Swiss Institute of Bioinformatics

INTRODUCTION TO SEQUENCING-BASED TRANSCRIPTOMICS DATA ANALYSIS

Clustering

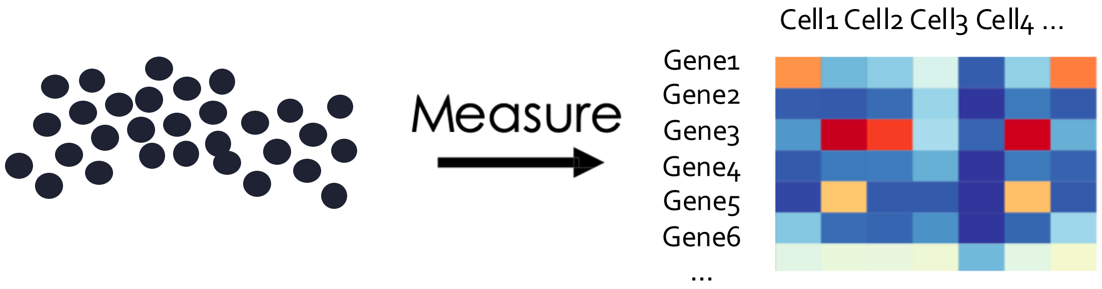
Joana Carlevaro Fita

December 9-10, 2025

elixir SWITZERLAND

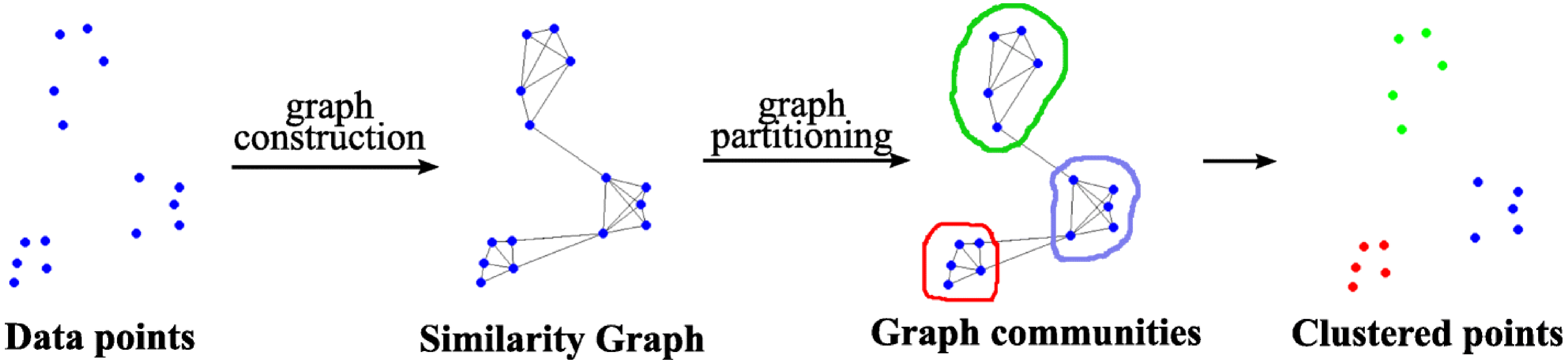
Standard scRNAseq methods appliec

Non-spatially aware clustering



Standard scRNAseq methods applied

Graph-based clustering Cells within the same community are assigned to the same cluster



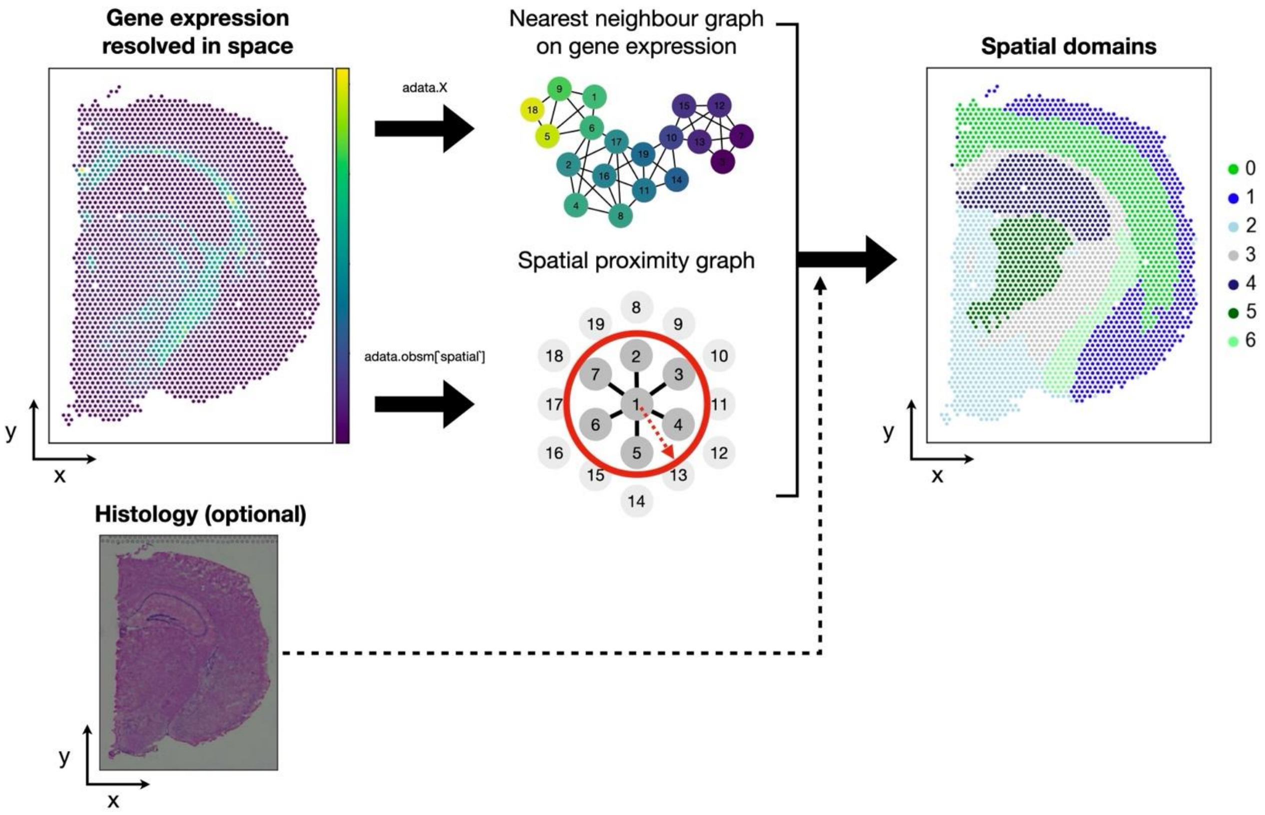
Spatially unaware graph-based clustering: based on a shared nearest neighbor (SNN) graph and the Leiden or Louvain algorithm for community detection

