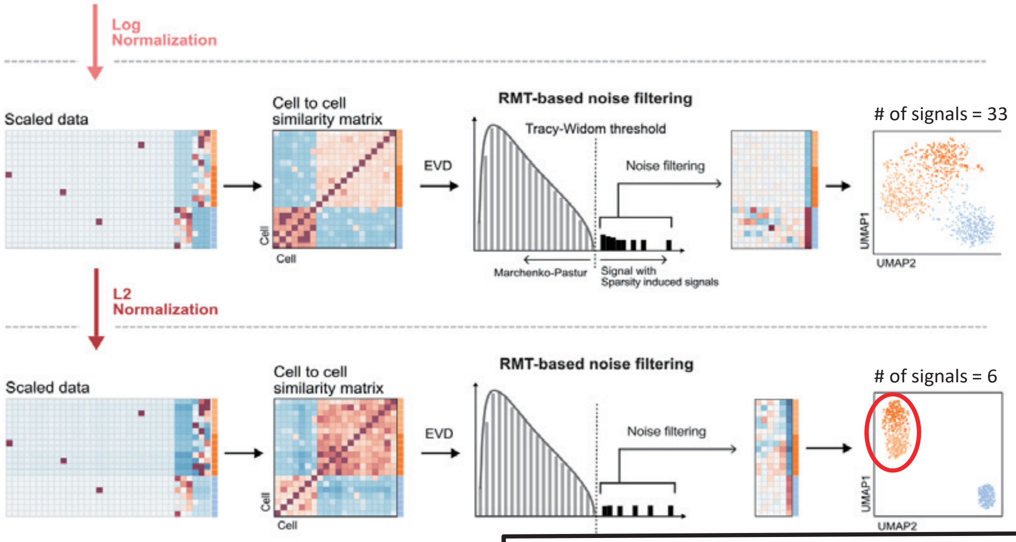
## Additional L2 normalization eliminates false signals

Fake signals!

20

Kim et al. 2024. Nat. Commun.

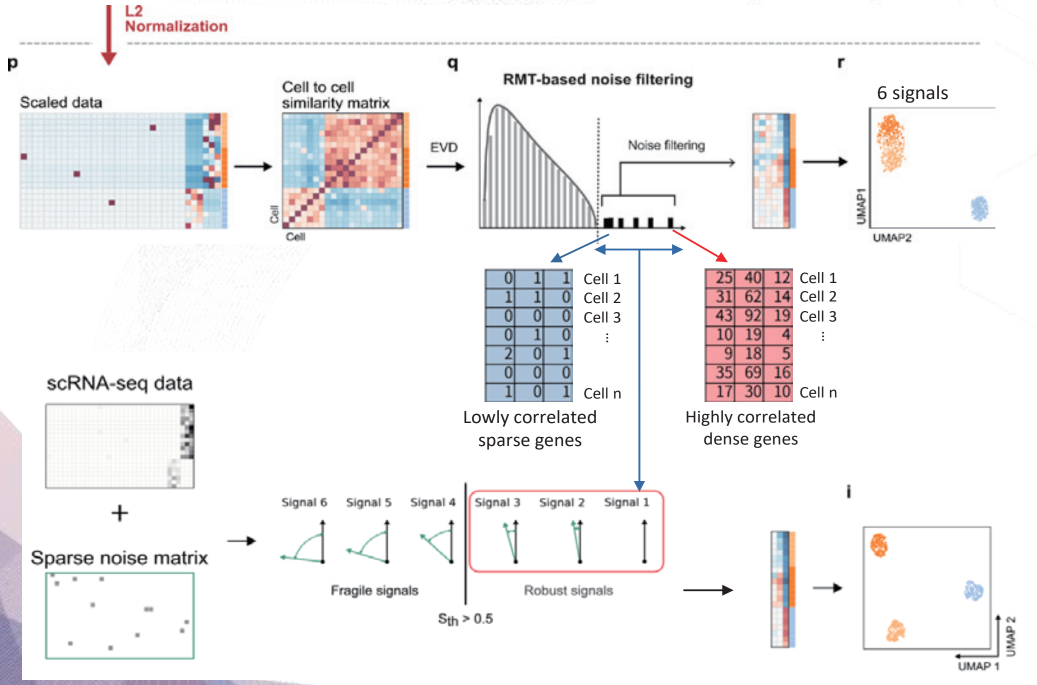
## L2-norm effectively corrects the signal distortion.



**However, this result is still unsatisfactory**

Kim et al. 2024. Nat. Commun.

## Low-quality signal filtering using a signal robustness test enhances the quality of outcomes.



| | | | |
|---|---|---|---|
| 0 | 1 | 1 | Cell 1 |
| 1 | 1 | 0 | Cell 2 |
| 0 | 0 | 0 | Cell 3 |
| 0 | 1 | 0 | ⋮ |
| 2 | 0 | 1 | |
| 0 | 0 | 0 | |
| 1 | 0 | 1 | Cell n |

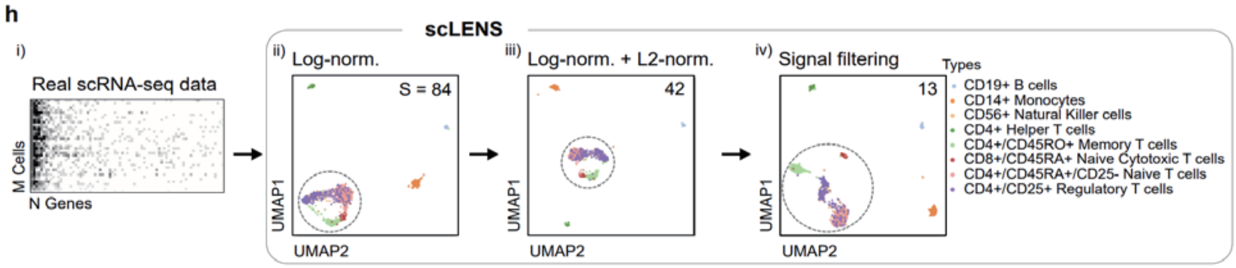
Lowly correlated sparse genes

| | | | |
|---|---|---|---|
| 25 | 40 | 12 | Cell 1 |
| 31 | 62 | 14 | Cell 2 |
| 43 | 92 | 19 | Cell 3 |
| 10 | 19 | 4 | ⋮ |
| 9 | 18 | 5 | |
| 35 | 69 | 16 | |
| 17 | 30 | 10 | Cell n |

Highly correlated dense genes

Kim et al. 2024. Nat. Commun.

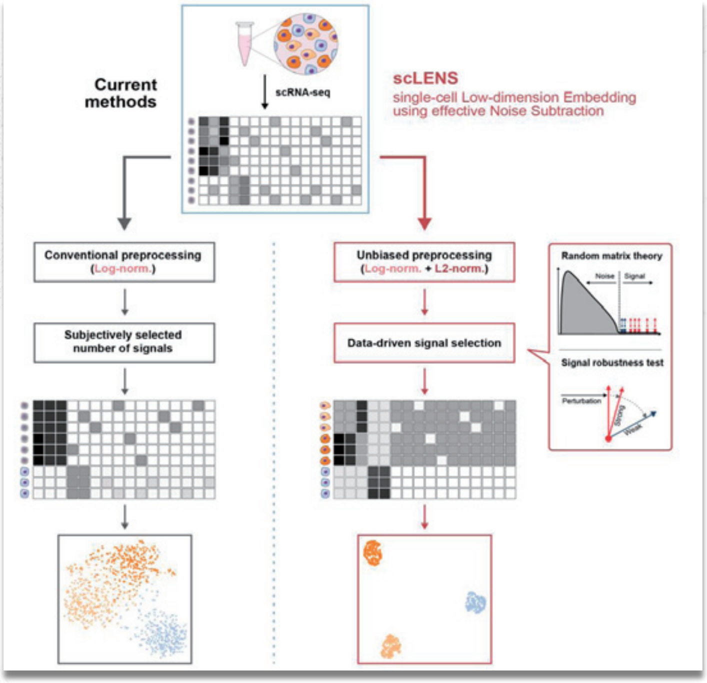
## single-cell Low-dimension Embedding using effective Nose Subtraction (scLENS)



23

Kim et al. 2024. Nat. Commun.

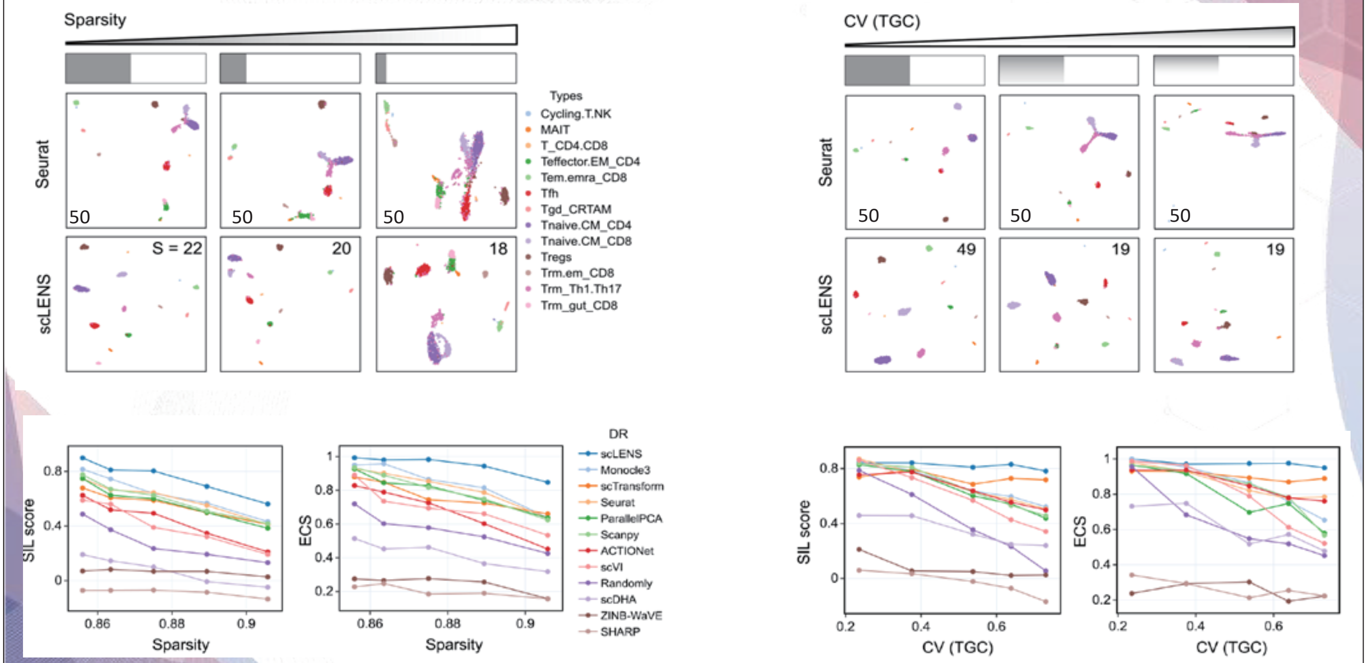
## single-cell Low-dimension Embedding using effective Nose Subtraction (scLENS)


Current methods
scRNA-seq
scLENS
single-cell Low-dimension Embedding using effective Noise Subtraction
Conventional preprocessing (Log-norm.)
Subjectively selected number of signals
Unbiased preprocessing (Log-norm. + L2-norm.)
Data-driven signal selection
Random matrix theory
Noise
Signal
Signal robustness test
Perturbation
Strong
Weak

24

Kim et al. 2024. Nat. Commun.

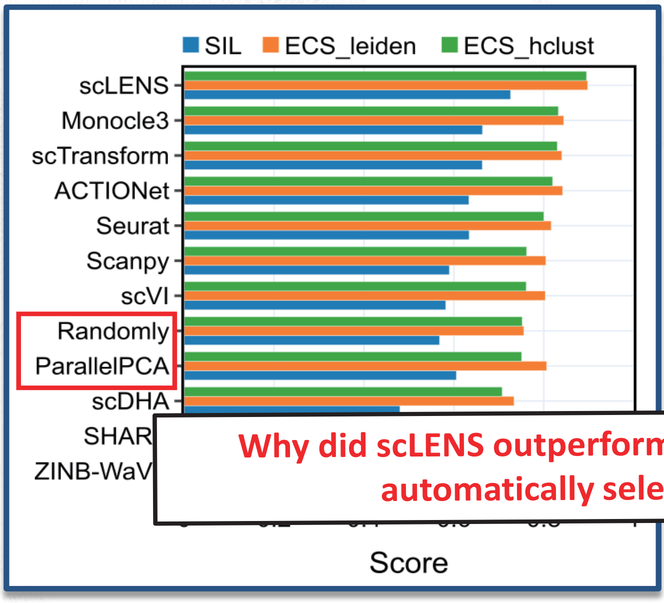
## scLENS outperforms others on the data with the high level of sparsity and skewness of data distribution



25

Kim et al. 2024. Nat. Commun.

