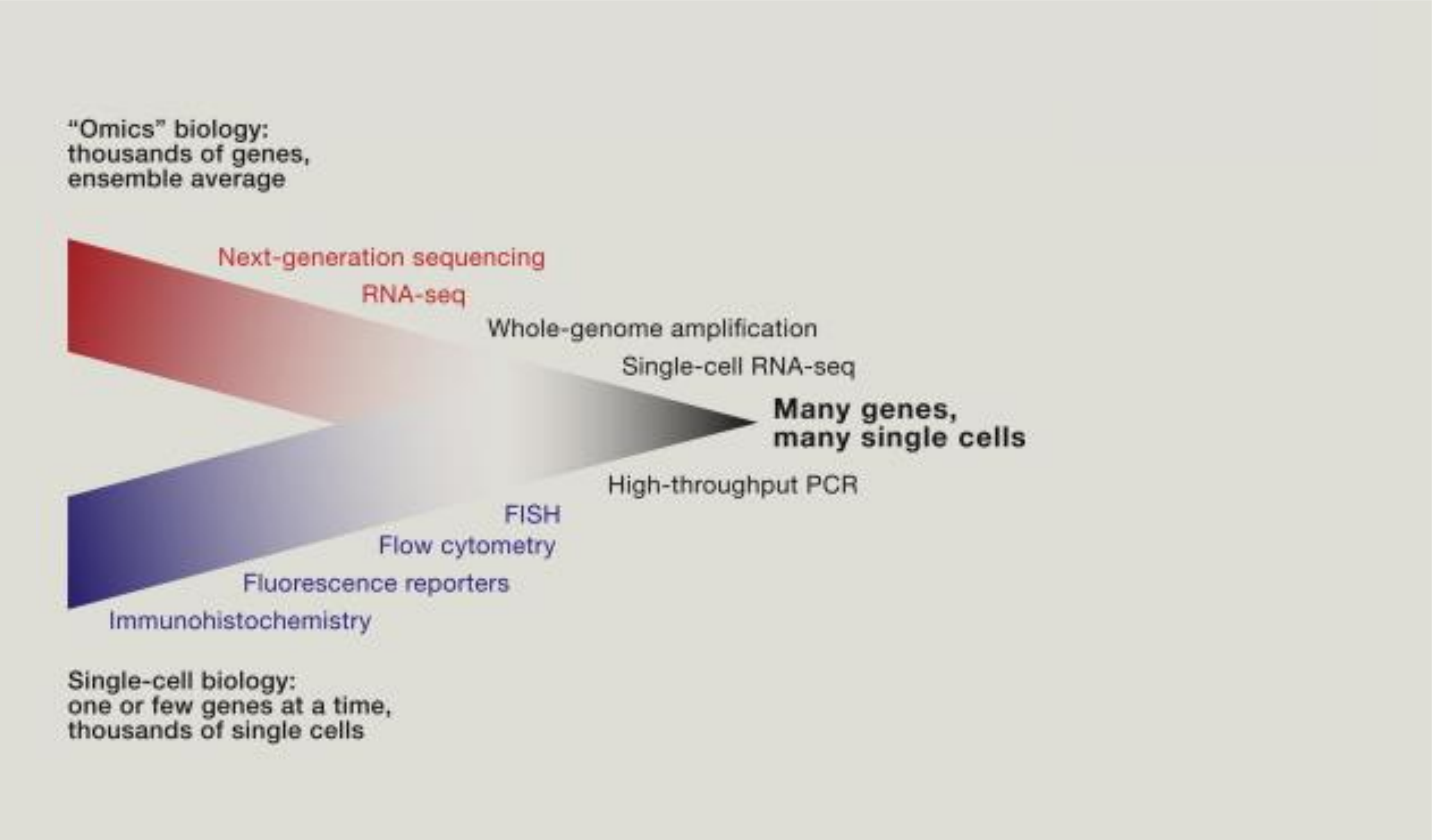
SpatialData: an open and universal data framework for spatial omics

Huan Xu

Single-cell biology: genomics & microscopy

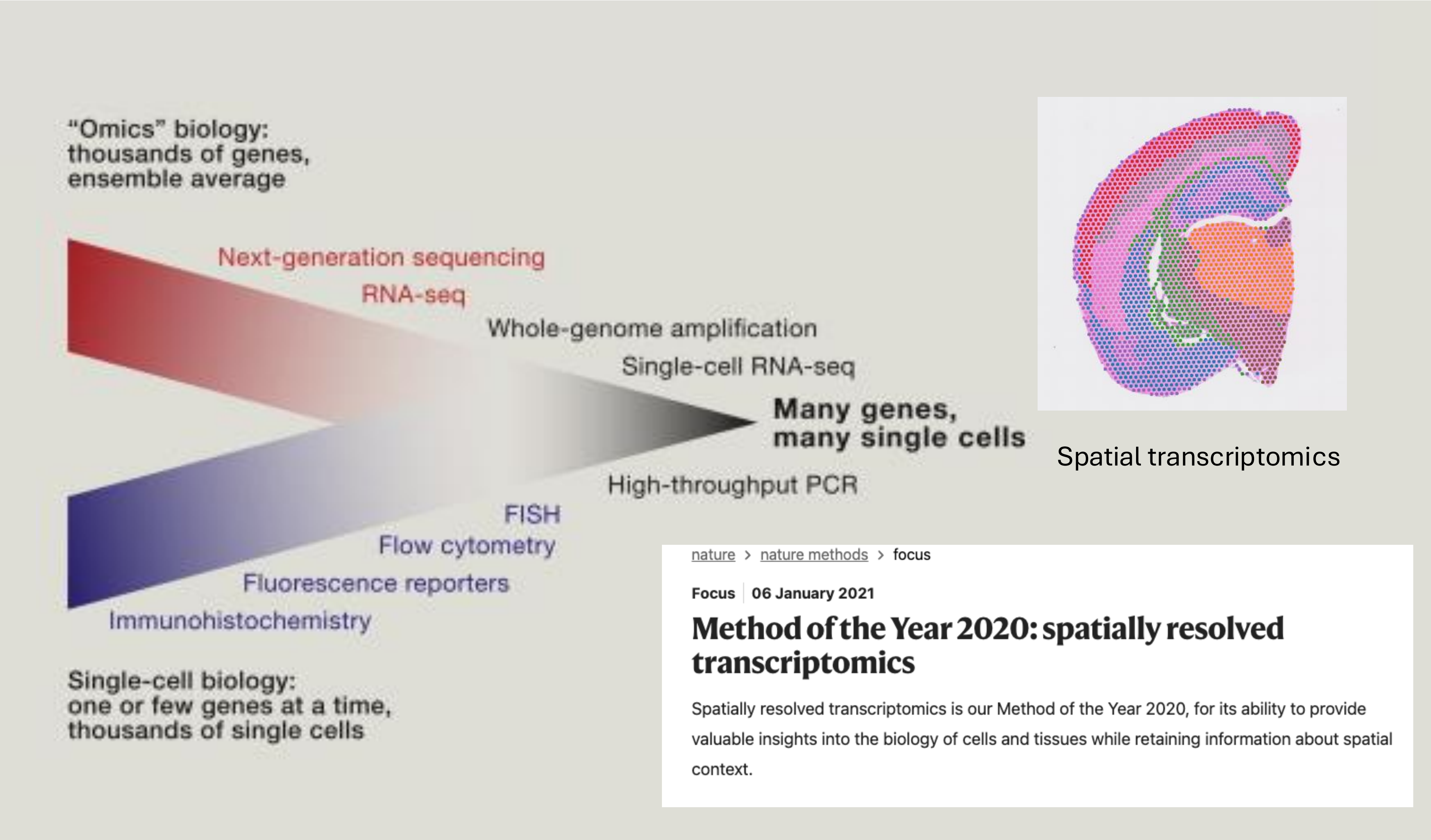


“Omics” biology:
thousands of genes,
ensemble average

Single-cell biology:
one or few genes at a time,
thousands of single cells

Junker & Oudenaarden, Every cell is special, Cell, 2014

Single-cell biology: genomics & microscopy



Spatial transcriptomics

nature › nature methods › focus

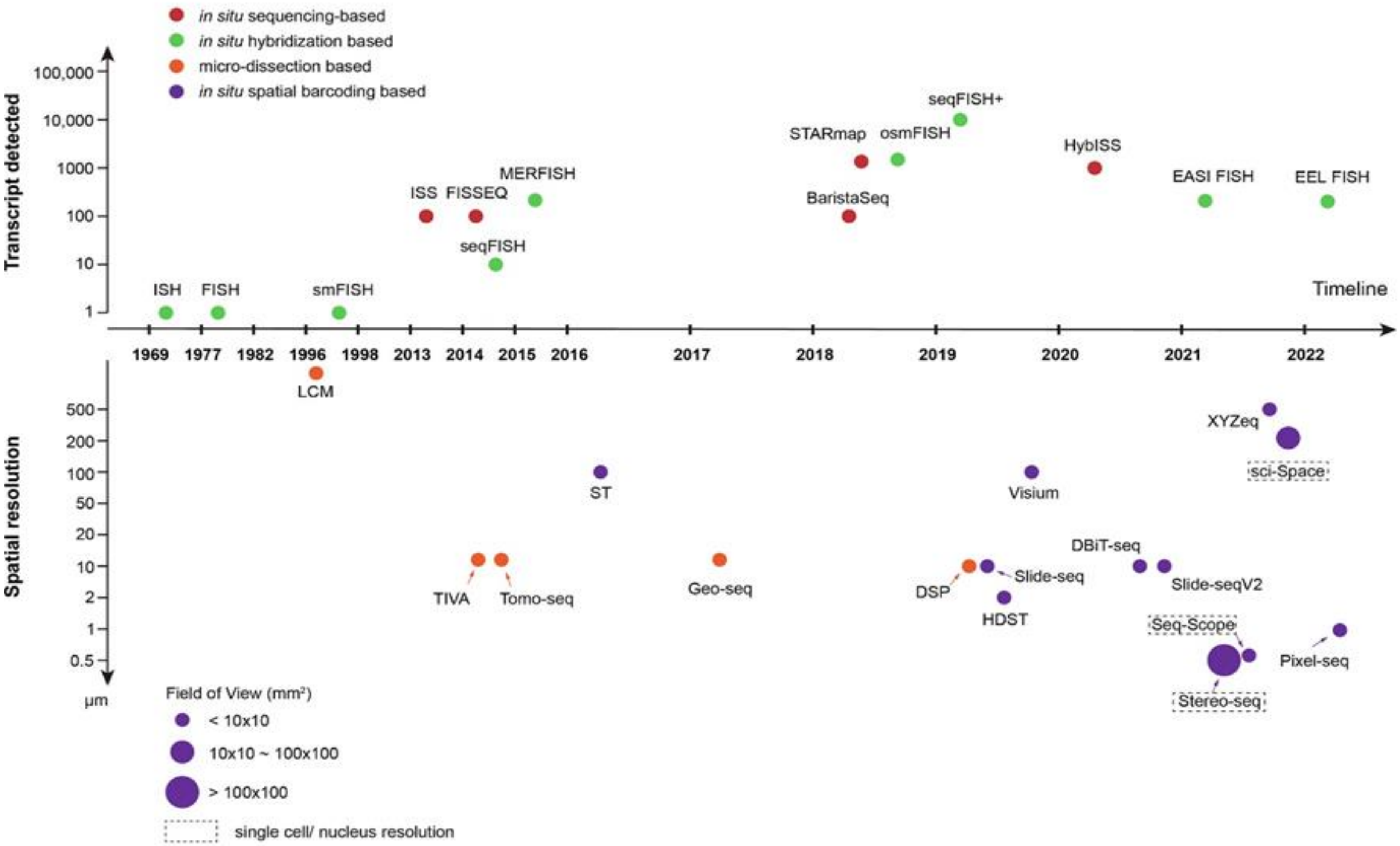
Focus | 06 January 2021

Method of the Year 2020: spatially resolved transcriptomics

Spatially resolved transcriptomics is our Method of the Year 2020, for its ability to provide valuable insights into the biology of cells and tissues while retaining information about spatial context.