Clustering

Non-spatially aware clustering

- Use only gene expression
- Can give discontinous results, they don't use information about neighbours
- Can be coupled to spatially-aware pre-processing methods

Spatially aware clustering



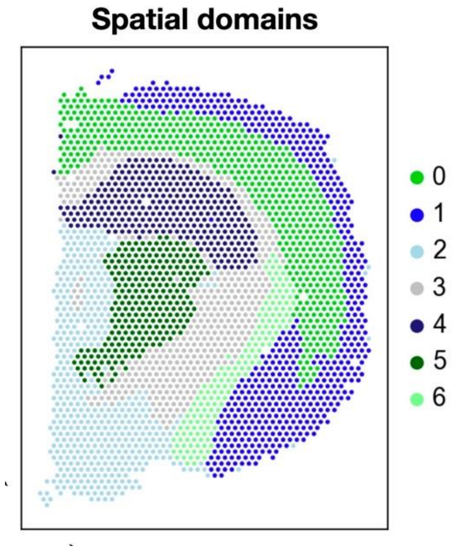
Cell type clustering vs tissue domain identification
Different algorithmic problem

Heumos, L. et. Al. Nat Rev Genet, 2023

Spatially-aware clustering

Spatial domains can be identified based on morphology/biological knowladge or using clustering (spatially aware methods).

Several methods developed for **spatial domain identification**



Goals

- Use gene expression **and** spatial information
- Define spatial regions with similar spatial gene expression / cell composition --> downstream analysis (ie. Annotation, DE...)
- Understand biological processes in terms of gene expression across spatial localisation

Assumptions

- Cell type / cell state is influenced by interactions with **neighbouring** cells.
- Spatial domains may be composed of different cell types
- Cells from same type/state may be spatially far a part (ie. Cerebral hemispheres, epithelial layers, blood vessels...)

Spatially-aware clustering methods

Methods have different assumptions, methodologies and computational trade-offs

Following OSTA book classification (Crowell et al. Biorxiv, 2025) :

- **Probabilistic** (HMRF, BayesSpace)

Examine each cell and its surrounding cell's expression to define domains
Encourages neighbouring spots to have the same label

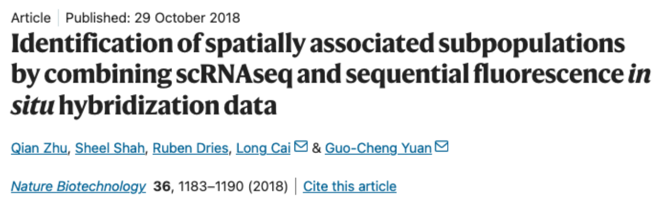
- **Encoder-based** (CellCharter, STAGATE)

Uses encoder architectures to generate latent embedding
Model spatial neighbourhoods as graphs, or jointly model gene expression and spatial coordinates.
May be a good fit for imaging-based data.

- **Neighbourhood-based** (BANKSY)

Use augmented features and embed cells in a product space containing information on the own cell and its local microenvironment. Clustered using standard algorithms

Spatially-aware clustering methods

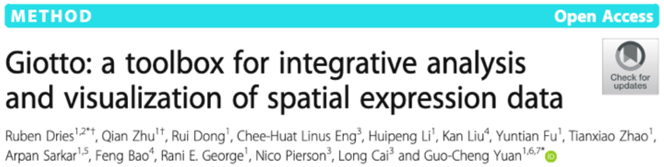


Article | Published: 29 October 2018

Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data

Qian Zhu, Sheel Shah, Ruben Dries, Long Cai & Guo-Cheng Yuan

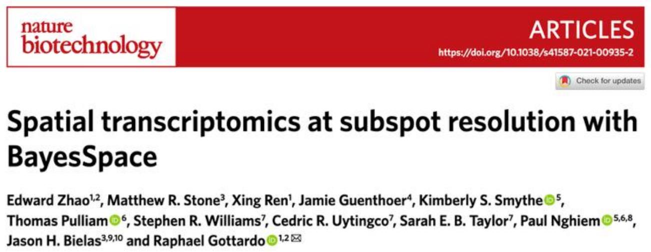
Nature Biotechnology **36**, 1183–1190 (2018) | Cite this article



