# Spatially-resolved transcriptomic data integration

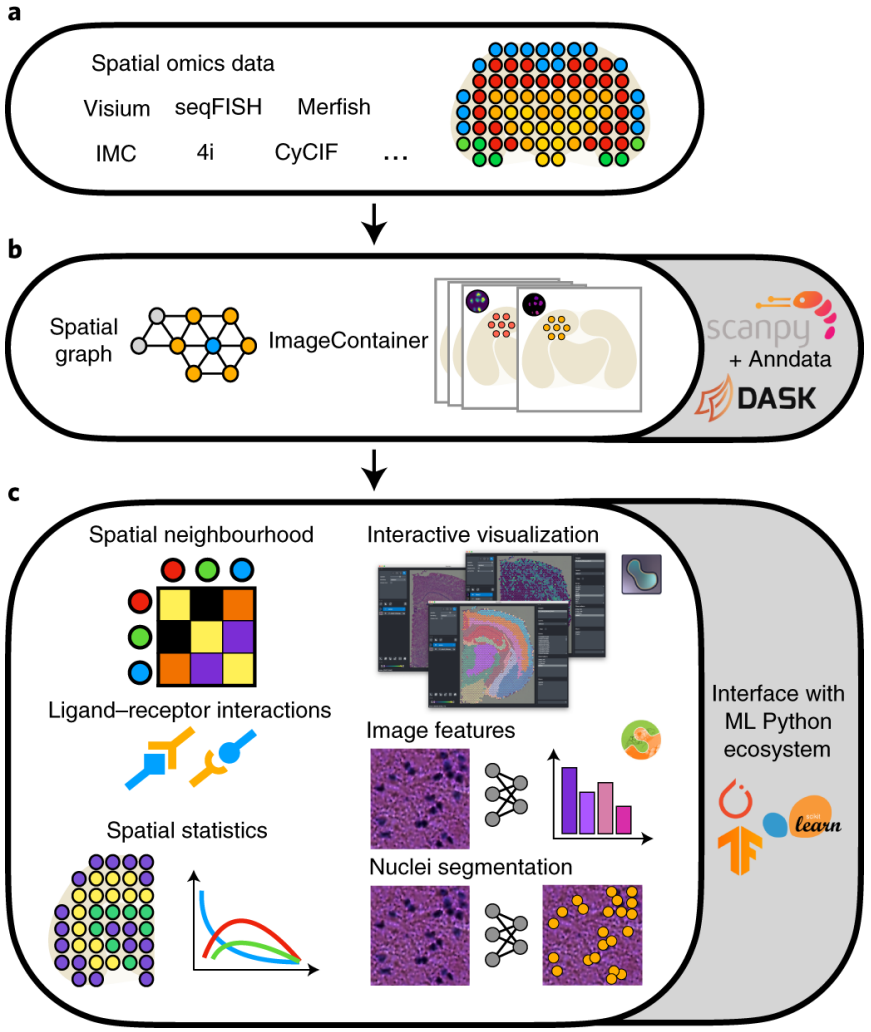
Ahmed Mahfouz

Human Genetics, Leiden University Medical Center
Pattern Recognition and Bioinformatics, TU Delft

# Spatial data analysis pipelines

- Image preprocessing
- Cell segmentation / segmentation-free
- Quality control / filtering / normalization
- Downstream analysis:
  - Cell annotation
  - Spatially-aware clustering (aka. Domain identification)
  - Cell-cell communication
  - Deconvolution
  - Predicting unmeasured genes
  - Consecutive sections alignment
  - …

**Pipelines:** Squidpy [Palla,…Theis, 2022], SPArrOW [Pollaris*, Vanneste*,…,Saeys 2024], MCMICRO [Shapiro,…, Sorger 2021], SOPA [Balmpey,…, Cournède 2024],…




**Squidpy:** Palla,…Theis (2022)

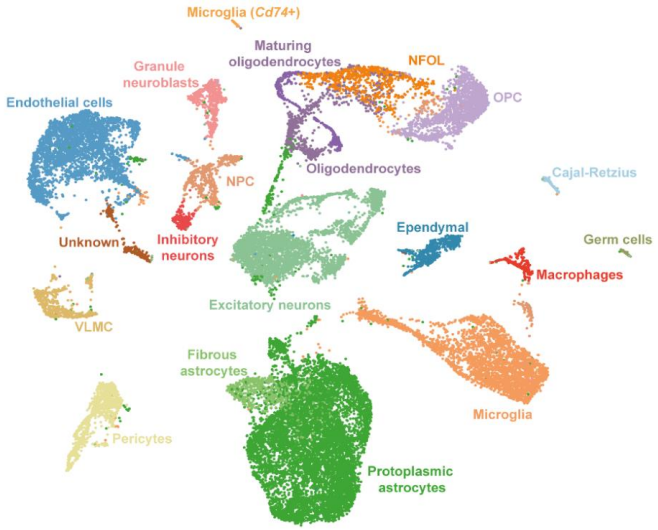
# Integration challenges in SRT

- Integrating SRT and sc/snRNA-seq data
- SRT data alignment (pseudo 3D, virtual block,…)
- Spatial multiomics

# Different technologies -> different questions

## Single-cell/nuclei RNA-seq

(droplet, combinatorial indexing, plate-based)

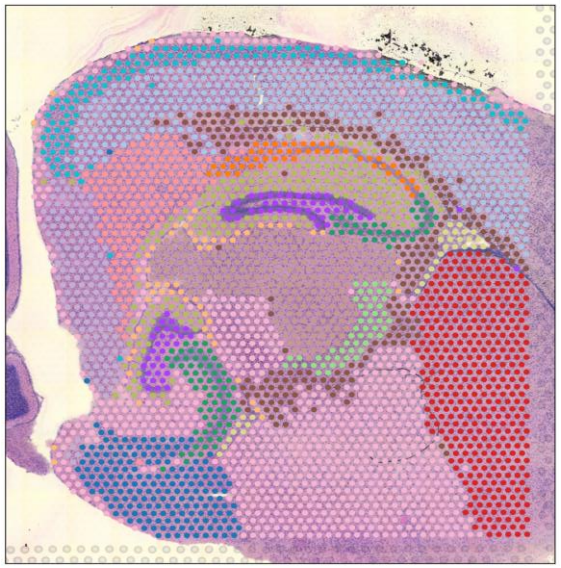



+ "Untargeted"

- No spatial information

## NGS-based

(Visium, Slide-seq, Stereo-seq, Open-ST,…)

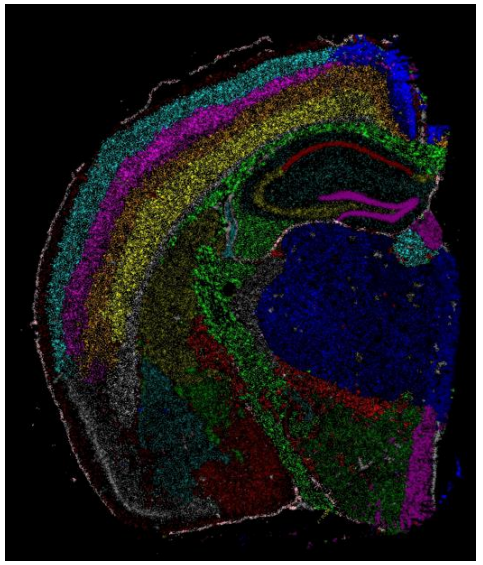


+ "Untargeted"

- Low resolution (spots)

## In-situ sequencing / x-FISH

(Xenium, EEL-FISH, MERFISH, seqFISH)

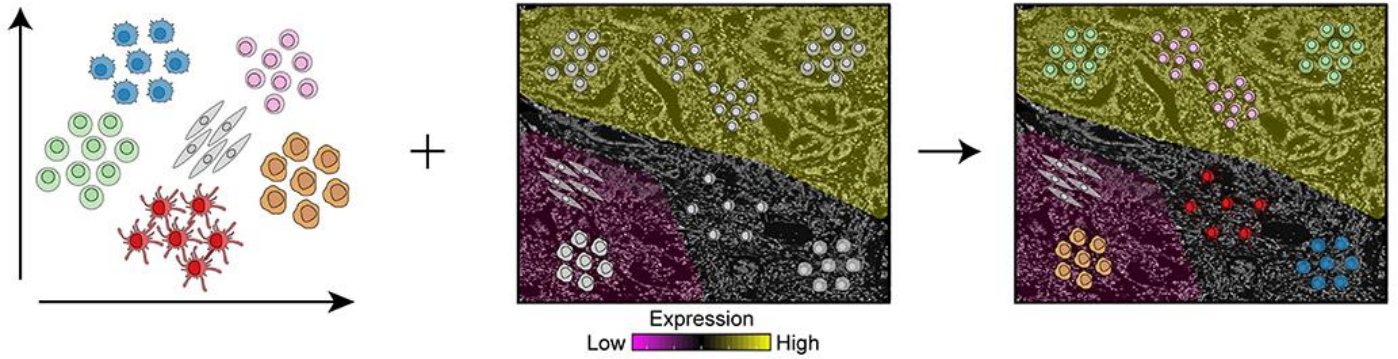


+ High resolution

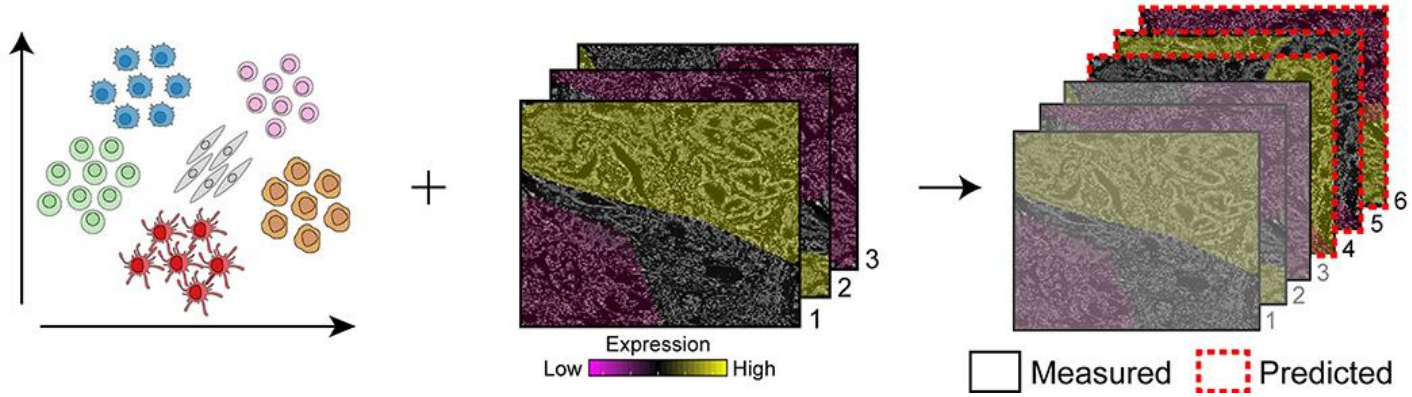
- Limited genomic features

# Single-cell + spatial

- Predict the location of dissociated cells
- Deconvolution




- Predicting spatial expression of unmeasured genes






de Vries,…, de Miranda (2020)

# Approaches for scRNA-seq & spatial data integration

| | Deconvolution | Imputation |
|---|---|---|
| **Joint embedding**<br>e.g. SpaGE, Seurat, LIGER, gimVI, ENVI, stPlus DSTG, SD |  |  |
| **Probabilistic modelling**<br>e.g. CARD, cell2location, RCTD, stereoscope, SpatialDecon, STRIDE, NMFreg, SpatialDWLS, SPOTlight, Stdeconvolve, SpiceMix, Berglund |  | |
| **Probabilistic mapping**<br>e.g. Tangram, novoSpaRc, SpaOTsc |  |  |

# Approaches for scRNA-seq & spatial data integration

| | Deconvolution | Imputation |
|---|---|---|
| **Joint embedding**<br>e.g. SpaGE, Seurat, LIGER, gimVI, ENVI, stPlus DSTG, SD |  |  |
| **Probabilistic modelling**<br>e.g. CARD, cell2location, RCTD, stereoscope, SpatialDecon, STRIDE, NMFreg, SpatialDWLS, SPOTlight, Stdeconvolve, SpiceMix, Berglund |  | |
| **Probabilistic mapping**<br>e.g. Tangram, novoSpaRc, SpaOTsc |  |  |

# Tangram

- Given scRNAseq ($S_{n_{cells} \times n_{genes}}$), and spatial data ($G_{n_{voxels} \times n_{genes}}$)
- Learn a mapping matrix $M_{n_{cells} \times n_{voxels}}$, $M_{ij} \geq 0$ is the probability of cell $i$ of being in voxel $j$, $\sum_{j}^{n_{voxels}} M_{ij} = 1$.
- $M^T S$: predicted spatial gene expression, $m_j = \sum_{i}^{n_{cells}} \frac{M_{ij}}{n_{cells}}$: predicted cell density in voxel $j$

Biancalani… Regev (2021)

# Tangram

- Given scRNAseq ($S_{n_{cells} \times n_{genes}}$), and spatial data ($G_{n_{voxels} \times n_{genes}}$)
- Learn a mapping matrix $M_{n_{cells} \times n_{voxels}}$, $M_{ij} \geq 0$ is the probability of cell $i$ of being in voxel $j$, $\sum_{j}^{n_{voxels}} M_{ij} = 1$.
- $M^T S$: predicted spatial gene expression, $m_j = \sum_{i}^{n_{cells}} \frac{M_{ij}}{n_{cells}}$: predicted cell density in voxel $j$

- Objective function:

$$\Phi(\widetilde{M}) = KL(\vec{m}, \vec{d}) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T S)_{*,k}, G_{*,k}\right) - \sum_{k}^{n_{voxels}} cos_{sim}\left((M^T S)_{j,*}, G_{j,*}\right)$$

Density term

gene/voxel expression term

voxel/gene expression term

Biancalani… Regev (2021)