Junker & Oudenaarden, Every cell is special, Cell, 2014

Spatial transcriptomics timeline



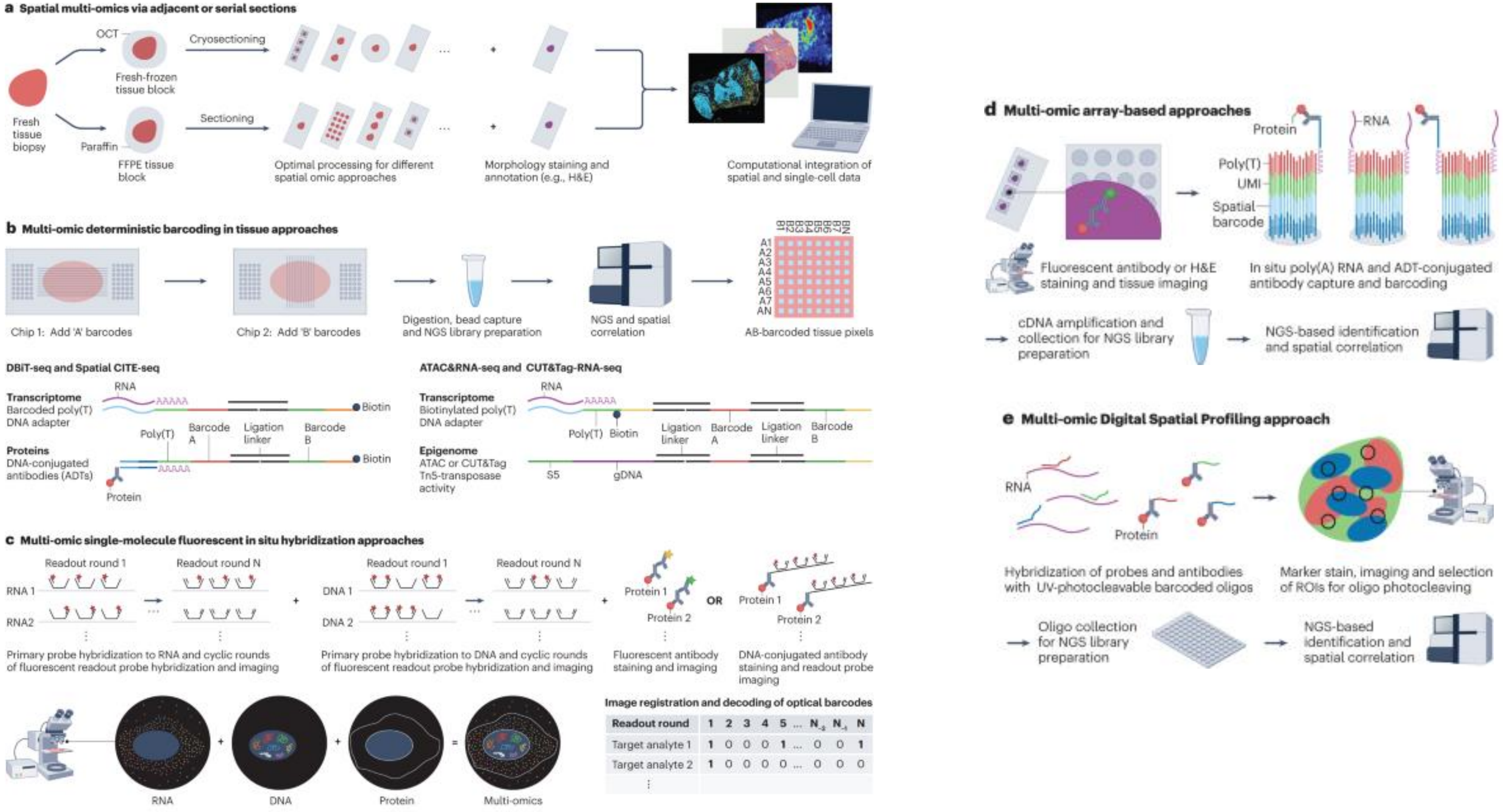
High-multiplexing, single cell resolution

2024

Xenium, CosMx

Cheng, M. *et al. Journal of Genetics and Genomics* (2023).

Spatial multi-omics approaches



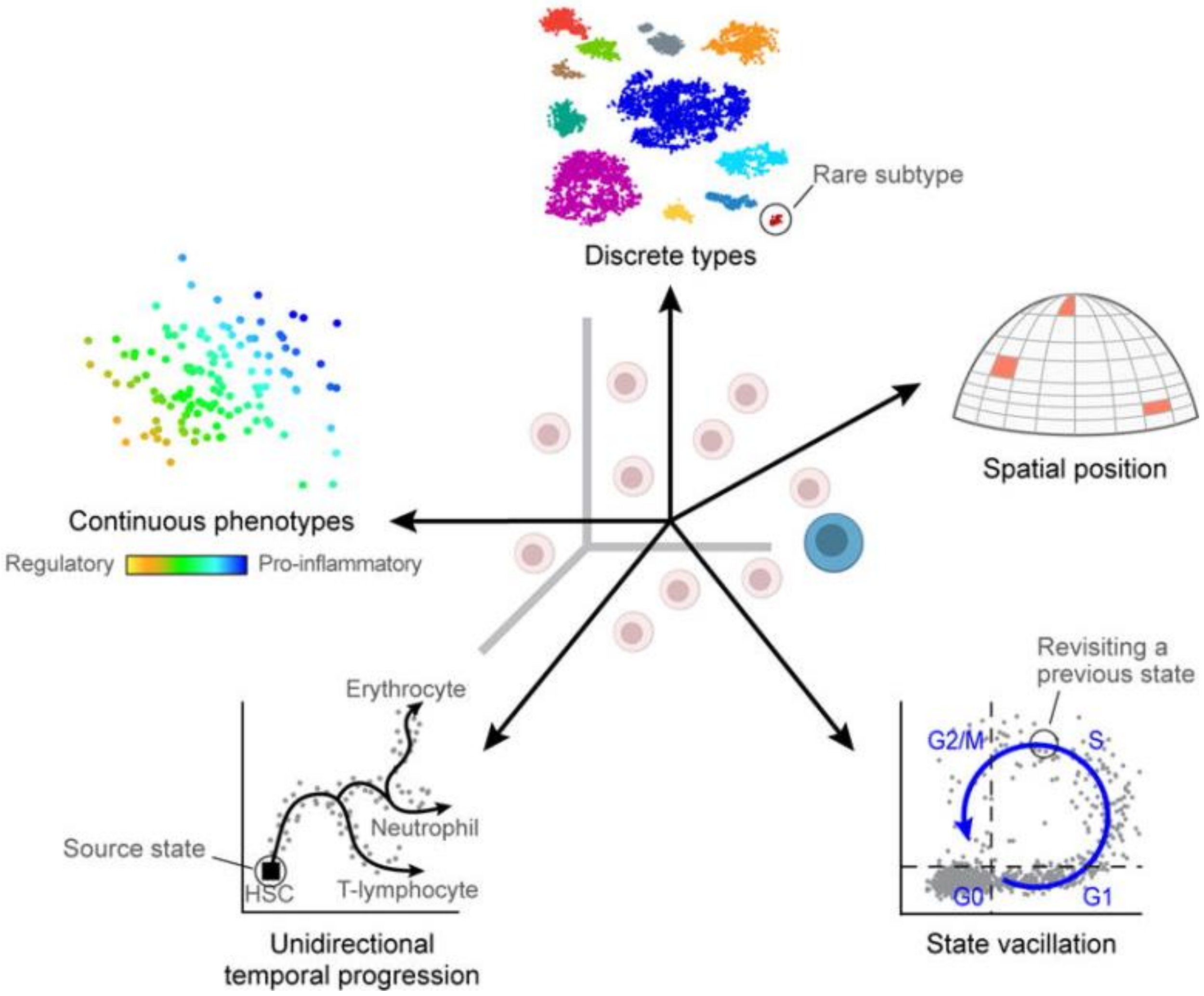
a Spatial multi-omics via adjacent or serial sections
OCT
Cryosectioning
Fresh-frozen tissue block
Fresh tissue biopsy
Paraffin
Sectioning
FFPE tissue block
Optimal processing for different spatial omic approaches
Morphology staining and annotation (e.g., H&E)
Computational integration of spatial and single-cell data
b Multi-omic deterministic barcoding in tissue approaches
Chip 1: Add 'A' barcodes
Chip 2: Add 'B' barcodes
Digestion, bead capture and NGS library preparation
NGS and spatial correlation
B1 B2 B3 B4 B5 B6 B7 BN
A1 A2 A3 A4 A5 A6 A7 AN
AB-barcoded tissue pixels
DBiT-seq and Spatial CITE-seq
RNA
Transcriptome
Barcoded poly(T) DNA adapter
AAAAA
Biotin
Poly(T)
Barcode A
Ligation linker
Barcode B
Proteins
DNA-conjugated antibodies (ADTs)
AAAAA
Biotin
Protein
ATAC&RNA-seq and CUT&Tag-RNA-seq
RNA
Transcriptome
Biotinylated poly(T) DNA adapter
AAAAA
Poly(T)
Biotin
Ligation linker
Barcode A
Ligation linker
Barcode B
Epigenome
ATAC or CUT&Tag Tn5-transposase activity
S5
gDNA
c Multi-omic single-molecule fluorescent in situ hybridization approaches
Readout round 1
Readout round N
RNA 1
RNA2
Primary probe hybridization to RNA and cyclic rounds of fluorescent readout probe hybridization and imaging
Readout round 1
Readout round N
DNA 1
DNA 2
Primary probe hybridization to DNA and cyclic rounds of fluorescent readout probe hybridization and imaging
Protein 1
Protein 2
Fluorescent antibody staining and imaging
OR
Protein 1
Protein 2
DNA-conjugated antibody staining and readout probe imaging
RNA
DNA
Protein
Multi-omics
d Multi-omic array-based approaches
Protein
RNA
Poly(T)
UMI
Spatial barcode
Fluorescent antibody or H&E staining and tissue imaging
In situ poly(A) RNA and ADT-conjugated antibody capture and barcoding
cDNA amplification and collection for NGS library preparation
NGS-based identification and spatial correlation
e Multi-omic Digital Spatial Profiling approach
RNA
Protein
Hybridization of probes and antibodies with UV-photocleavable barcoded oligos
Marker stain, imaging and selection of ROIs for oligo photocleaving
Oligo collection for NGS library preparation
NGS-based identification and spatial correlation

Image registration and decoding of optical barcodes

Readout round	1	2	3	4	5	...	N_{-2}	N_{-1}	N
Target analyte 1	1	0	0	0	1	...	0	0	1
Target analyte 2	1	0	0	0	0	...	0	0	0
⋮									

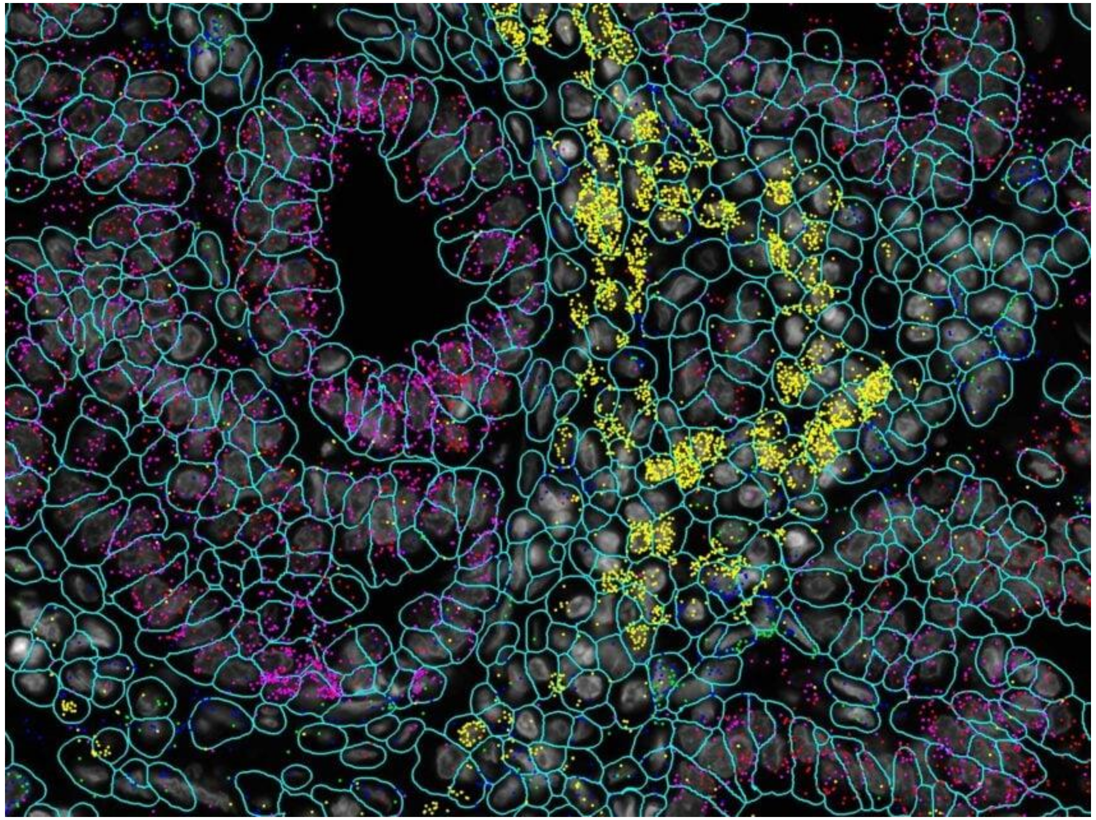
Katy et al, Nature Reviews Genetics (2023)