METHOD Open Access

Giotto: a toolbox for integrative analysis and visualization of spatial expression data

Ruben Dries[1,2*†], Qian Zhu[1†], Rui Dong[1], Chee-Huat Linus Eng[3], Huipeng Li[1], Kan Liu[4], Yuntian Fu[1], Tianxiao Zhao[1], Arpan Sarkar[1,5], Feng Bao[4], Rani E. George[1], Nico Pierson[3], Long Cai[3] and Guo-Cheng Yuan[1,6,7*]



nature biotechnology ARTICLES

https://doi.org/10.1038/s41587-021-00935-2

Check for updates

Spatial transcriptomics at subspot resolution with BayesSpace

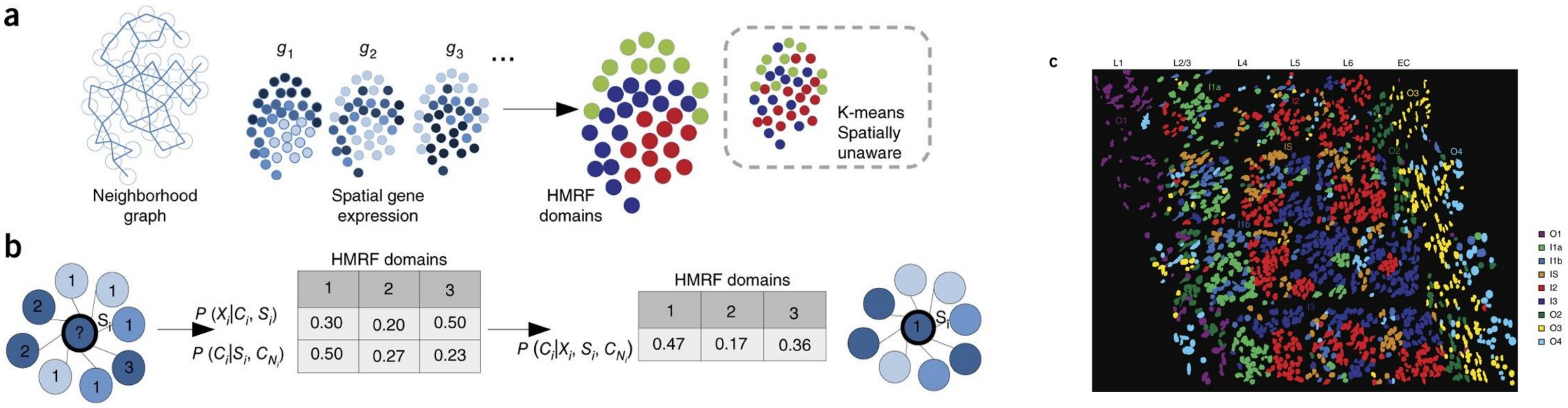
Edward Zhao[1,2], Matthew R. Stone[3], Xing Ren[1], Jamie Guenthoer[4], Kimberly S. Smythe[5], Thomas Pulliam[6], Stephen R. Williams[7], Cedric R. Uytingco[7], Sarah E. B. Taylor[7], Paul Nghiem[5,6,8], Jason H. Bielas[3,9,10] and Raphael Gottardo[1,2]

Hidden Markov Random Field (HMRF)

Method: Graph-based model commonly used for pattern recognition in image data analyses

Assumption: Cell type / cell state is influenced by (interactions with) **neighbouring** cells.

Intuition: Examine each cell and its surrounding cell's expression to define domains



HMRF domains

1	2	3
0.30	0.20	0.50
0.50	0.27	0.23

HMRF domains

1	2	3
0.47	0.17	0.36

BayesSpace

Intuition:

“BayesSpace enables spatial clustering by modelling a low-dimensional representation of the gene expression matrix and encouraging neighbouring spots to belong to the same cluster via a spatial prior“

Assumptions:

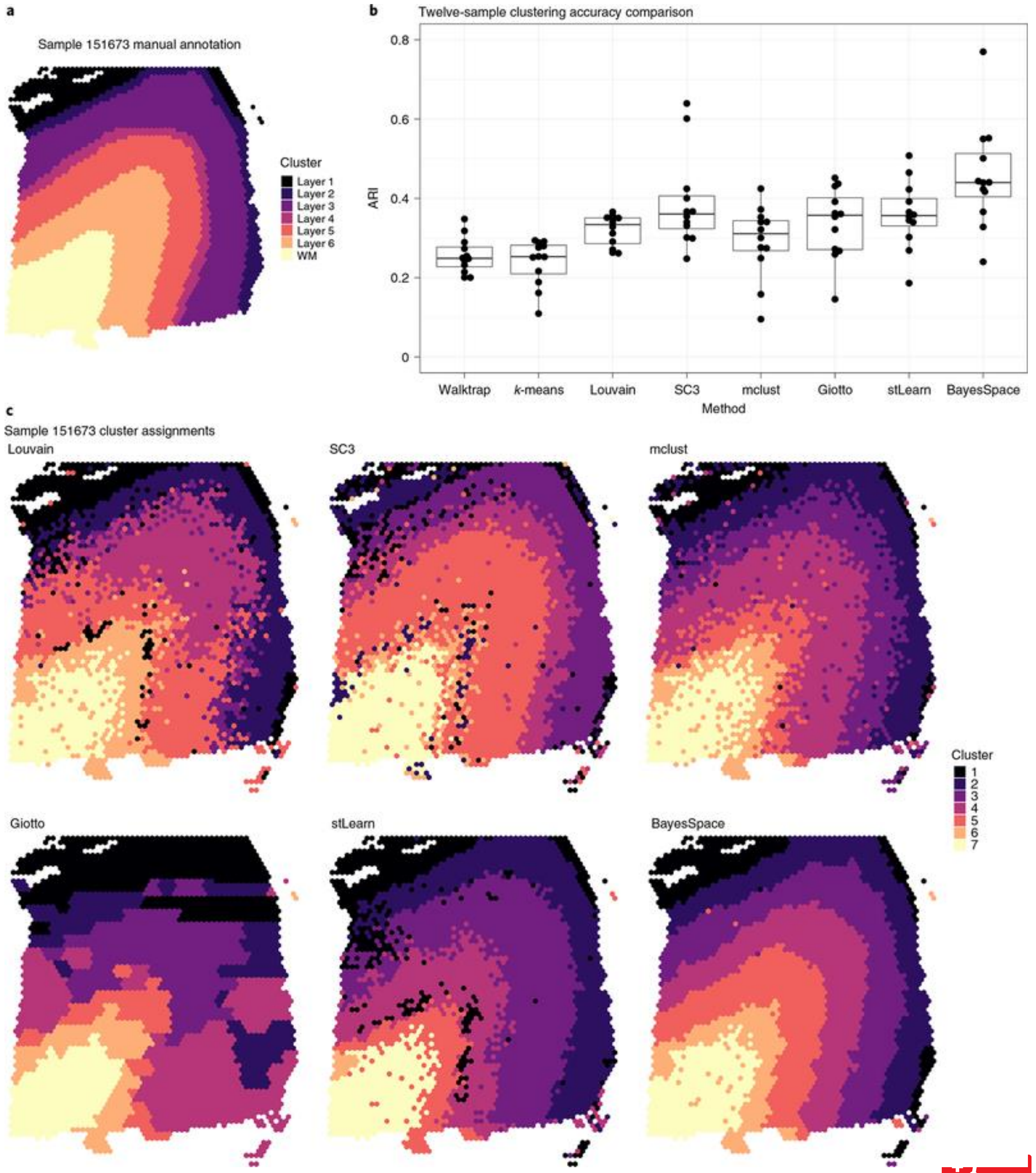
Nieghbouring cells are more likely to have similar transcriptomes

Doesn’t require pre-selection of marker genes

Goals:

Doesn’t require independent scRNAseq datasets to perform deconvolution

Adress noise and sparcity getting a smoother separation (good domain segmentation)



BANKSY

Building Aggregates with a Neighborhood Kernel and Spatial Yardstick (BANKSY)

Intuition:

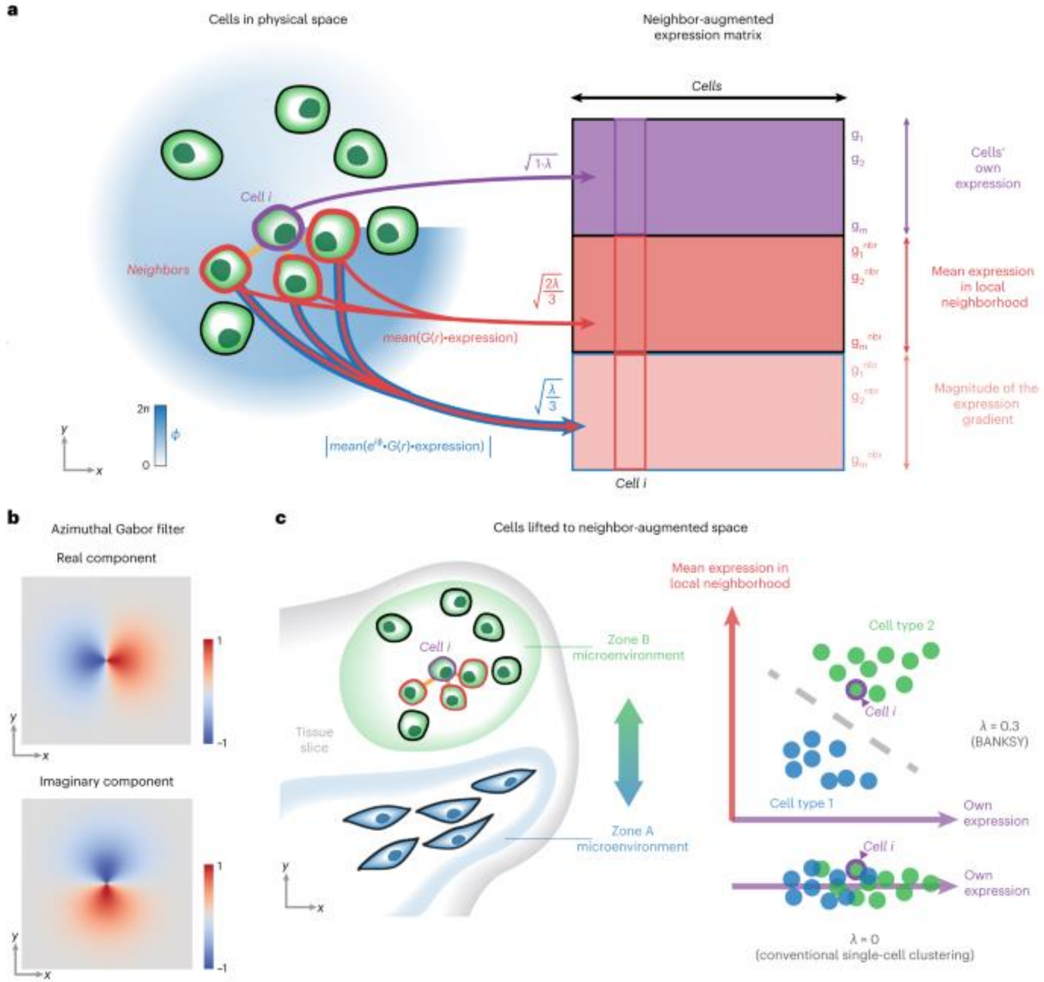
> "BANKSY uses a pair of spatial kernels to encode the transcriptomic texture of the microenvironment, one constructed using the weighted mean of gene expression in each cell's neighborhood and the other using an azimuthal Gabor filter (AGF)