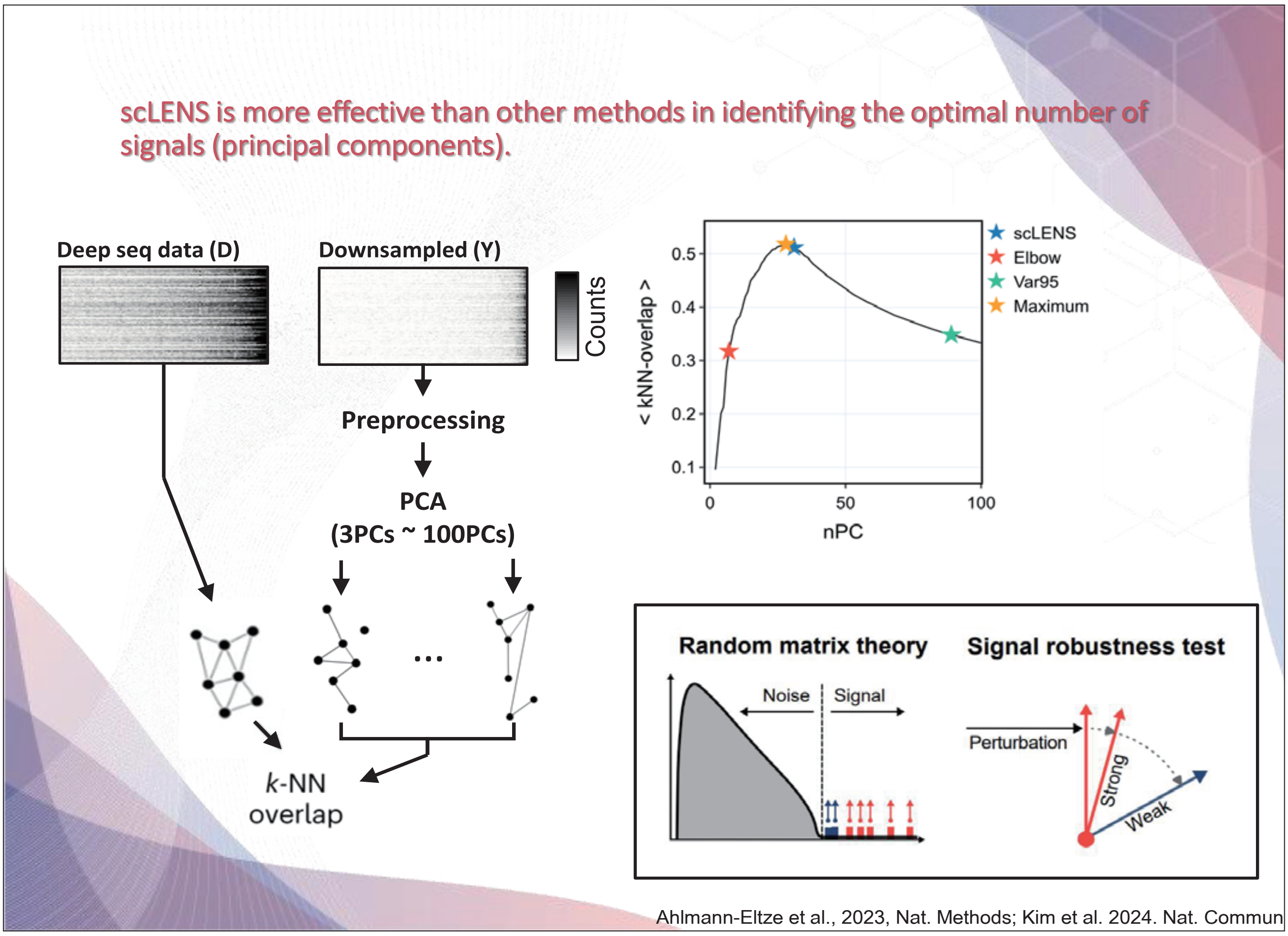
## scLENS outperforms other 11 packages.



Why did scLENS outperform others, which also automatically select signals?

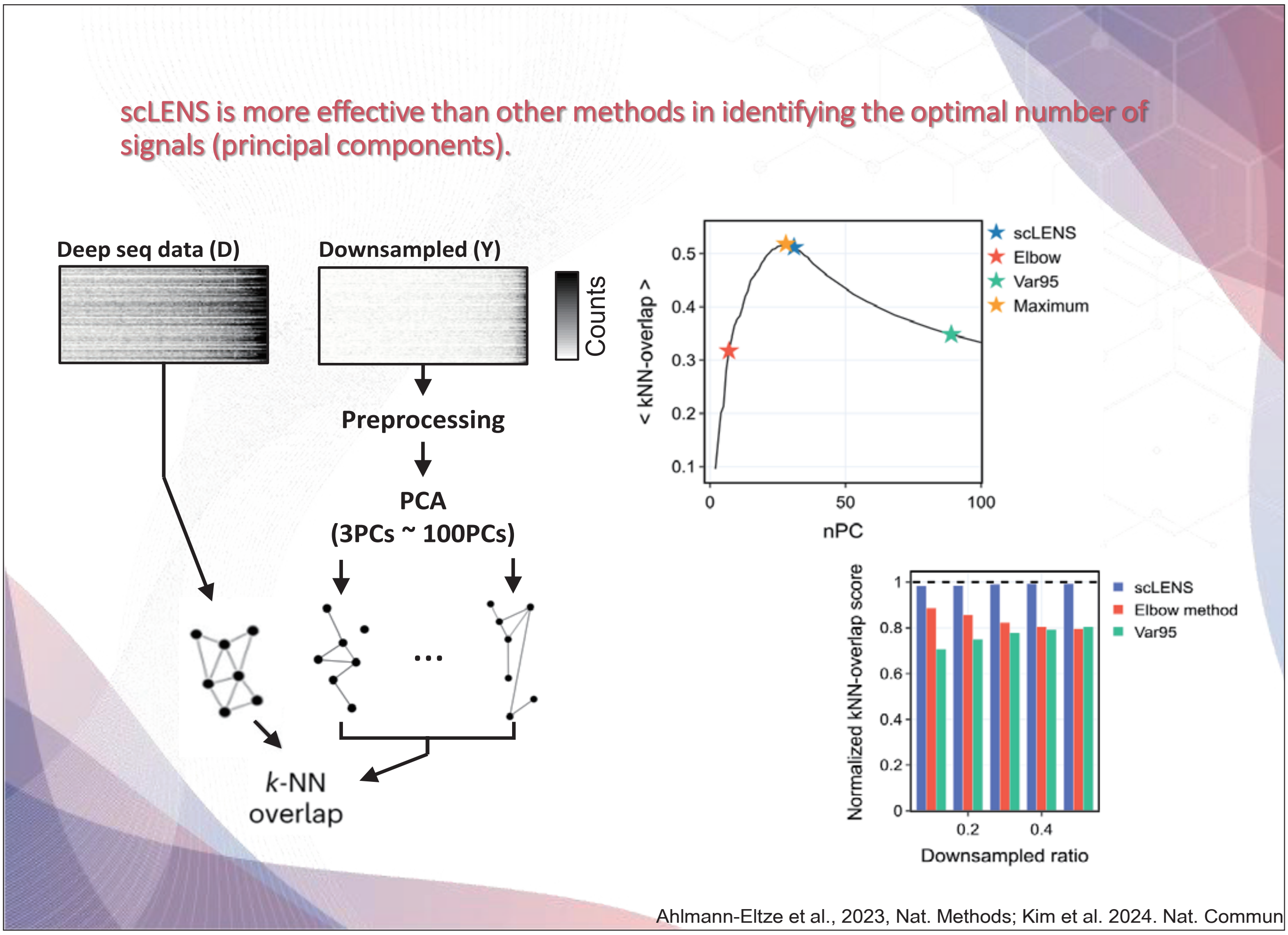
26

Kim et al. 2024. Nat. Commun.



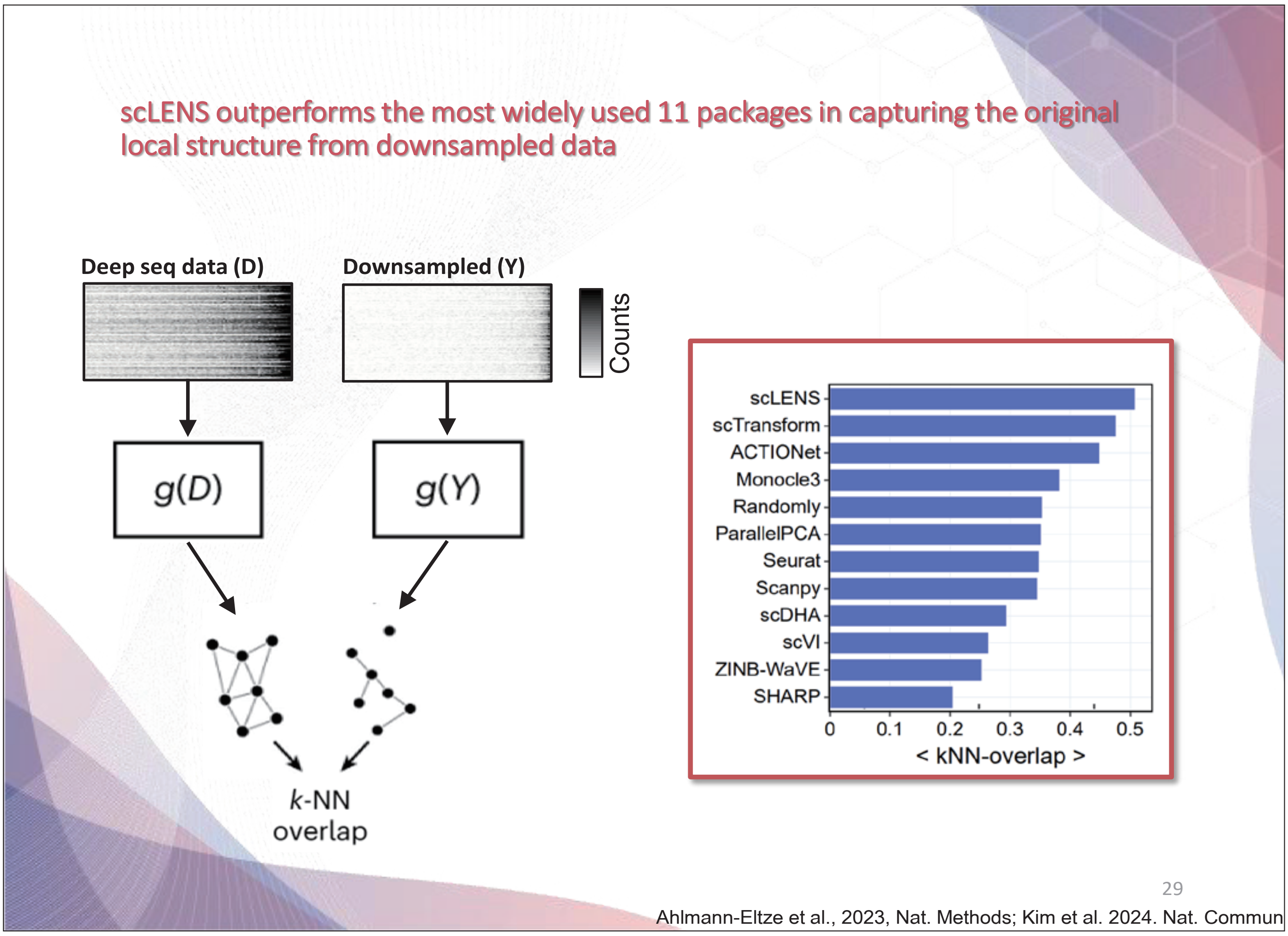
## scLENS is more effective than other methods in identifying the optimal number of signals (principal components).

Ahlmann-Eltze et al., 2023, Nat. Methods; Kim et al. 2024. Nat. Commun



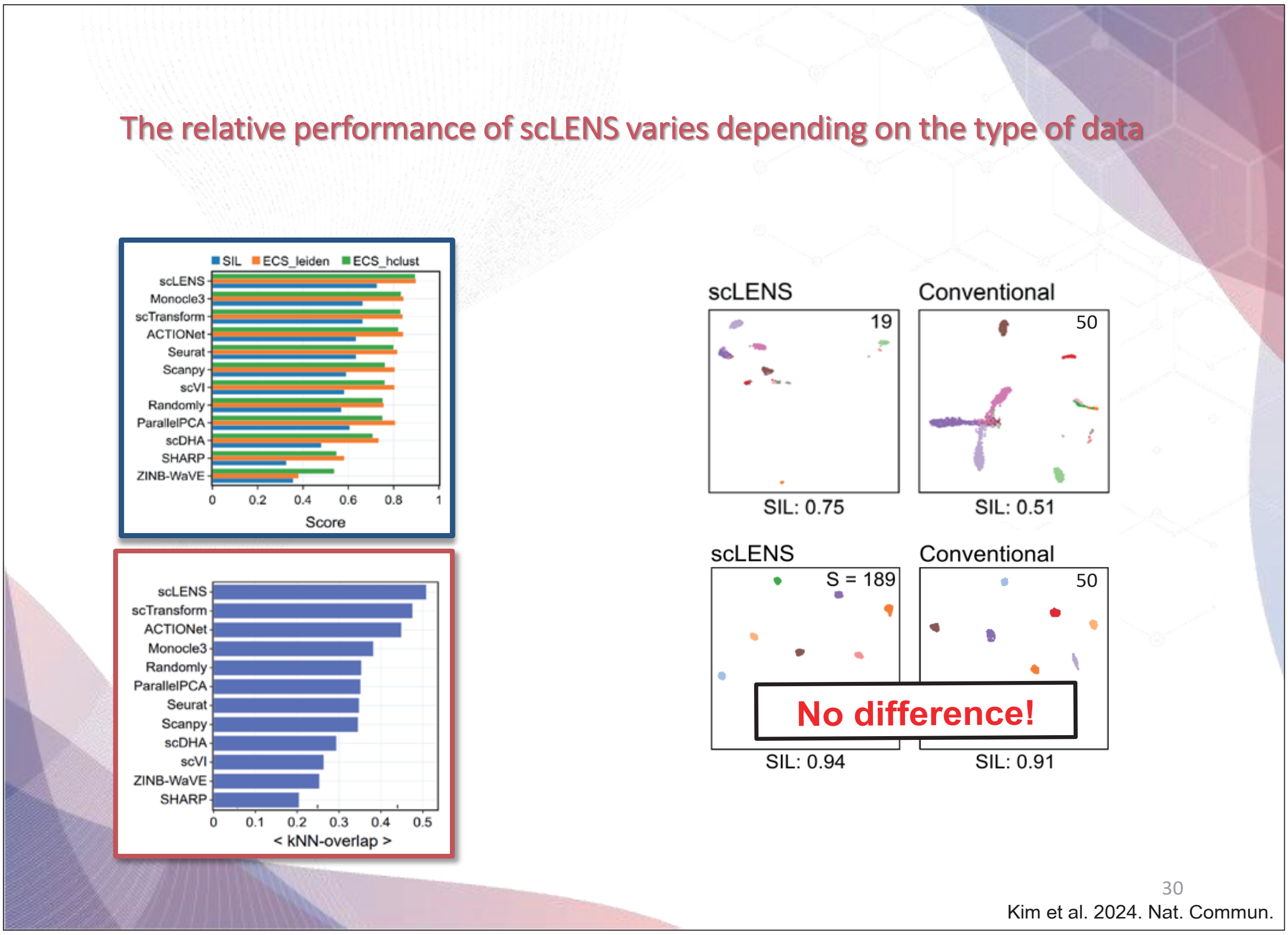
## scLENS is more effective than other methods in identifying the optimal number of signals (principal components).