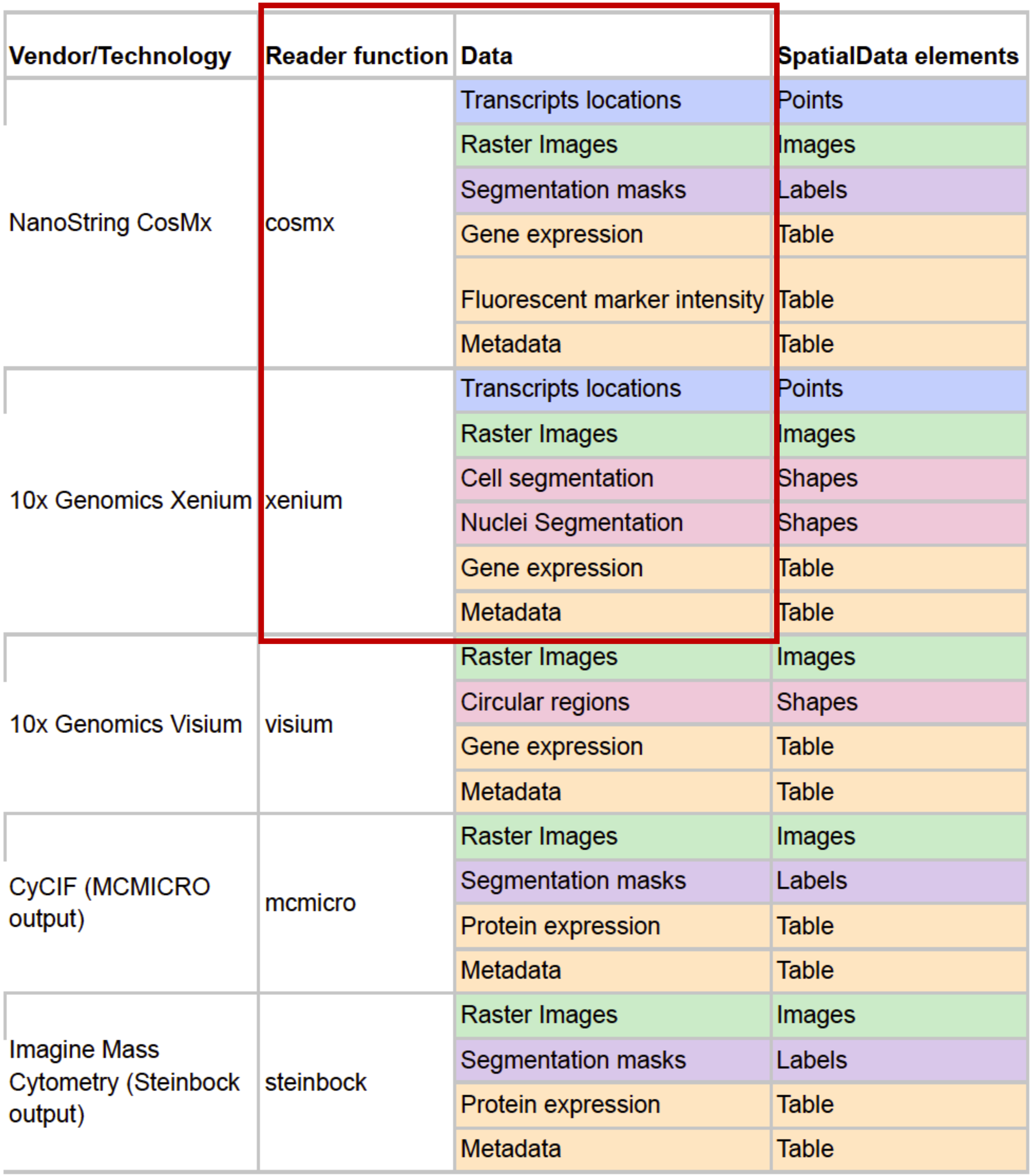
Spatial components of molecular tissue biology



Wagner et al. Nat Biotechnol. 2016

Spatial omics data elements





Vendor/Technology	Reader function	Data	SpatialData elements
NanoString CosMx	cosmx	Transcripts locations	Points
		Raster Images	Images
		Segmentation masks	Labels
		Gene expression	Table
		Fluorescent marker intensity	Table
		Metadata	Table
10x Genomics Xenium	xenium	Transcripts locations	Points
		Raster Images	Images
		Cell segmentation	Shapes
		Nuclei Segmentation	Shapes
		Gene expression	Table
		Metadata	Table
10x Genomics Visium	visium	Raster Images	Images
		Circular regions	Shapes
		Gene expression	Table
		Metadata	Table
CyCIF (MCMICRO output)	mcmicro	Raster Images	Images
		Segmentation masks	Labels
		Protein expression	Table
		Metadata	Table
Imagine Mass Cytometry (Steinbock output)	steinbock	Raster Images	Images
		Segmentation masks	Labels
		Protein expression	Table
		Metadata	Table

Analytical challenges of Multi-omics spatial data

1. Integration of large image data

   --> Lazy loading the required details

2. Spatial alignment of multimodal spatial omics data

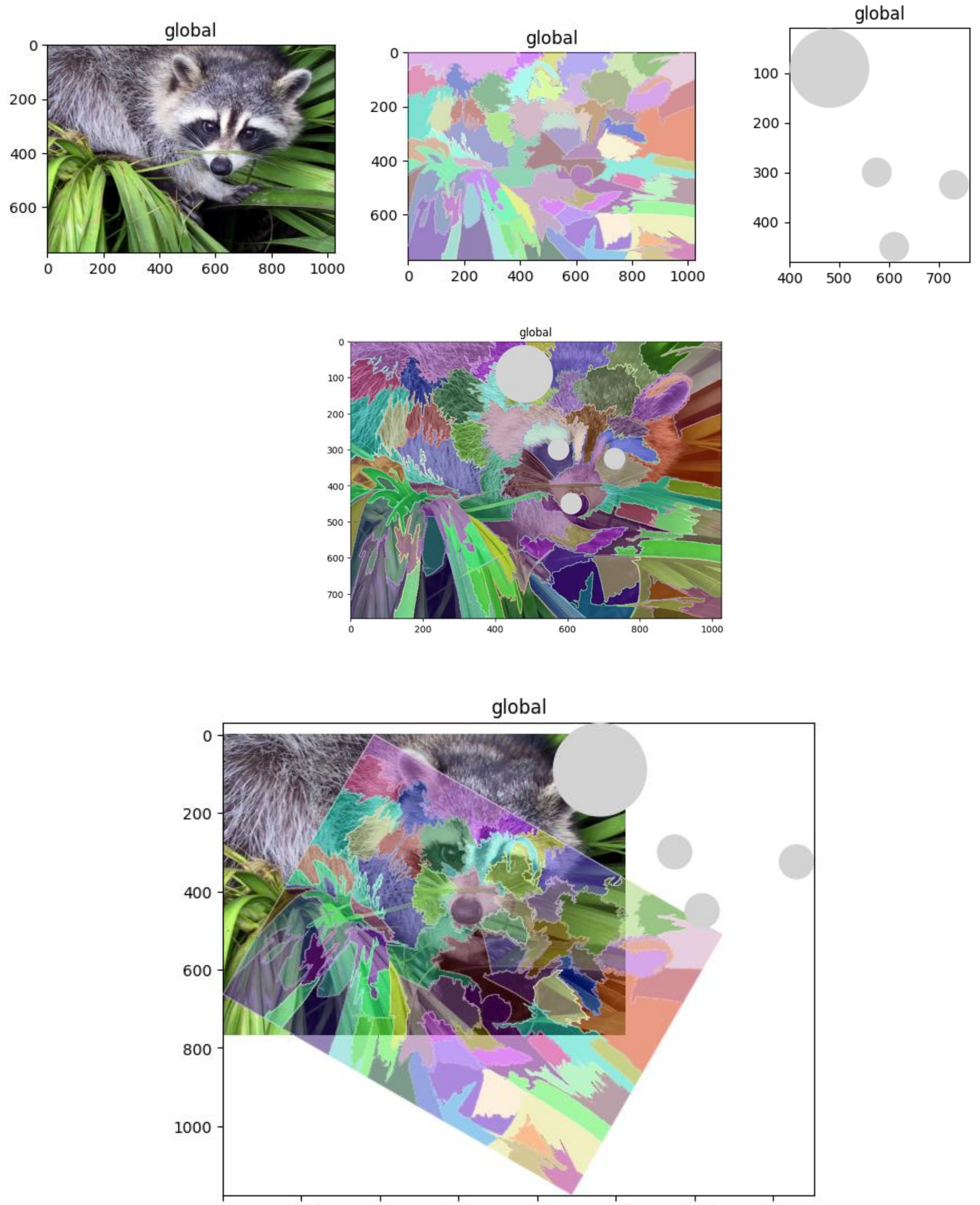
   --> Data transformation, Common coordinate system

3. Cross-modality aggregation

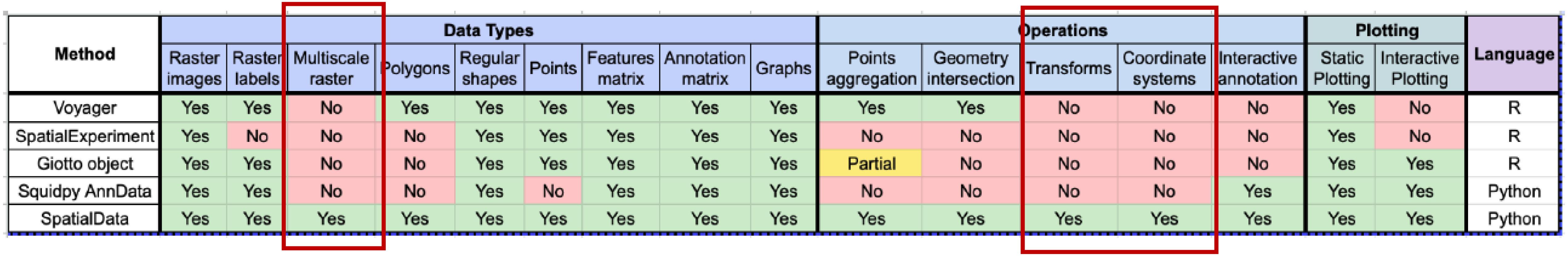
   --> Uniform interface for aggregating all data types

4. Interactive annotation

   --> Interactive digital viewer/analyzer

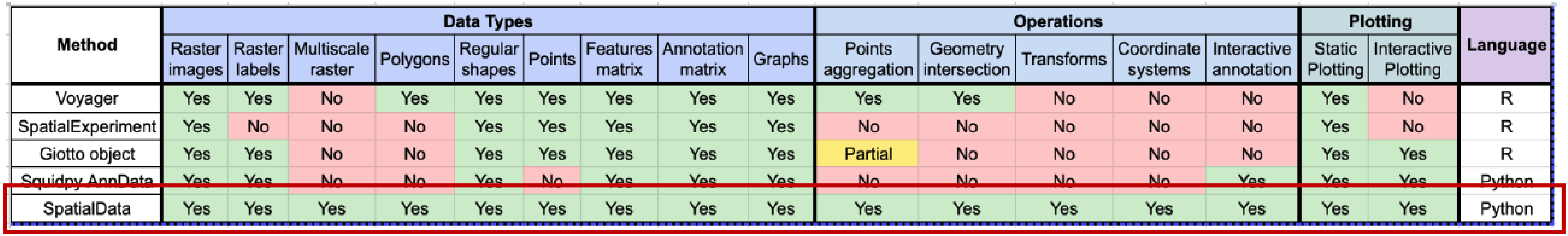


Existing spatial multi-omics tools

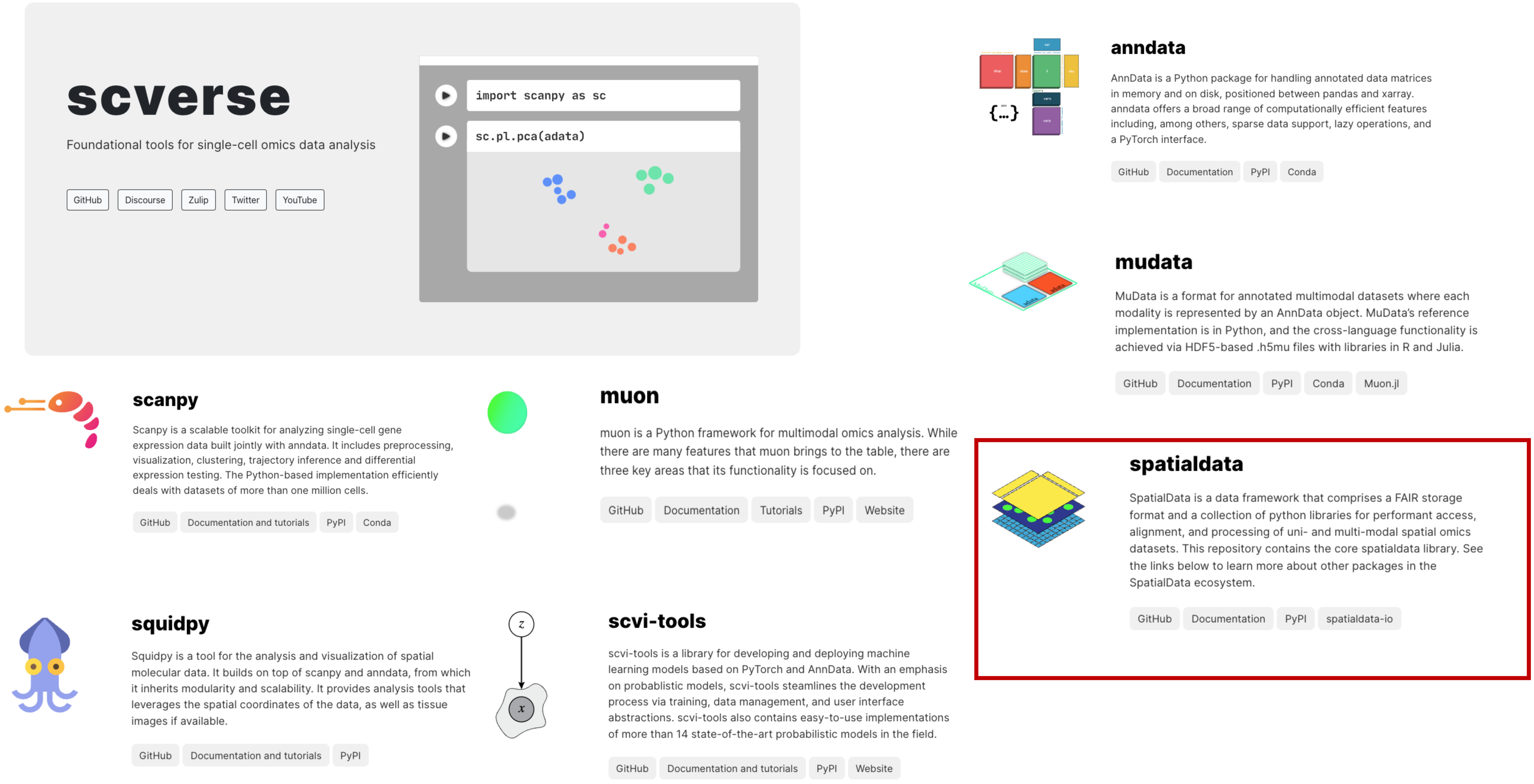


Method	Data Types									Operations					Plotting		Language
	Raster images	Raster labels	Multiscale raster	Polygons	Regular shapes	Points	Features matrix	Annotation matrix	Graphs	Points aggregation	Geometry intersection	Transforms	Coordinate systems	Interactive annotation	Static Plotting	Interactive Plotting	
Voyager	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	R
SpatialExperiment	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	R
Giotto object	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Partial	No	No	No	No	Yes	Yes	R
Squidpy AnnData	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Python
SpatialData	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Python

Existing spatial omics tools



Method	Data Types									Operations					Plotting		Language
	Raster images	Raster labels	Multiscale raster	Polygons	Regular shapes	Points	Features matrix	Annotation matrix	Graphs	Points aggregation	Geometry intersection	Transforms	Coordinate systems	Interactive annotation	Static Plotting	Interactive Plotting	
Voyager	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	R
SpatialExperiment	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	R
Giotto object	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Partial	No	No	No	No	Yes	Yes	R
Squidpy AnnData	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Python
SpatialData	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Python



scverse

Foundational tools for single-cell omics data analysis

GitHub Discourse Zulip Twitter YouTube

```
import scanpy as sc
```

```
sc.pl.pca(adata)
```

anndata

AnnData is a Python package for handling annotated data matrices in memory and on disk, positioned between pandas and xarray. anndata offers a broad range of computationally efficient features including, among others, sparse data support, lazy operations, and a PyTorch interface.

GitHub Documentation PyPI Conda

mudata

MuData is a format for annotated multimodal datasets where each modality is represented by an AnnData object. MuData's reference implementation is in Python, and the cross-language functionality is achieved via HDF5-based .h5mu files with libraries in R and Julia.

GitHub Documentation PyPI Conda Muon.jl

scanpy

Scanpy is a scalable toolkit for analyzing single-cell gene expression data built jointly with anndata. It includes preprocessing, visualization, clustering, trajectory inference and differential expression testing. The Python-based implementation efficiently deals with datasets of more than one million cells.

GitHub Documentation and tutorials PyPI Conda

muon

muon is a Python framework for multimodal omics analysis. While there are many features that muon brings to the table, there are three key areas that its functionality is focused on.

GitHub Documentation Tutorials PyPI Website

spatialdata

SpatialData is a data framework that comprises a FAIR storage format and a collection of python libraries for performant access, alignment, and processing of uni- and multi-modal spatial omics datasets. This repository contains the core spatialdata library. See the links below to learn more about other packages in the SpatialData ecosystem.

GitHub Documentation PyPI spatialdata-io

squidpy

Squidpy is a tool for the analysis and visualization of spatial molecular data. It builds on top of scanpy and anndata, from which it inherits modularity and scalability. It provides analysis tools that leverages the spatial coordinates of the data, as well as tissue images if available.

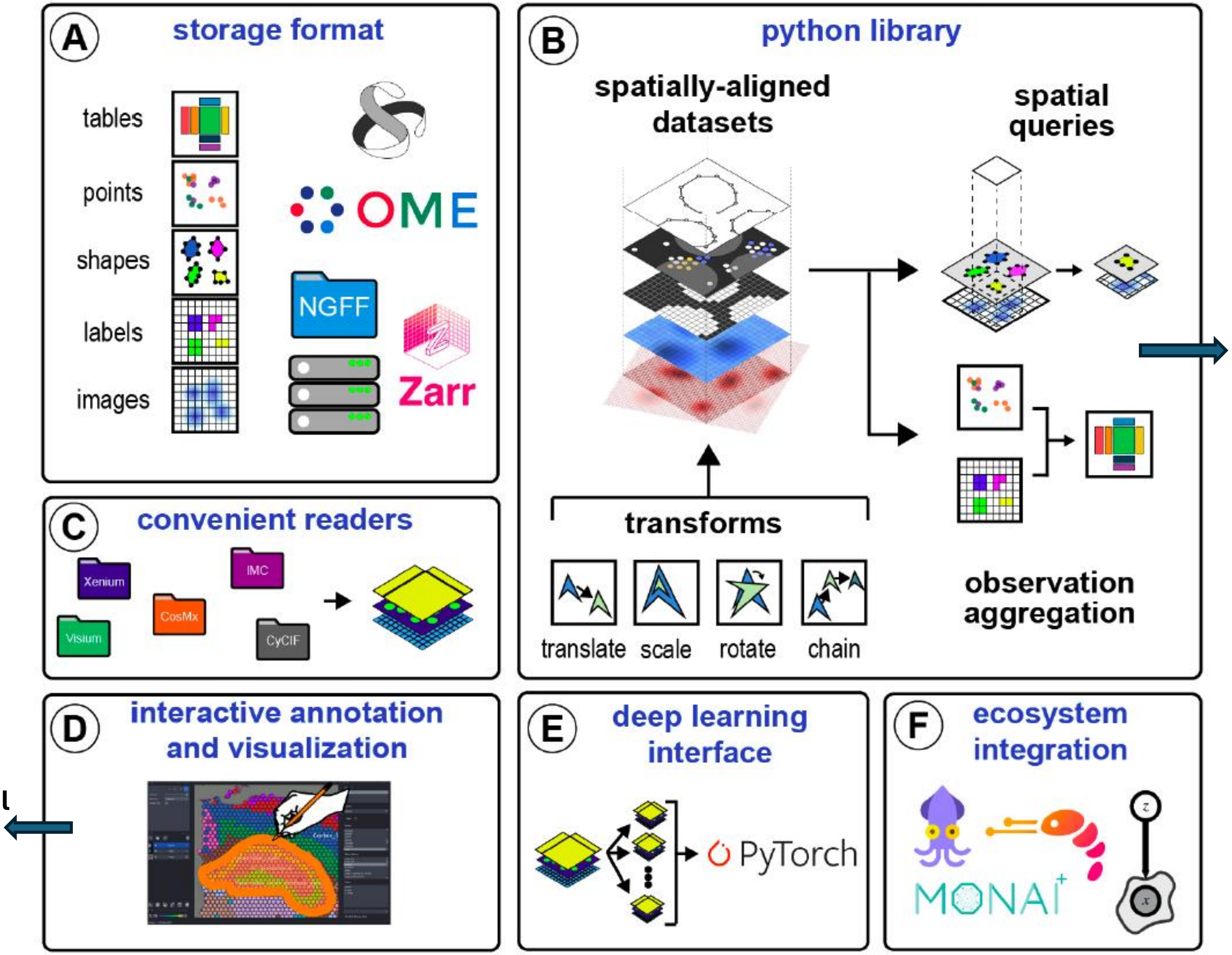
GitHub Documentation and tutorials PyPI

scvi-tools

scvi-tools is a library for developing and deploying machine learning models based on PyTorch and AnnData. With an emphasis on probablistic models, scvi-tools steamlines the development process via training, data management, and user interface abstractions. scvi-tools also contains easy-to-use implementations of more than 14 state-of-the-art probabilistic models in the field.

GitHub Documentation and tutorials PyPI Website

SpatialData Framework

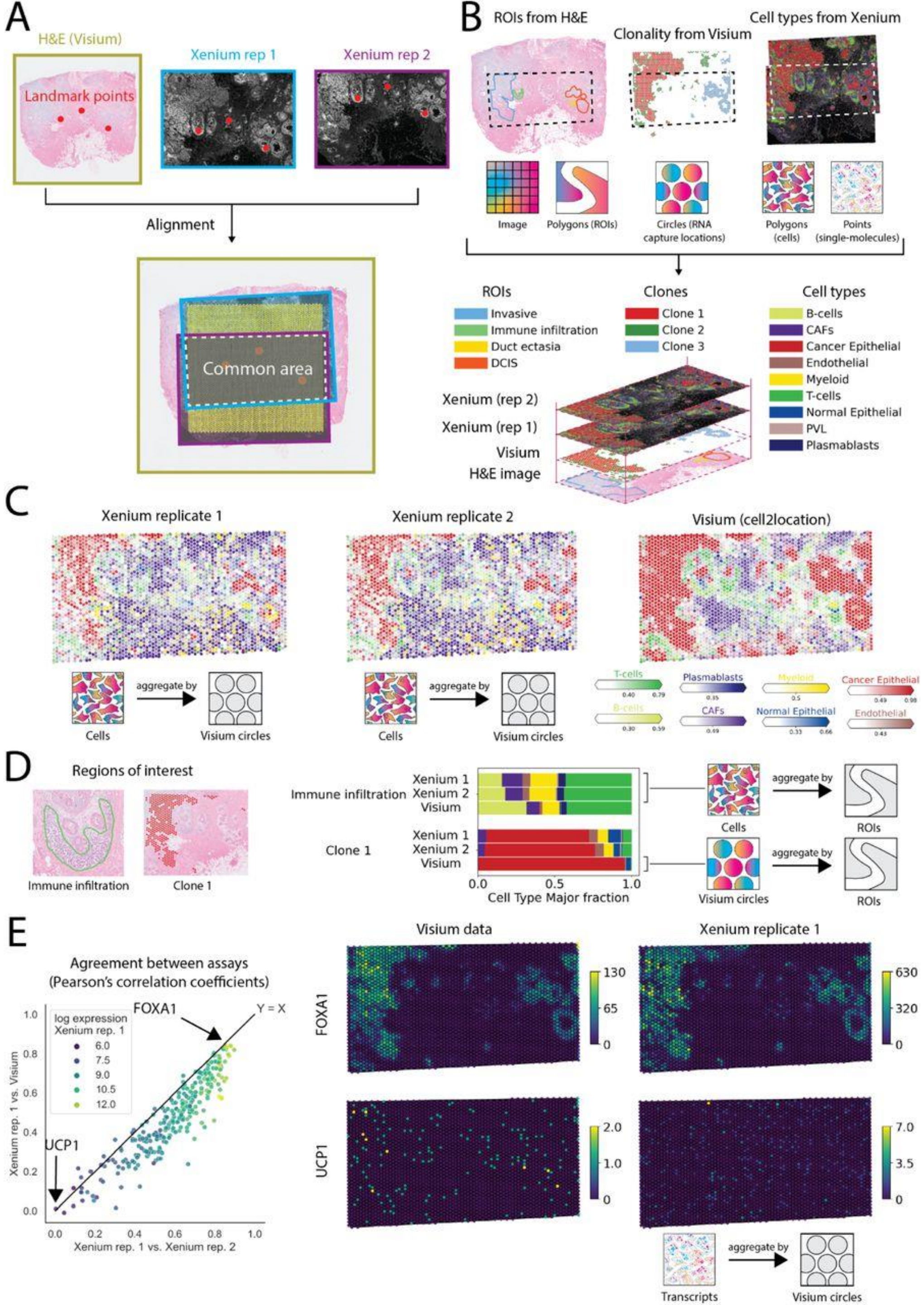


Generate new datasets for exploration

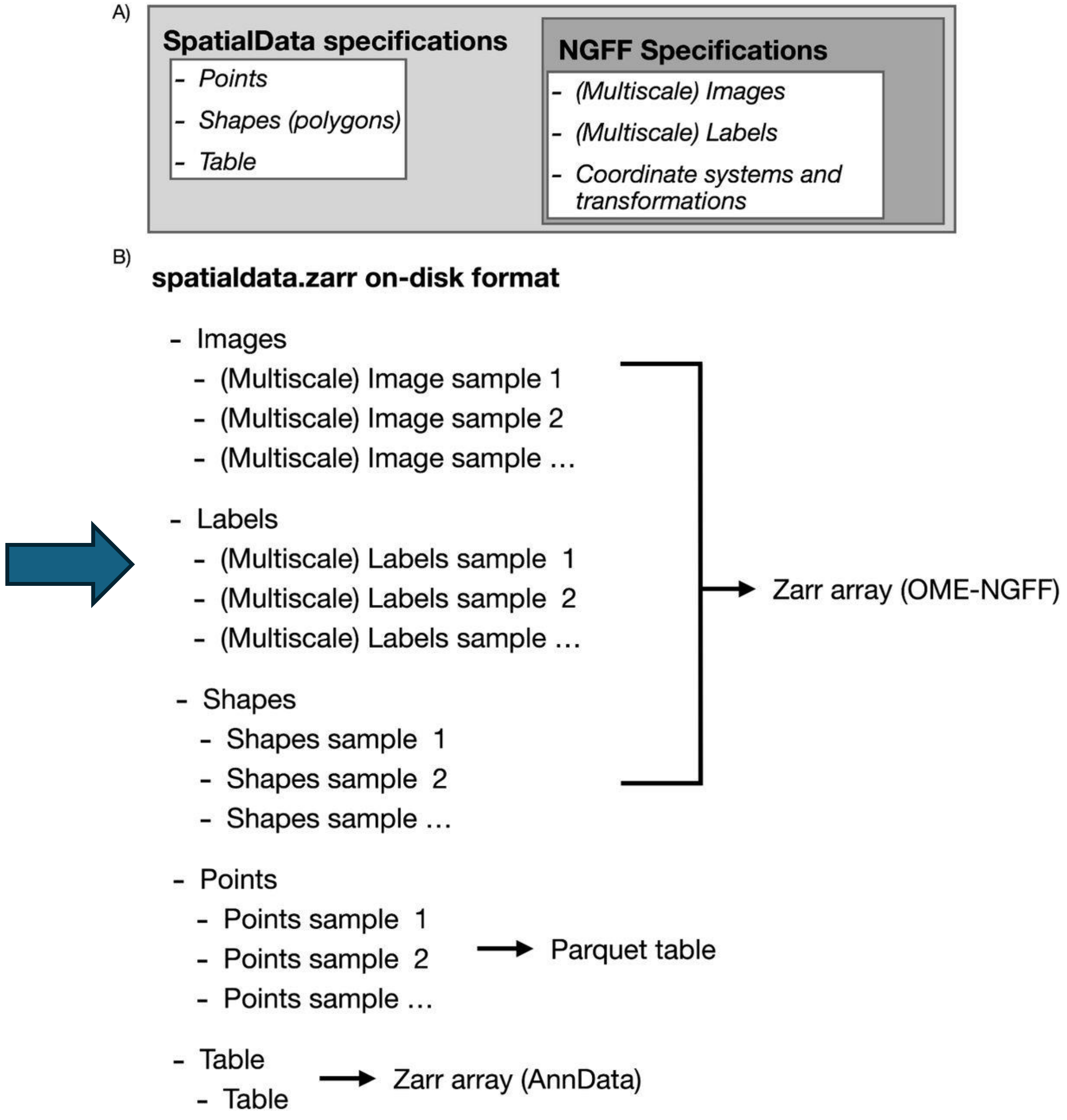
Napari spatial data viewer

Integrative analysis of breast cancer spatial data:

H&E, Xenium and Visium



Vendor/Technology	Reader function	Data	SpatialData elements
NanoString CosMx	cosmx	Transcripts locations	Points
		Raster Images	Images
		Segmentation masks	Labels
		Gene expression	Table
		Fluorescent marker intensity	Table
		Metadata	Table
10x Genomics Xenium	xenium	Transcripts locations	Points
		Raster Images	Images
		Cell segmentation	Shapes
		Nuclei Segmentation	Shapes
		Gene expression	Table
		Metadata	Table
10x Genomics Visium	visium	Raster Images	Images
		Circular regions	Shapes
		Gene expression	Table
		Metadata	Table
CyCIF (MCMICRO output)	mcmicro	Raster Images	Images
		Segmentation masks	Labels
		Protein expression	Table
		Metadata	Table
Imagine Mass Cytometry (Steinbock output)	steinbock	Raster Images	Images
		Segmentation masks	Labels
		Protein expression	Table
		Metadata	Table



A)

SpatialData specifications

- *Points*
- *Shapes (polygons)*
- *Table*

NGFF Specifications

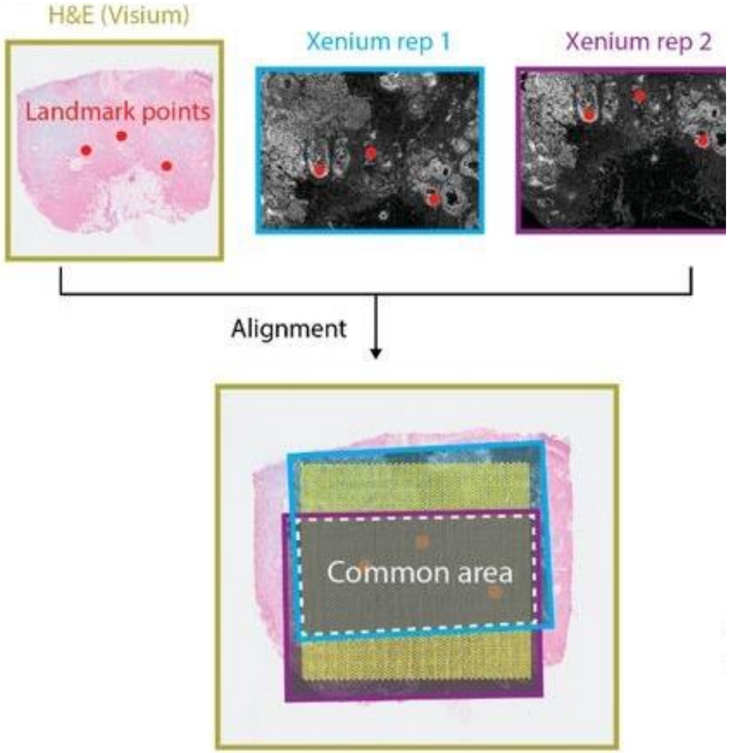
- *(Multiscale) Images*
- *(Multiscale) Labels*
- *Coordinate systems and transformations*

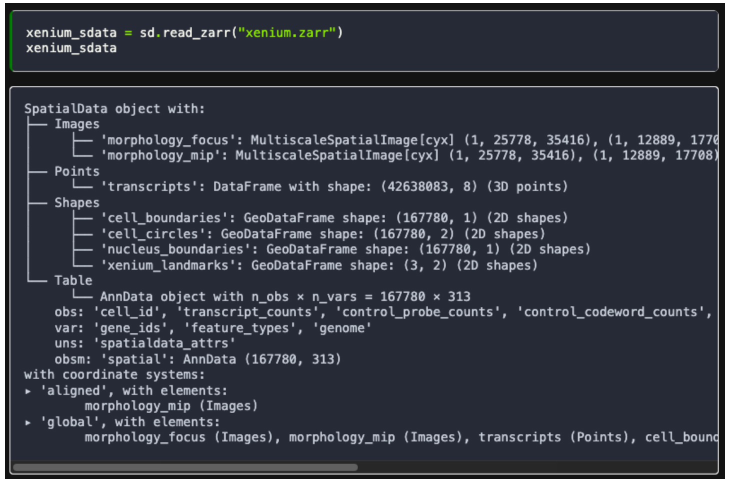
B)

spatialdata.zarr on-disk format

- Images
  - (Multiscale) Image sample 1
  - (Multiscale) Image sample 2
  - (Multiscale) Image sample …
- Labels
  - (Multiscale) Labels sample 1
  - (Multiscale) Labels sample 2
  - (Multiscale) Labels sample …
- Shapes
  - Shapes sample 1
  - Shapes sample 2
  - Shapes sample …
- Points
  - Points sample 1
  - Points sample 2
  - Points sample …
- Table
  - Table

1: -omics layer alignment

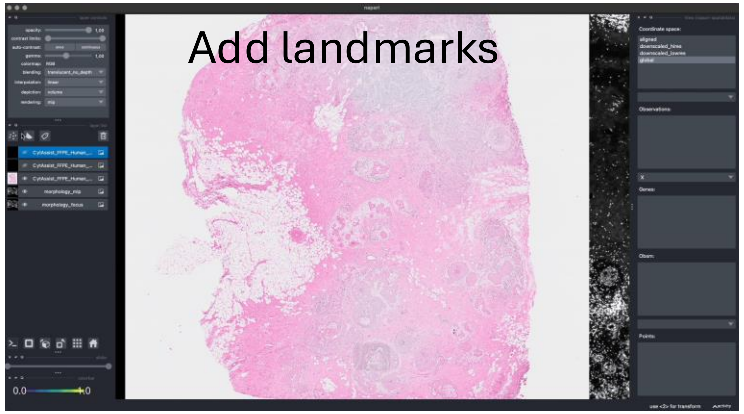




```
xenium_sdata = sd.read_zarr("xenium.zarr")
xenium_sdata
```

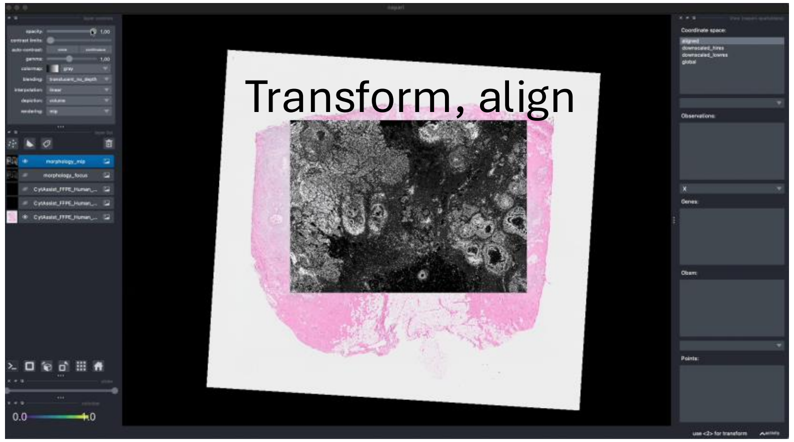
```
SpatialData object with:
├── Images
│     ├── 'morphology_focus': MultiscaleSpatialImage[cyx] (1, 25778, 35416), (1, 12889, 1770
│     └── 'morphology_mip': MultiscaleSpatialImage[cyx] (1, 25778, 35416), (1, 12889, 17708)
├── Points
│     └── 'transcripts': DataFrame with shape: (42638083, 8) (3D points)
├── Shapes
│     ├── 'cell_boundaries': GeoDataFrame shape: (167780, 1) (2D shapes)
│     ├── 'cell_circles': GeoDataFrame shape: (167780, 2) (2D shapes)
│     ├── 'nucleus_boundaries': GeoDataFrame shape: (167780, 1) (2D shapes)
│     └── 'xenium_landmarks': GeoDataFrame shape: (3, 2) (2D shapes)
└── Table
      └── AnnData object with n_obs × n_vars = 167780 × 313
    obs: 'cell_id', 'transcript_counts', 'control_probe_counts', 'control_codeword_counts',
    var: 'gene_ids', 'feature_types', 'genome'
    uns: 'spatialdata_attrs'
    obsm: 'spatial': AnnData (167780, 313)
with coordinate systems:
▸ 'aligned', with elements:
        morphology_mip (Images)
▸ 'global', with elements:
        morphology_focus (Images), morphology_mip (Images), transcripts (Points), cell_bound
```

- Set python env
- Load libraries
- Prepare Xenium raw data into Zarr data
- Read Zarr data in Napari

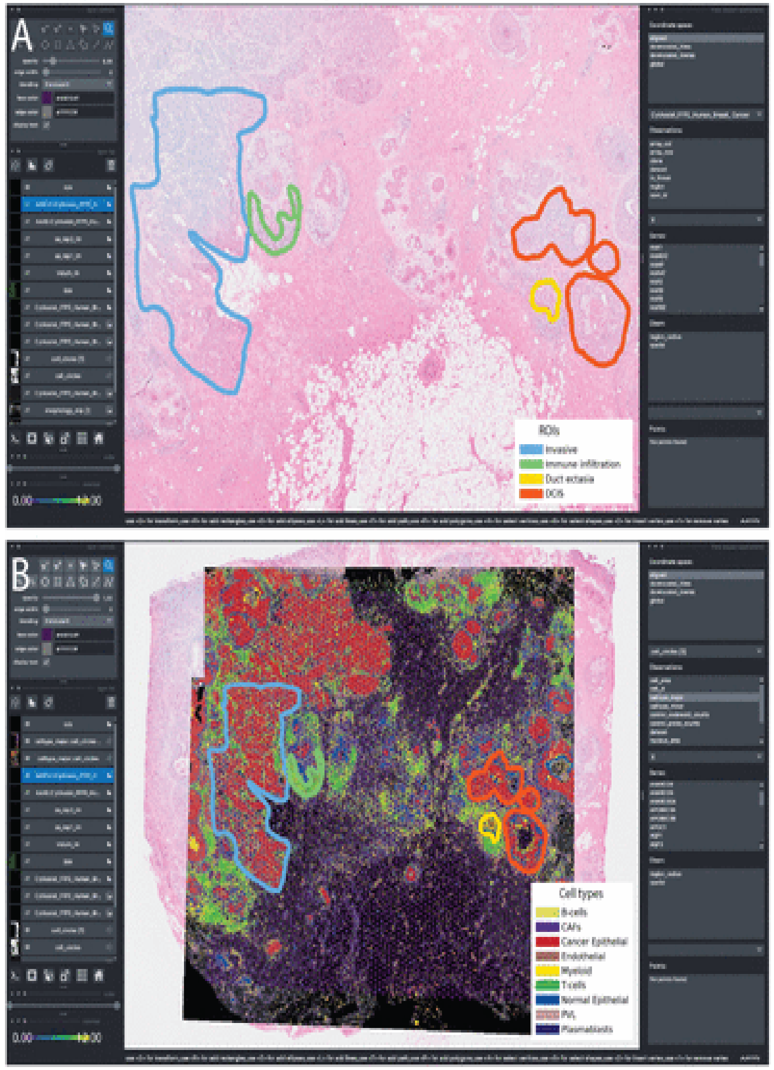


- Add landmarks in different image layers
- Affine similarity transformation
- Align the –omics images layers
  - Align the rest of the SpatialElements
- Save back as Zarr data (lightweight)

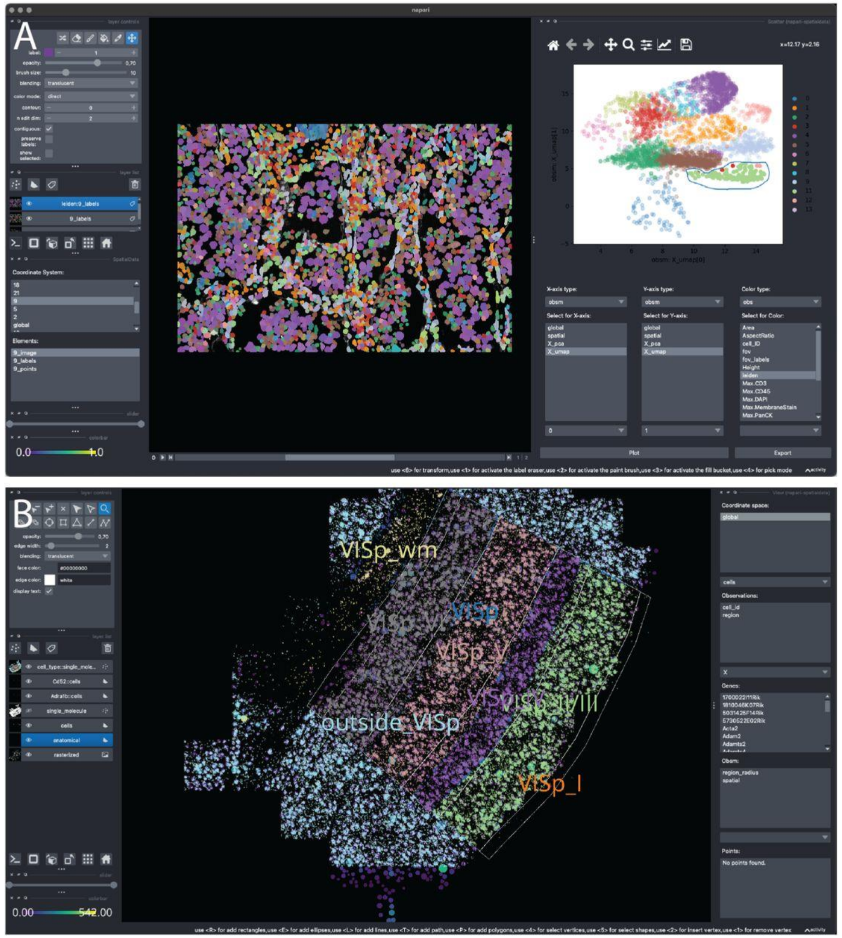
Napari is available in Orion!
Access through Thinlinc



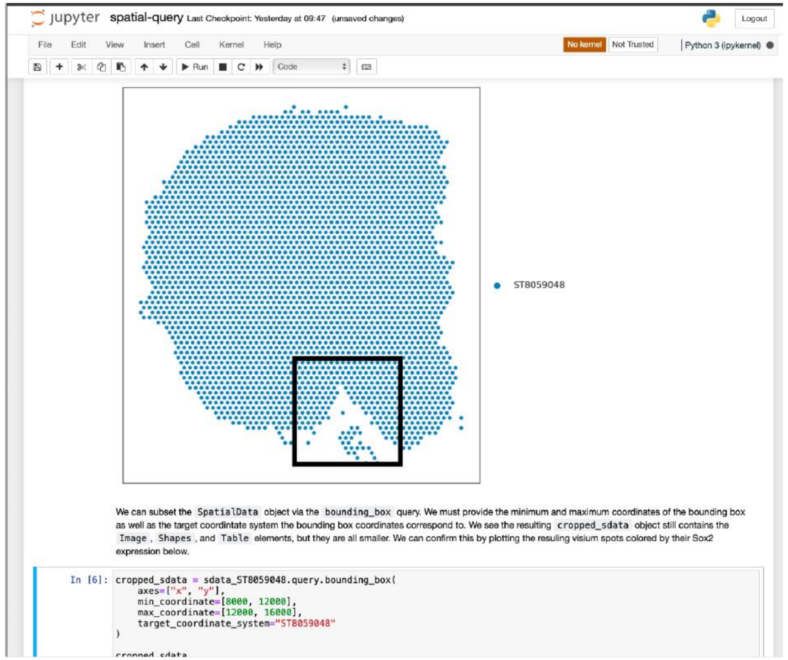
2: Data query



Napari- interactive query



Jupyter notebook query



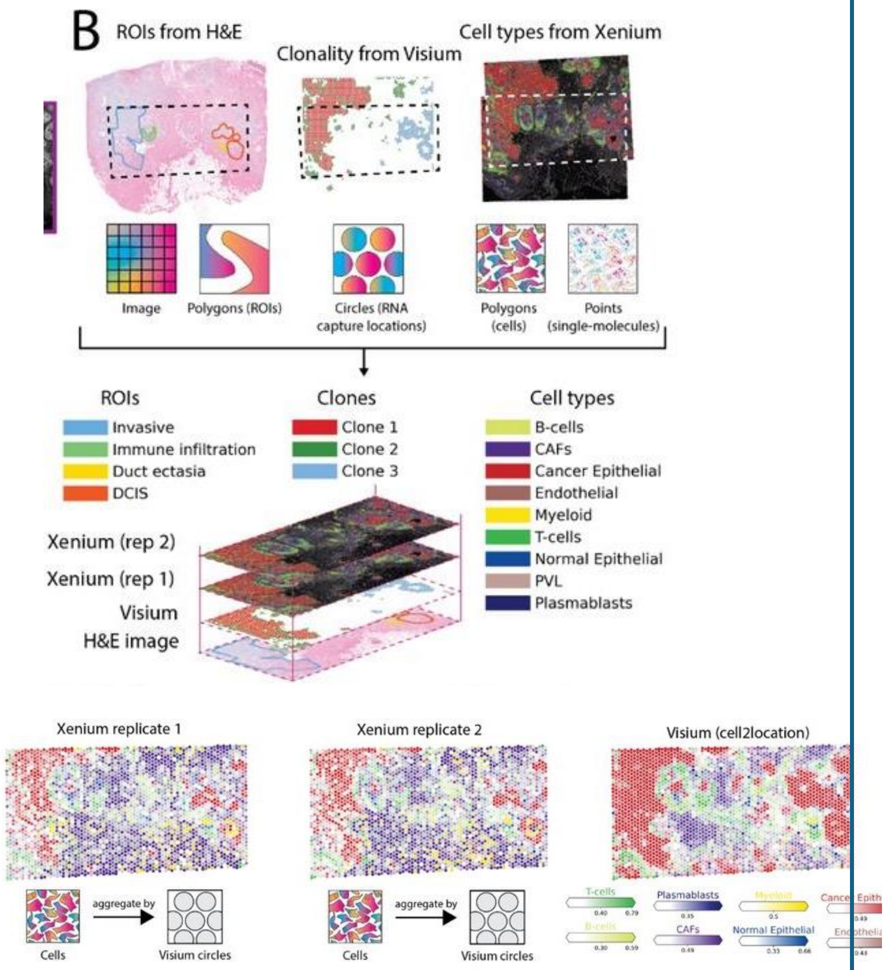
spatial-query Last Checkpoint: Yesterday at 09:47 (unsaved changes)

We can subset the `SpatialData` object via the `bounding_box` query. We must provide the minimum and maximum coordinates of the bounding box as well as the target coordintate system the bounding box coordinates correspond to. We see the resulting `cropped_sdata` object still contains the `Image`, `Shapes`, and `Table` elements, but they are all smaller. We can confirm this by plotting the resuling visium spots colored by their Sox2 expression below.

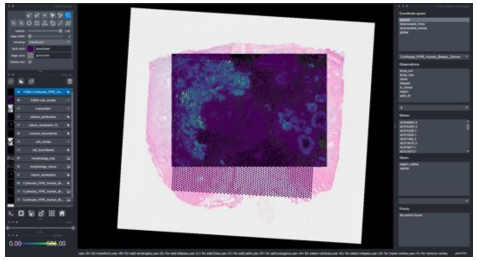
```
In [6]: cropped_sdata = sdata_ST8059048.query.bounding_box(
            axes=["x", "y"],
            min_coordinate=[8000, 12000],
            max_coordinate=[12000, 16000],
            target_coordinate_system="ST8059048"
        )

        cropped_sdata
```

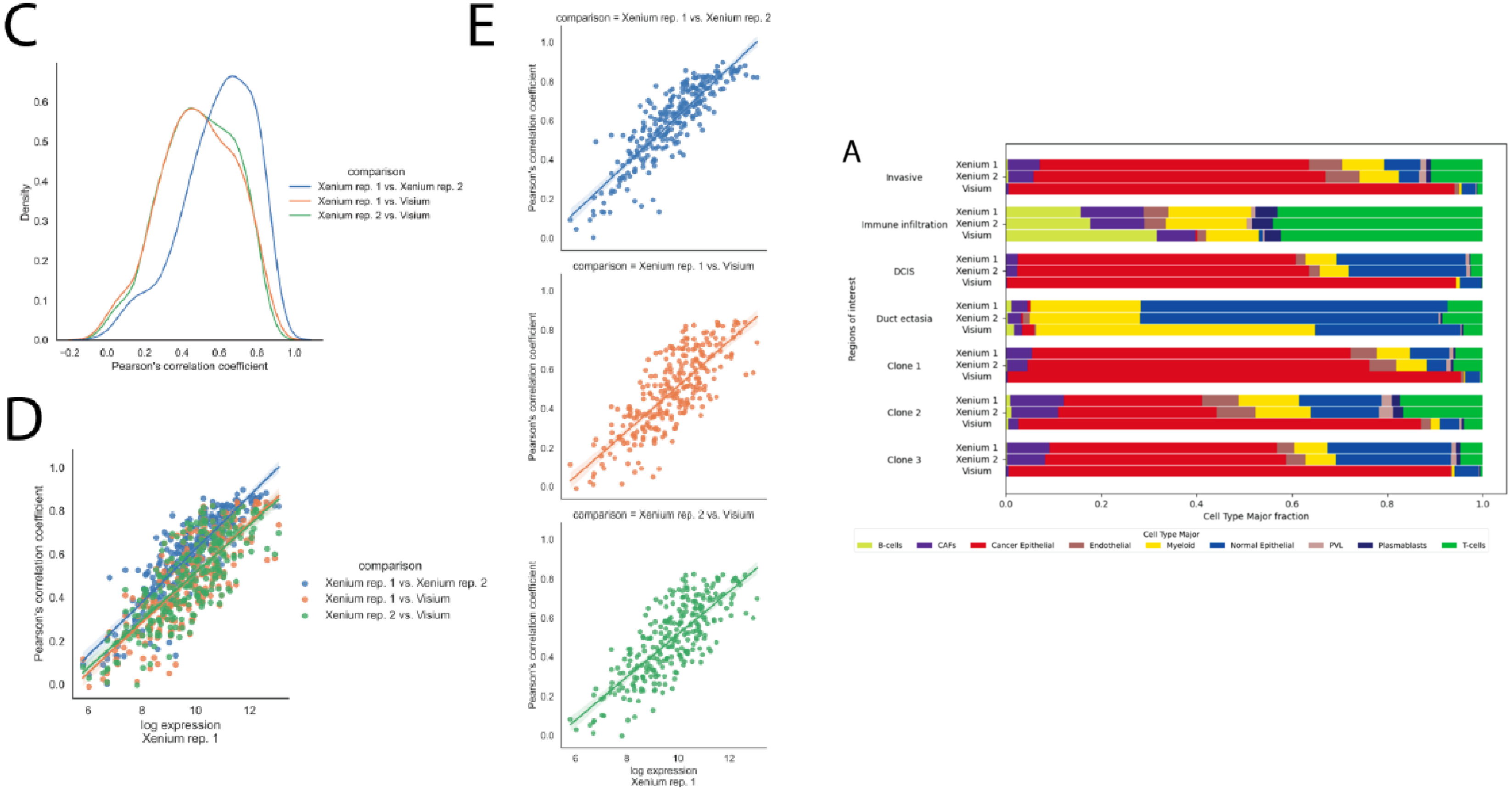
3: Aggregate signals across spatial layers



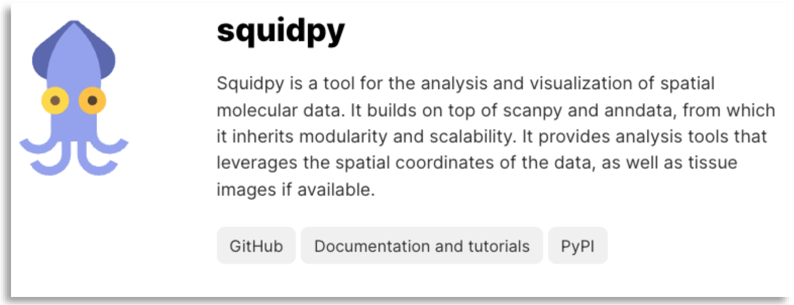
Aligned multi-omics layers



- Select ROIs (aggregate shapes by shapes)
- Aggregate SpatialElements
- Annotation layers, benchmarking
- Determine major subclones (Visium data, CopyKat)
- Xenium sc Annotation – Scanpy label transfer
- Visium sc-deconvolution – Cell2location

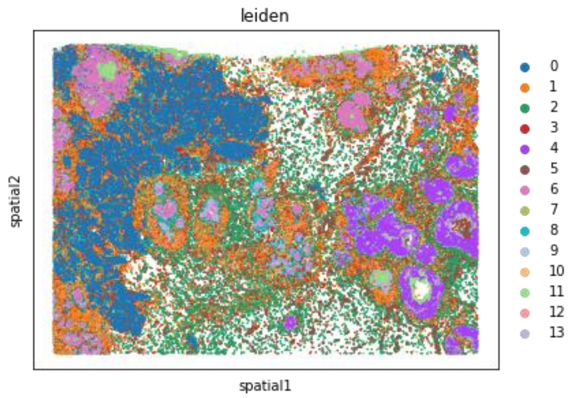


4: Spatial data analysis & visualization

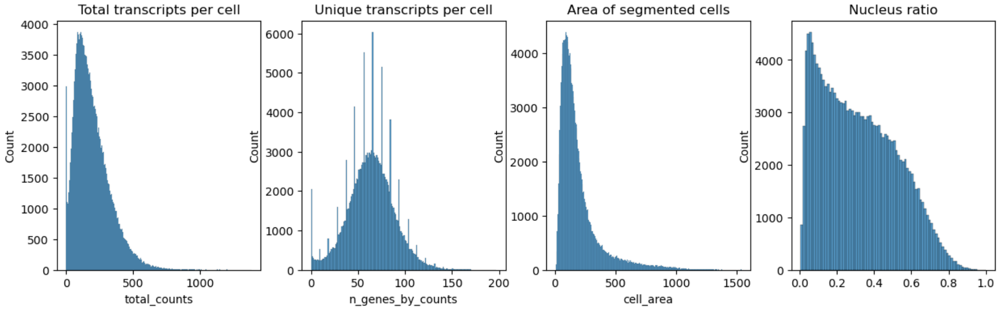


squidpy

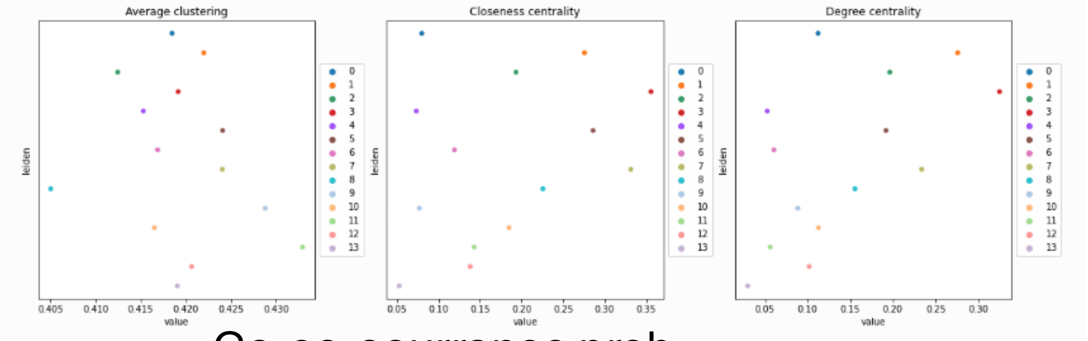
Squidpy is a tool for the analysis and visualization of spatial molecular data. It builds on top of scanpy and anndata, from which it inherits modularity and scalability. It provides analysis tools that leverages the spatial coordinates of the data, as well as tissue images if available.