Assumes

- A cell's transcriptome doesn't necessarily resemble the average transcriptome of its domain

Goals:

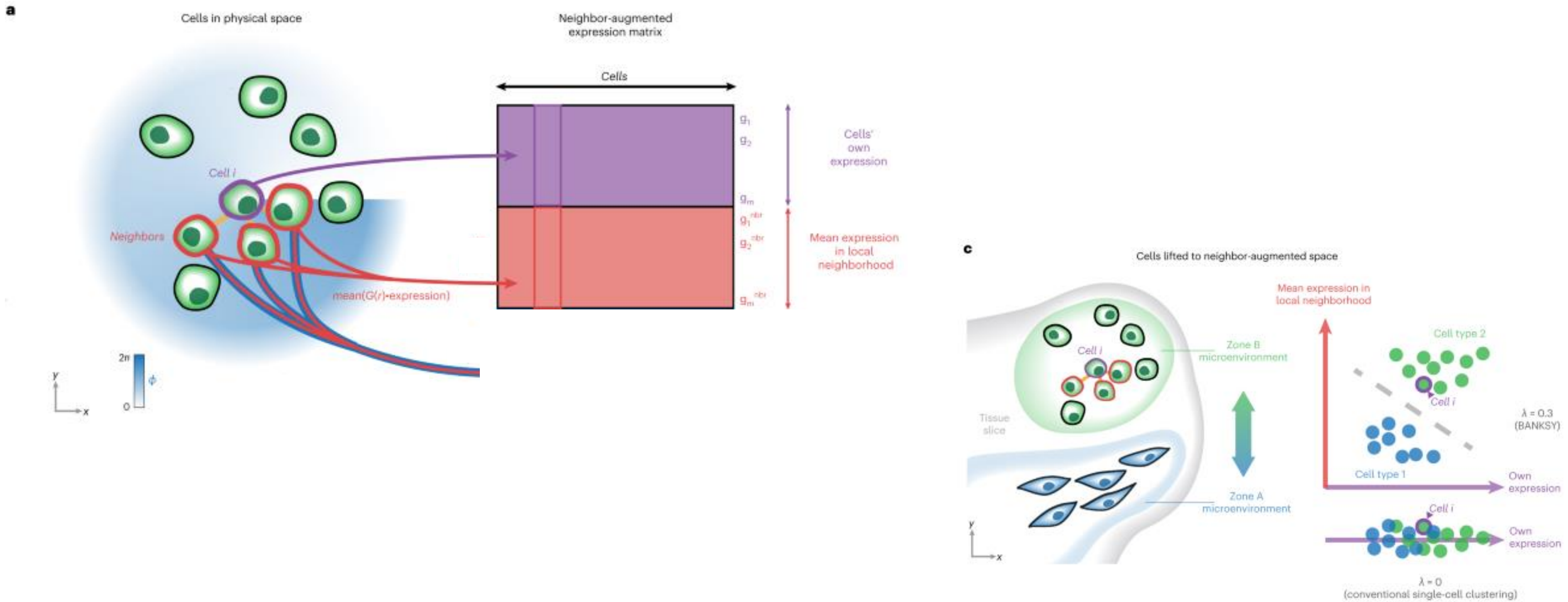
Take into account that similar cells may be far apart (intermingled, repeated patterns..) → solve both cell typing and domain segmentation