# Tangram

- Given scRNAseq ($S_{n_{cells} \times n_{genes}}$), and spatial data ($G_{n_{voxels} \times n_{genes}}$)
- Learn a mapping matrix $M_{n_{cells} \times n_{voxels}}$, $M_{ij} \geq 0$ is the probability of cell $i$ of being in voxel $j$, $\sum_{j}^{n_{voxels}} M_{ij} = 1$.
- $M^T S$: predicted spatial gene expression, $m_j = \sum_{i}^{n_{cells}} \frac{M_{ij}}{n_{cells}}$: predicted cell density in voxel $j$

- Objective function:

$$\Phi(\tilde{M}) = KL(\vec{m}, \vec{d}) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T S)_{*,k}, G_{*,k}\right) - \sum_{k}^{n_{voxels}} cos_{sim}\left((M^T S)_{j,*}, G_{j,*}\right)$$

Density term

gene/voxel expression term

voxel/gene expression term

Biancalani… Regev (2021)

# Tangram

- Given scRNAseq ($S_{n_{cells} \times n_{genes}}$), and spatial data ($G_{n_{voxels} \times n_{genes}}$)
- Learn a mapping matrix $M_{n_{cells} \times n_{voxels}}$, $M_{ij} \geq 0$ is the probability of cell $i$ of being in voxel $j$, $\sum_{j}^{n_{voxels}} M_{ij} = 1$.
- $M^T S$: predicted spatial gene expression, $m_j = \sum_{i}^{n_{cells}} \frac{M_{ij}}{n_{cells}}$: predicted cell density in voxel $j$

- Objective function:

$$\Phi(\widetilde{M}) = KL(\vec{m}, \vec{d}) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T S)_{*,k}, G_{*,k}\right) - \sum_{k}^{n_{voxels}} cos_{sim}\left((M^T S)_{j,*}, G_{j,*}\right)$$

Density term | gene/voxel expression term | voxel/gene expression term

Biancalani… Regev (2021)

# Tangram

- Actual objective function:

$$\Phi\left(\tilde{M}, \vec{\tilde{f}}\right) = KL\left(\vec{\mathbf{m}}^{\mathbf{f}}, \vec{\mathbf{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left(\left(M^T S^f\right)_{*,k}, G_{*,k}\right)$$

$$- \sum_{j}^{n_{voxels}} cos_{sim}\left(\left(M^T S^f\right)_{j,*}, G_{j,*}\right) - \lambda_{r_1} \sum_{i,j}^{n_{cells},\, n_{voxels}} M_{ij} log\left(M_{ij}\right)$$

$$+ abs\left(\sum_{i}^{n_{cells}} f_i - \mathrm{n}_{\mathrm{target\_cells}}\right) + \sum_{i}^{n_{cells}} \left(f_i - f_i^2\right).$$

Biancalani… Regev (2021)

# Tangram

- Actual objective function:

$$\Phi\left(\tilde{M}, \vec{\tilde{f}}\right) = KL\left(\vec{\mathbf{m}^{\mathbf{f}}}, \vec{\mathbf{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left(\left(M^T S^f\right)_{*,k}, G_{*,k}\right)$$

$$- \sum_{j}^{n_{voxels}} cos_{sim}\left(\left(M^T S^f\right)_{j,*}, G_{j,*}\right) - \lambda_{r_1} \sum_{i,j}^{n_{cells},\, n_{voxels}} M_{ij} log\left(M_{ij}\right)$$

Entropy regularizer

$$+ abs\left(\sum_{i}^{n_{cells}} f_i - \mathrm{n}_{\mathrm{target\_cells}}\right) + \sum_{i}^{n_{cells}} \left(f_i - f_i^2\right).$$

Biancalani… Regev (2021)

# Tangram

- Actual objective function:

$$\Phi\left(\tilde{M}, \vec{\tilde{f}}\right) = KL\left(\vec{\mathbf{m}^{\mathbf{f}}}, \vec{\mathbf{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left(\left(M^T S^f\right)_{*,k}, G_{*,k}\right)$$

$$- \sum_{j}^{n_{voxels}} cos_{sim}\left(\left(M^T S^f\right)_{j,*}, G_{j,*}\right) - \lambda_{r_1} \sum_{i,j}^{n_{cells},\, n_{voxels}} M_{ij} log\left(M_{ij}\right)$$

$$+ abs\left(\sum_{i}^{n_{cells}} f_i - \mathrm{n}_{\mathrm{target\_cells}}\right) + \sum_{i}^{n_{cells}} \left(f_i - f_i^2\right).$$

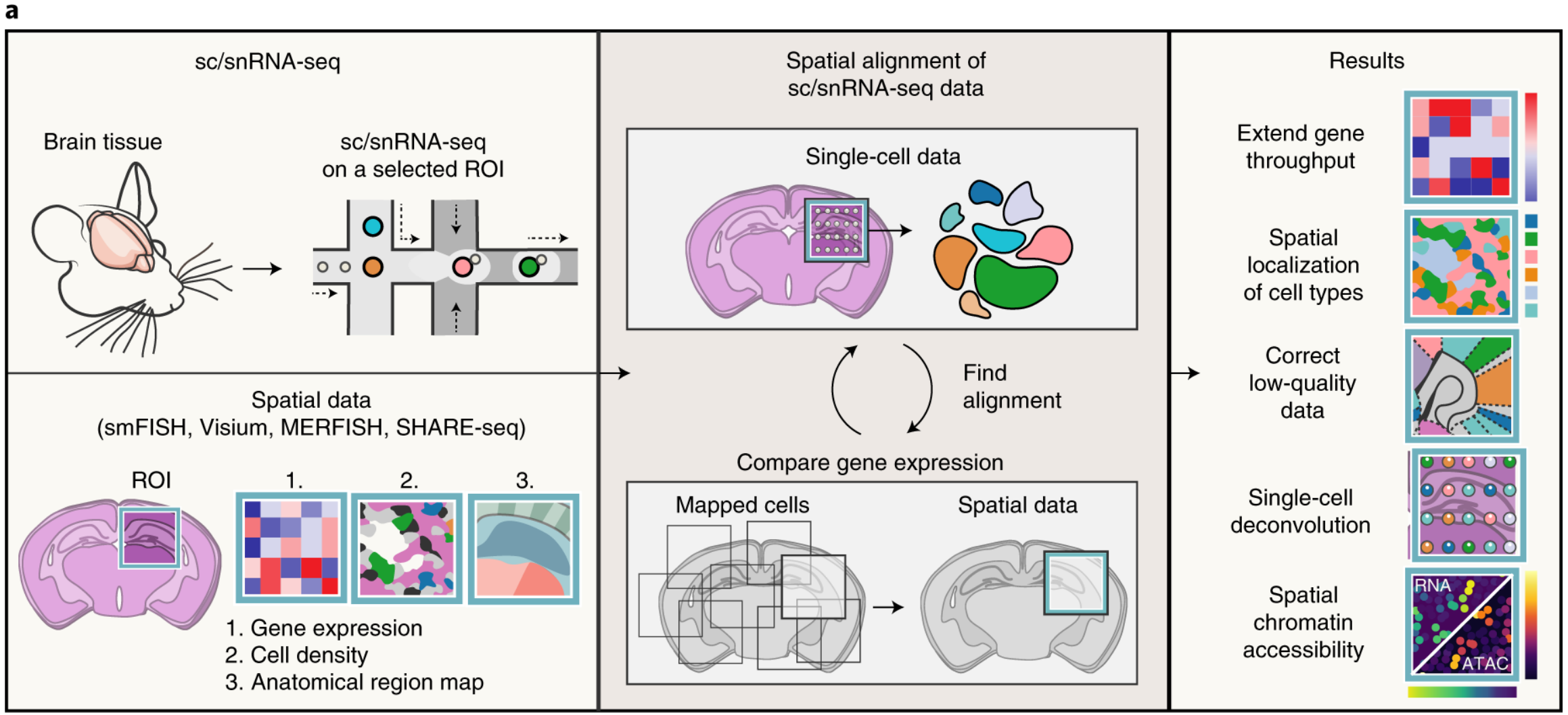
Count term

Filter regularizer
(promote Boolean values)

$\overrightarrow{f_{n_{cells}}}$: filter vector

Biancalani… Regev (2021)

# Tangram




Biancalani… Regev (2021)

# cell2location

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

$s: spatial\ location$
$g: gene$
$e: batch\ (section, slide, \dots)$

Kleshchevnikov… Stegle, Baryaktar (2022)

# cell2location

$s$: *spatial location*
$g$: *gene*
$e$: *batch (section, slide, ...)*

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

$$\mu_{s,g} = \left( \underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\ g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Kleshchevnikov… Stegle, Baryaktar (2022)

# cell2location

$s$: *spatial location*
$g$: *gene*
$e$: *batch (section, slide, ...)*

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

Reference signature from scRNA-seq data

$$\mu_{s,g} = \left( \underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Cell type abundance. Prior: similarity in cell type between locations

Kleshchevnikov … Stegle, Baryaktar (2022)

# cell2location

$s$: *spatial location*
$g$: *gene*
$e$: *batch (section, slide, …)*

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

$$\mu_{s,g} = \left( \underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Reference signature from scRNA-seq data

To account for differences between scRNA-seq and spatial

Cell type abundance. Prior: similarity in cell type between locations

Kleshchevnikov… Stegle, Baryaktar (2022)

# cell2location

$s$: *spatial location*
$g$: *gene*
$e$: *batch (section, slide, …)*

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

$$\mu_{s,g} = \left( \underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Reference signature from scRNA-seq data

To account for differences between scRNA-seq and spatial

Cell type abundance. Prior: similarity in cell type between locations

To account for free floating RNA (background noise)

Kleshchevnikov … Stegle, Baryaktar (2022)

# cell2location

$s$: *spatial location*
$g$: *gene*
$e$: *batch (section, slide, …)*

$$Spatial\ counts \sim NB(\mu_{s,g}, \alpha_{e,g})$$

$$\mu_{s,g} = \left( \underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Reference signature from scRNA-seq data

To account for differences between scRNA-seq and spatial

Cell type abundance. Prior: similarity in cell type between locations

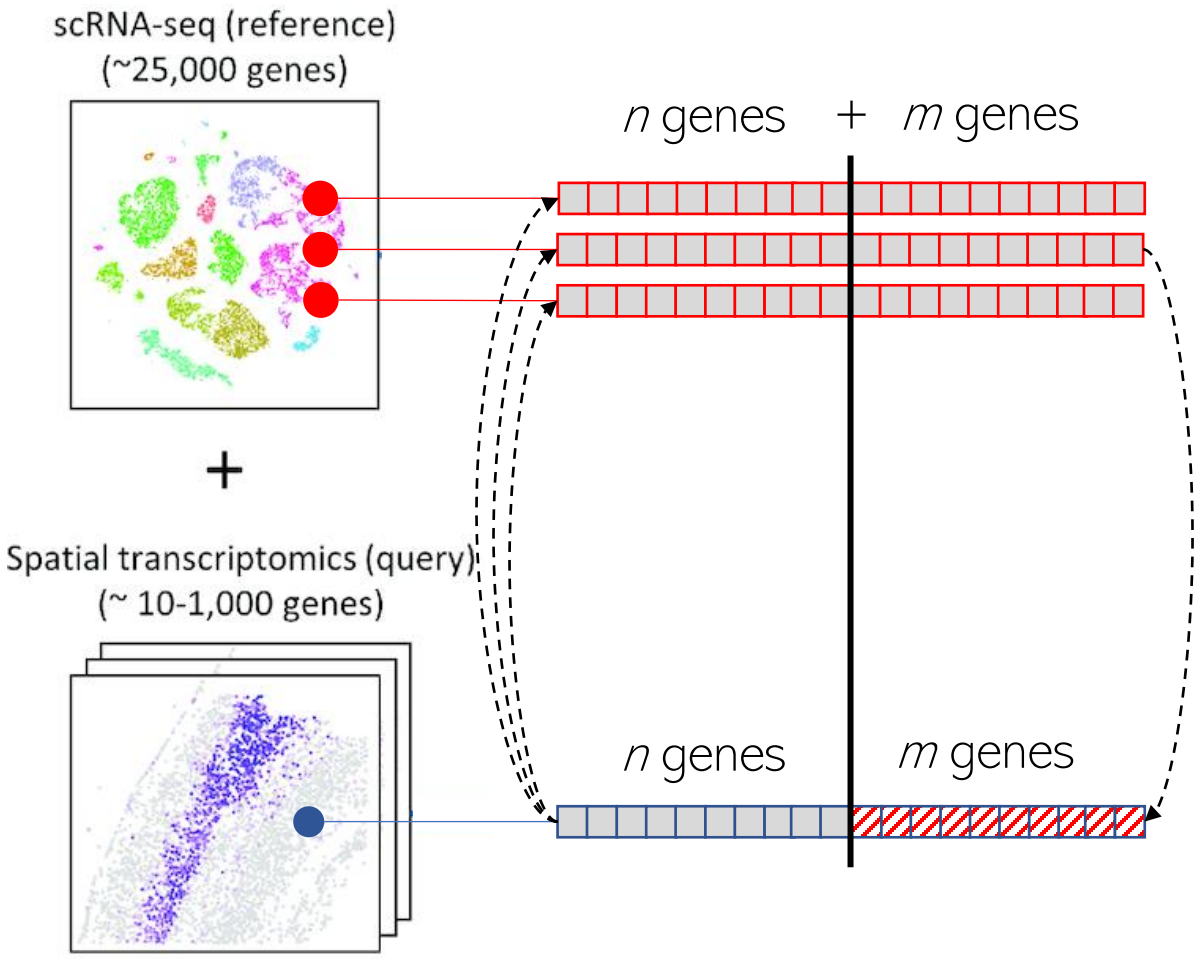
To account for free floating RNA (background noise)