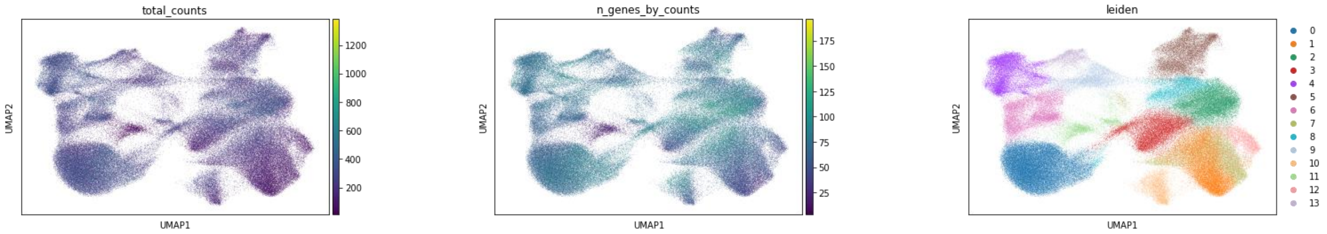
QC



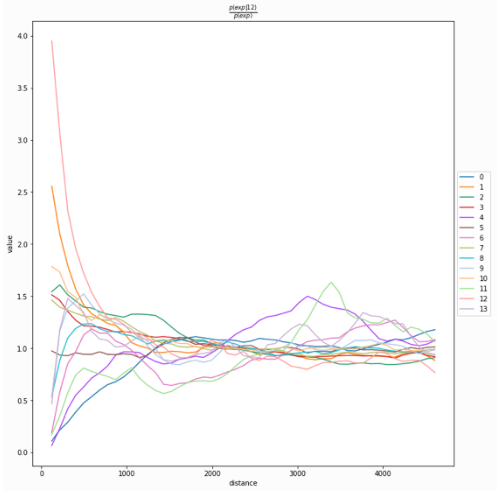
Connectivity



UMAP

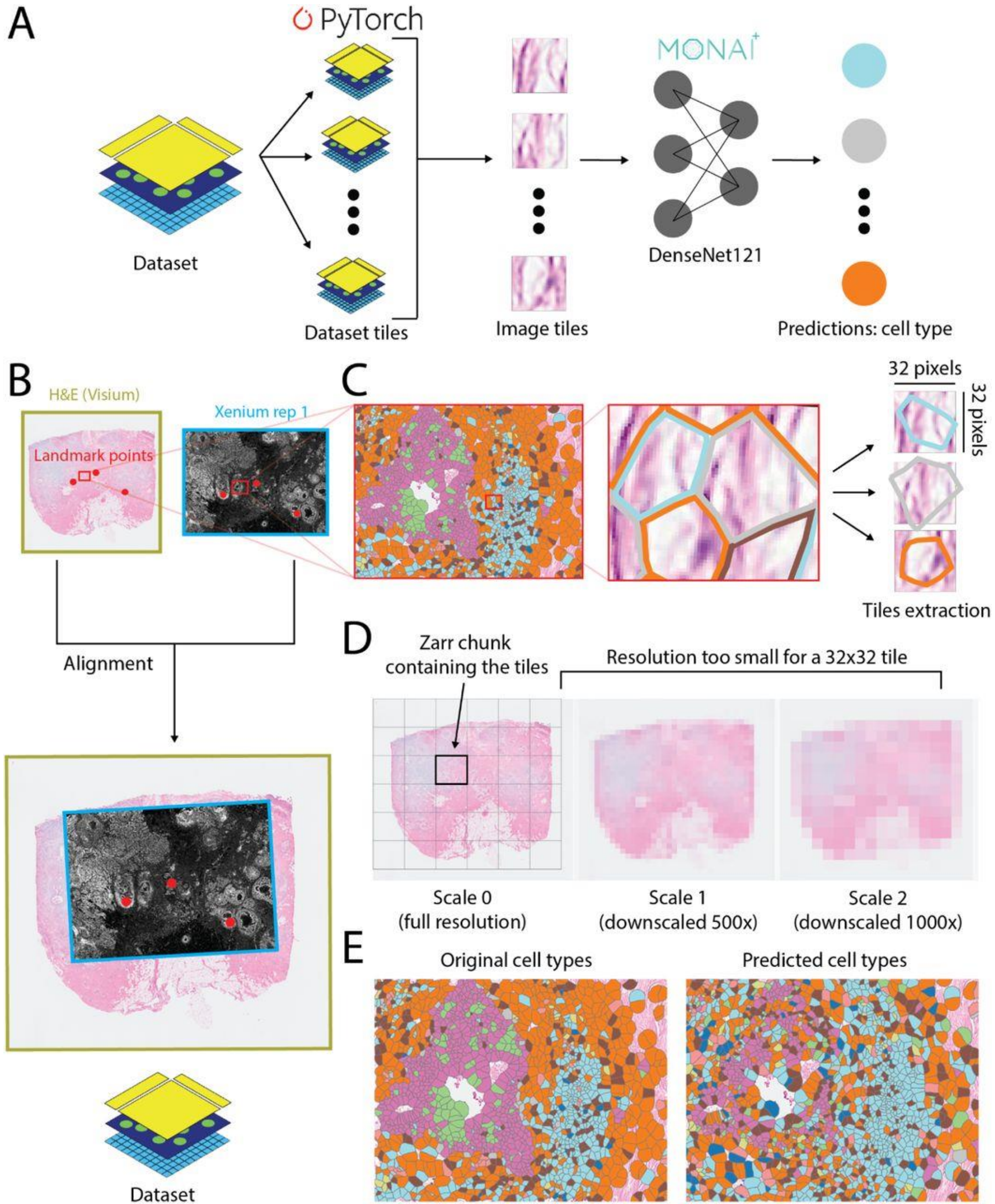


Co-co-ocurrance prob



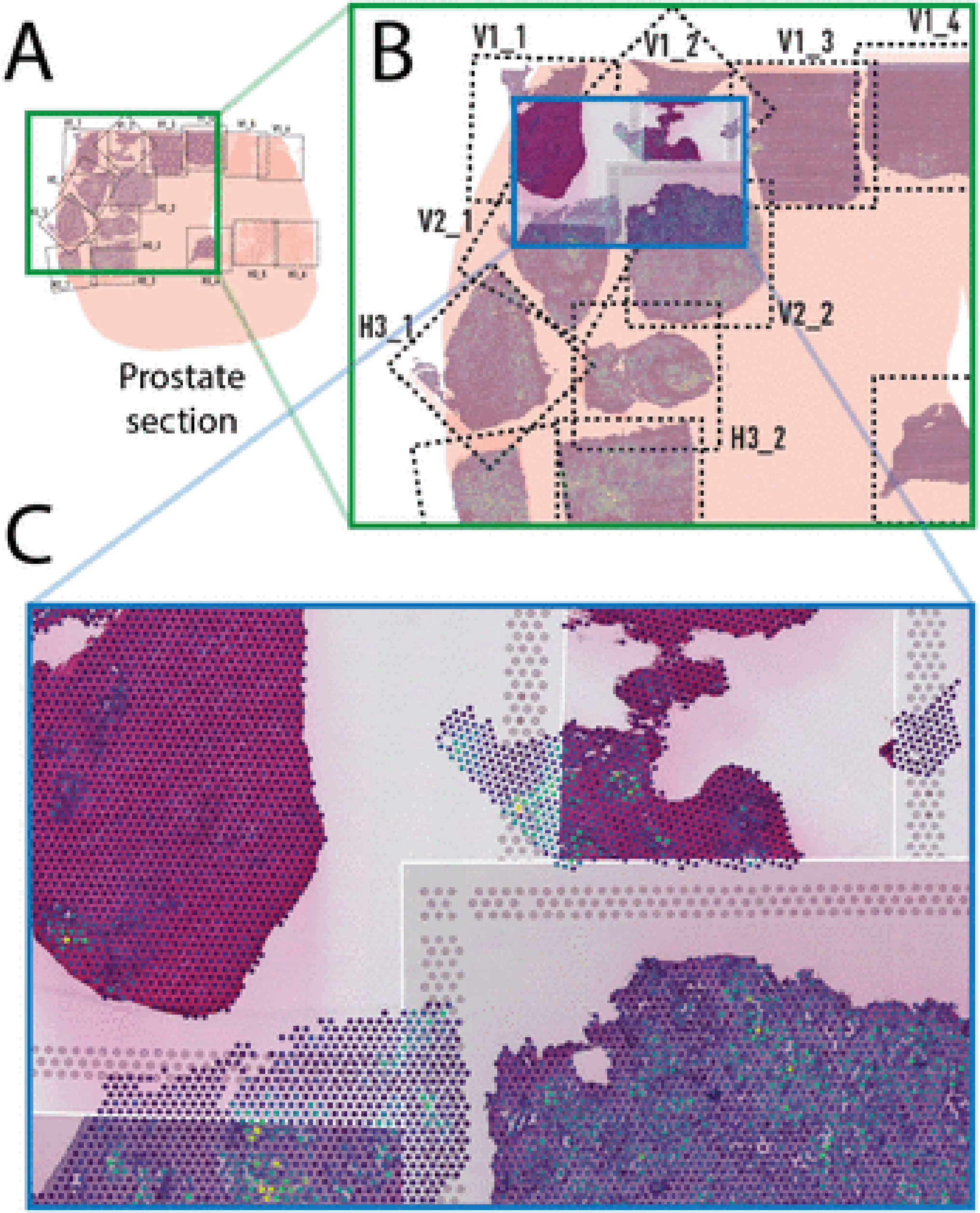
Many more…

5: Deep-learning and predictions of spatial –omics data

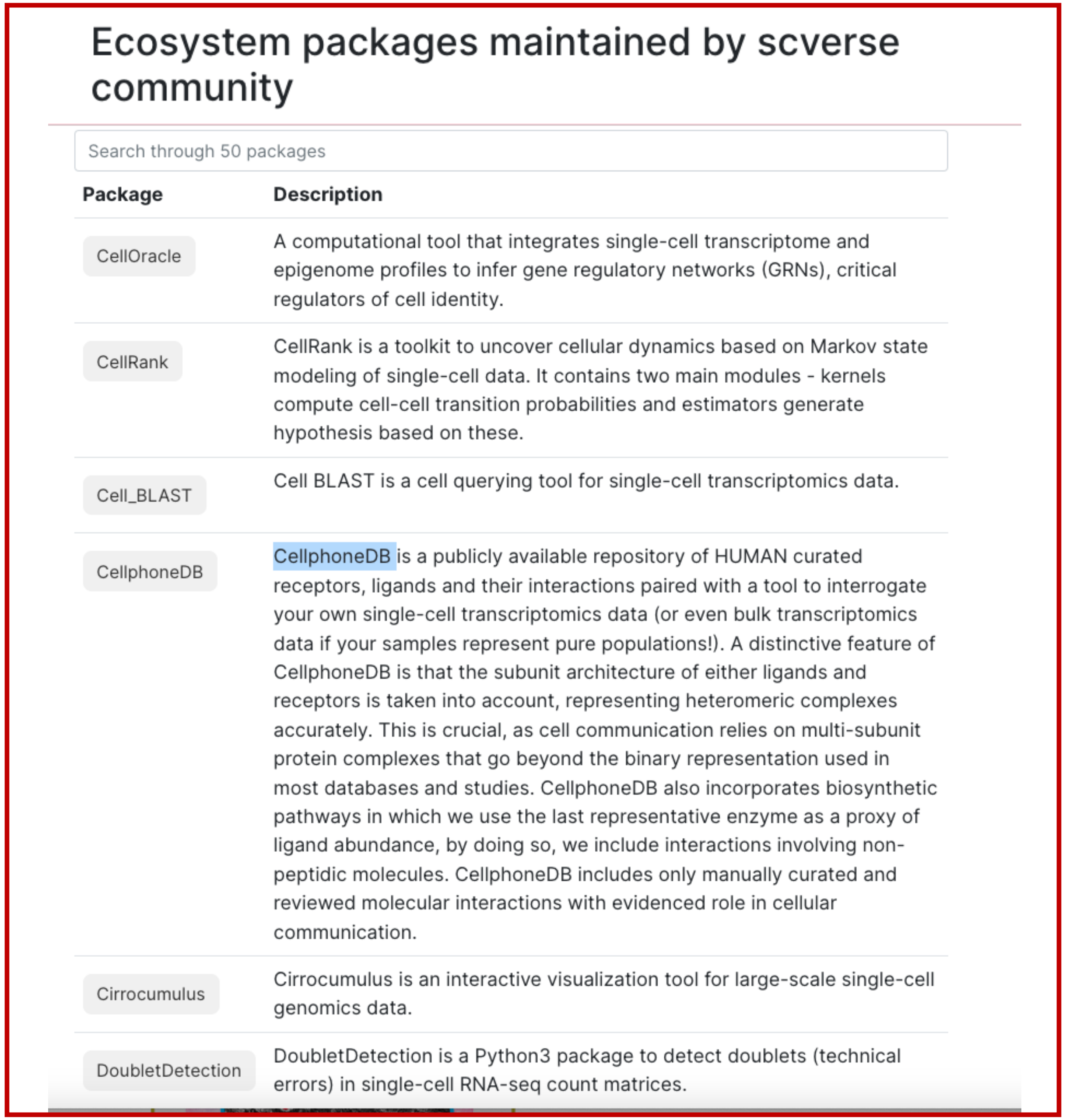


Additional aligned modality training data improves annotation accuracy

6: Others: Combing multiple spatial datasets



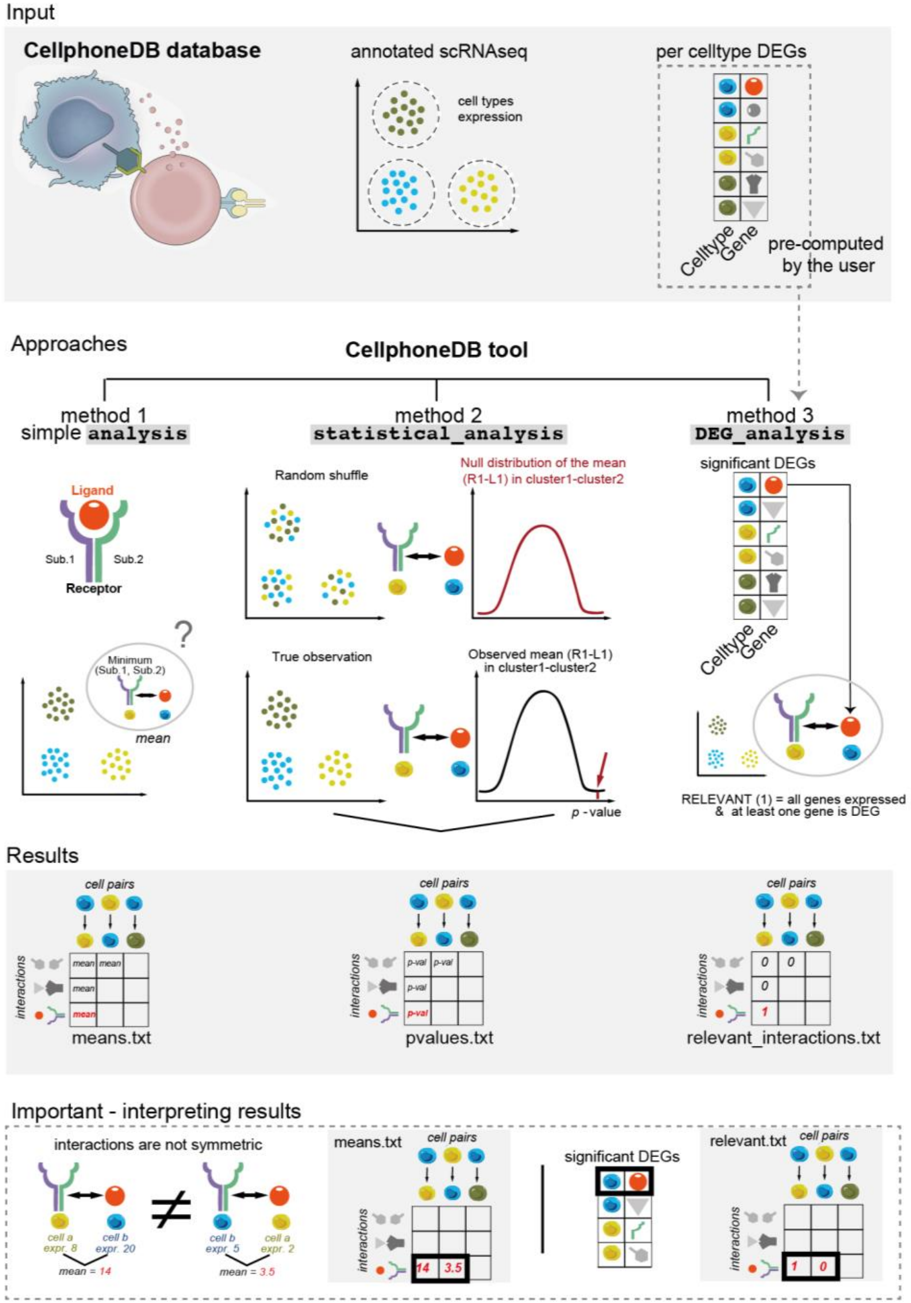
6: Others: Cell-cell communication



Ecosystem packages maintained by scverse community

Search through 50 packages

Package	Description
CellOracle	A computational tool that integrates single-cell transcriptome and epigenome profiles to infer gene regulatory networks (GRNs), critical regulators of cell identity.
CellRank	CellRank is a toolkit to uncover cellular dynamics based on Markov state modeling of single-cell data. It contains two main modules - kernels compute cell-cell transition probabilities and estimators generate hypothesis based on these.
Cell_BLAST	Cell BLAST is a cell querying tool for single-cell transcriptomics data.
CellphoneDB	CellphoneDB is a publicly available repository of HUMAN curated receptors, ligands and their interactions paired with a tool to interrogate your own single-cell transcriptomics data (or even bulk transcriptomics data if your samples represent pure populations!). A distinctive feature of CellphoneDB is that the subunit architecture of either ligands and receptors is taken into account, representing heteromeric complexes accurately. This is crucial, as cell communication relies on multi-subunit protein complexes that go beyond the binary representation used in most databases and studies. CellphoneDB also incorporates biosynthetic pathways in which we use the last representative enzyme as a proxy of ligand abundance, by doing so, we include interactions involving non-peptidic molecules. CellphoneDB includes only manually curated and reviewed molecular interactions with evidenced role in cellular communication.
Cirrocumulus	Cirrocumulus is an interactive visualization tool for large-scale single-cell genomics data.
DoubletDetection	DoubletDetection is a Python3 package to detect doublets (technical errors) in single-cell RNA-seq count matrices.



Summary

SpatialData framework:

- Represents and manipulates multi-omics spatial data formats
- Aligns datasets across modalities using coordinate transformations
- Enables spatial querying, aggregation, and annotation
- Integrates with deep learning (PyTorch) and analysis (scverse) ecosystems, which allows in depth interrogation of the –omics spatial data