Singhal et al. Nat Gen, 2024

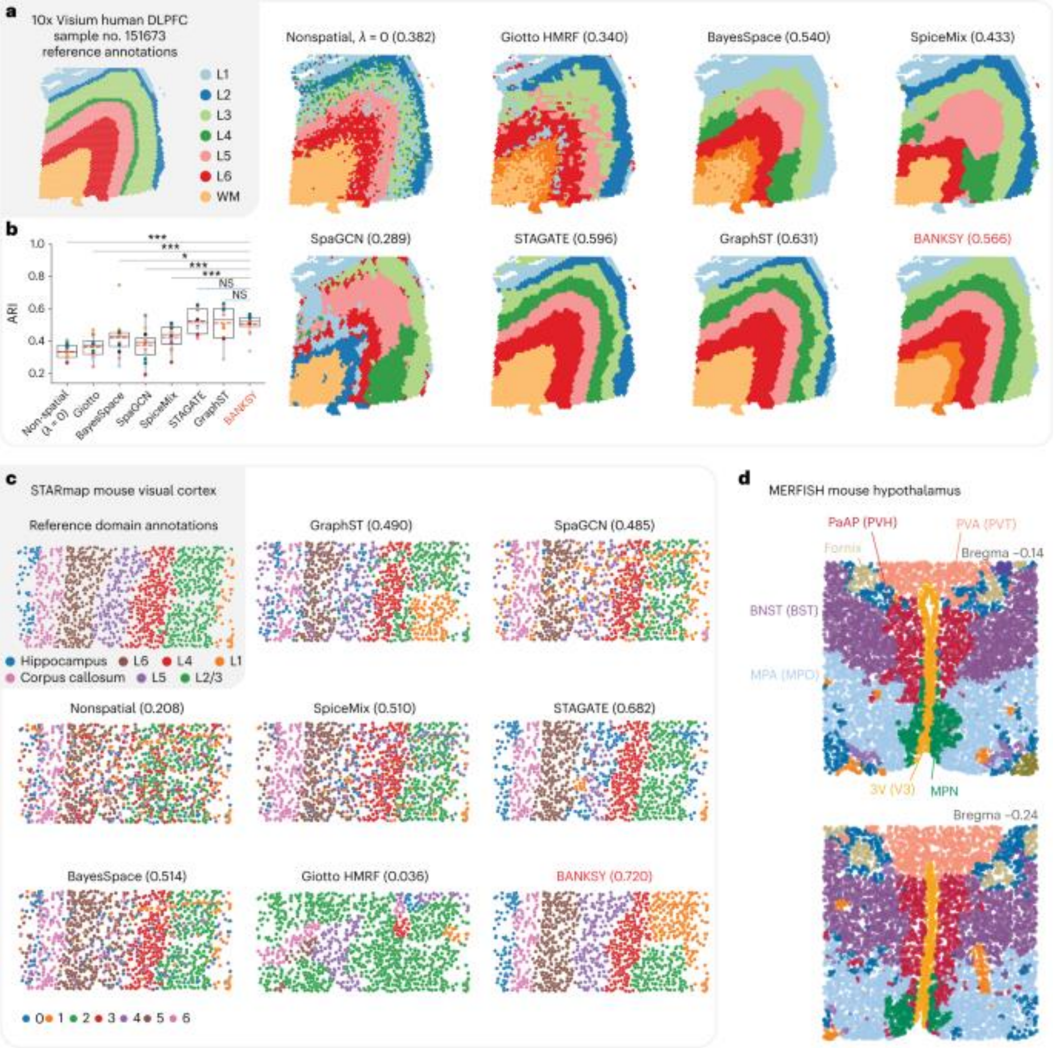
BANKSY

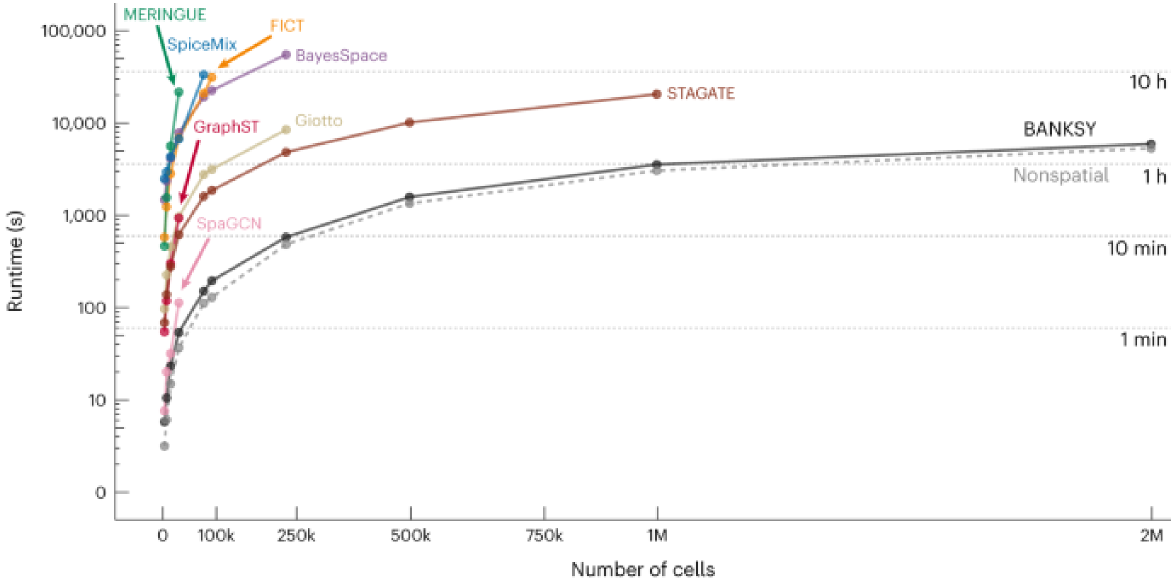
Building Aggregates with a Neighborhood Kernel and Spatial Yardstick (BANKSY)