Ahlmann-Eltze et al., 2023, Nat. Methods; Kim et al. 2024. Nat. Commun



## scLENS outperforms the most widely used 11 packages in capturing the original local structure from downsampled data

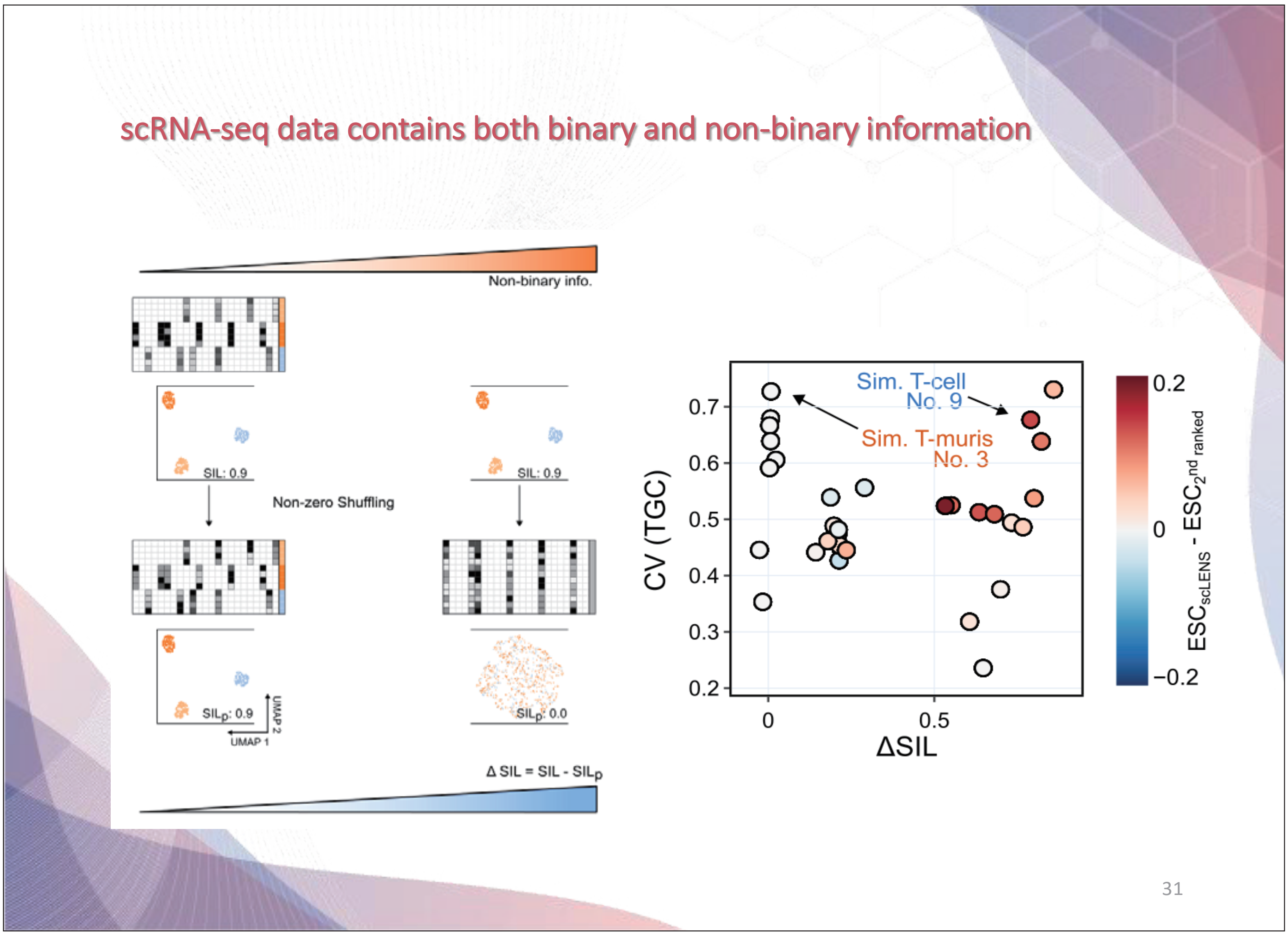
Ahlmann-Eltze et al., 2023, Nat. Methods; Kim et al. 2024. Nat. Commun



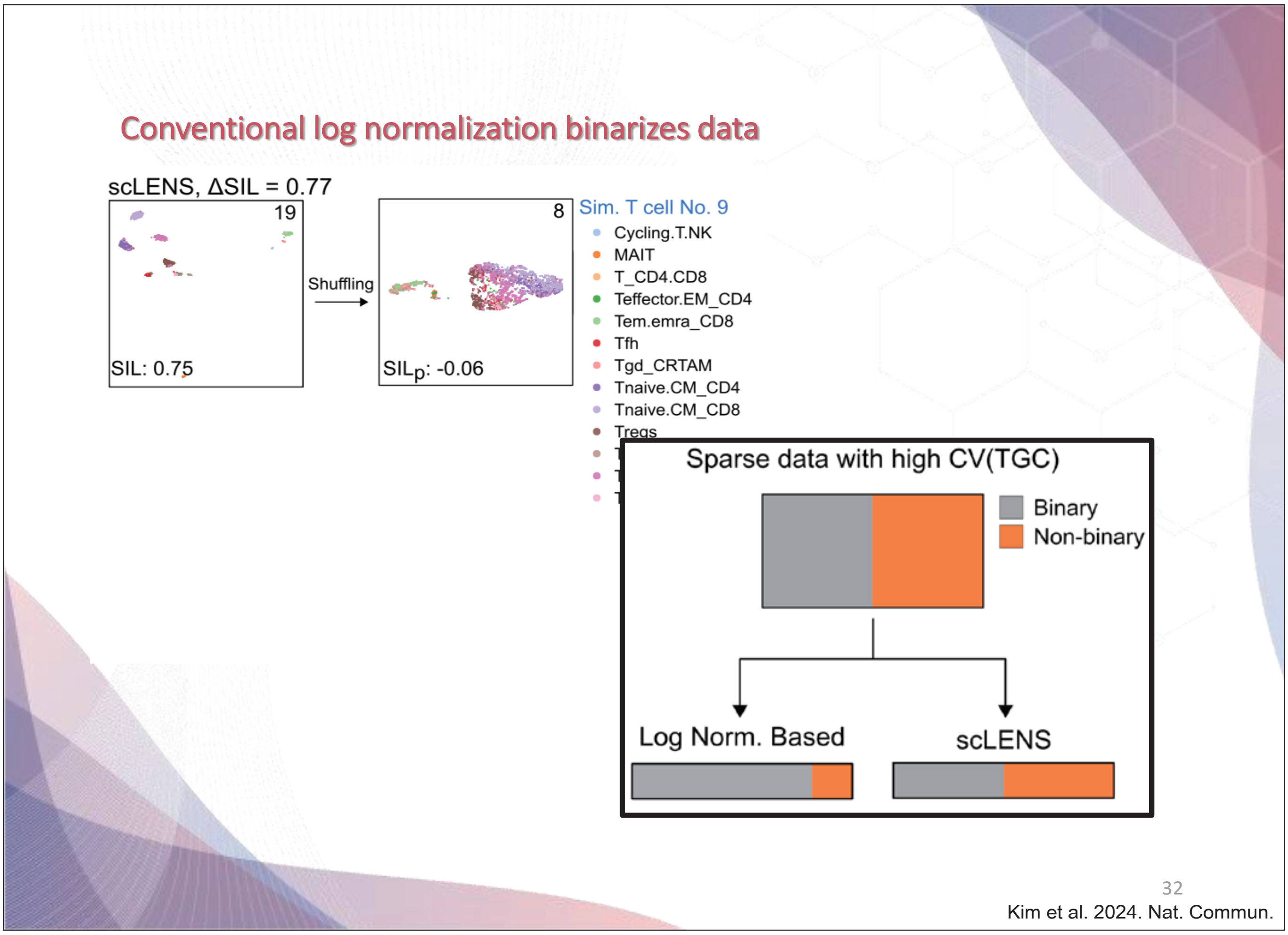
## The relative performance of scLENS varies depending on the type of data

No difference!

Kim et al. 2024. Nat. Commun.



## scRNA-seq data contains both binary and non-binary information



## Conventional log normalization binarizes data

Kim et al. 2024. Nat. Commun.

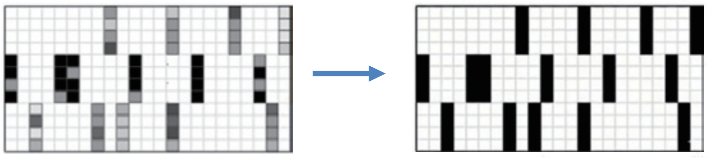
## The choice of the number of PCs has a considerable impact on the outcomes.

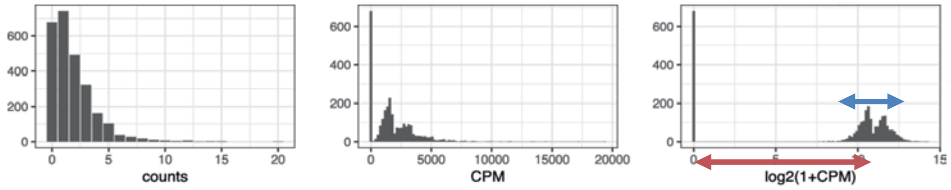
$$X_{ij}^{\text{log-trans}} = \log\left(1 + \frac{X_{ij}^{\text{raw}}}{\sum_j X_{ij}^{\text{raw}}} L\right)$$

***L*: Scaling factor**

Seurat: 10,000

Scanpy: median(TGC)





(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Kim et al., Nat. Commun.; F. William Townes et al., Genome Biol , 2019

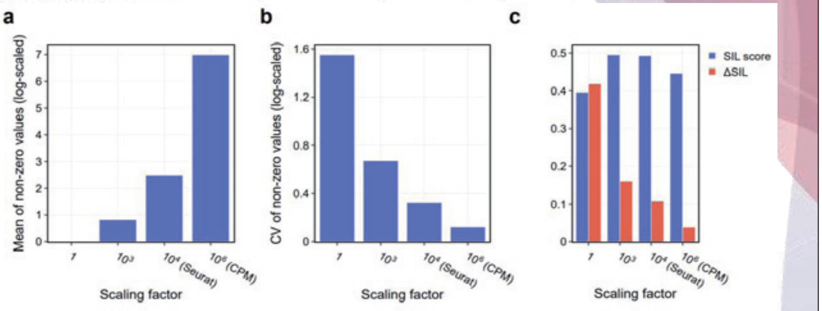
## During log normalization, multiplying a large scaling factor binarizes data

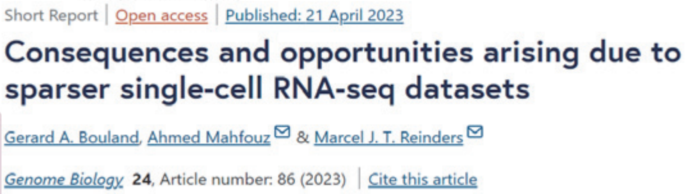
$$X_{ij}^{\text{log-trans}} = \log\left(1 + \frac{X_{ij}^{\text{raw}}}{\sum_j X_{ij}^{\text{raw}}} L\right)$$

***L*: Scaling factor**

Seurat: 10,000

Scanpy: median(TGC)



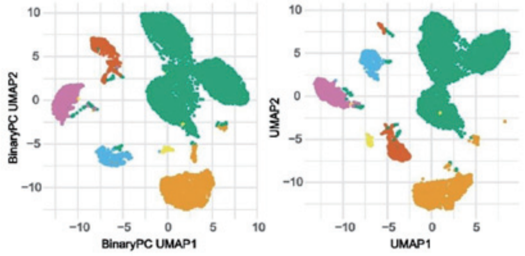


Short Report | Open access | Published: 21 April 2023

Consequences and opportunities arising due to sparser single-cell RNA-seq datasets

Gerard A. Bouland, Ahmed Mahfouz ✉ & Marcel J. T. Reinders ✉

Genome Biology 24, Article number: 86 (2023) | Cite this article

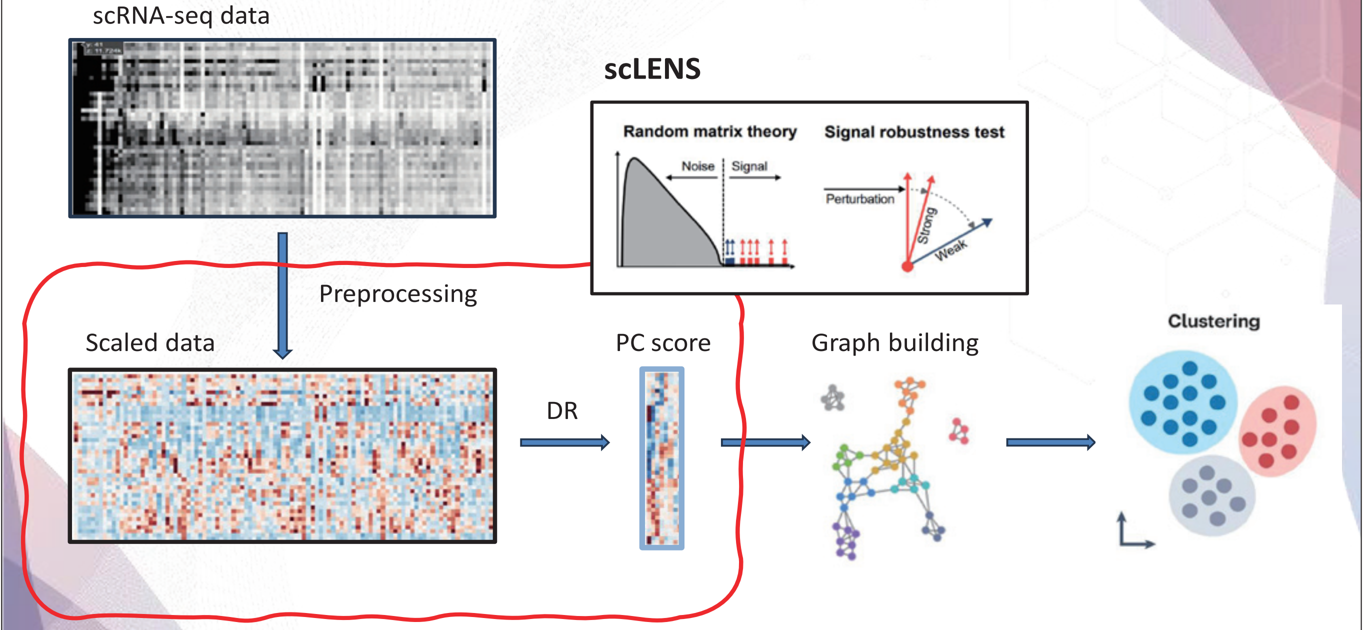


Kim et al., Nat. Commun.; Bouland et al., Genome Biol , 2023

## Summary

- When using PCA, the selection of meaningful principal components has traditionally been based on user experience, but this can lead to issues with the reproducibility and reliability of results.
- Log-normalization can cause signal distortion in sparse data with large differences between samples.
- The large scaling factor used in log-normalization tends to binarize the data.
- scLENS, through appropriate preprocessing, corrects for signal distortion that may arise due to data properties and eliminates user bias in PC determination, thereby providing more accurate and reliable results.

## After accurate dimensionality reduction, the next crucial step applied in scRNA-seq data analysis is clustering analysis.



scRNA-seq data

scLENS

Random matrix theory

Signal robustness test

Preprocessing

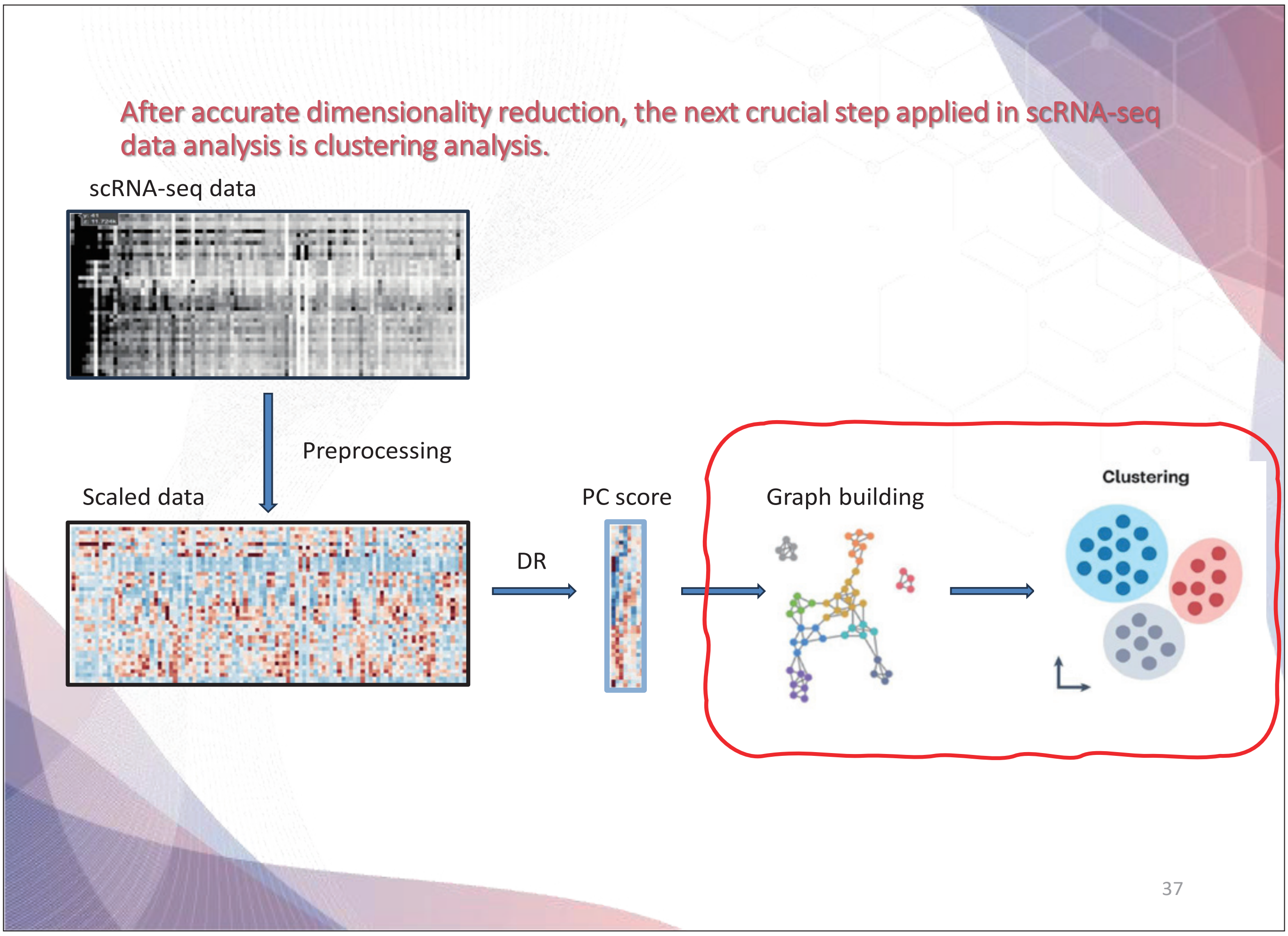
Scaled data

DR

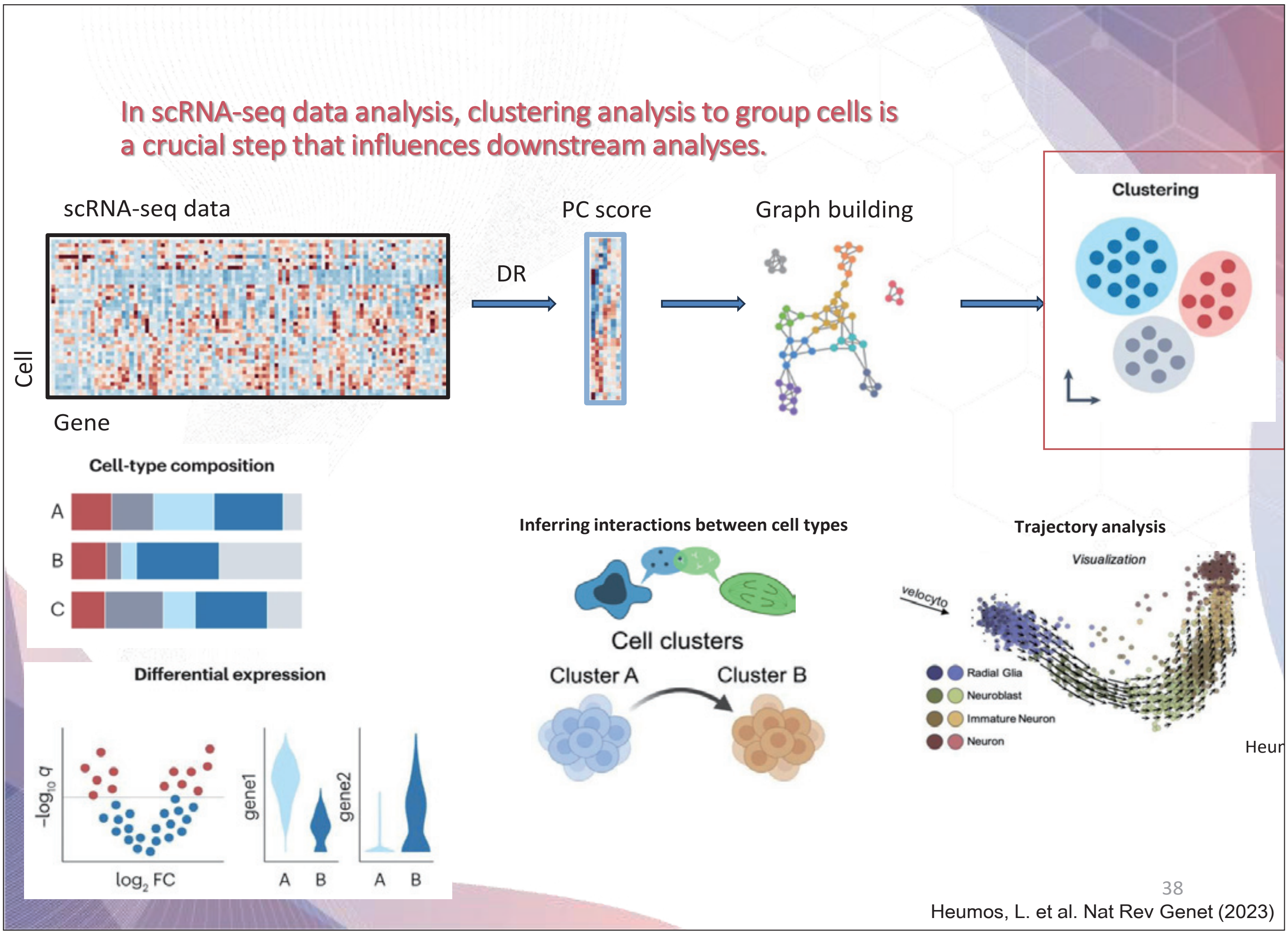
PC score

Graph building

Clustering



After accurate dimensionality reduction, the next crucial step applied in scRNA-seq data analysis is clustering analysis.

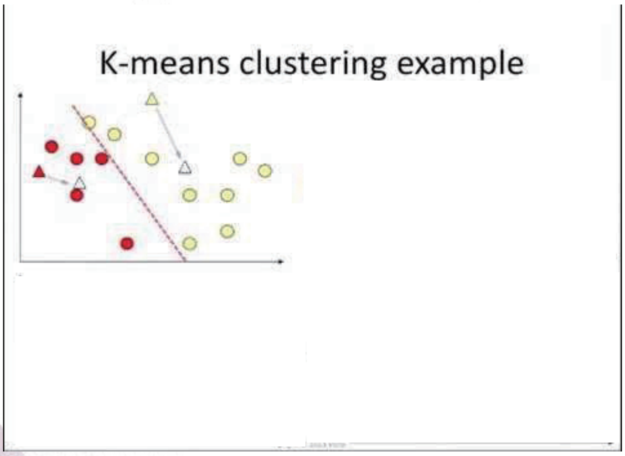


In scRNA-seq data analysis, clustering analysis to group cells is a crucial step that influences downstream analyses.

Heumos, L. et al. Nat Rev Genet (2023)

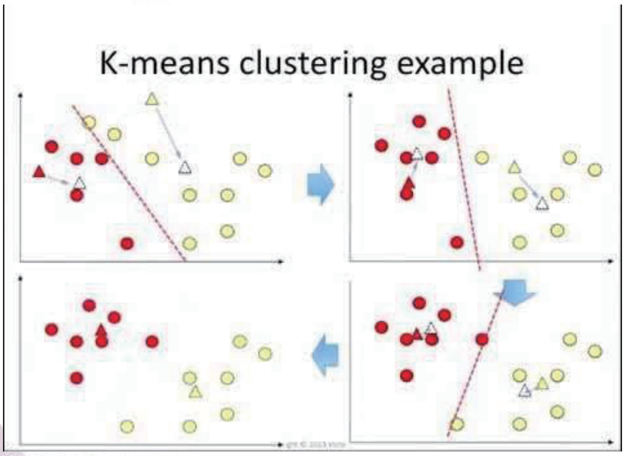
## The Basic Tool for Clustering Analysis: K-means Clustering

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

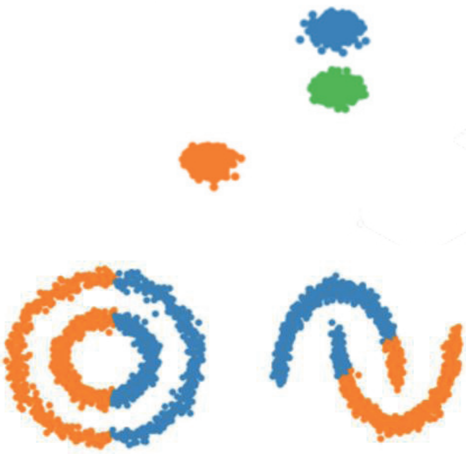


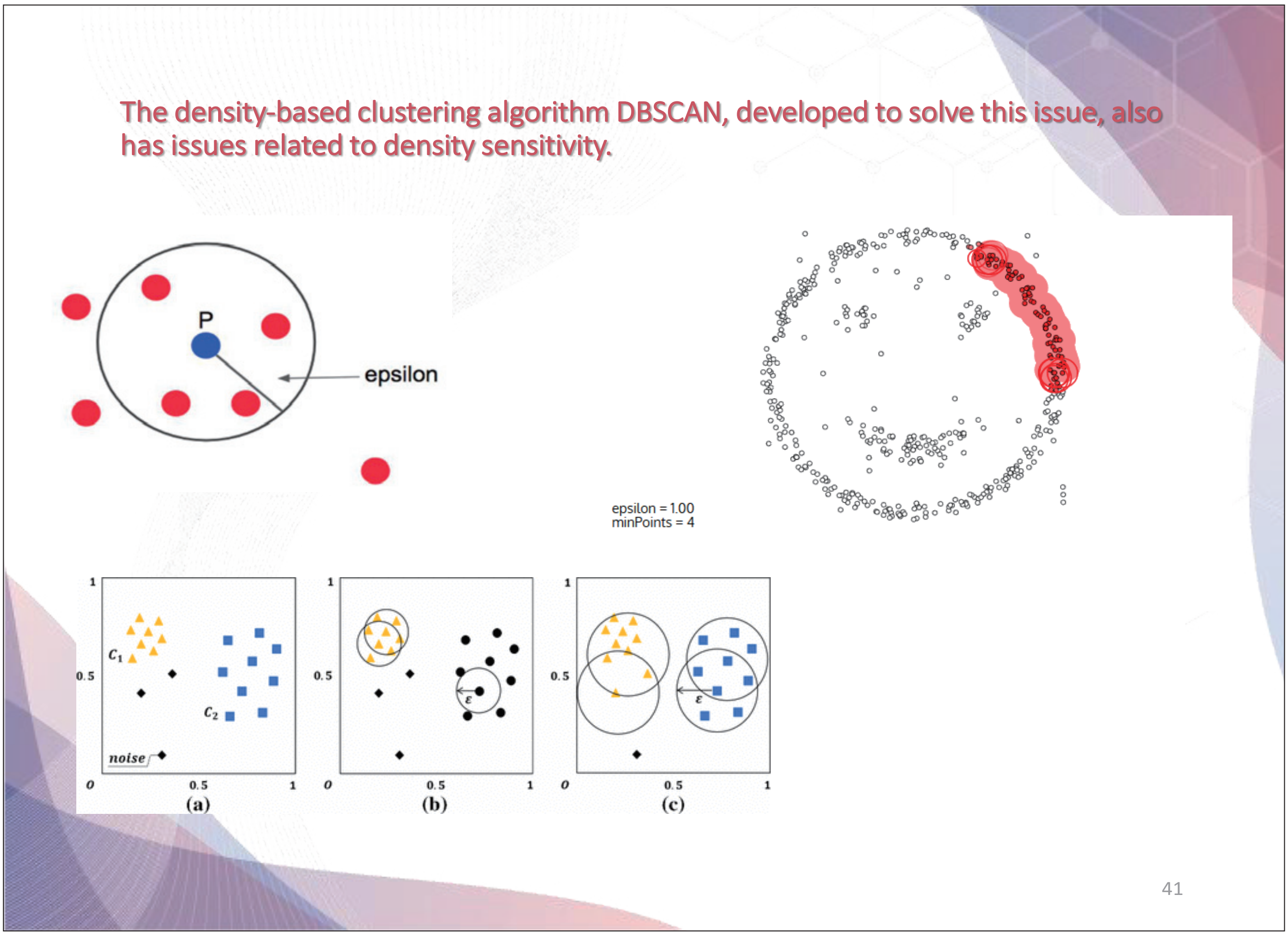
## K-means clustering fails to capture clusters of arbitrary shape, as it assumes a simple cluster distribution.

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

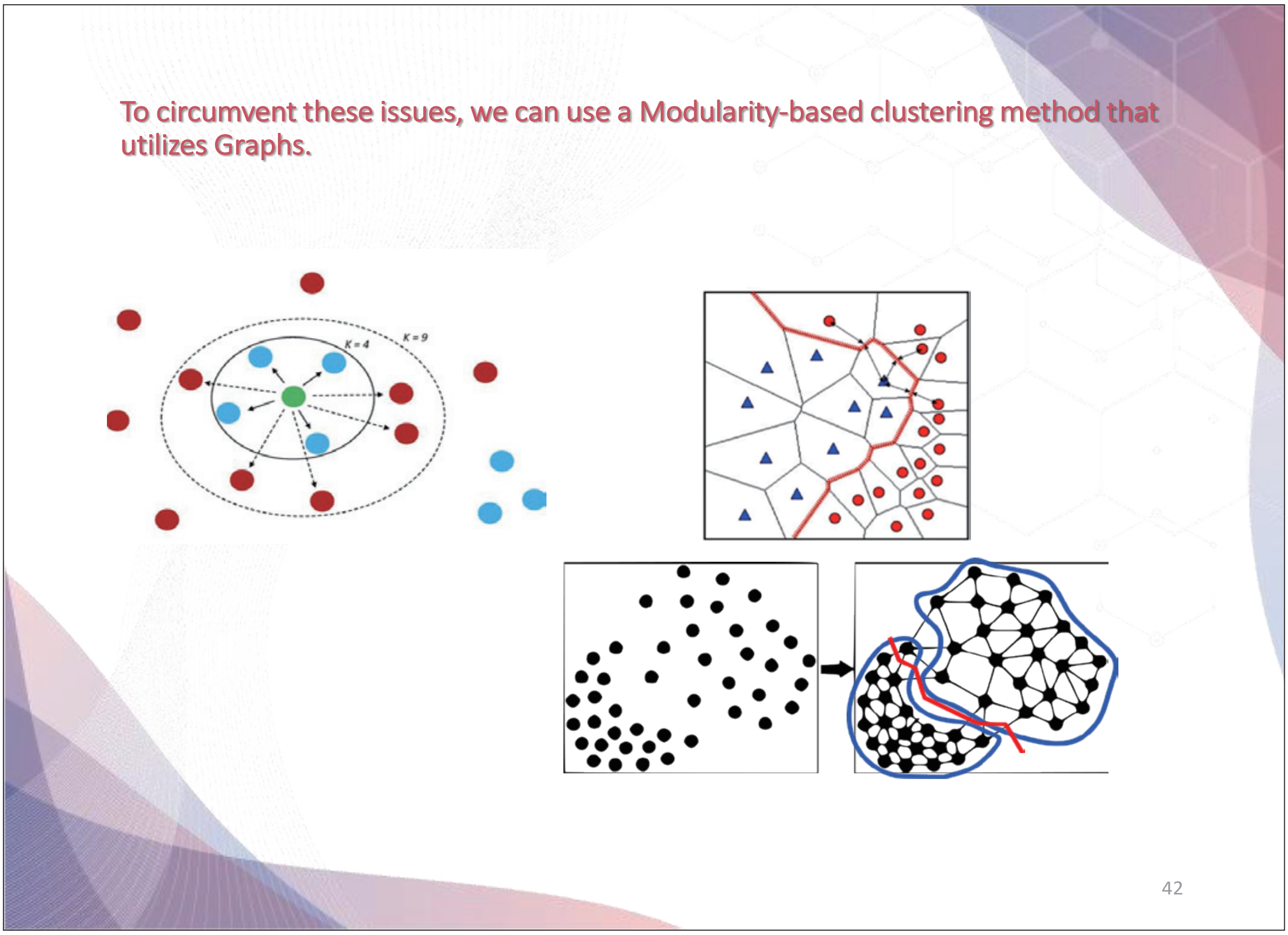


$$P(x_i|\mu_k) \propto \exp(-\frac{||x_i-\mu_k||^2}{2\sigma^2})$$

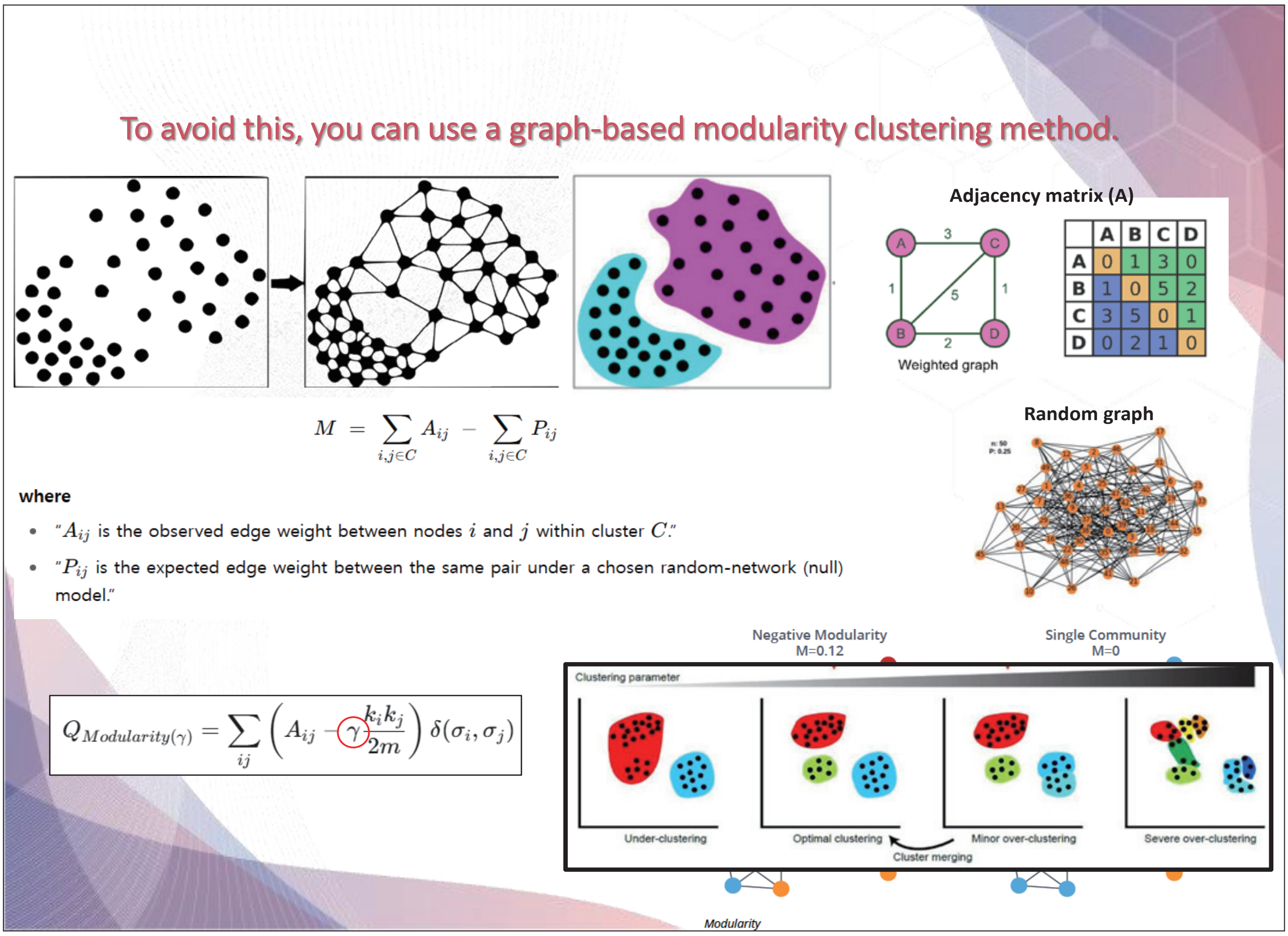




The density-based clustering algorithm DBSCAN, developed to solve this issue, also has issues related to density sensitivity.



To circumvent these issues, we can use a Modularity-based clustering method that utilizes Graphs.



## To avoid this, you can use a graph-based modularity clustering method.

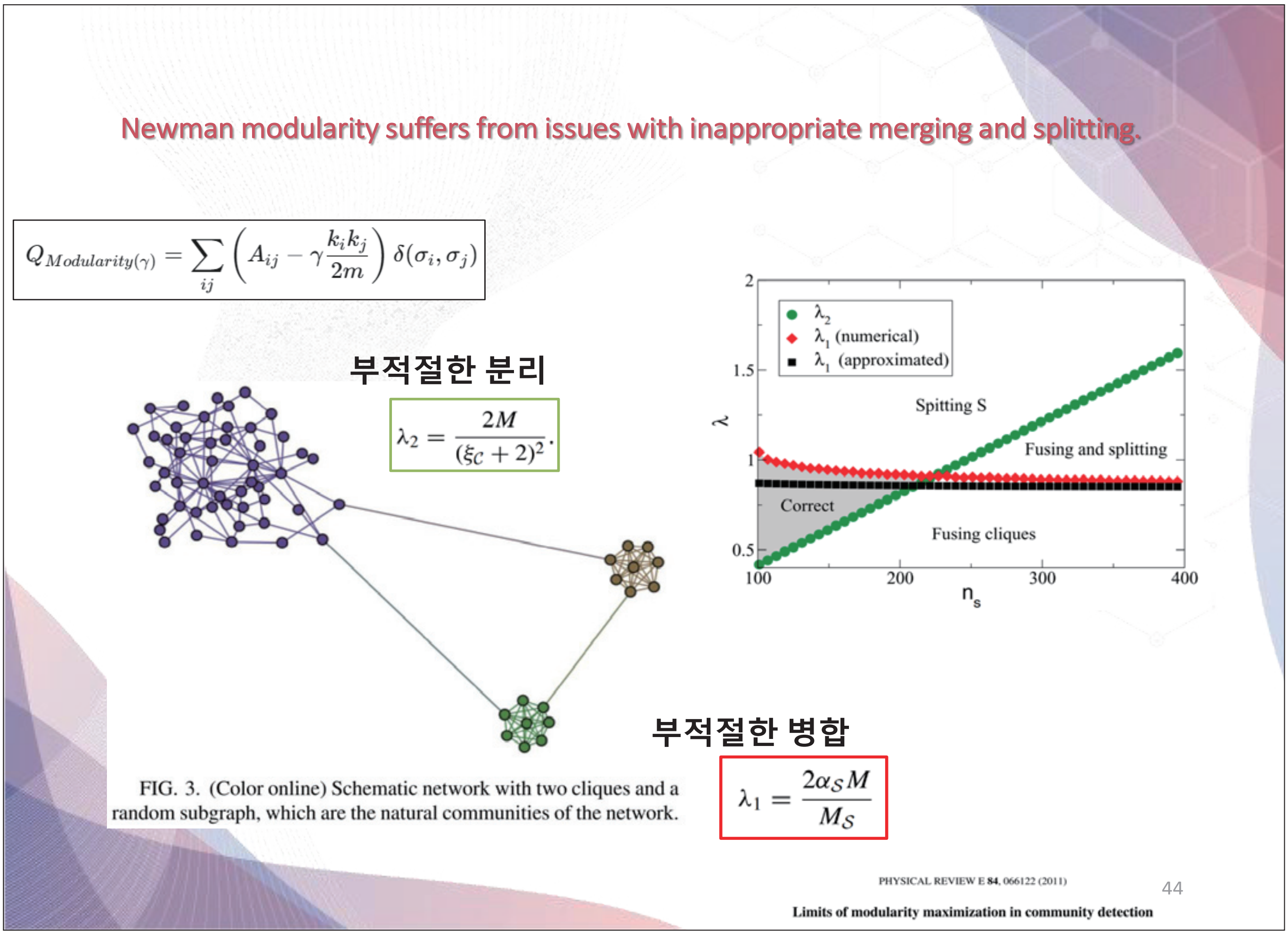
|   | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 1 | 3 | 0 |
| B | 1 | 0 | 5 | 2 |
| C | 3 | 5 | 0 | 1 |
| D | 0 | 2 | 1 | 0 |

$$M = \sum_{i,j \in C} A_{ij} - \sum_{i,j \in C} P_{ij}$$

where

- "$A_{ij}$ is the observed edge weight between nodes $i$ and $j$ within cluster $C$."
- "$P_{ij}$ is the expected edge weight between the same pair under a chosen random-network (null) model."

$$Q_{Modularity(\gamma)} = \sum_{ij} \left( A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$$



## Newman modularity suffers from issues with inappropriate merging and splitting.

$$Q_{Modularity(\gamma)} = \sum_{ij} \left( A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$$

부적절한 분리

$$\lambda_2 = \frac{2M}{(\xi_C + 2)^2}.$$

FIG. 3. (Color online) Schematic network with two cliques and a random subgraph, which are the natural communities of the network.

부적절한 병합

$$\lambda_1 = \frac{2\alpha_S M}{M_S}$$

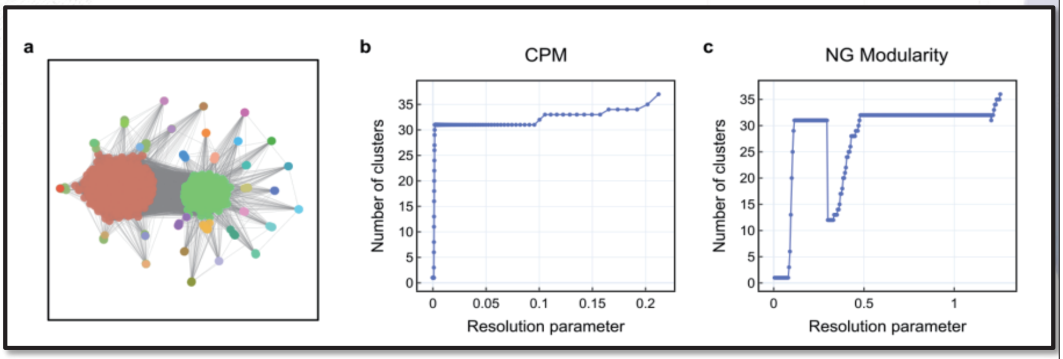
PHYSICAL REVIEW E **84**, 066122 (2011)

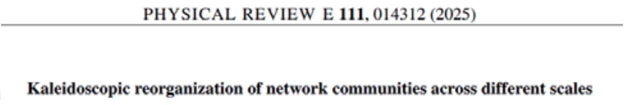
**Limits of modularity maximization in community detection**

## Newman modularity suffers from issues with inappropriate merging and splitting.

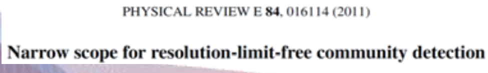
$$Q_{Modularity(\gamma)} = \sum_{ij} \left( A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$$

$$Q_{CPM} = \sum_{ij} (A_{ij} - \gamma) \delta(\sigma_i, \sigma_j)$$




PHYSICAL REVIEW E **111**, 014312 (2025)

**Kaleidoscopic reorganization of network communities across different scales**


PHYSICAL REVIEW E **84**, 016114 (2011)

**Narrow scope for resolution-limit-free community detection**

Kim et al. 2025. Nat. Commun.

## However, these algorithms exhibit inconsistencies in their results, which can undermine their reliability.

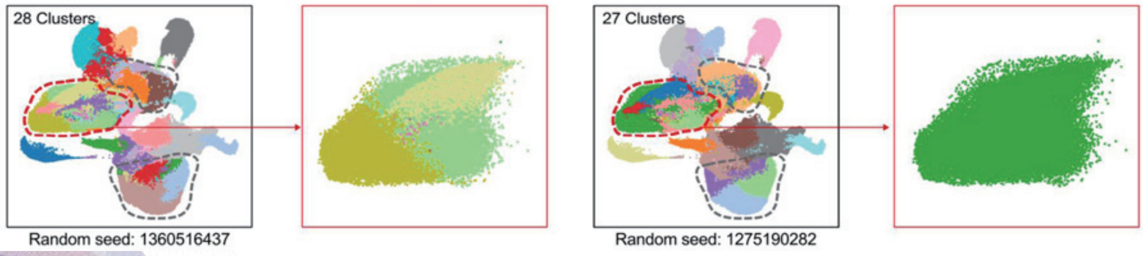
Article | Open access | Published: 23 May 2024

**Single-cell and spatial transcriptomics analysis of non-small cell lung cancer**

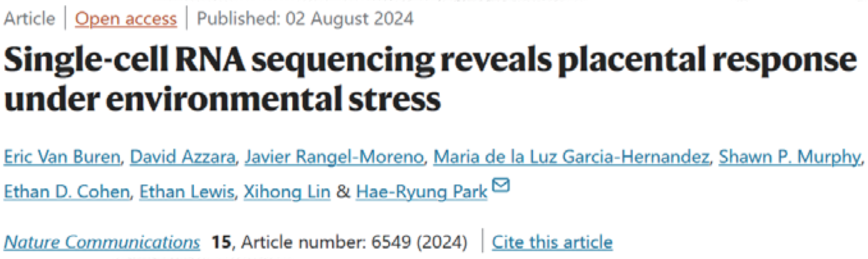
Marco De Zuani, Haoliang Xue, Jun Sung Park, Stefan C. Dentro, Zaira Seferbekova, Julien Tessier, Sandra Curras-Alonso, Angela Hadjipanayis, Emmanouil I. Athanasiadis, Moritz Gerstung, Omer Bayraktar & Ana Cvejic

*Nature Communications* **15**, Article number: 4388 (2024) | Cite this article

**Healthy cells from the NSCLC single-cell transcriptomics study**



However, these algorithms exhibit inconsistencies in their results, which can undermine their reliability.



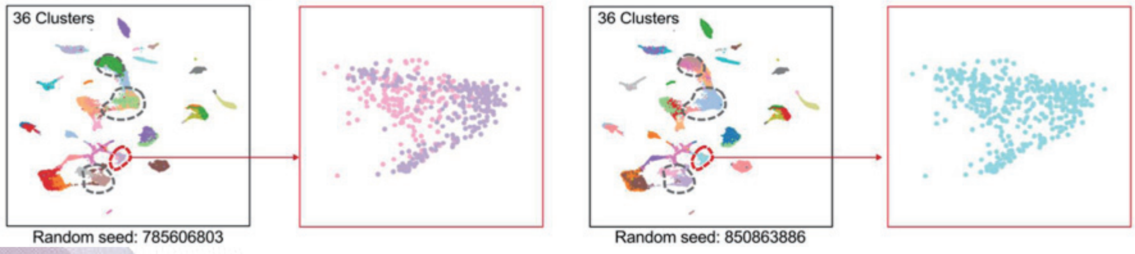
Article | Open access | Published: 02 August 2024

## Single-cell RNA sequencing reveals placental response under environmental stress

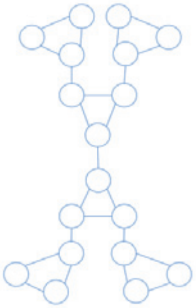
Eric Van Buren, David Azzara, Javier Rangel-Moreno, Maria de la Luz Garcia-Hernandez, Shawn P. Murphy, Ethan D. Cohen, Ethan Lewis, Xihong Lin & Hae-Ryung Park

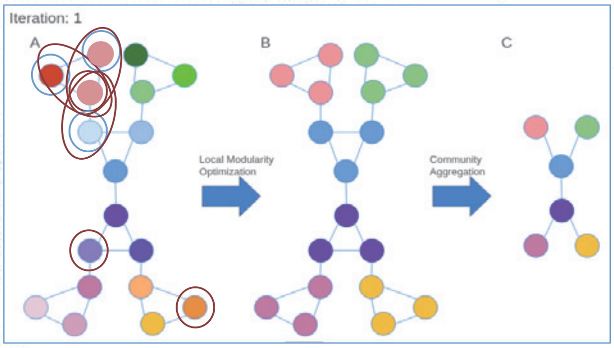
*Nature Communications* **15**, Article number: 6549 (2024) | Cite this article

Cells from the mouse placenta

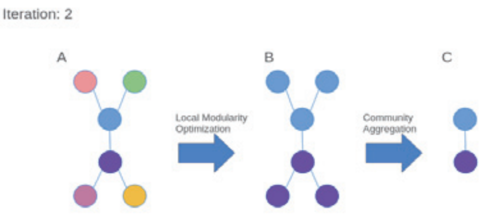


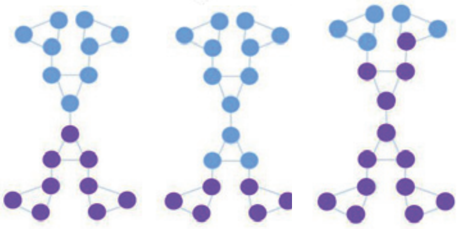
These inconsistencies stem from the random partition search.





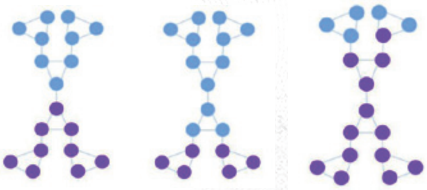
$$M = \sum_{i,j\in C} A_{ij} - \sum_{i,j\in C} P_{ij} \uparrow\downarrow$$





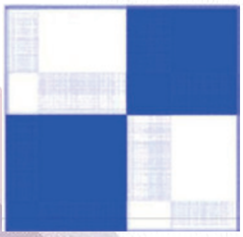
Traag, V. A. et al. Sci Rep (2019)

To estimate this inconsistency, a PAC score based on the consensus matrix was proposed; however, slow computation limits its practical use.

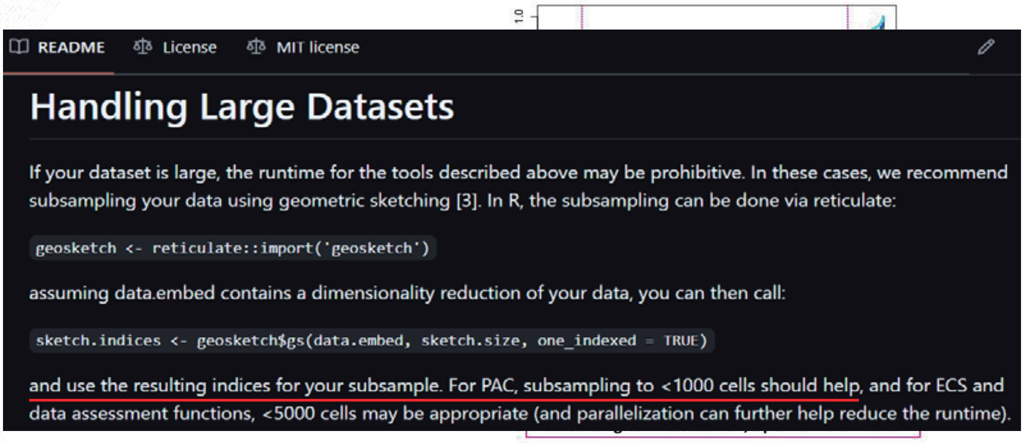


$$M^h(i,j) = \begin{cases} 1, & \text{if points i and j belong to the same cluster} \\ 0, & \text{otherwise} \end{cases}$$

Consensus matrix (C)



Consensus matrix CDFs from K=2 to 6



## Handling Large Datasets

If your dataset is large, the runtime for the tools described above may be prohibitive. In these cases, we recommend subsampling your data using geometric sketching [3]. In R, the subsampling can be done via reticulate:

```
geosketch <- reticulate::import('geosketch')
```

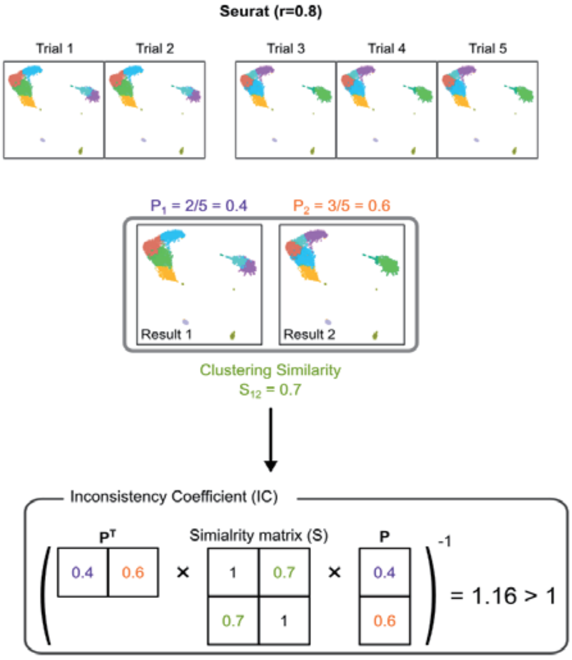
assuming data.embed contains a dimensionality reduction of your data, you can then call:

```
sketch.indices <- geosketch$gs(data.embed, sketch.size, one_indexed = TRUE)
```

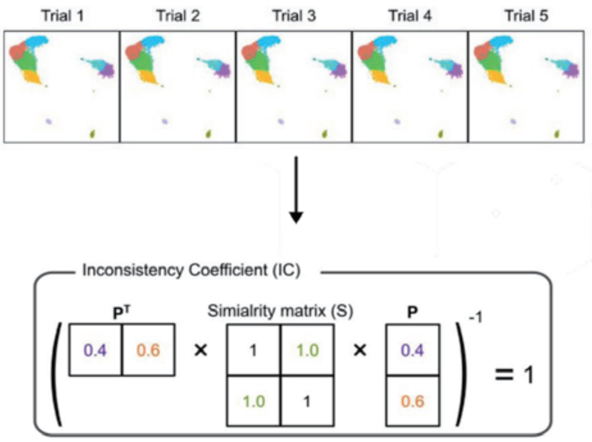
and use the resulting indices for your subsample. For PAC, subsampling to <1000 cells should help, and for ECS and data assessment functions, <5000 cells may be appropriate (and parallelization can further help reduce the runtime).

Arash Shahsavari et al. bioRxiv (2022)

In contrast, the recently proposed Inconsistency Coefficient (IC) offers rapid quantification of clustering inconsistencies.



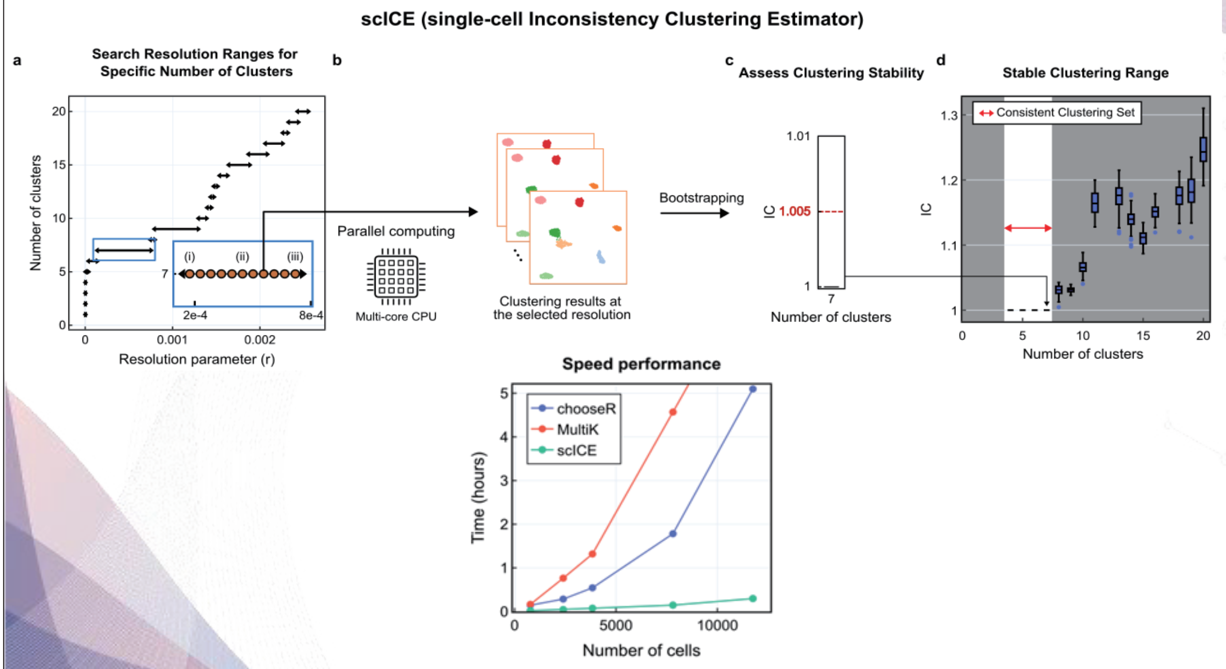
$$\left( \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \times \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \times \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \right)^{-1} = 1.16 > 1$$



$$\left( \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \times \begin{pmatrix} 1 & 1.0 \\ 1.0 & 1 \end{pmatrix} \times \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \right)^{-1} = 1$$

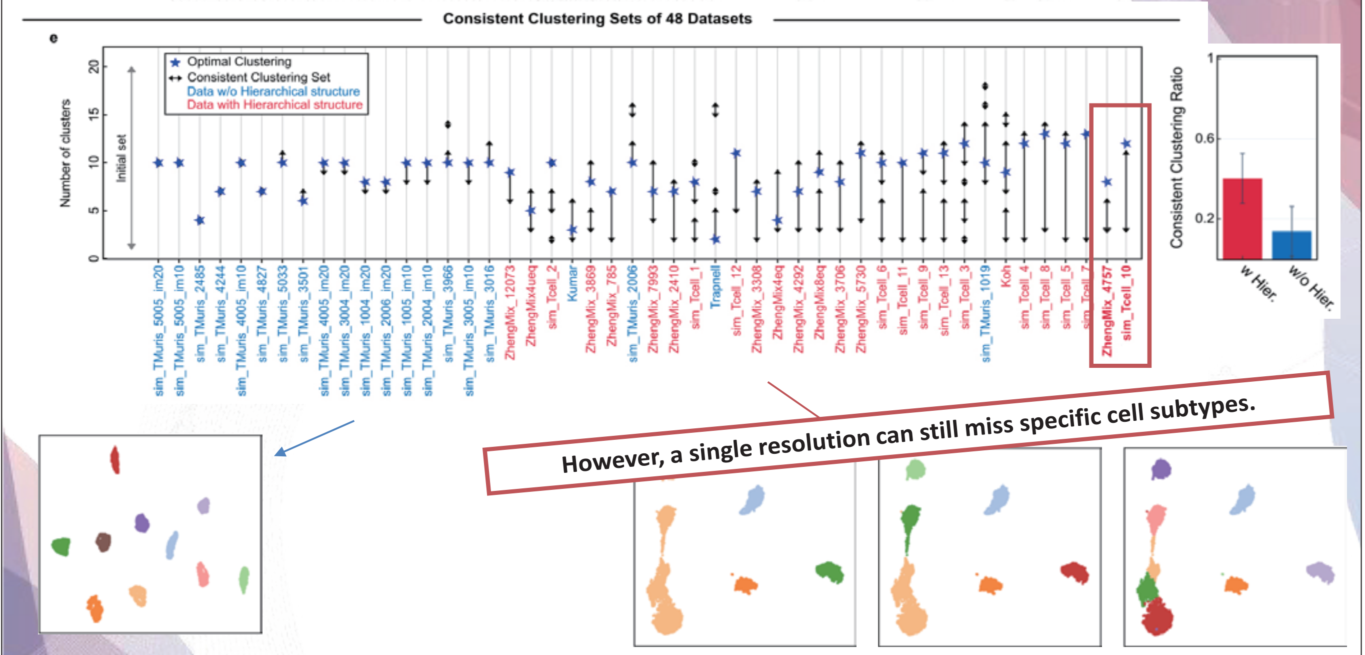
Lee, D. et al. PRE (2021)

## Leveraging IC and parallel computing, scICE achieves up to 30x faster speed performance than conventional methods.



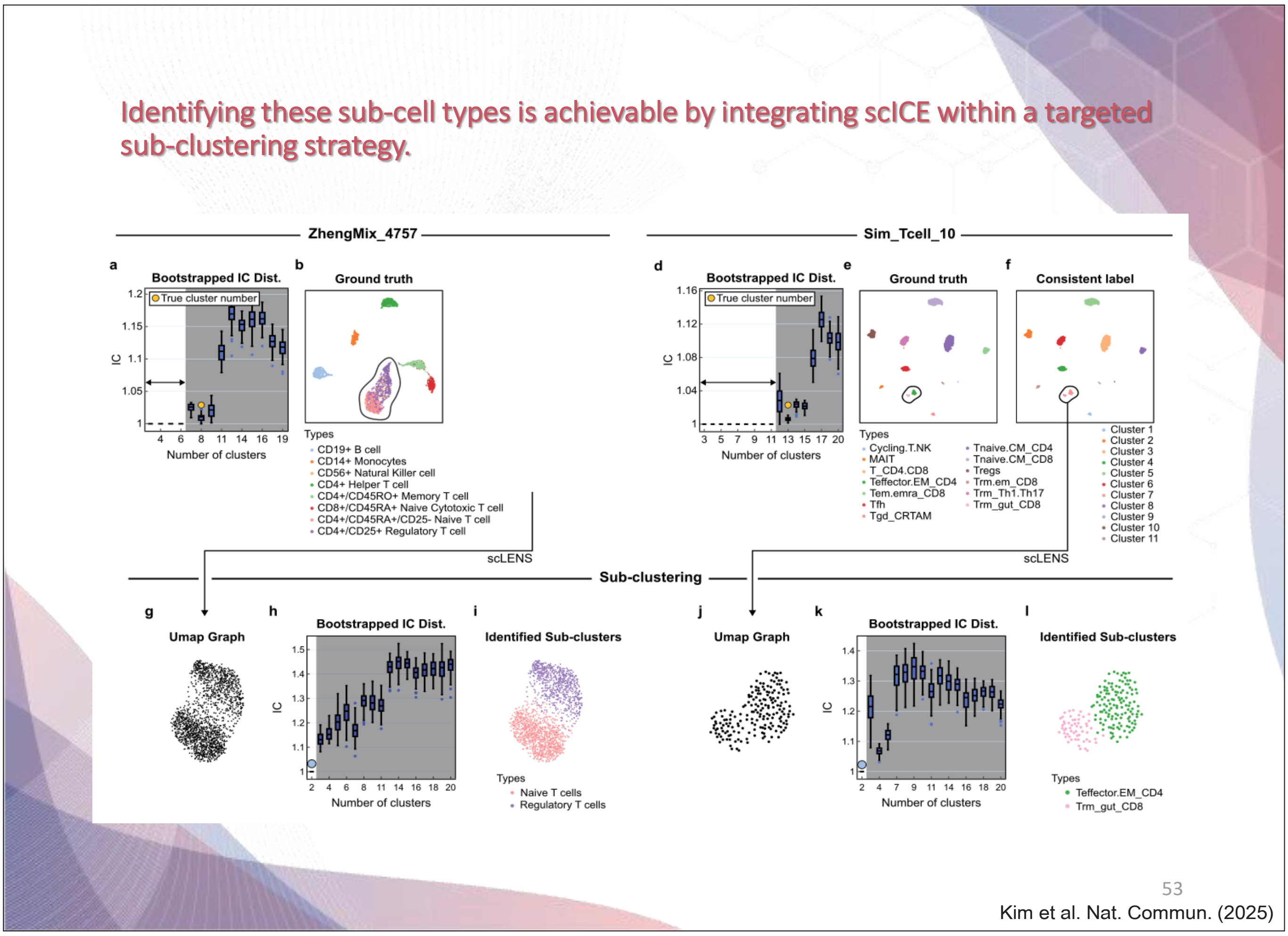
Traag, V. A. et al. Sci Rep (2019)

## Applying scICE revealed that hierarchical datasets yield more consistent cluster labels.



However, a single resolution can still miss specific cell subtypes.

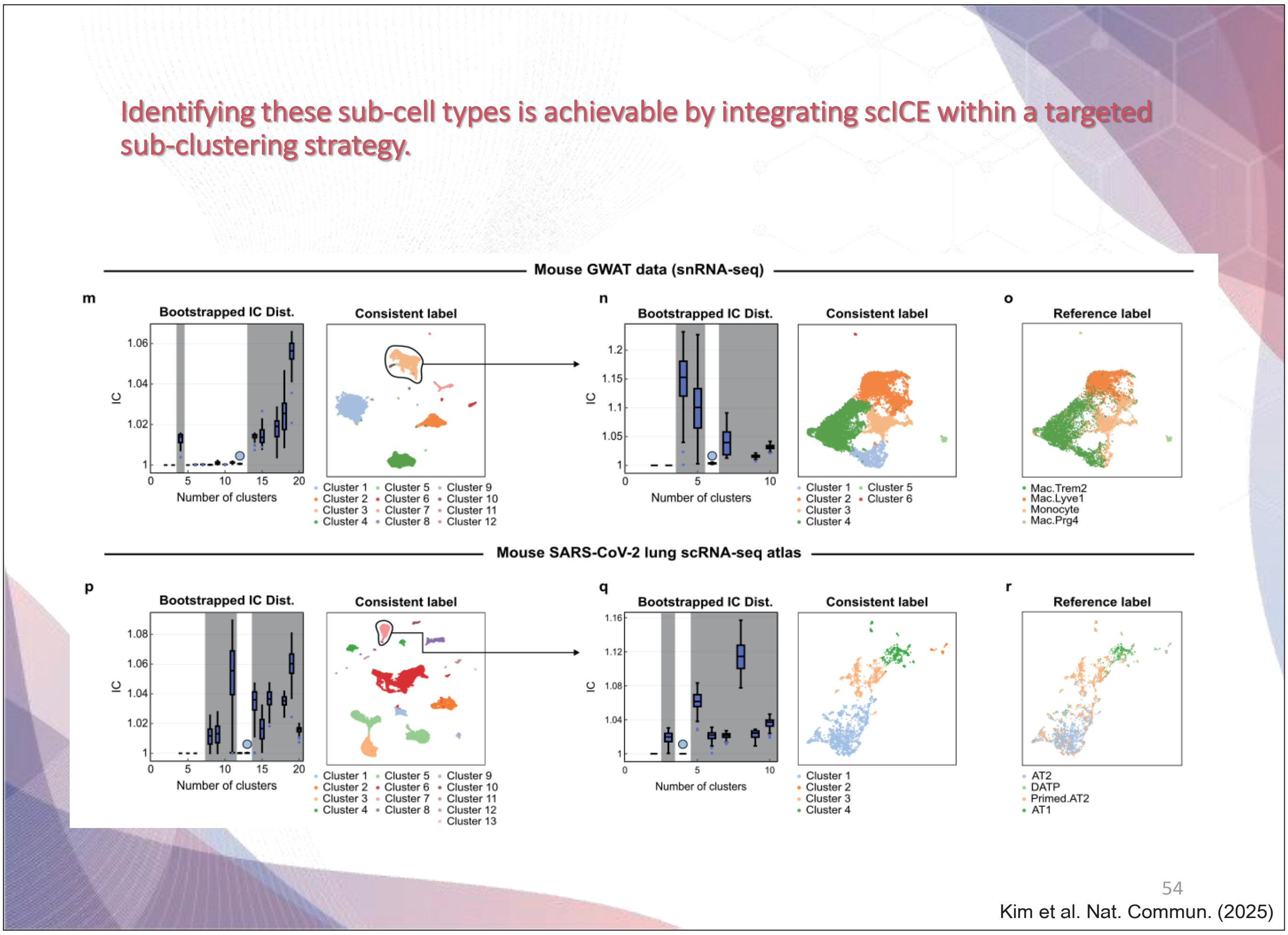
Kim et al. Nat. Commun. (2025)



Identifying these sub-cell types is achievable by integrating scICE within a targeted sub-clustering strategy.

53

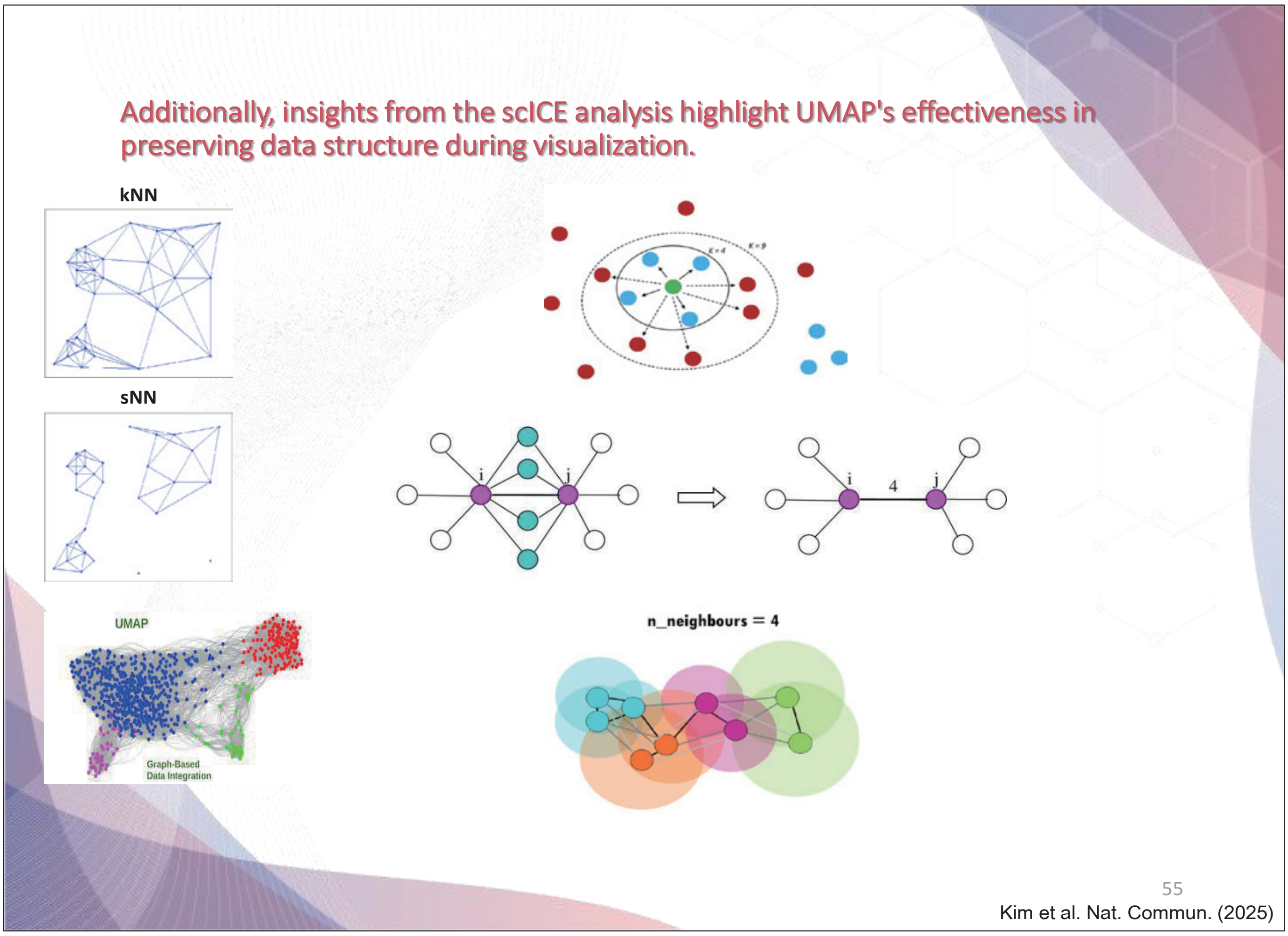
Kim et al. Nat. Commun. (2025)



Identifying these sub-cell types is achievable by integrating scICE within a targeted sub-clustering strategy.

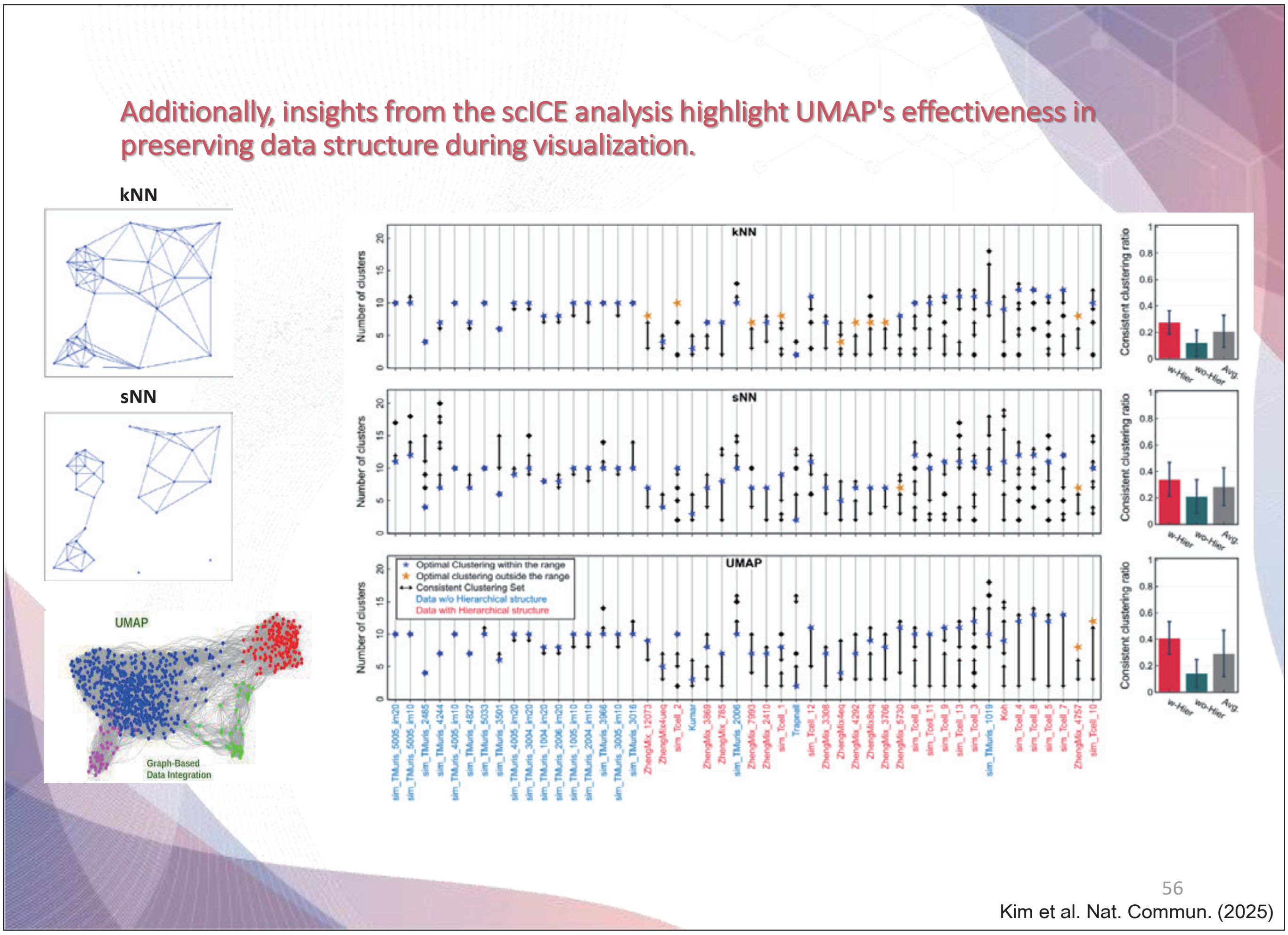
54

Kim et al. Nat. Commun. (2025)



Additionally, insights from the scICE analysis highlight UMAP's effectiveness in preserving data structure during visualization.

Kim et al. Nat. Commun. (2025)



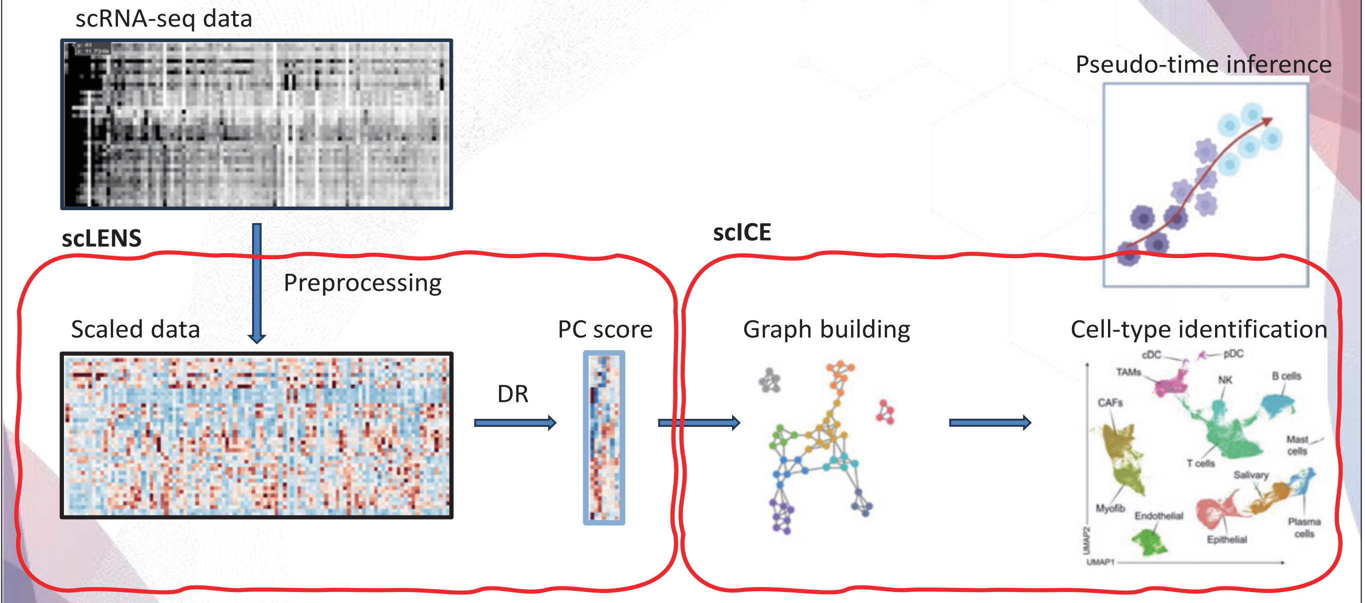
Additionally, insights from the scICE analysis highlight UMAP's effectiveness in preserving data structure during visualization.

Kim et al. Nat. Commun. (2025)

## Summary

- When using modularity-based clustering algorithms, the inconsistency of results, which affects reliability, must be considered, and this can be efficiently measured using the Inconsistency Coefficient (IC).
- The reliable cluster results provided through scICE are few in number, including the actual cluster structure, thus helping users find the optimal clusters.
- To accurately and efficiently find more cell subtypes, sub-clustering with scICE can be utilized.
- The UMAP graph captures the cluster structures more accurately than kNN or sNN graphs.

## The powerful combination of scLENS and scICE increases analytical convenience while delivering reliable and accurate results.



scRNA-seq data

scLENS

Preprocessing

Scaled data

DR

PC score

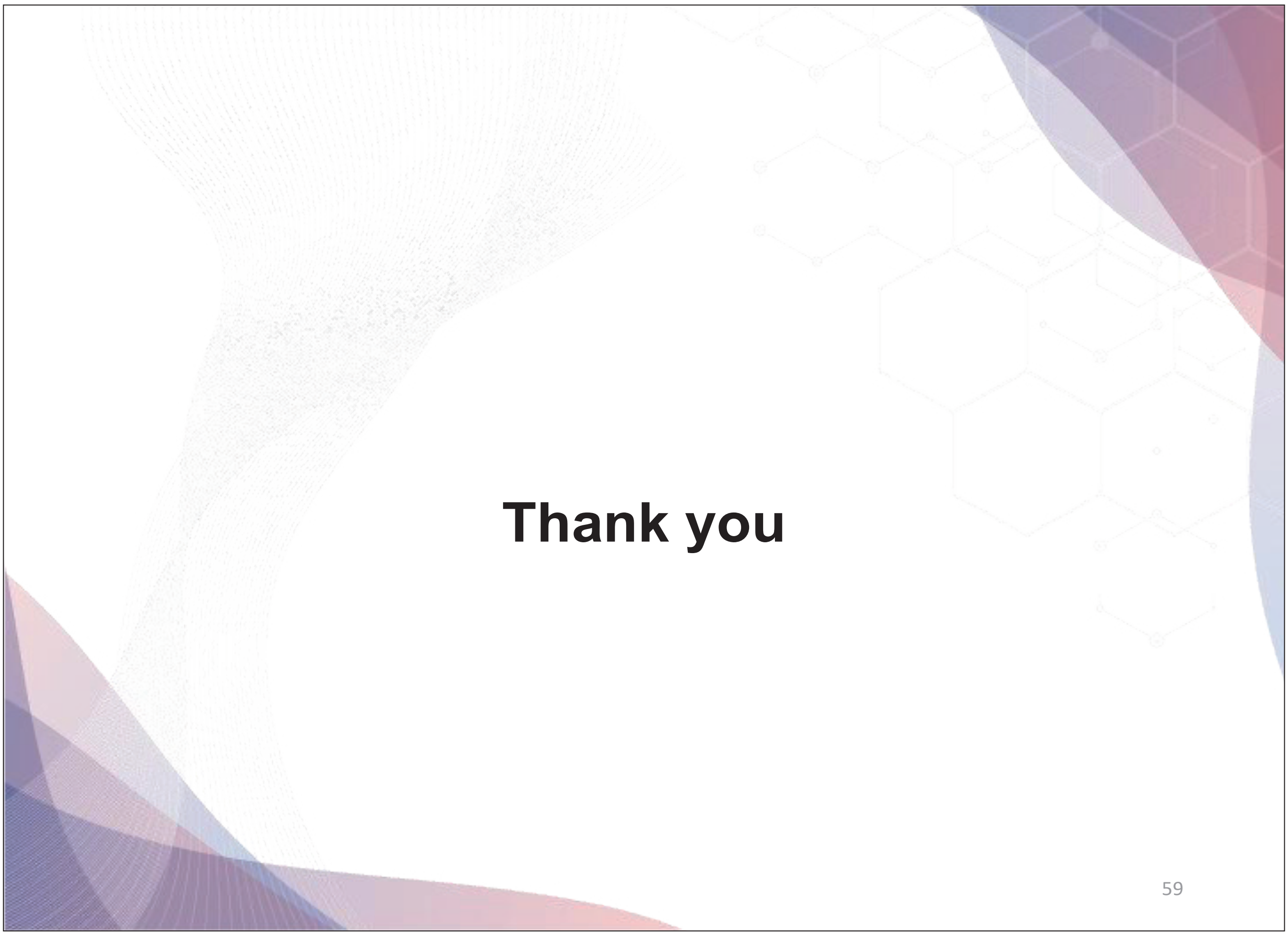
scICE

Graph building

Cell-type identification

Pseudo-time inference

Traag, V. A. et al. Sci Rep (2019)



# Thank you