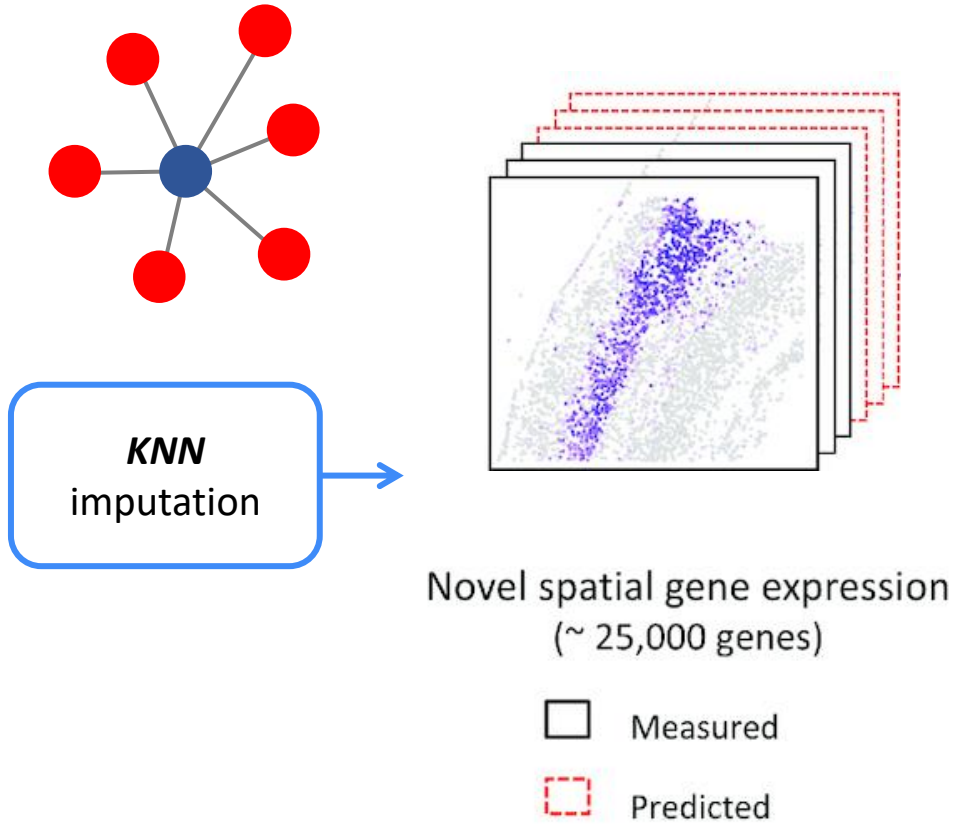
To account for differences in sensitivity within a section

Kleshchevnikov … Stegle, Baryaktar (2022)

Tamim Abdelaal

# SpaGE: Spatial Gene Expression Enhancement



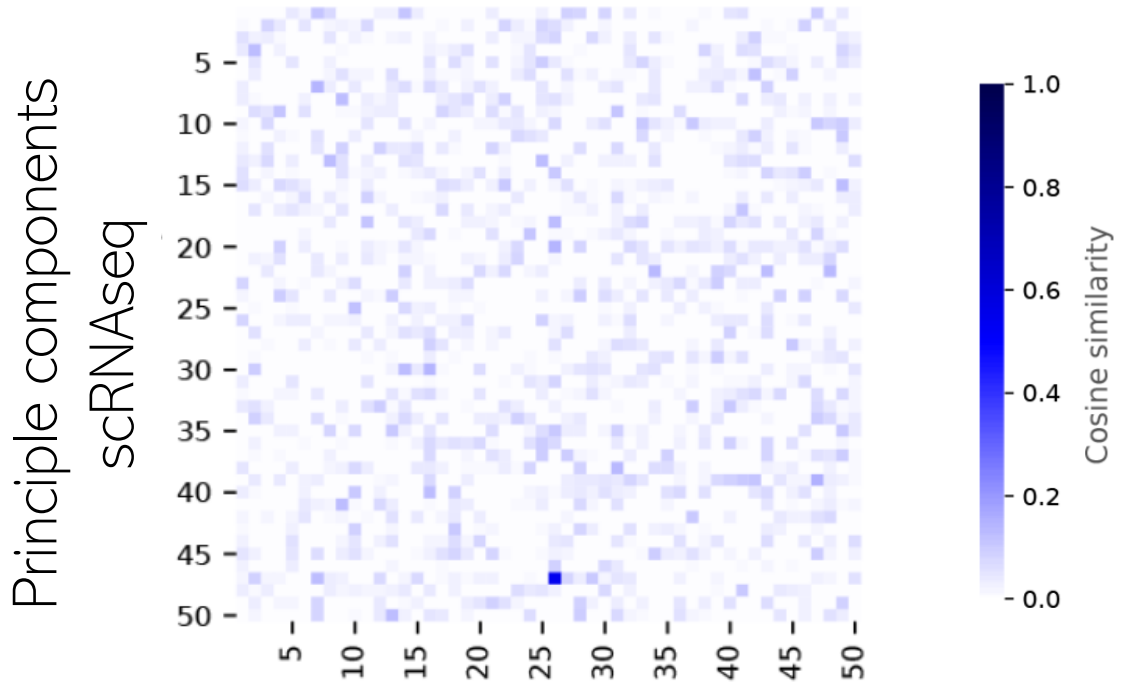






Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# Problem: single-cell and spatial data don't align

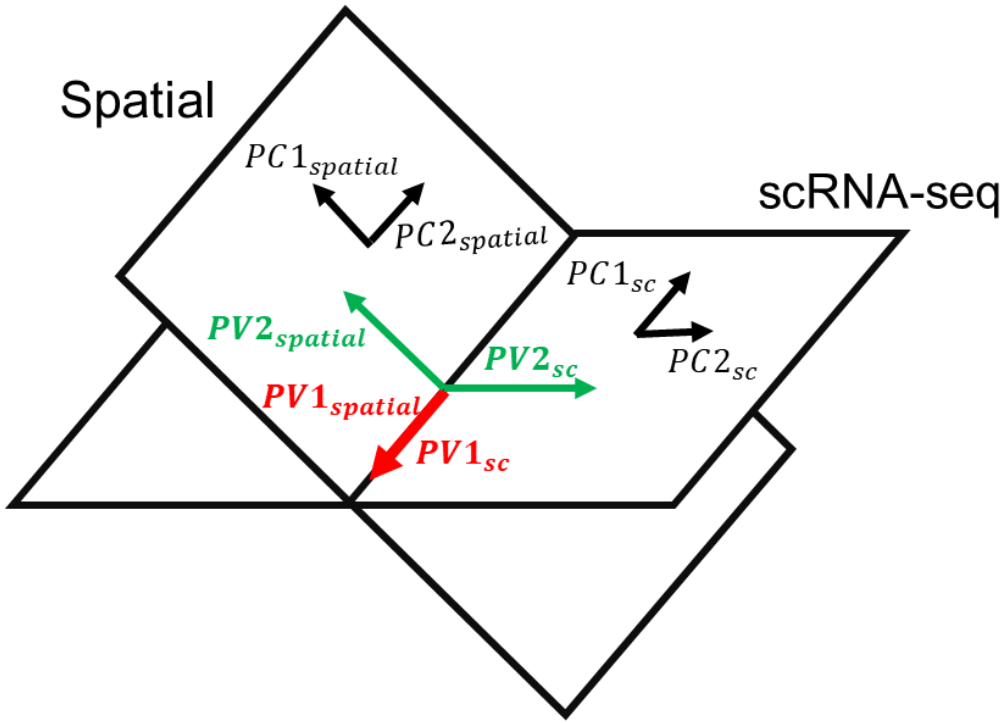
Similarity between principal components

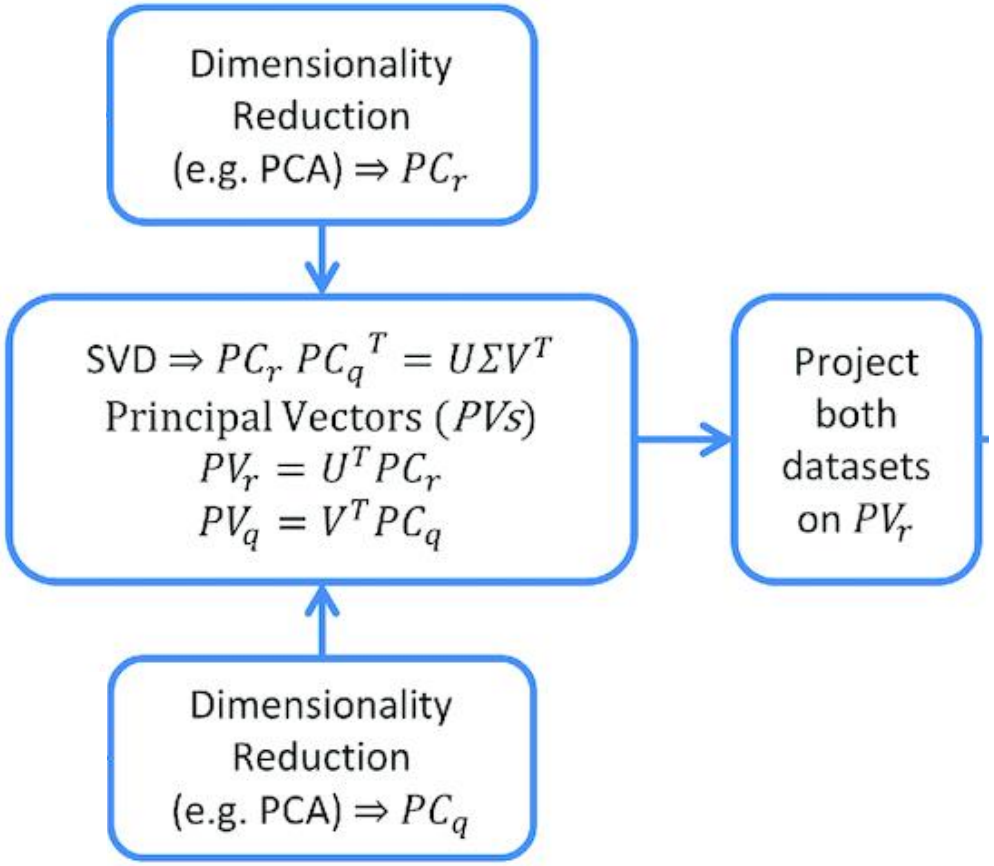



Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# Aligning single-cell and spatial data

Domain Adaptation using PRECISE

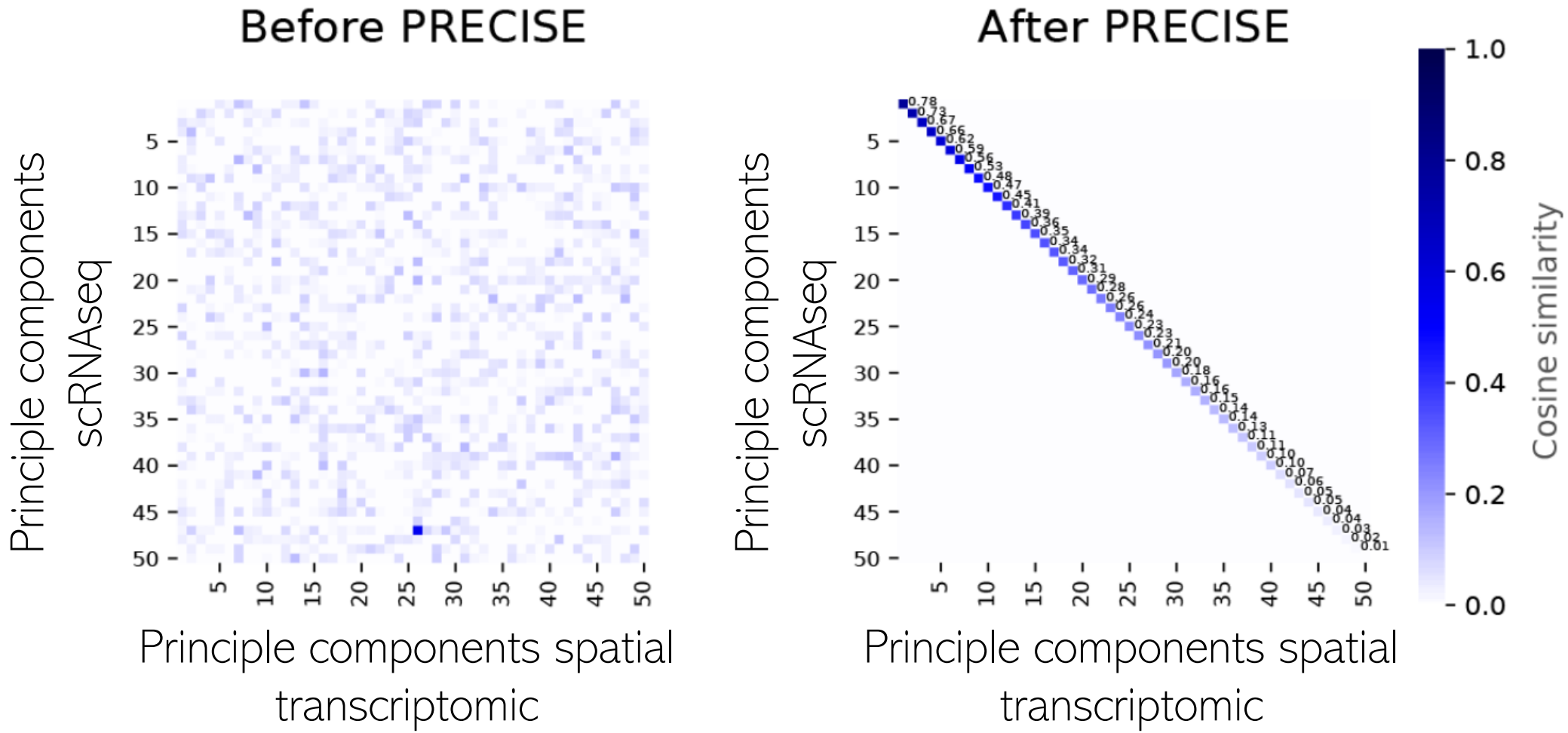




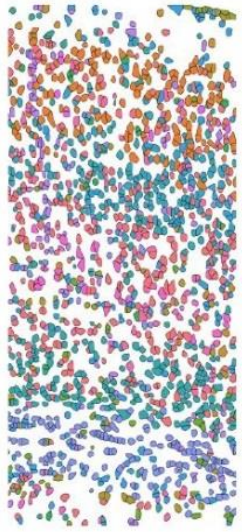

Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# Aligning single-cell and spatial data




Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# SpaGE in primary visual cortex (VISp)

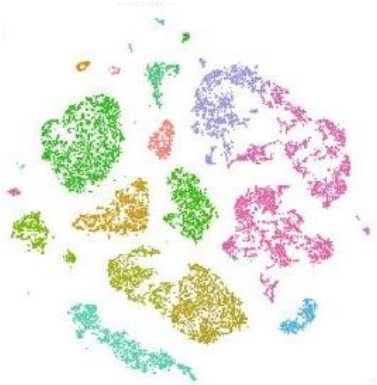


**STARmap**

1,549 cells

1,020 genes

Wang et al. Science 2018



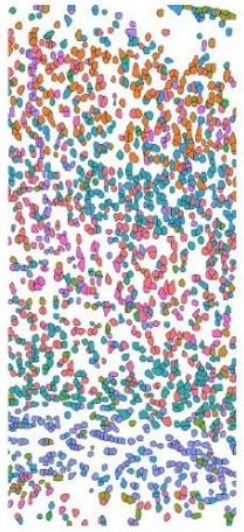
**scRNA-seq**

14,249 cells

34,617 transcripts

Tasic et al. Nature 2018

# SpaGE in primary visual cortex (VISp)

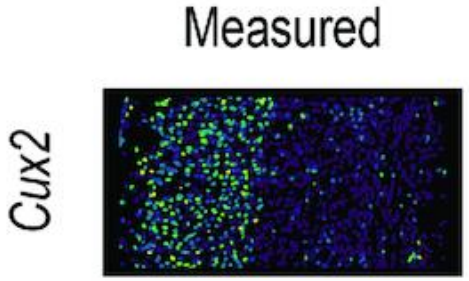


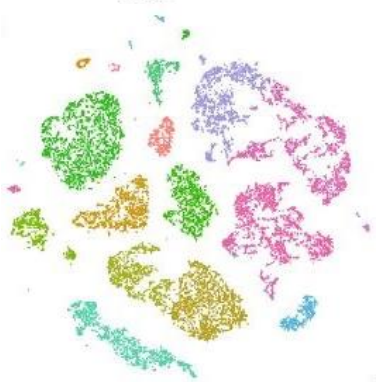
**STARmap**

1,549 cells

1,020 genes

Wang et al. Science 2018



**scRNA-seq**

14,249 cells

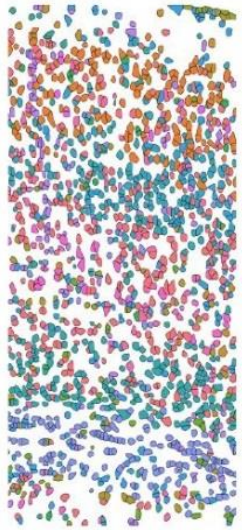
34,617 transcripts

Tasic et al. Nature 2018



Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# SpaGE in primary visual cortex (VISp)

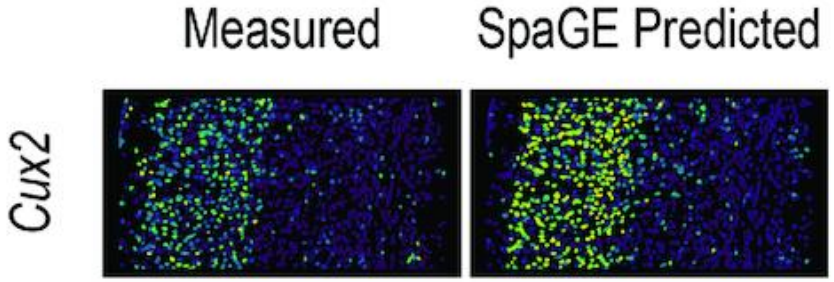


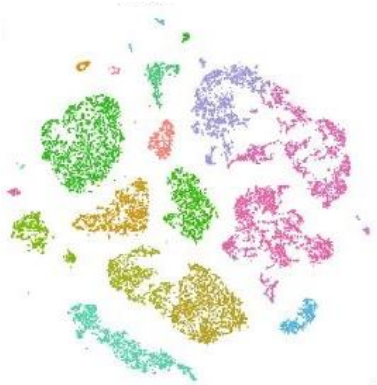
**STARmap**

1,549 cells

1,020 genes

Wang et al. Science 2018



**scRNA-seq**

14,249 cells

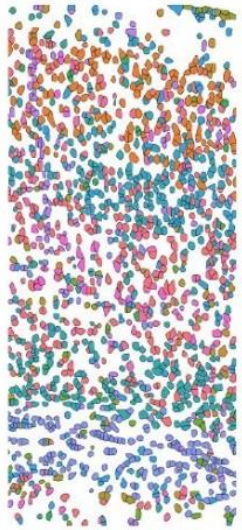
34,617 transcripts

Tasic et al. Nature 2018



Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# SpaGE in primary visual cortex (VISp)

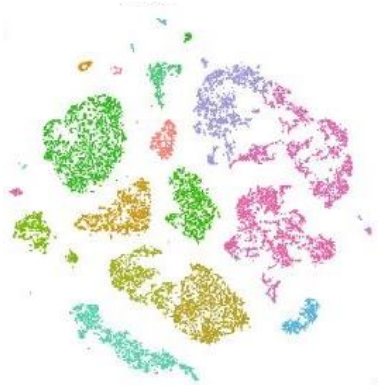


**STARmap**

1,549 cells

1,020 genes

Wang et al. Science 2018

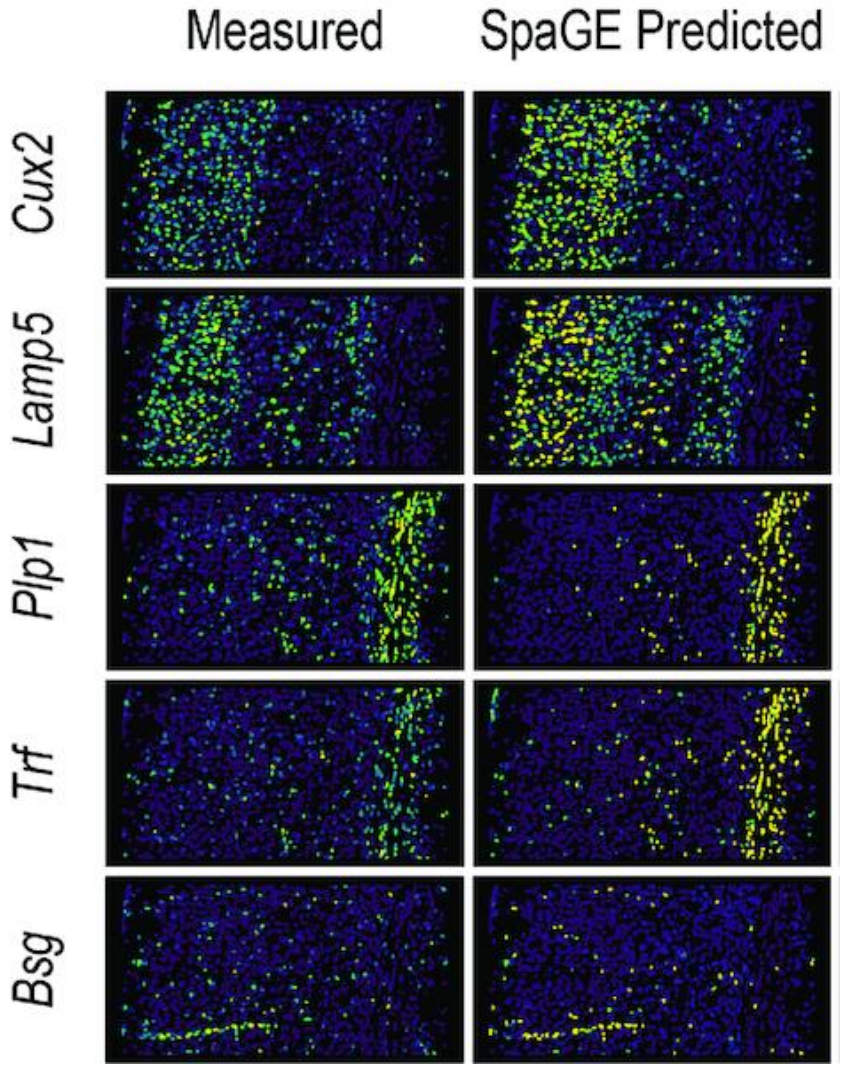


**scRNA-seq**

14,249 cells

34,617 transcripts

Tasic et al. Nature 2018






Abdelaal, Mourragui, Mahfouz*, Reinders* (2020)

# How do all these methods compare to each other?

nature communications

Article

https://doi.org/10.1038/s41467-023-37168-7

## A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics

Received: 30 September 2022

Accepted: 3 March 2023

Published online: 21 March 2023

Check for updates

Haoyang Li[1,2,6], Juexiao Zhou[1,2,6], Zhongxiao Li[1,2], Siyuan Chen[1,2], Xingyu Liao[1,2], Bin Zhang[1,2], Ruochi Zhang[3], Yu Wang[3], Shiwei Sun[4,5] & Xin Gao[1,2]

Spatial transcriptomics technologies are used to profile transcriptomes while preserving spatial information, which enables high-resolution characterization of transcriptional patterns and reconstruction of tissue architecture. Due to the existence of low-resolution spots in recent spatial transcriptomics technologies, uncovering cellular heterogeneity is crucial for disentangling the spatial patterns of cell types, and many related methods have been proposed. Here, we benchmark 18 existing methods resolving a cellular deconvolution task with 50 real-world and simulated datasets by evaluating the accuracy, robustness, and usability of the methods. We compare these methods comprehensively using different metrics, resolutions, spatial transcriptomics technologies, spot numbers, and gene numbers. In terms of performance, CARD, Cell2location, and Tangram are the best methods for conducting the cellular deconvolution task. To refine our comparative results, we provide decision-tree-style guidelines and recommendations for method selection and their additional features, which will help users easily choose the best method for fulfilling their concerns.

Li*, Zhou*..., Gao(2022)

ANALYSIS

https://doi.org/10.1038/s41592-022-01480-9

nature methods

Check for updates

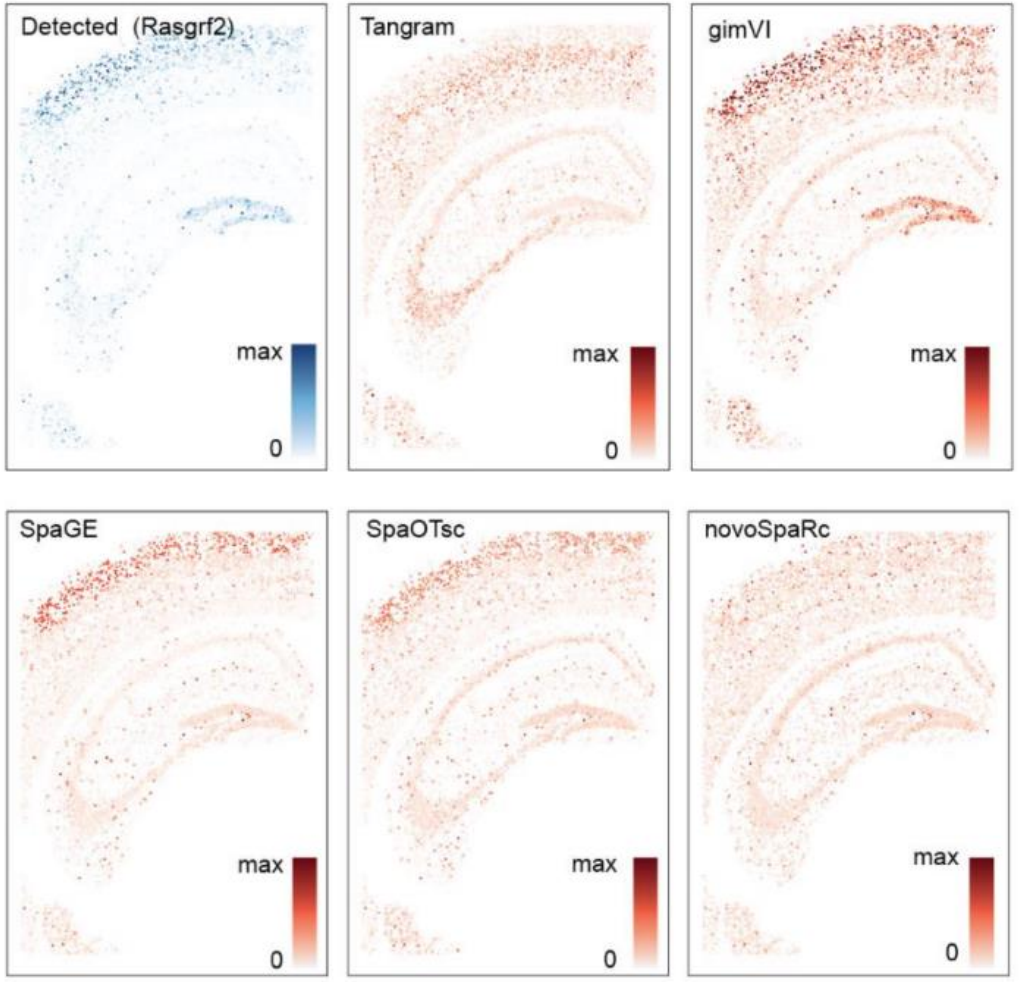
## Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution

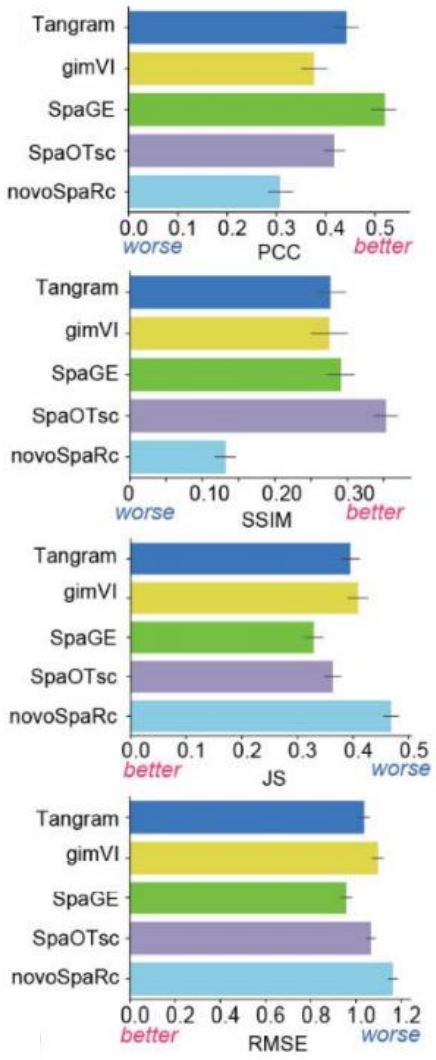
Bin Li[1,7], Wen Zhang[1,2,7], Chuang Guo[1,7], Hao Xu[1,2], Longfei Li[3], Minghao Fang[3], Yinlei Hu[4], Xinye Zhang[3], Xinfeng Yao[1], Meifang Tang[1], Ke Liu[1], Xuetong Zhao[5], Jun Lin[1,2], Linzhao Cheng[3], Falai Chen[4], Tian Xue[3] and Kun Qu[1,2,6]

**Spatial transcriptomics approaches have substantially advanced our capacity to detect the spatial distribution of RNA transcripts in tissues, yet it remains challenging to characterize whole-transcriptome-level data for single cells in space. Addressing this need, researchers have developed integration methods to combine spatial transcriptomic data with single-cell RNA-seq data to predict the spatial distribution of undetected transcripts and/or perform cell type deconvolution of spots in histological sections. However, to date, no independent studies have comparatively analyzed these integration methods to benchmark their performance. Here we present benchmarking of 16 integration methods using 45 paired datasets (comprising both spatial transcriptomics and scRNA-seq data) and 32 simulated datasets. We found that Tangram, gimVI, and SpaGE outperformed other integration methods for predicting the spatial distribution of RNA transcripts, whereas Cell2location, SpatialDWLS, and RCTD are the top-performing methods for the cell type deconvolution of spots. We provide a benchmark pipeline to help researchers select optimal integration methods to process their datasets.**

Li*, Guo*,..., Qu (2022)

# Imputation performance on Xenium data



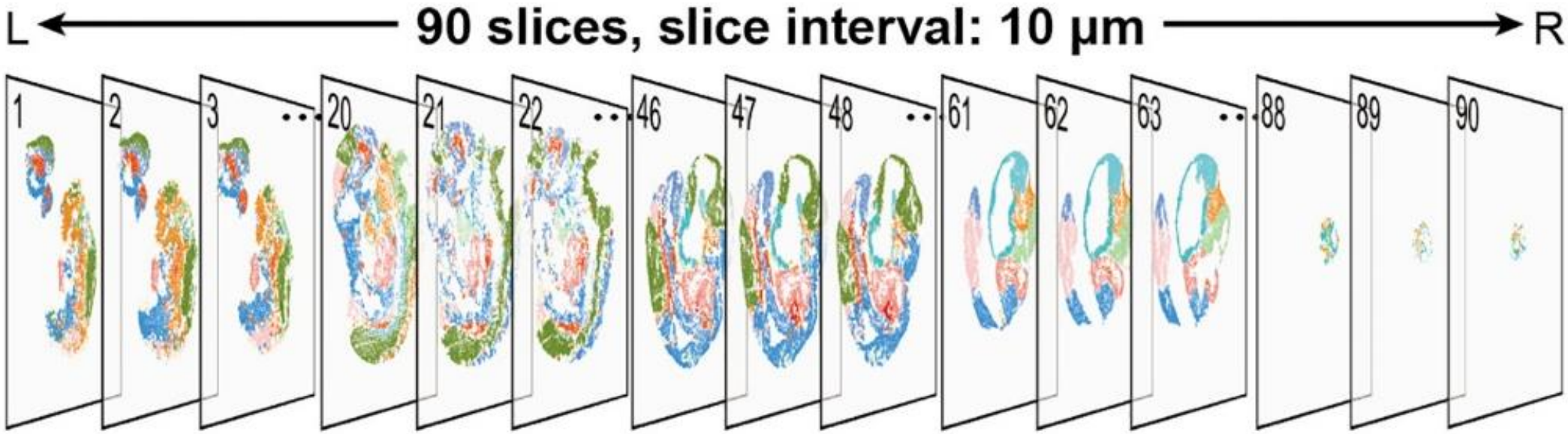




Salas,…, Nilsson (2023)

# SRT data alignment

Pseudo 3D / virtual blocks / …

# Scalable SRT allows whole tissue mapping using consecutive sections




Qiu,…, Lai (2024)

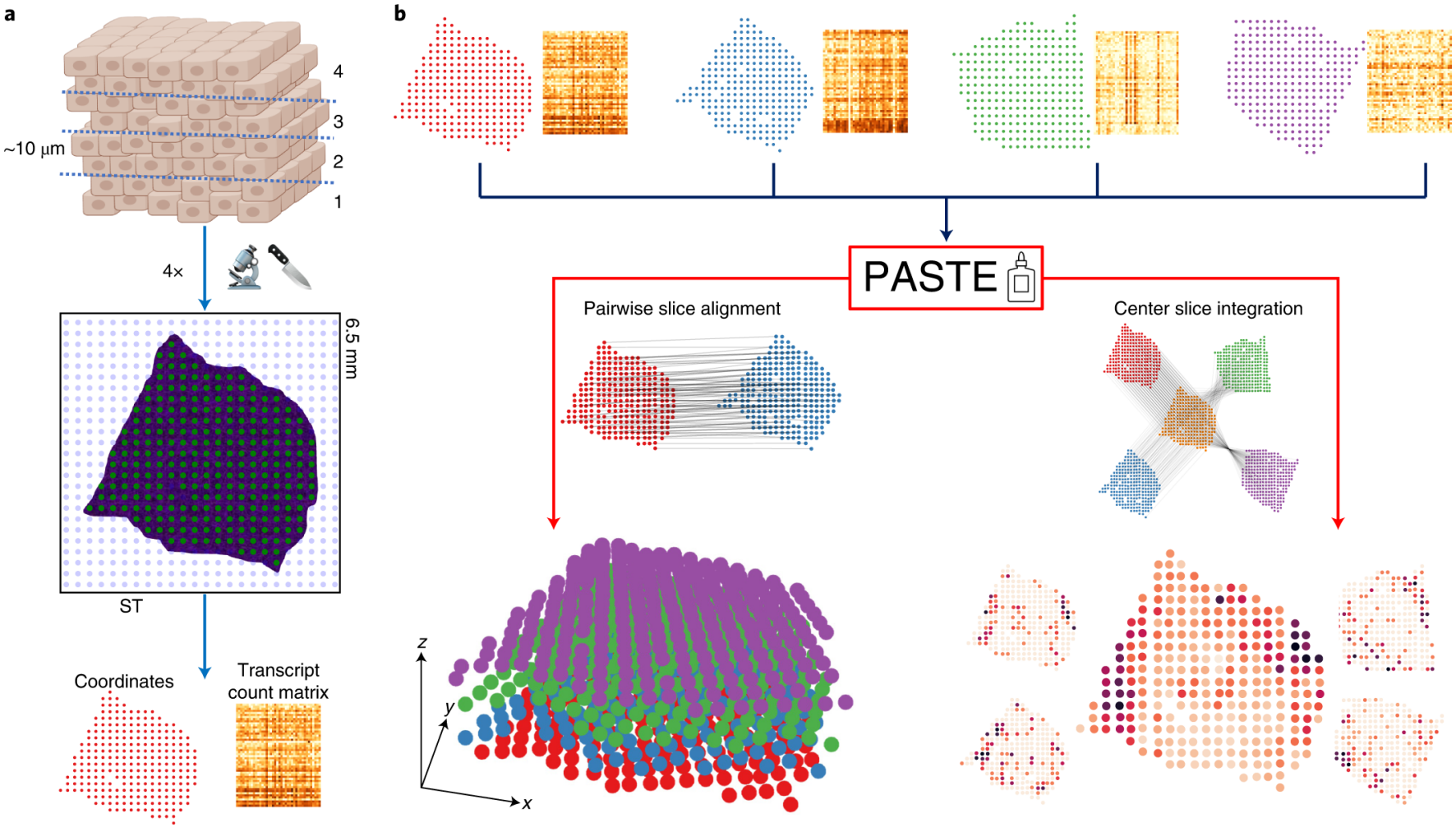
# SRT alignment approaches

- Alignment methods: designed to align or match spots or cells from different ST sections or datasets to a common spatial or anatomical reference
  - e.g. PASTE, PASTE2, SPACEL, STalign, GPSA, STIM, CAST

- Integration methods: learn shared latent spot embeddings
  - STAligner, DeepST, PRECAST, SPIRAL

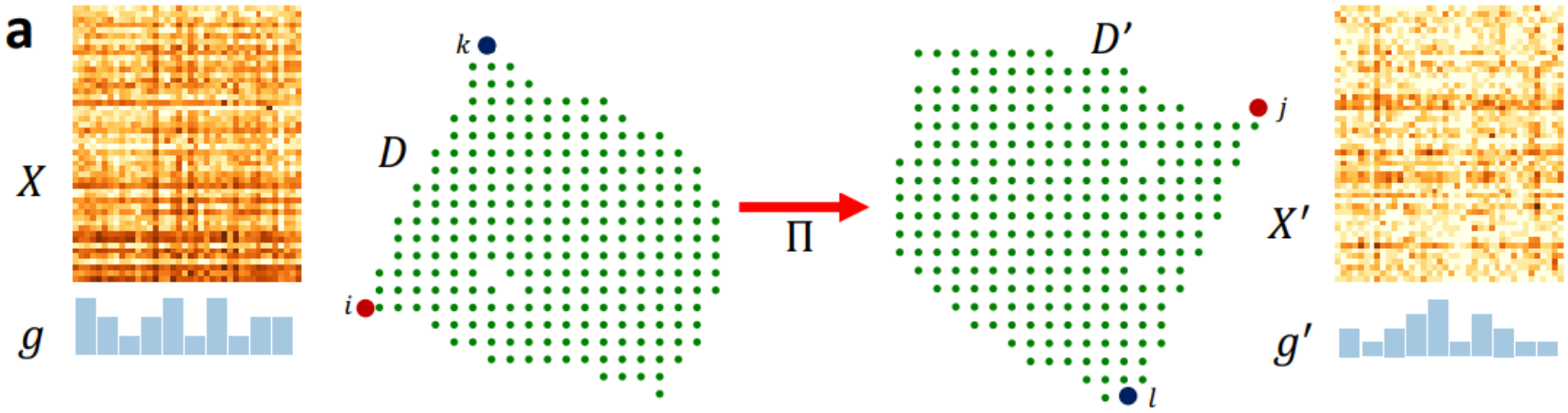
Hu,…, Zhou (2024)

# PASTE




Zeira,…, Raphael (2022)

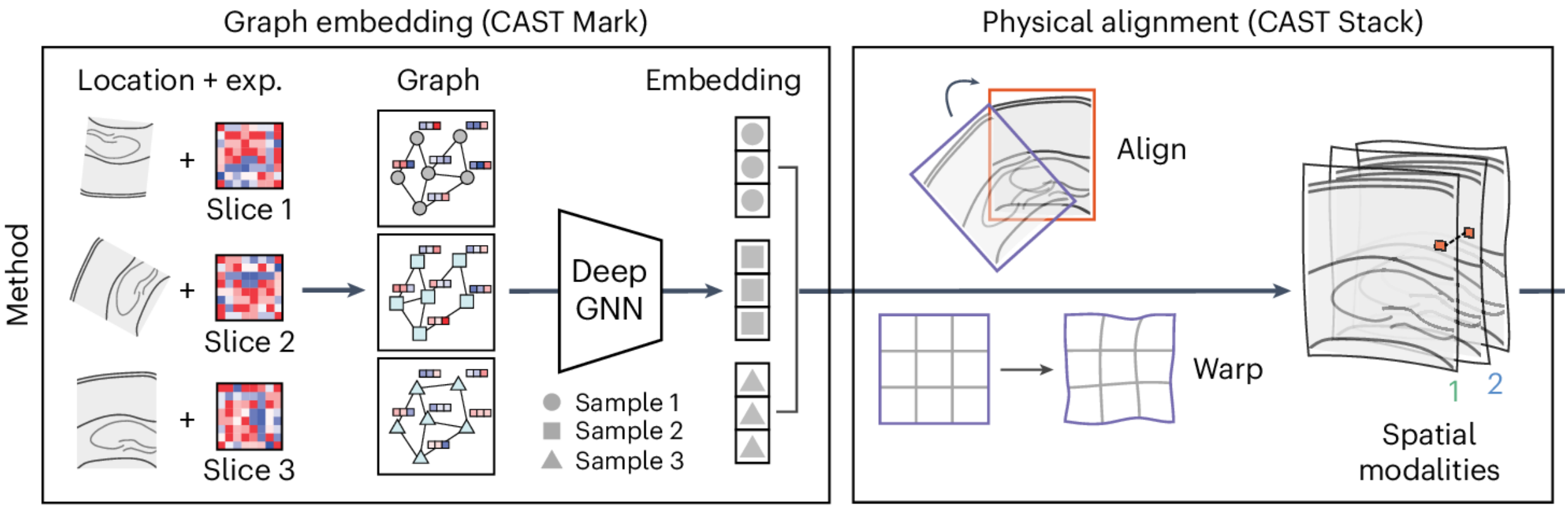
# PASTE




$$F(\Pi;\, X, D, X', D', c, \alpha) = (1-\alpha)\underbrace{\sum_{i,j} c(x_{\cdot i}, x'_{\cdot j})\pi_{ij}}_{\text{Gene expression similarity}} + \alpha \underbrace{\sum_{i,j,k,l} (d_{ik} - d'_{jl})^2 \pi_{ij}\pi_{kl}}_{\text{Spatial distance preservation}}.$$

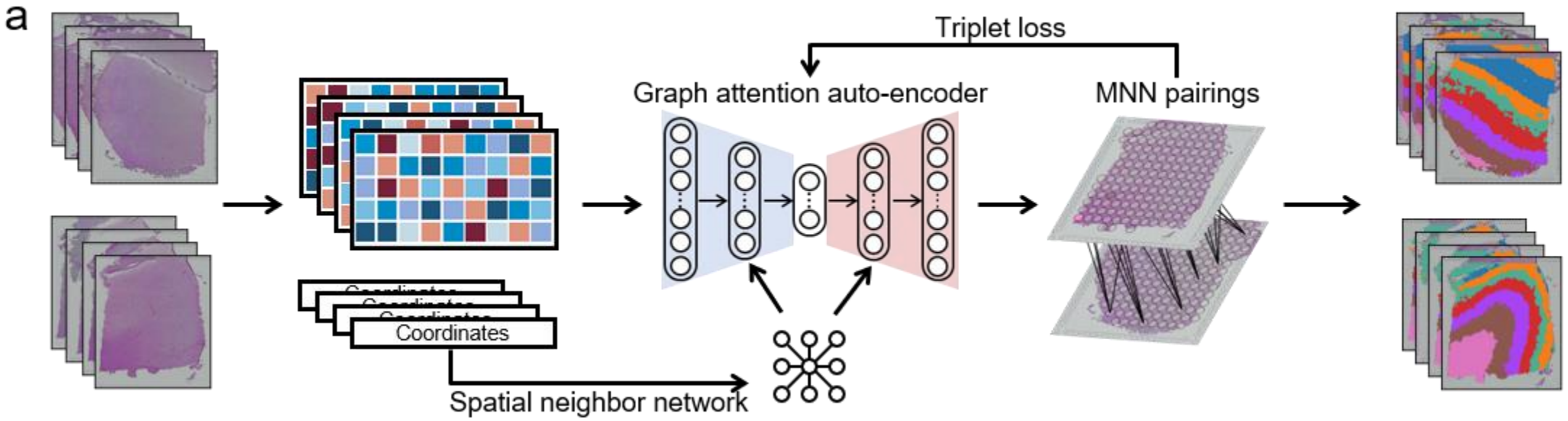
Zeira,…, Raphael (2022)

# CAST




Tang,…, Wang (2024)

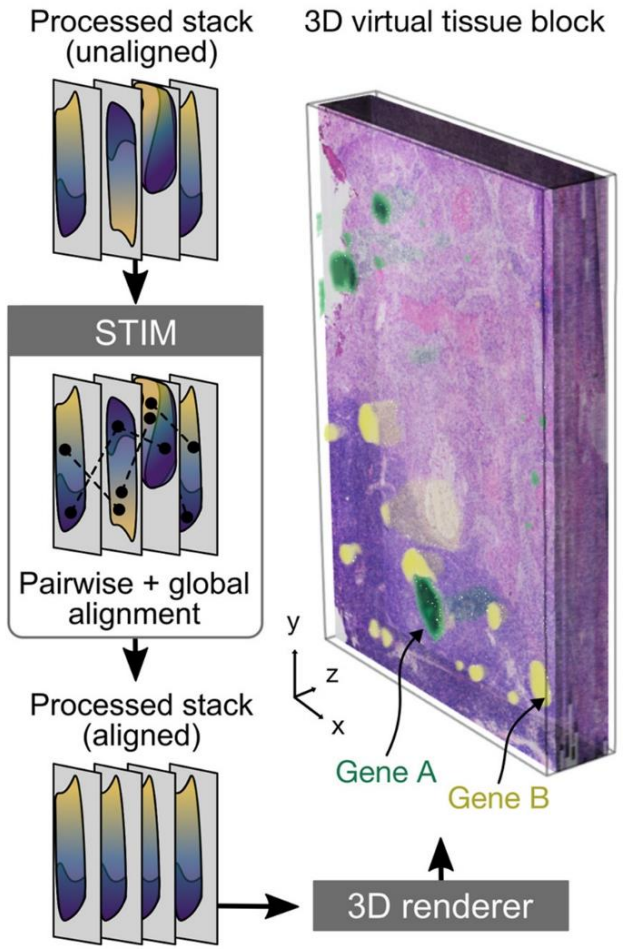
# STAligner

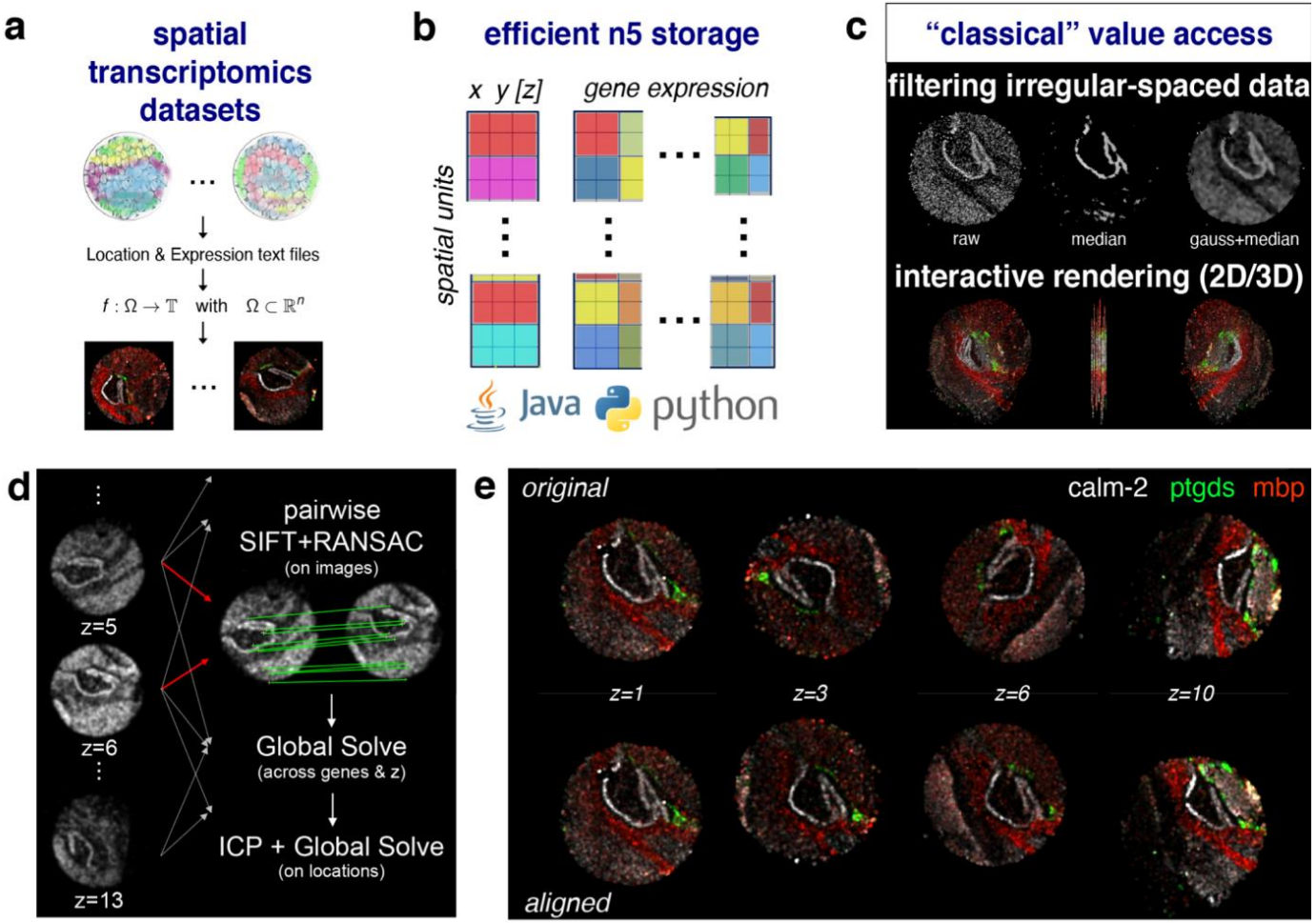



**Triplet:** anchor-positive and anchor-negative spot pairs
**Anchor-positive:** mutual nearest neighbors with similar gene expressions but belong to two different slices
**Anchor-negative:** a pair that belongs to the same slice with different spatial positions and dissimilar expressions

Zhou,…, Zhang (2023)

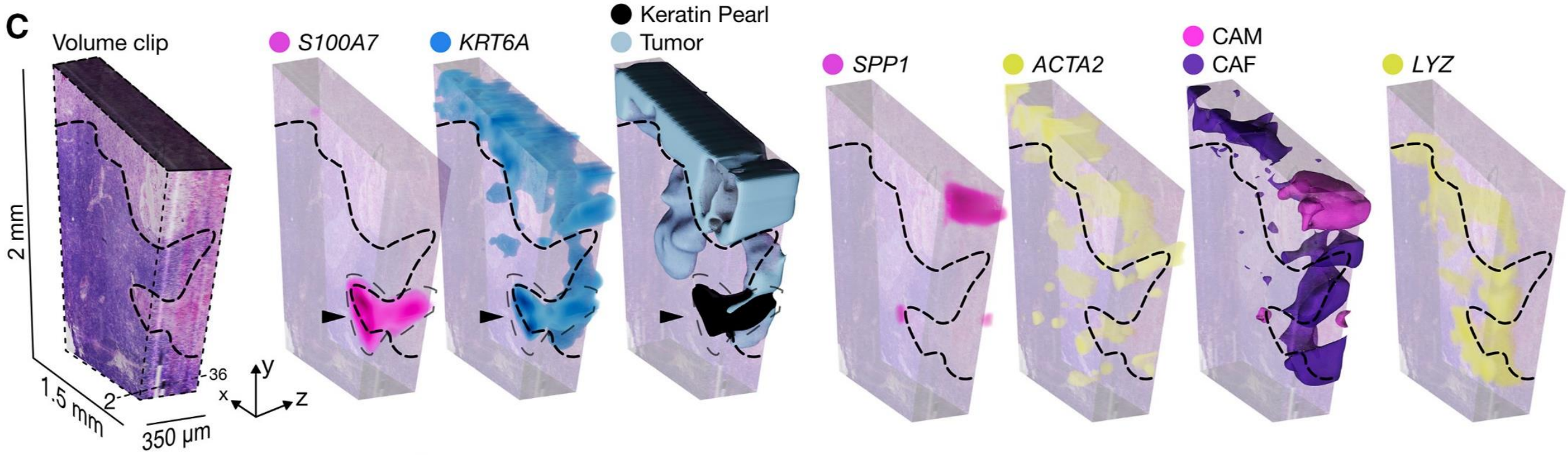
# STIM: Spatial Transcriptomics Imaging Framework



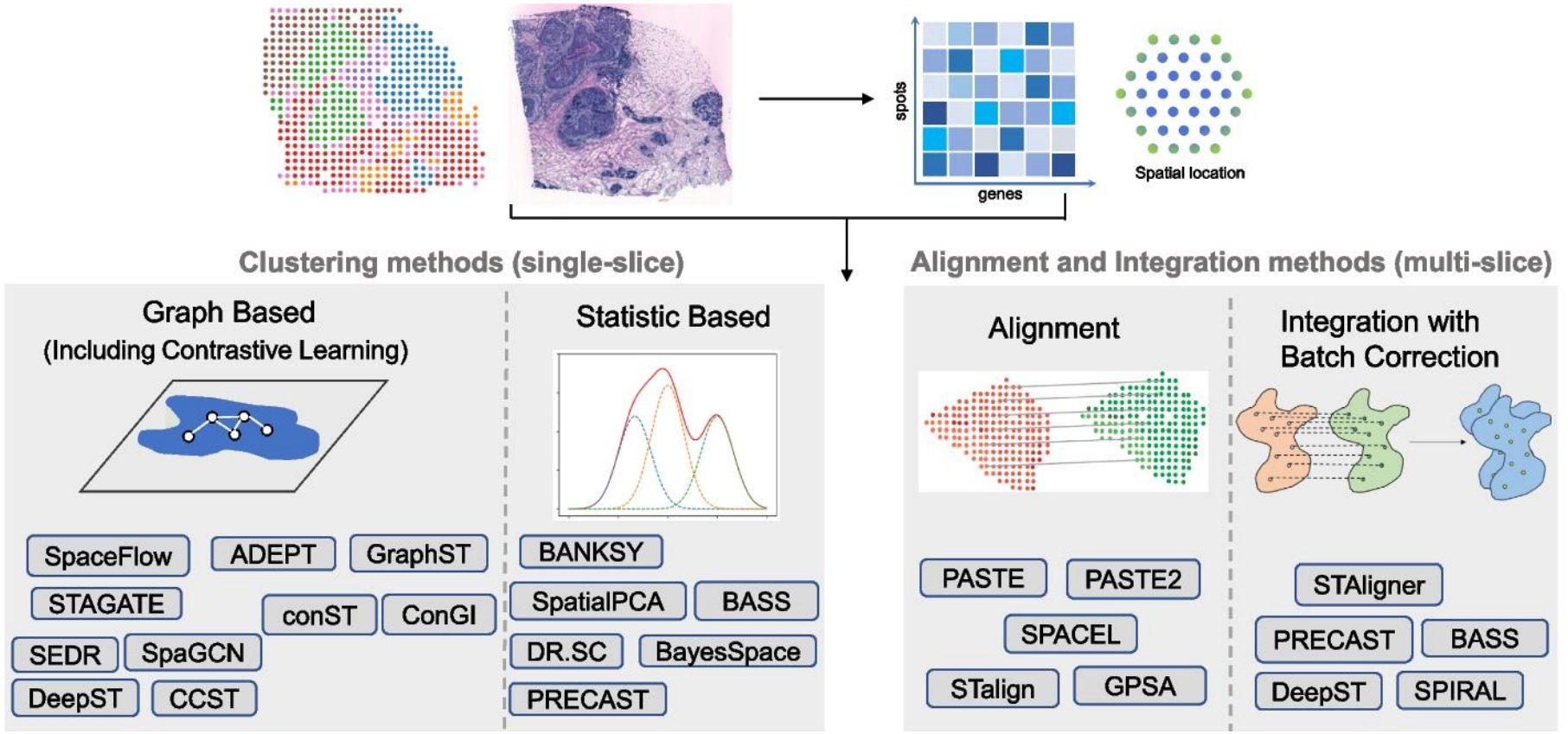

Schott*, León-Periñán*, Splendiani*,…, Macino, Karaiskos, Rajewsky (2024)
Preibisch*, Innerberger*,…, Karaiskos, Rajewsky (2024)

# 3D information is important




Schott*, León-Periñán*, Splendiani*,…, Macino, Karaiskos, Rajewsky (2024)

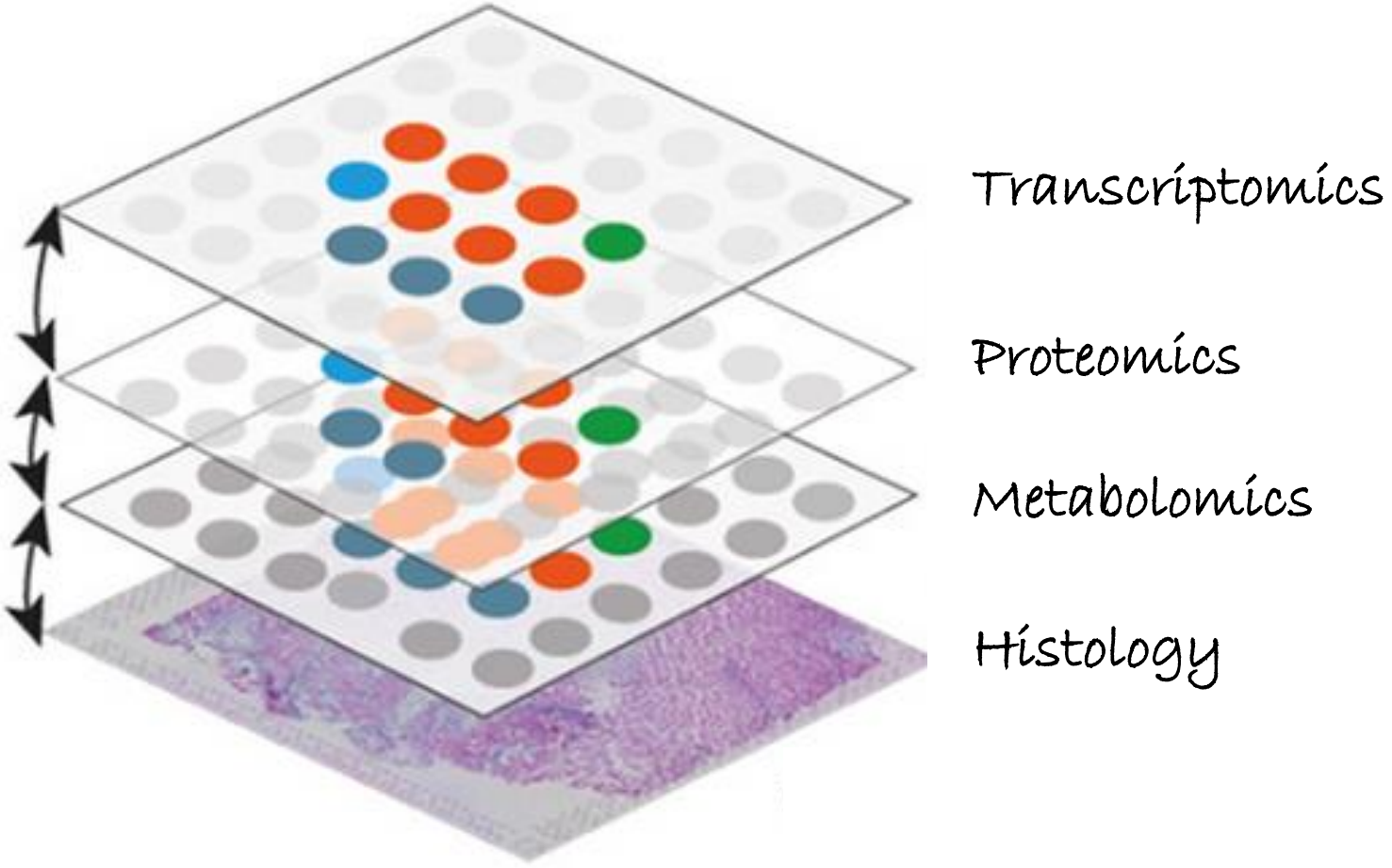
# Comparing SRT alignment and integration methods



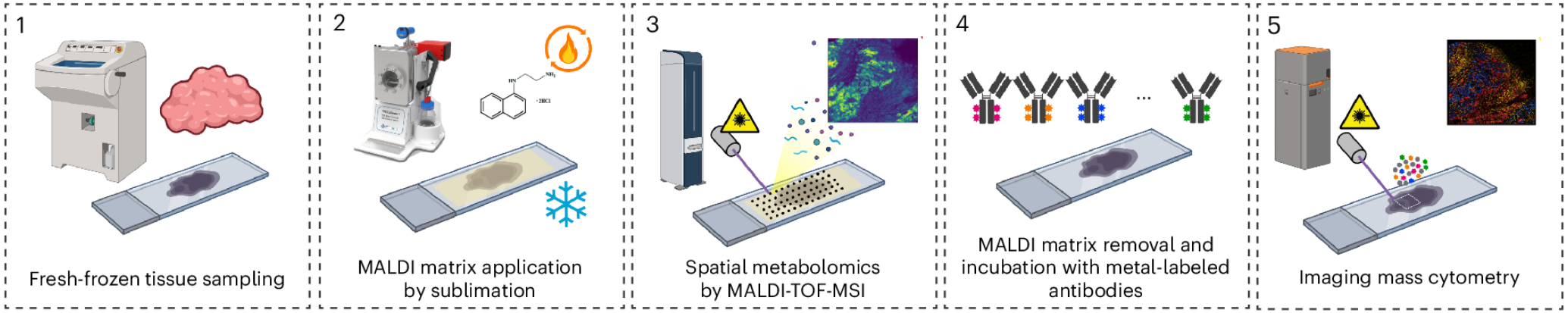

Hu,… Zhou (2023)

# Spatial multiomics

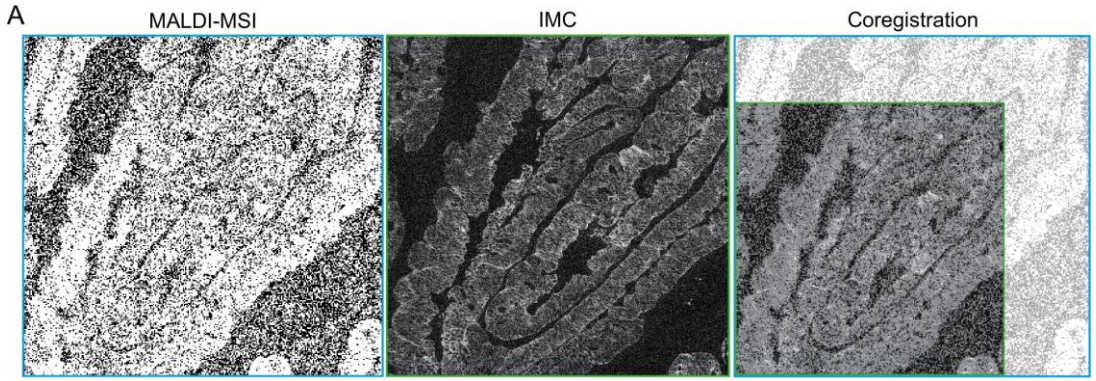
# Spatial multimodal

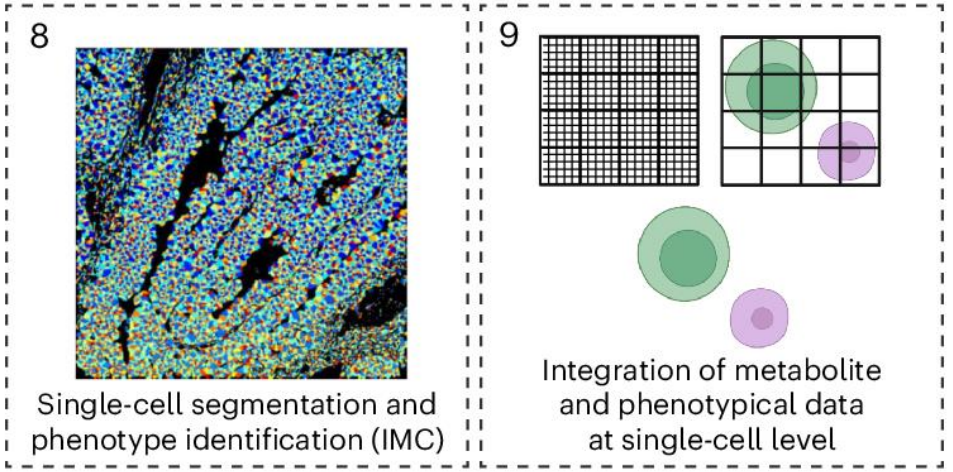


# Mass cytometry and mass spectrometry on the same section




MSI ro IMC image alignment

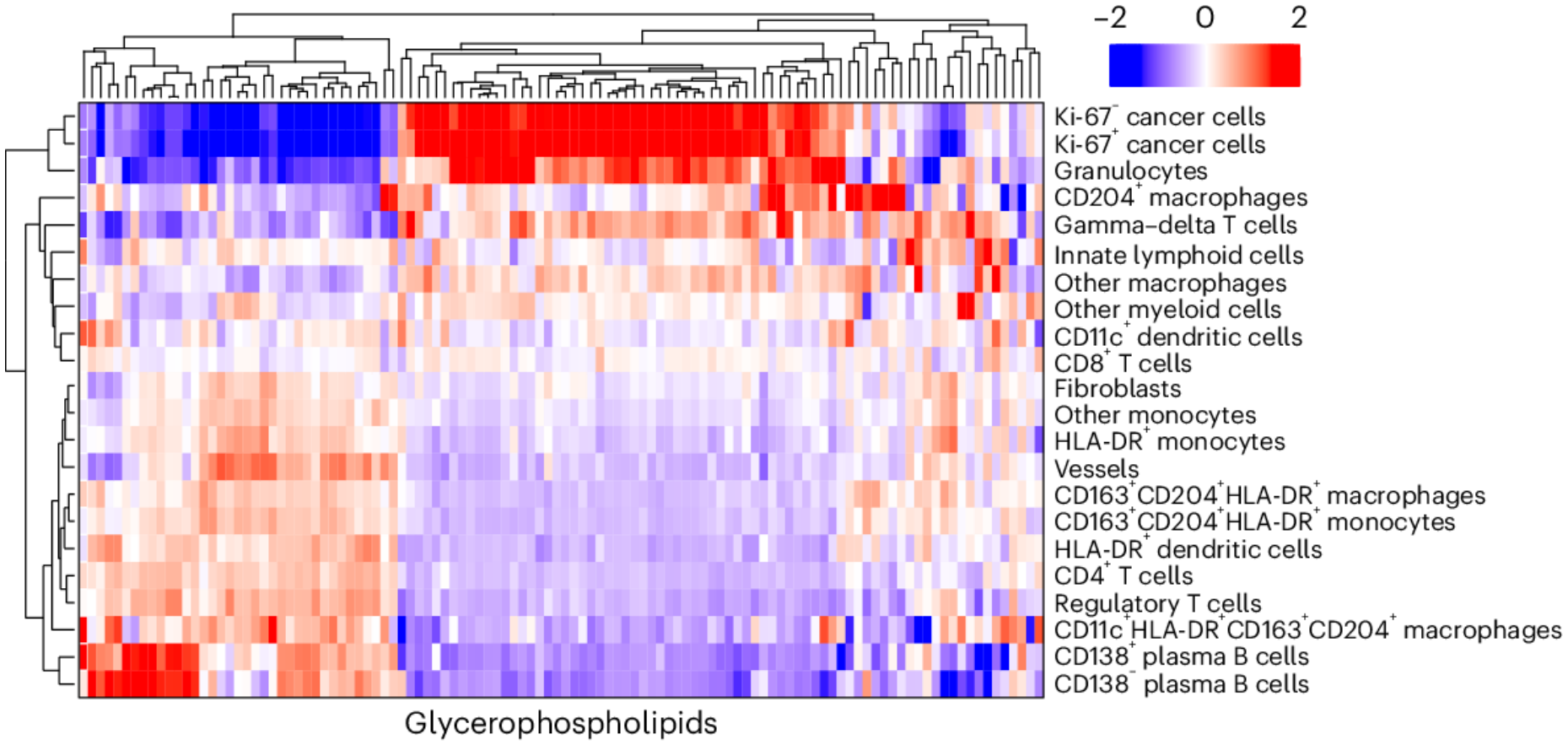


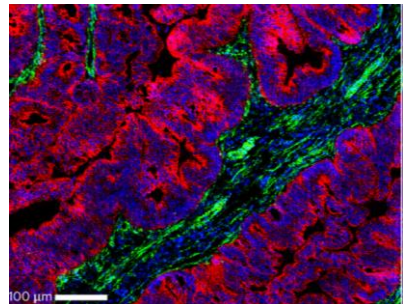



Nunes*, Ijsselsteijn*,…, de Miranda (2024)

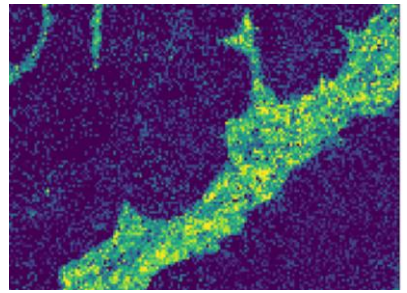
# Varying levels of glycerophospholipids across cell types



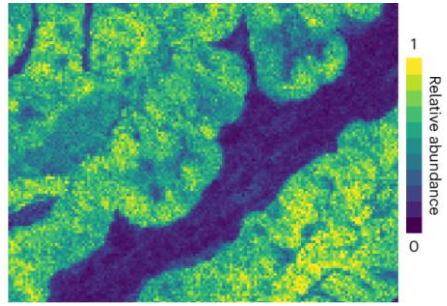

Keratin Vimentin DNA




Phosphatidylcholine PC(37:5)

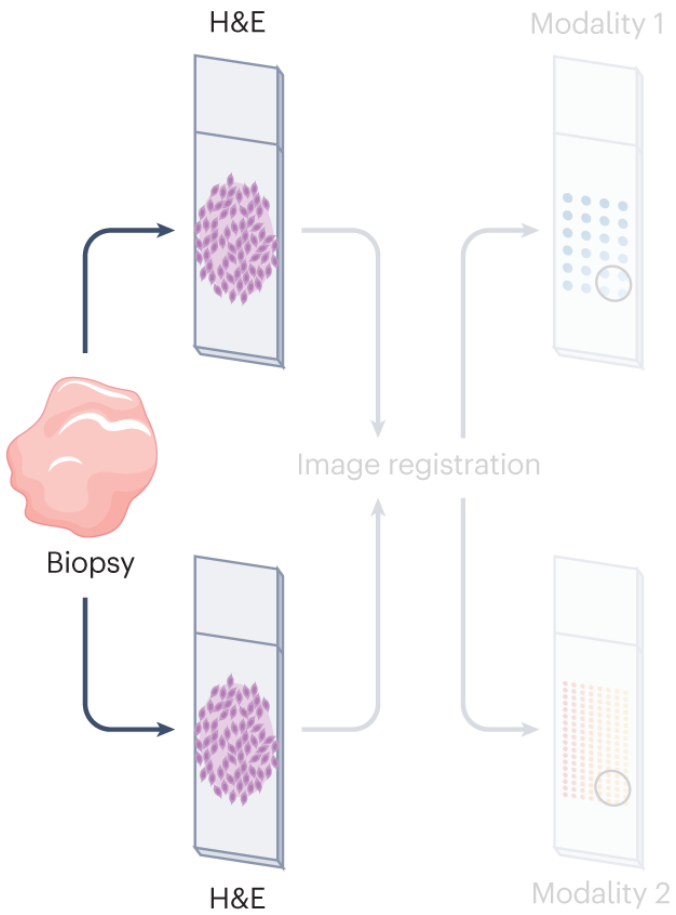


Phosphatidylinositol PI(34:1)



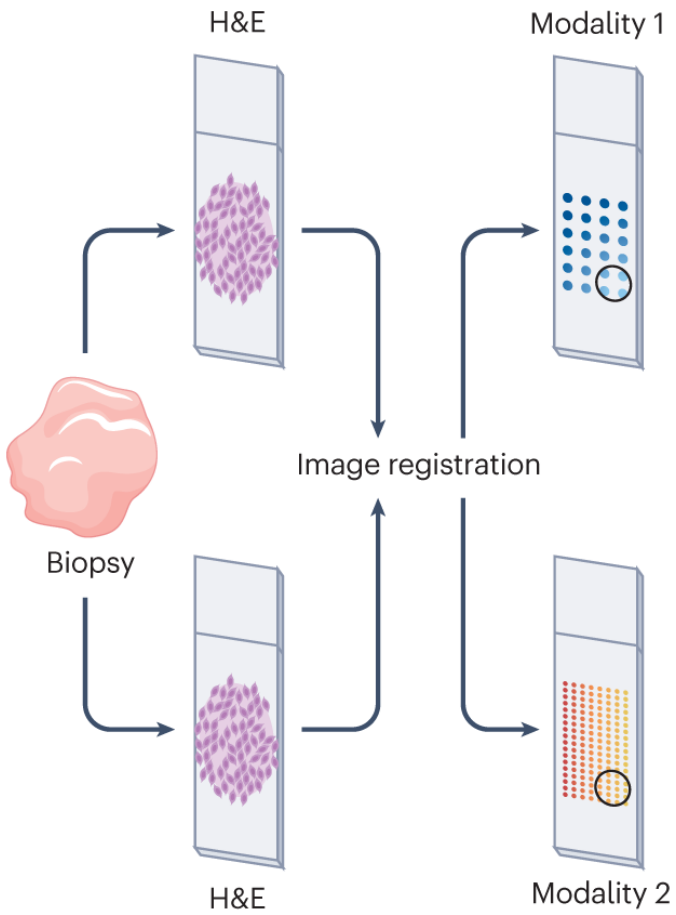

Nunes*, Ijsselsteijn*,…, de Miranda (2024)

# Spatial multiomics on consecutive sections



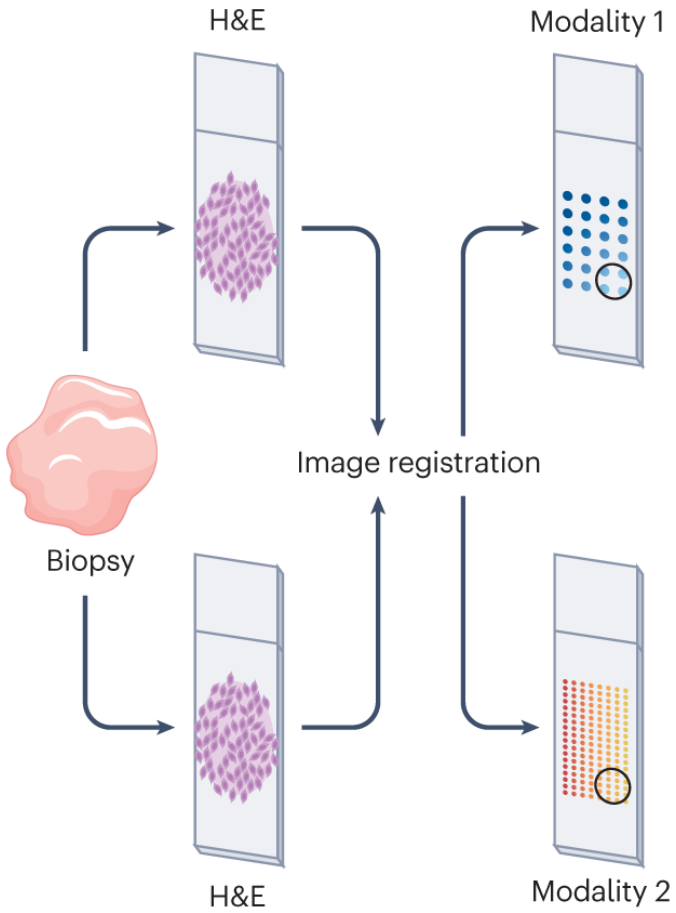

Vandereyken,,…, Voet (2023)

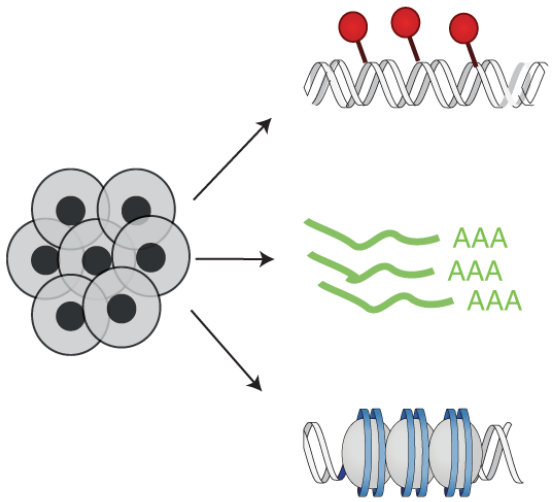
# Spatial multiomics on consecutive sections

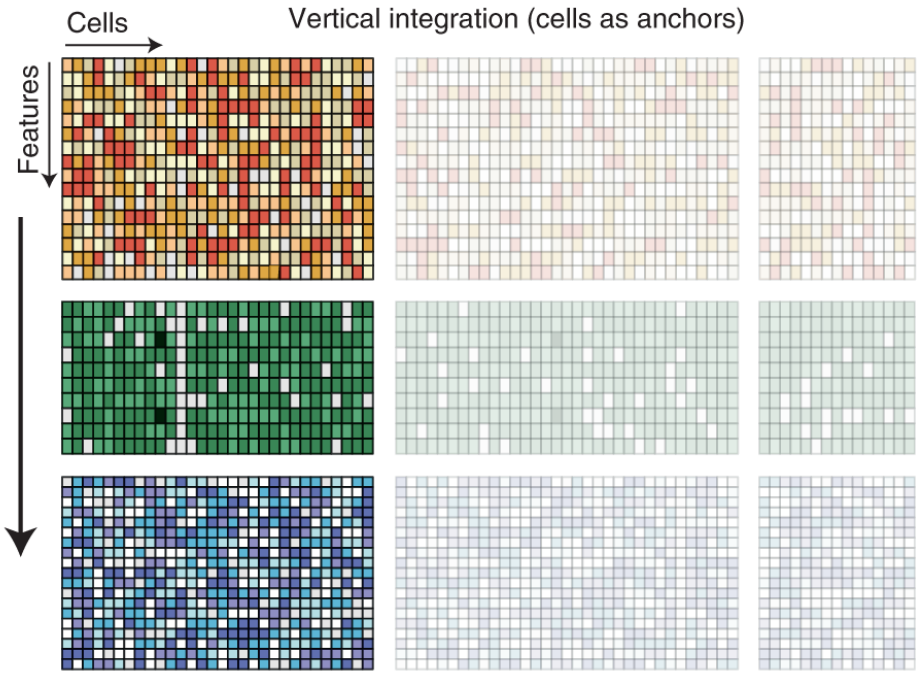



Vandereyken,,…, Voet (2023)

# Spatial multiomics on consecutive sections



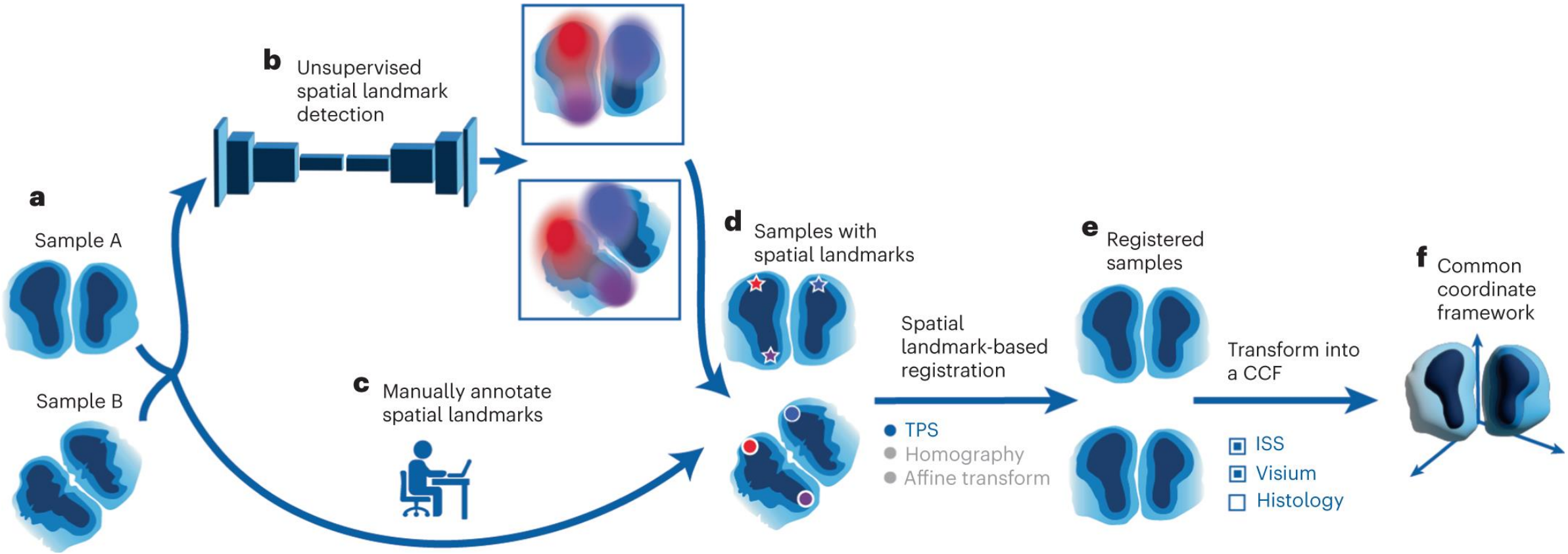




Once aligned can be solved as a vertical integration problem (e.g. MOFA+, CCA, JIVE, totalVI,…)

Vandereyken,,…, Voet (2023)
Argelaguet, Cuomo, Stegle, Marioni (2021)

# Spatial multimodal alignment using Effortless Landmark Detection (ELD)




Ekvall,…, Lundeberg (2024)

# Summary

- Integrating SRT and sc/snRNA-seq data
- SRT data alignment (pseudo 3D, virtual block,…)
- Spatial multiomics

# Thank You!

a.mahfouz@lumc.nl
mahfouzlab.org