Adapted from Singhal et al. Nat Gen, 2024

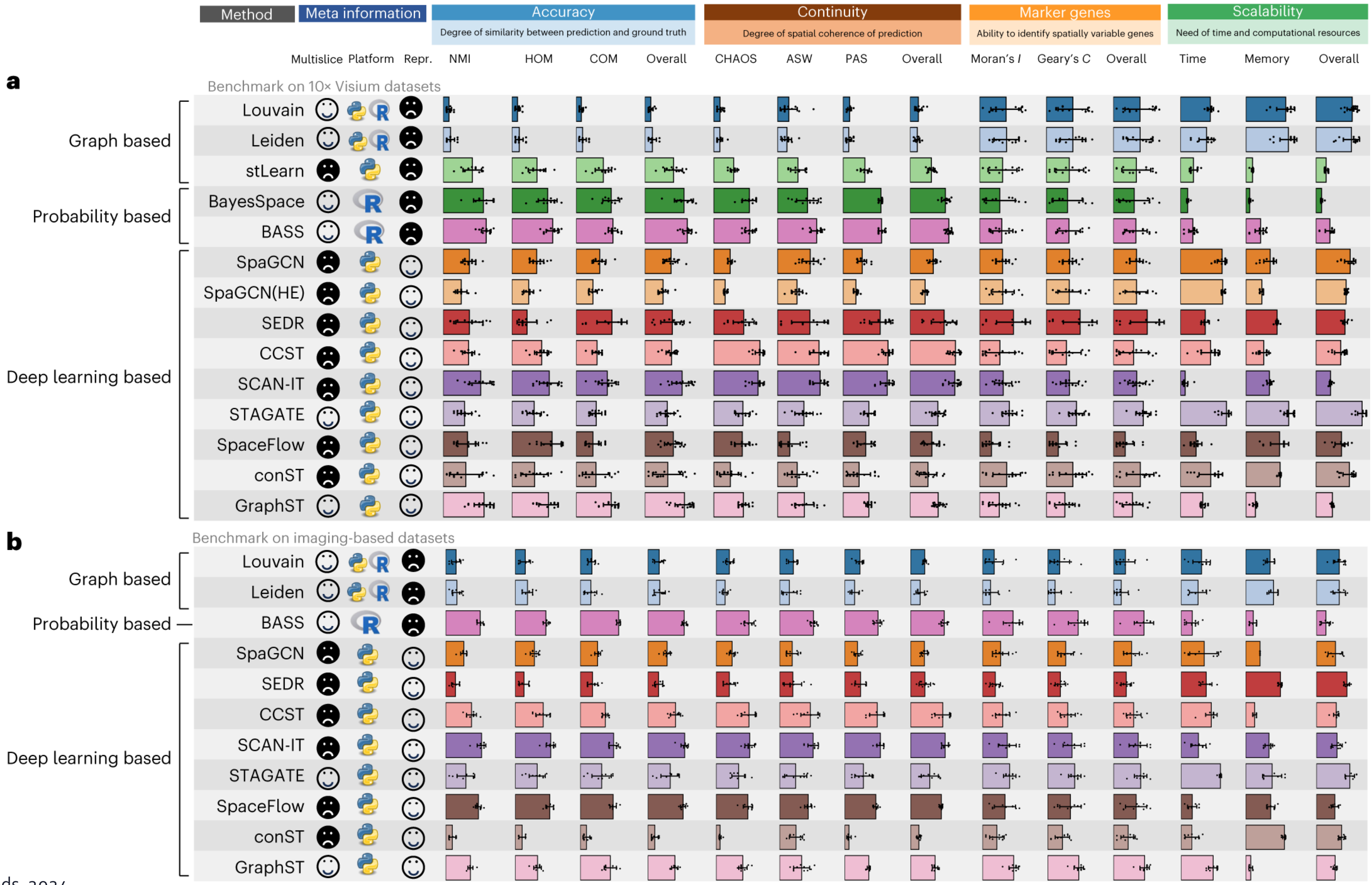
BANKSY



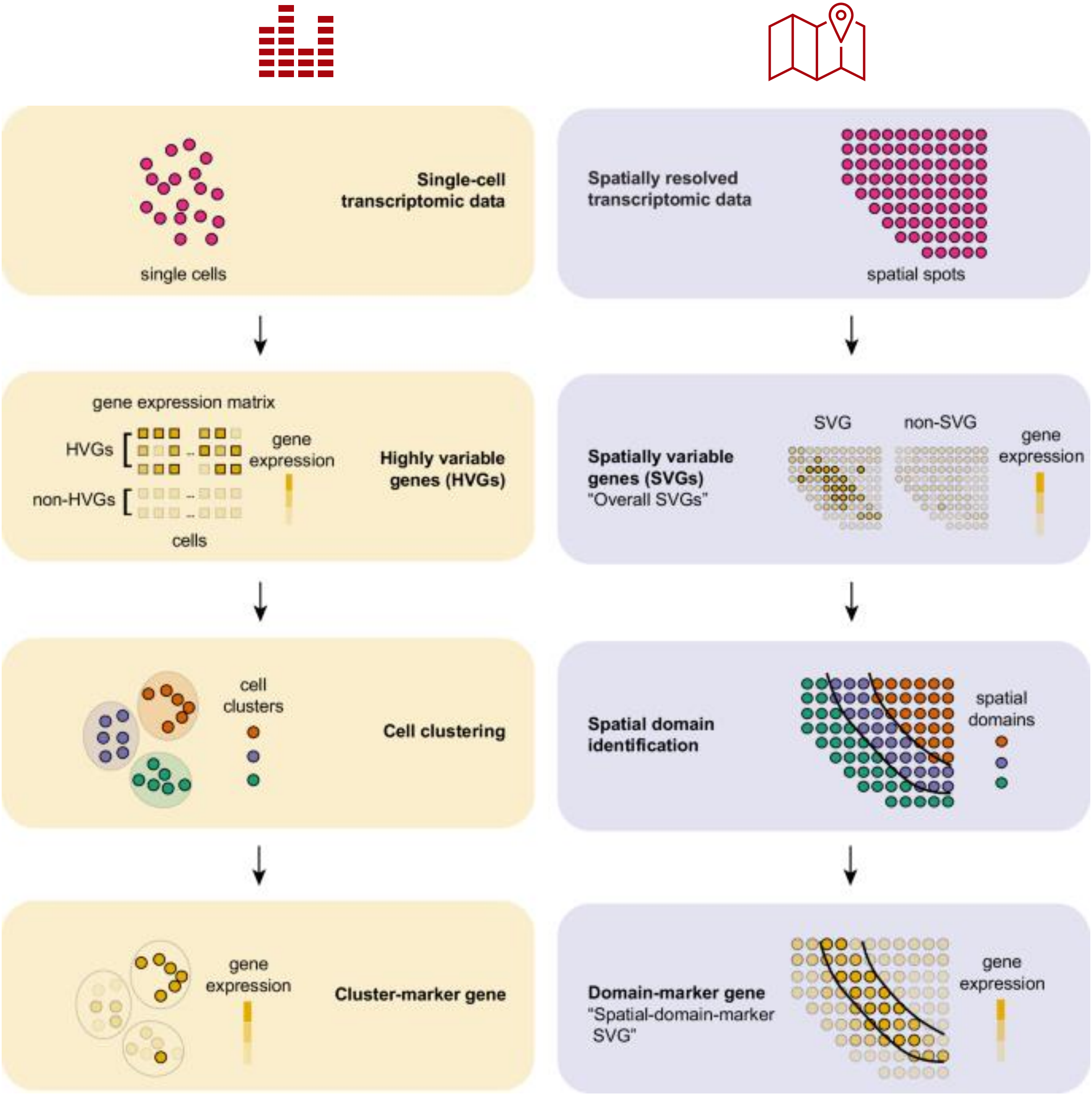


Singhal et al. Nat Gen, 2024

Benchmark on methods



Yuan et al. Nature Methods, 2024



Single-cell transcriptomic data

Highly variable genes (HVGs)

Cell clustering

Cluster-marker gene

Spatially resolved transcriptomic data

Spatially variable genes (SVGs)
"Overall SVGs"

Spatial domain identification

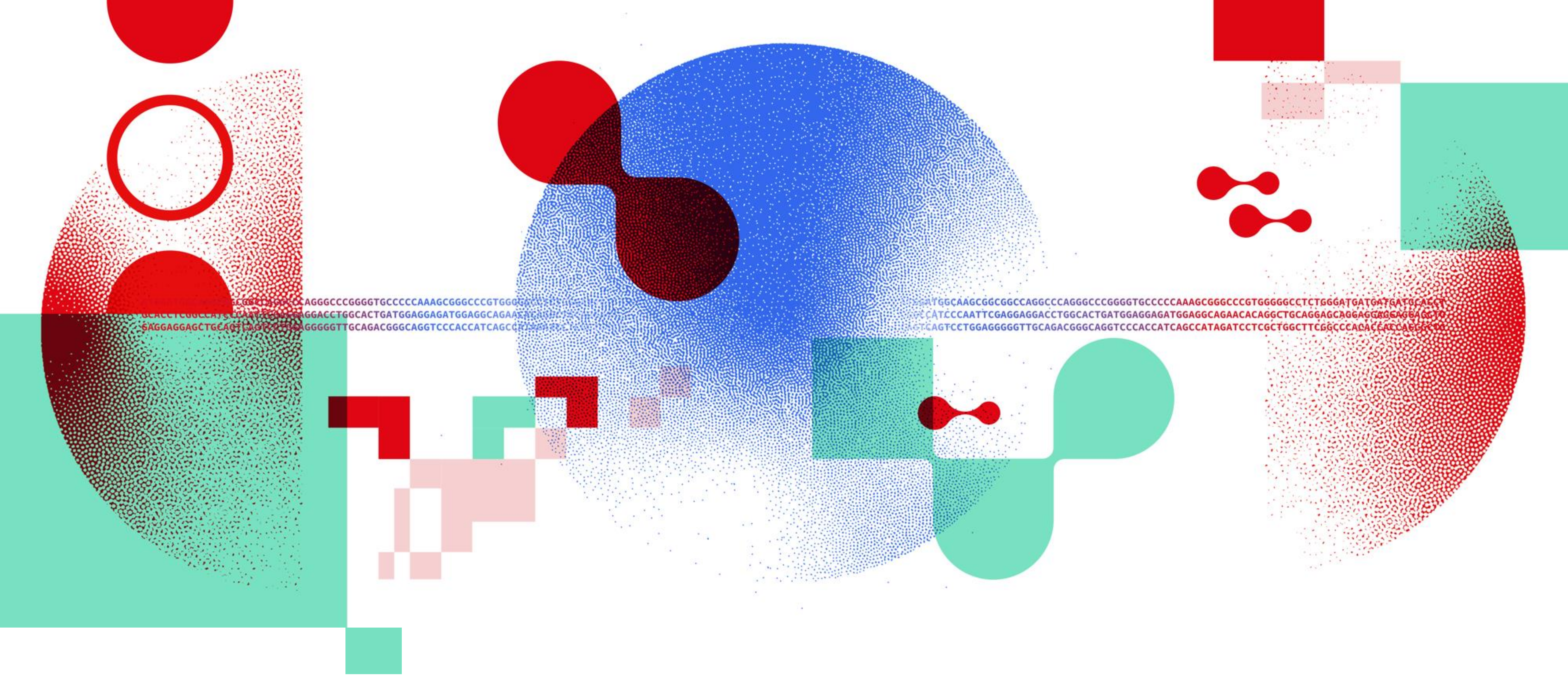
Domain-marker gene
"Spatial-domain-marker SVG"

Yan, Hua, Li. Nat Comm,, 2025

Conclusions

- Algorithms have been developed to incorporate spatial information in the process of clustering cells into spatial domains.
- Graph neural networks has integrated tissue images to improve the performance using anatomical information.

- Challenges remain
  - Methods tested in few datasets, mostly human dorsolateral prefrontal cortex (DLPFC) → problem with over-fitting
  - Scalability (tested in small datasets)



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss