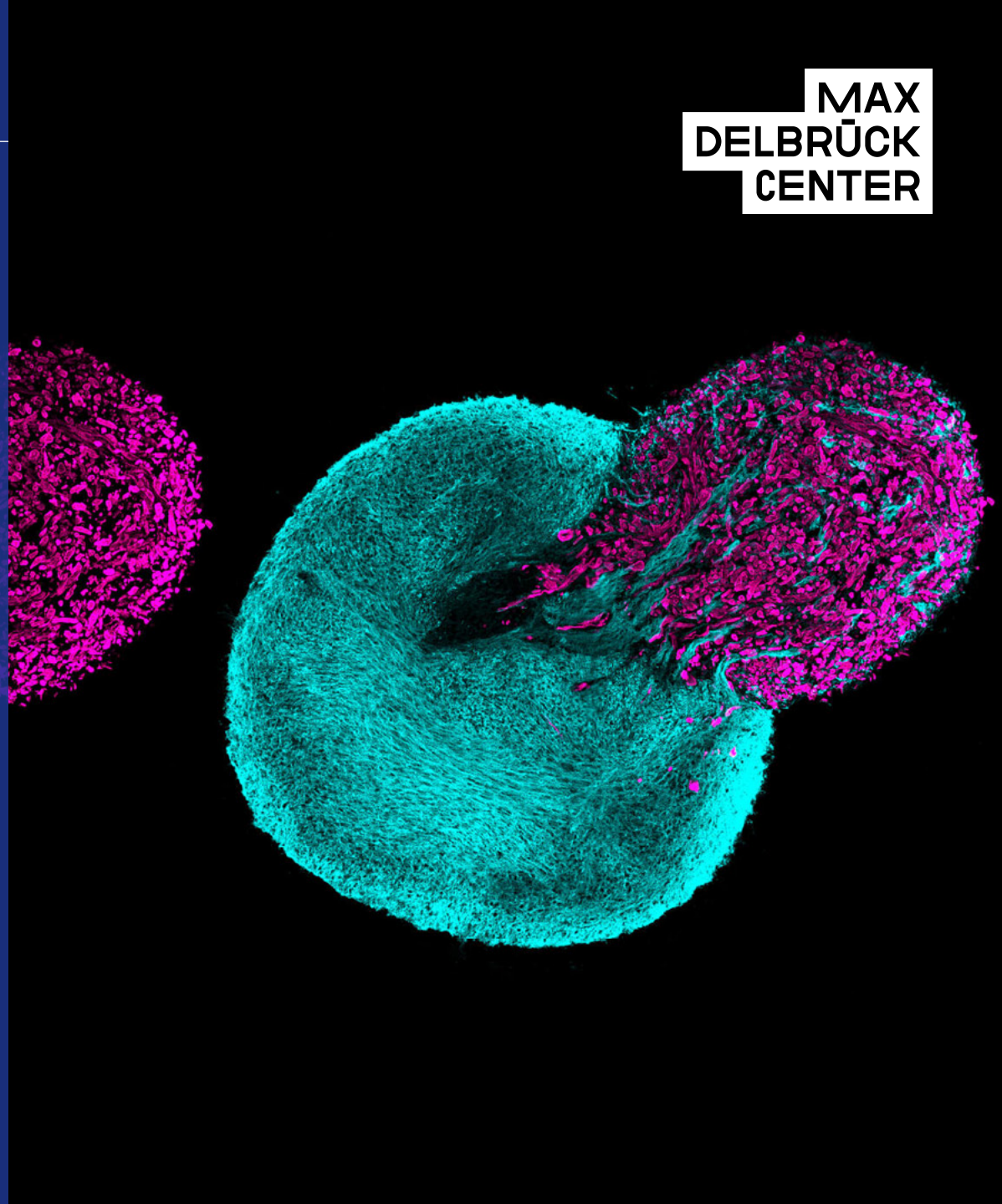
SPATIAL OMICS ANALYSIS WORKFLOWS (DAY 2)

Artür Manukyan
4 Mar. 2025

BIMSBBioinfo
CompGen2025 Module 2



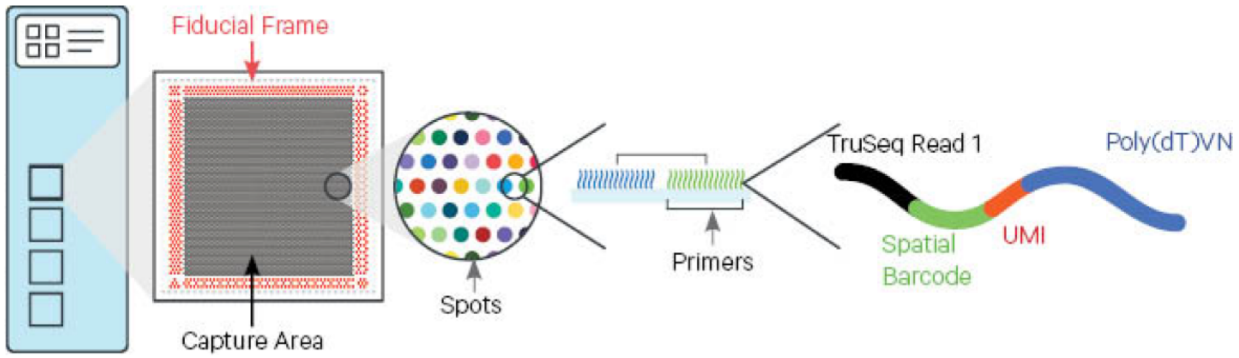
AGENDA

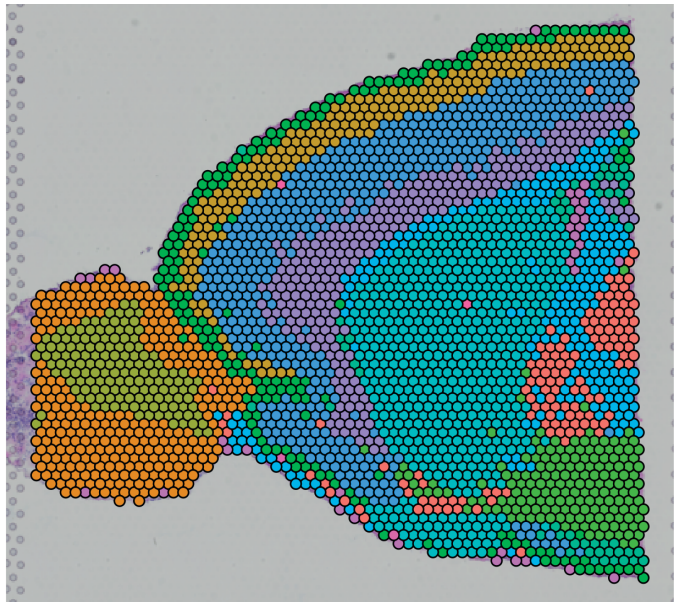
1. **Spot-level Data Analysis (Visium)**
    1. Unbiased Clustering
    2. Spot Deconvolution
    3. Niche Clustering of Spots

2. **Cell-level Data Analysis (Xenium)**
    1. Ondisk Support and Lazy Operations
    2. Unbiased Clustering
    3. Marker Analysis
    4. Niche Clustering of Cells
    5. Hot Spot Analysis

V
Spot-level Spatial Omics

10x Visium Gene Expression





https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

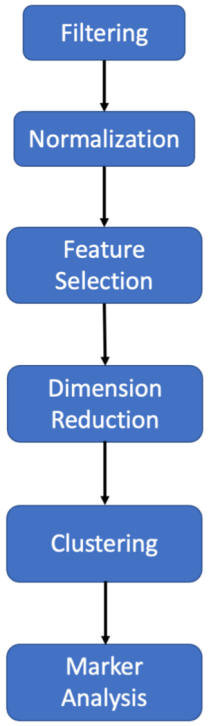
UNBIASED CLUSTERING OF OBSERVATIONS

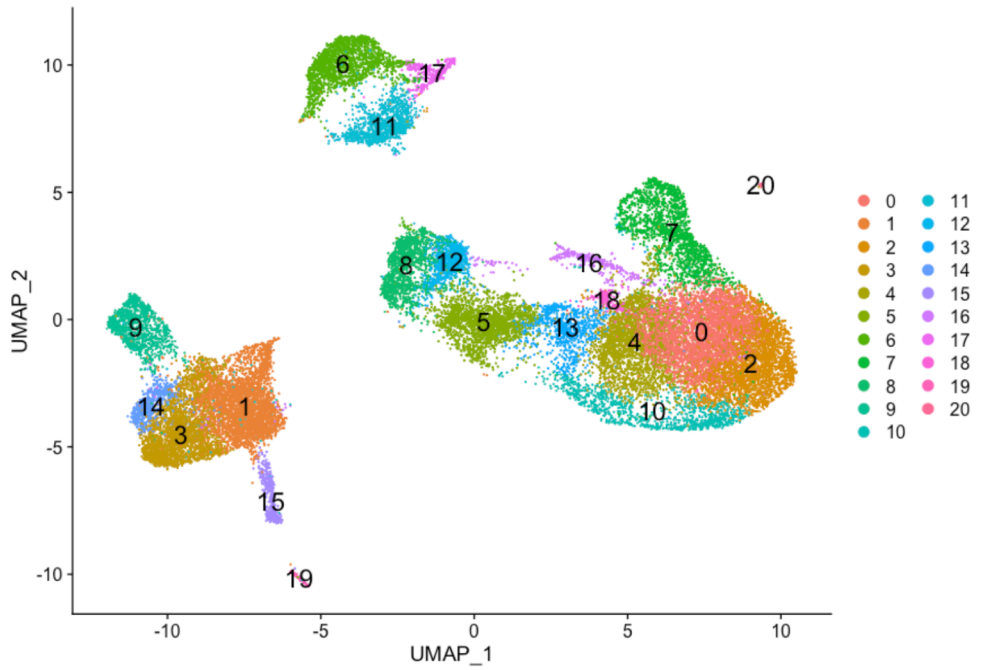
Clustering of spot/cell datasets

Cells

+10000 Genes

	Group A		Group B	
	P.1	P.2	P.3	P.4
Gene 1	42	43	10	9
Gene 2	25	24	2	3
Gene 3	10	9	100	98
Gene 4	40	39	4	5
Sum	117	115	116	115



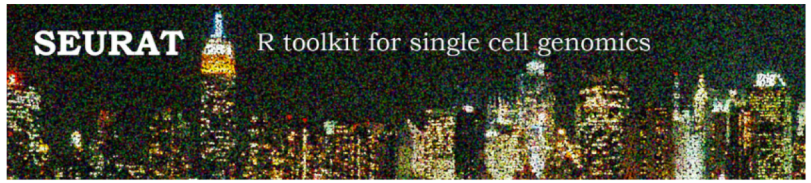


https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

FUNDAMENTALS OF SINGLE CELL ANALYSIS

https://satijalab.org/seurat/

Seurat 5.2.0 | Install | Get started | Vignettes | Extensions | FAQ | News | Reference | Archive



Seurat v5

We are excited to release Seurat v5! To install, please follow the instructions in our install page. This update brings the following new features and functionality:

- **Integrative multimodal analysis:** The cellular transcriptome is just one aspect of cellular identity, and recent technologies enable routine profiling of chromatin accessibility, histone modifications, and protein levels from single cells. In Seurat v5, we introduce 'bridge integration', a statistical method to integrate experiments measuring different modalities (i.e. separate scRNA-seq and scATAC-seq datasets), using a separate multiomic dataset as a molecular 'bridge'. For example, we demonstrate how to map scATAC-seq datasets onto scRNA-seq datasets, to assist users in interpreting and annotating data from new modalities.

  We recognize that while the goal of matching shared cell types across datasets may be important for many problems, users may also be concerned about which method to use, or that integration could result in a loss of biological resolution. In Seurat v5, we also introduce flexible and streamlined workflows for the integration of multiple scRNA-seq datasets. This makes it easier to explore the results of different integration methods, and to compare these results to a workflow that excludes integration steps.

  - Paper: Dictionary learning for integrative, multimodal, and scalable single-cell analysis
  - Vignette: Streamlined integration of scRNA-seq data
  - Vignette: Cross-modality bridge integration
  - Website: Azimuth-ATAC, reference-mapping for scATAC-seq datasets

- **Flexible, interactive, and highly scalable analsyis:** The size and scale of single-cell sequencing datasets is rapidly increasing, outpacing even Moore's law. In Seurat v5, we introduce new infrastructure and methods to analyze, interpret, and explore exciting datasets spanning millions of cells, even if they cannot be fully loaded into memory. We introduce support for 'sketch'-based analysis, where representative subsamples of a large dataset are stored in-memory to enable rapid and iterative analysis - while the full dataset remains accessible via on-disk storage.

  We enable high-performance via the BPCells package, developed by Ben Parks in the Greenleaf Lab. The BPCells package enables high-performance analysis via innovative bit-packing compression techniques, optimized C++ code, and use of streamlined and lazy operations.

Links

View on CRAN
Browse source code
Report a bug

License

Full license
MIT + file LICENSE

Community

Code of conduct

Citation

Citing Seurat

Developers

Rahul Satija
Author, maintainer
Satija Lab and Collaborators
Funder
More about authors...

Seurat 5.2.0 | Install | Get started | Vignettes | Extensions | FAQ | News | Reference | Archive

Getting Started with Seurat

Source: `vignettes/get_started_v5_new.Rmd`

We provide a series of vignettes, tutorials, and analysis walkthroughs to help users get started with Seurat. You can also check out our Reference page which contains a full list of functions available to users.

Our previous Get Started page for Seurat v4 is archived here.

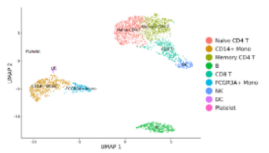
Contents

Introductory Vignettes
scRNA Data Integration
Multi-assay data
Flexible analysis of massively scalable datasets
Spatial analysis
Other
SeuratWrappers

Introductory Vignettes

For new users of Seurat, we suggest starting with a guided walk through of a dataset of 2,700 Peripheral Blood Mononuclear Cells (PBMCs) made publicly available by 10X Genomics. This tutorial implements the major components of a standard unsupervised clustering workflow including QC and data filtration, calculation of high-variance genes, dimensional reduction, graph-based clustering, and the identification of cluster markers. We provide additional vignettes introducing visualization techniques in Seurat, the sctransform normalization workflow, and storage/interaction with multimodal datasets. We also provide an 'essential commands cheatsheet' as a quick reference.

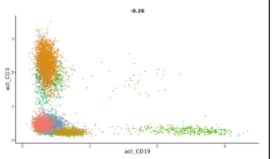
Guided tutorial — 2,700 PBMCs



A basic overview of Seurat that includes an introduction to common analytical workflows.

GO

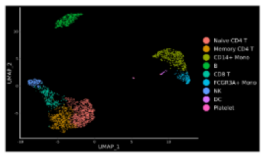
Multimodal analysis



An introduction to working with multi-modal datasets in Seurat.

GO

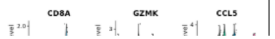
Visualization



An overview of the visualization capabilities within Seurat.

GO

SCTransform



Essential Commands Cheat Sheet

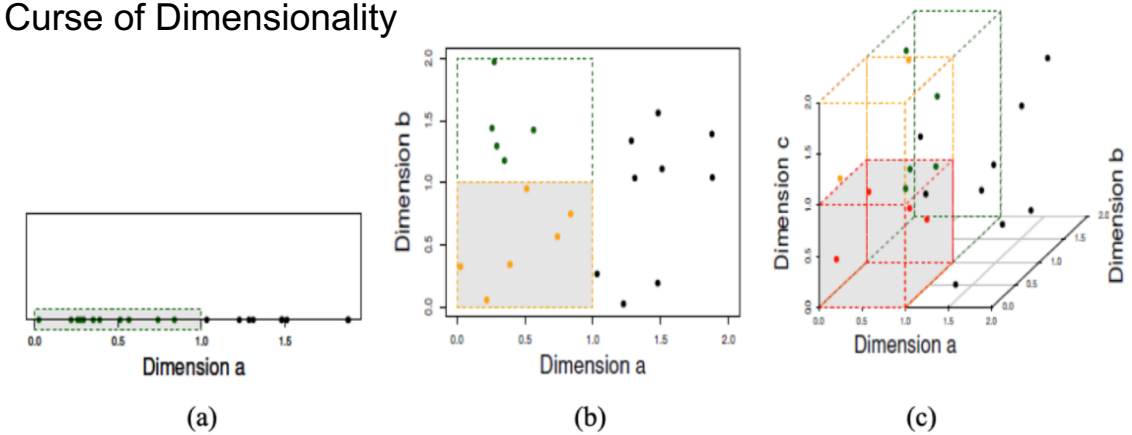
Standard `Seurat` workflow

https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

FEATURE SELECTION AND EXTRACTION

Clustering of spot/cell datasets

Curse of Dimensionality



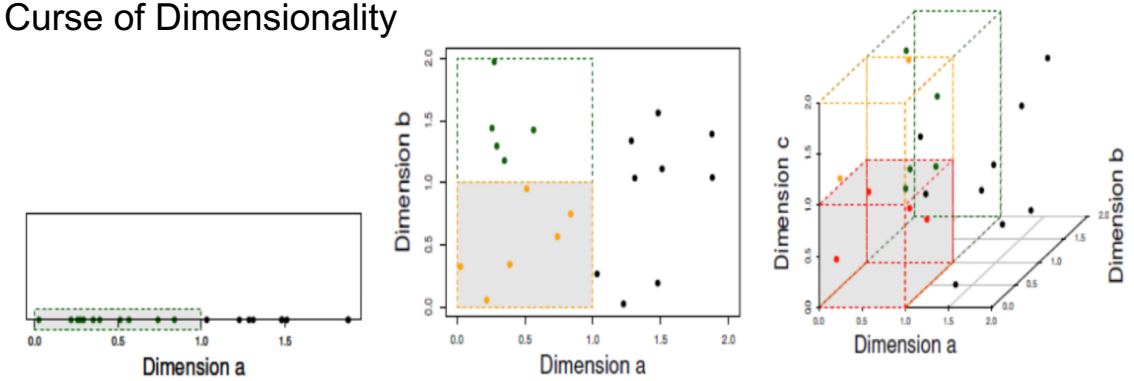
Source: Parsons et al. (2004)

https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

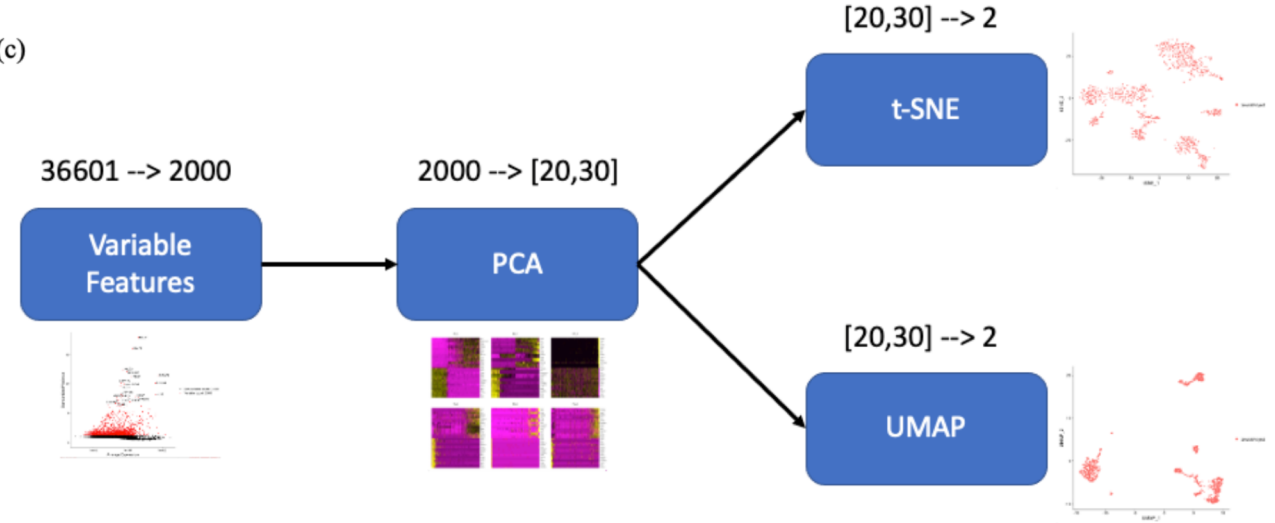
FEATURE SELECTION AND EXTRACTION

Clustering of spot/cell datasets

Curse of Dimensionality



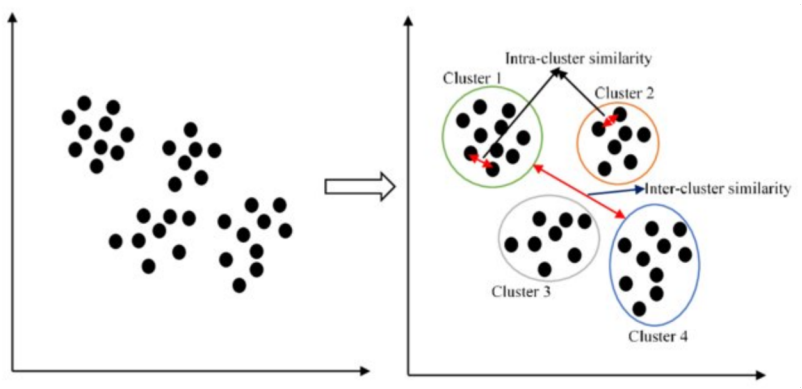
Source: Parsons et al. (2004)



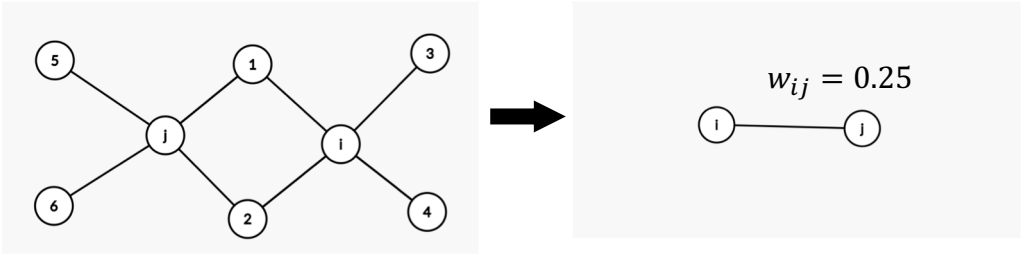
https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

GRAPH CLUSTERING

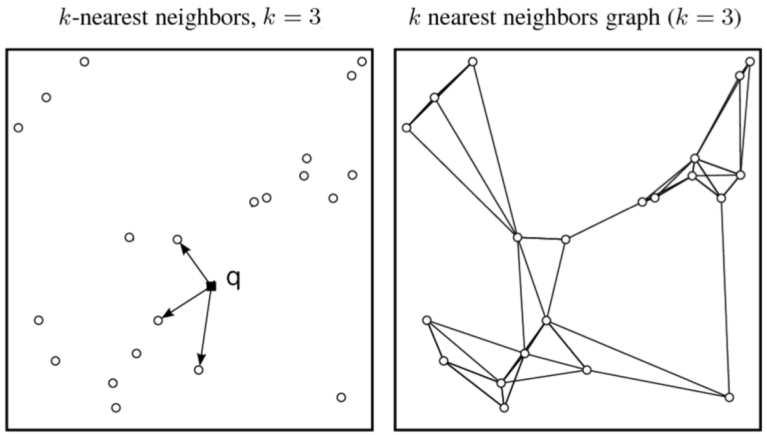
Partitioning cells using dense subgraphs



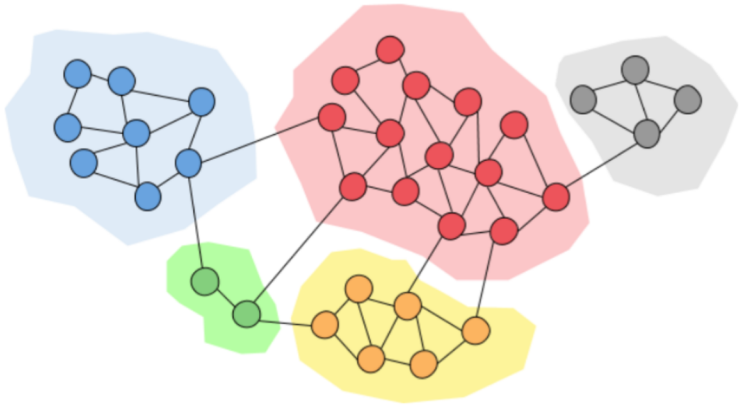
Shared Nearest Neighbor Graph and Jaccard Similarity



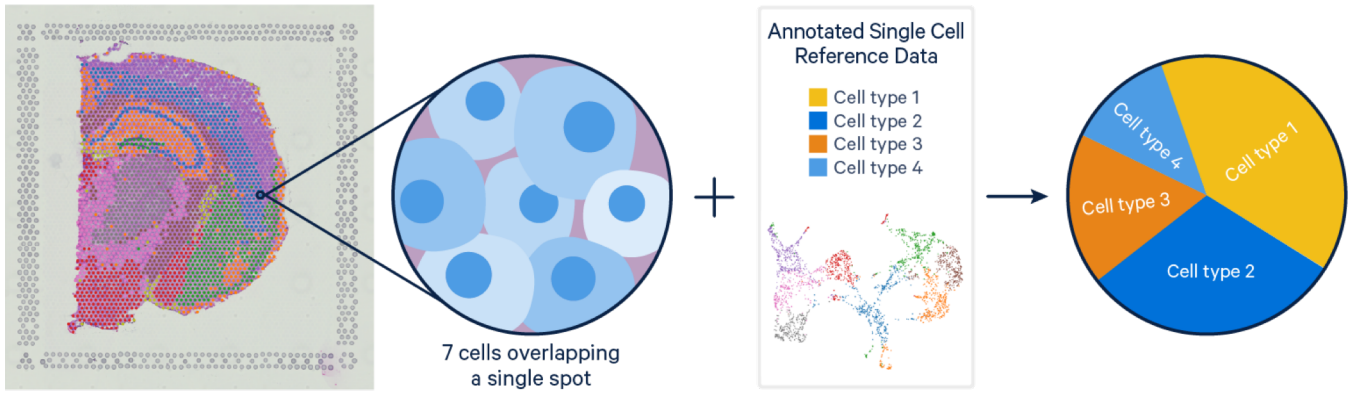
$$JS(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} = \frac{2}{8} = 0.25$$

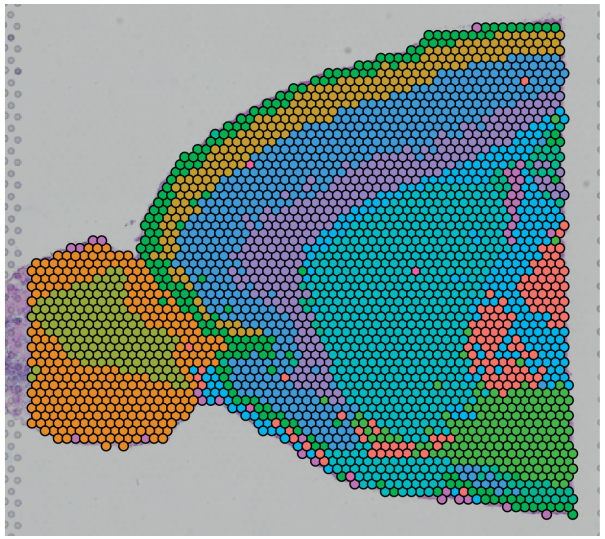


Leiden (Modularity-based) Clustering



Estimating and Incorporating Cell Mixtures of Spots





Niche Clustering (Spot-based)

Center log ratio transformation

Compositional data

3 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

In statistics, **compositional data** are quantitative descriptions of the parts of some whole, conveying relative information. Mathematically, compositional data is represented by points on a simplex. Measurements involving probabilities, proportions, percentages, and ppm can all be thought of as compositional data.

Simplicial sample space [edit]

In general, John Aitchison defined compositional data to be proportions of some whole in 1982.[1] In particular, a compositional data point (or *composition* for short) can be represented by a real vector with positive components. The sample space of compositional data is a simplex:

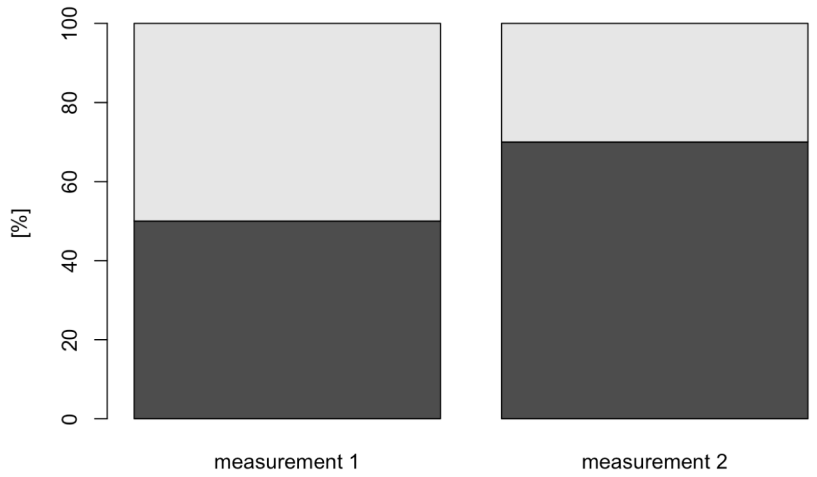
$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \in \mathbb{R}^D \;\middle|\; x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Center log ratio transform [edit]

The center log ratio (clr) transform is both an isomorphism and an isometry where $\mathrm{clr} : S^D \to U, \quad U \subset \mathbb{R}^D$

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)}, \cdots, \log \frac{x_D}{g(x)} \right]$$

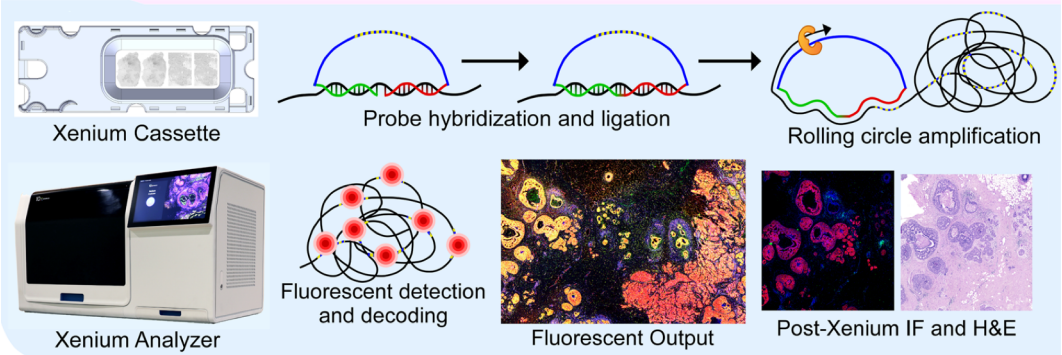
Where $g(x)$ is the geometric mean of x. The inverse of this function is also known as the softmax function.



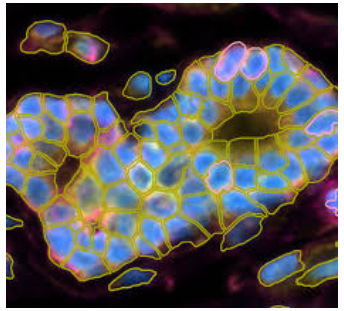
XENIUM (10X GENOMICS)

Cell (and subcellular)-level Spatial Omics

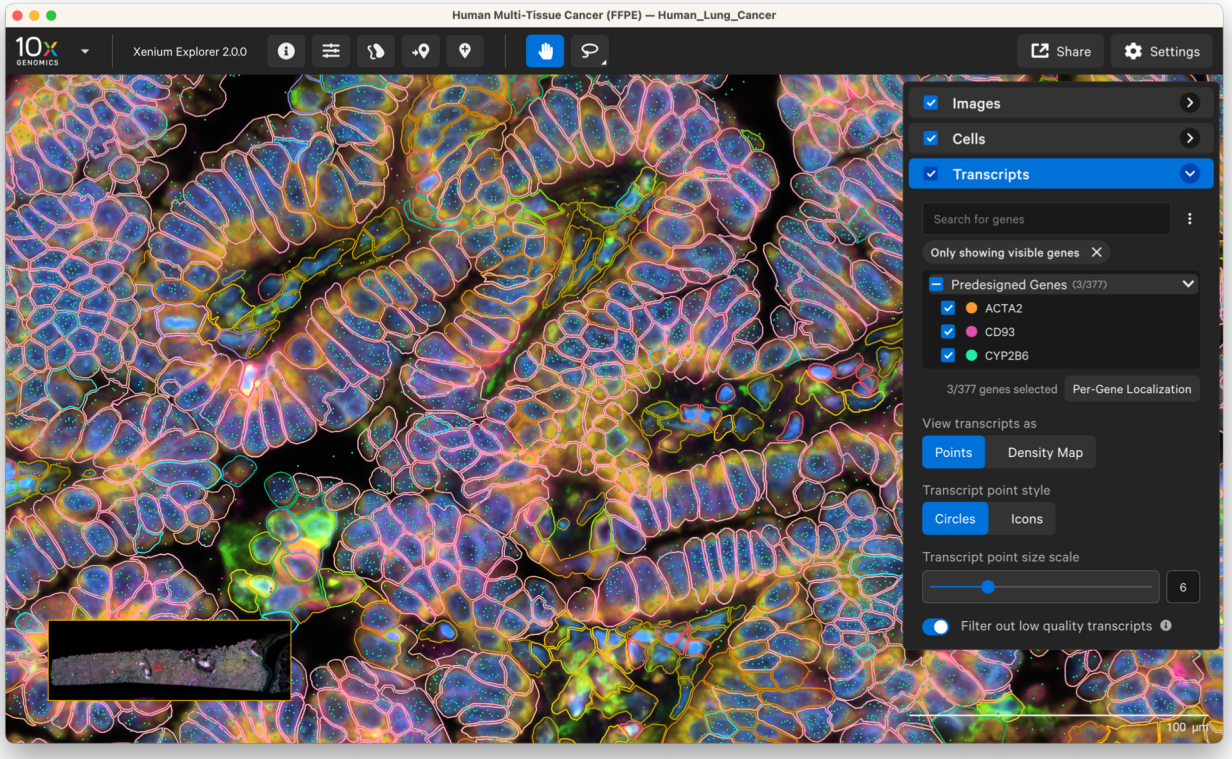
High-throughput single molecule FISH



Cell Segmentation



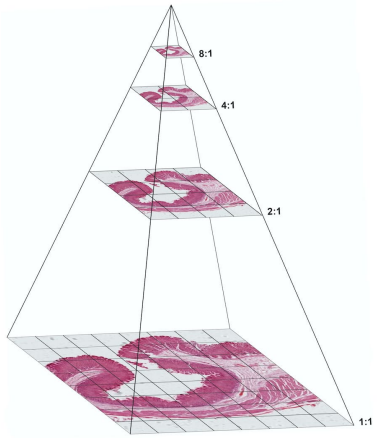
Single Cell Assay



https://www.10xgenomics.com/support/software/xenium-explorer/latest/resources/xenium-explorer-demos

- analysis
- analysis_summary.html
- analysis.zarr.zip
- cell_boundaries.csv.gz
- cell_boundaries.parquet
- cell_feature_matrix
- cell_feature_matrix.h5
- cell_feature_matrix.zarr.zip
- cells.csv.gz
- cells.parquet
- cells.zarr.zip
- experiment.xenium
- gene_panel.json
- metrics_summary.csv
- morphology_focus.ome.tif
- morphology_mip.ome.tif
- morphology.ome.tif
- nucleus_boundaries.csv.gz
- nucleus_boundaries.parquet
- transcripts.csv.gz
- transcripts.parquet
- transcripts.zarr.zip

Pyramid Representation



BioFormats (Open Microscopy Environment)

OME

Bio-Formats
Downloads by version
Documentation by version
Licensing
Previous topic
Dataset Structure Table
Next topic
3i SlideBook
Quick search
Go
This Page
Show Source
Show on GitHub
Edit on GitHub

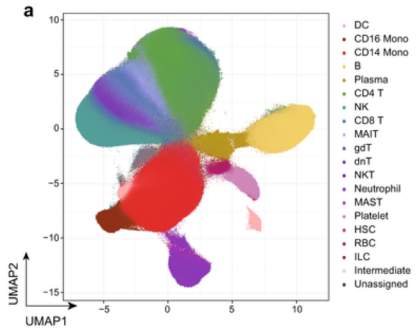
Supported Formats

Ratings legend and definitions

You can sort this table by clicking on any of the headings.

Format	Extensions
3i SlideBook	.sld
Andor Bio-Imaging Division (ABD) TIFF	.tif
AIM	.aim
Alicona 3D	.al3d
Amersham Biosciences Gel	.gel
Amira Mesh	.am, .amiramesh, .grey, .hx, .labels
Amnis FlowSight	.cif
Analyze 7.5	.img, .hdr
Andor SIF	.sif
Animated PNG	.png
Aperio AFI	.afi, .svs
Aperio SVS TIFF	.svs
Applied Precision CellWorX	.htd, .pnl
AVI (Audio Video Interleave)	.avi
Axon Raw Format	.arf
BD Pathway	.exp, .tif

Cell Populations In Millions

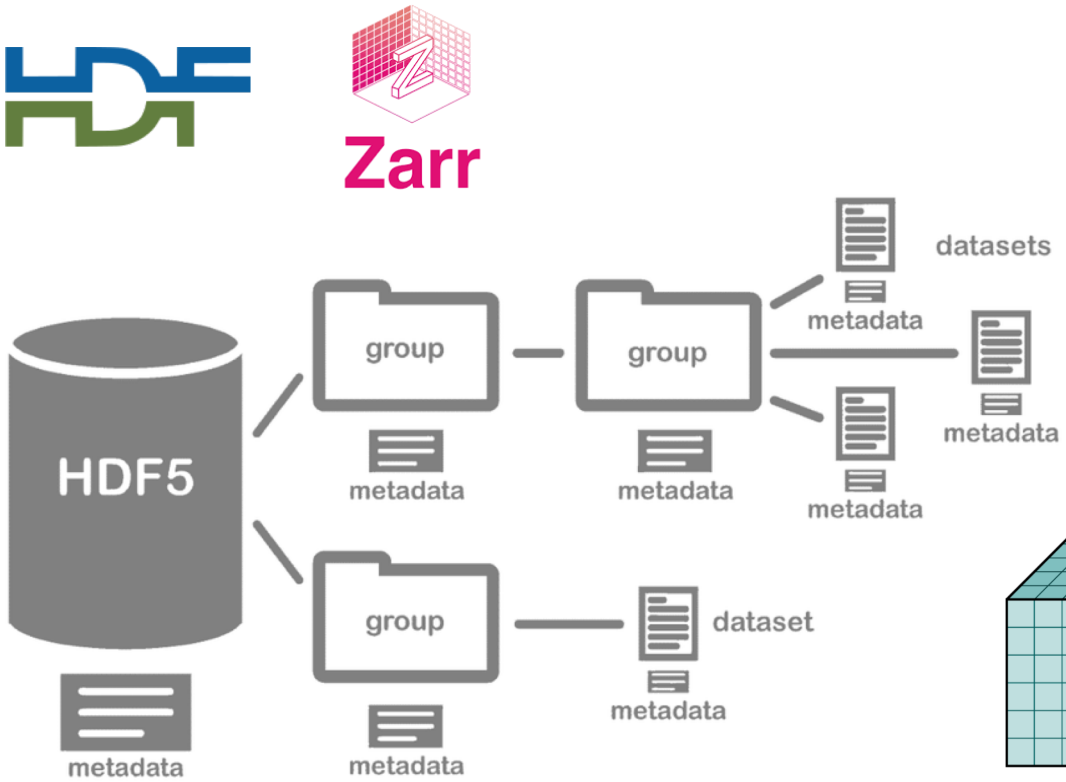


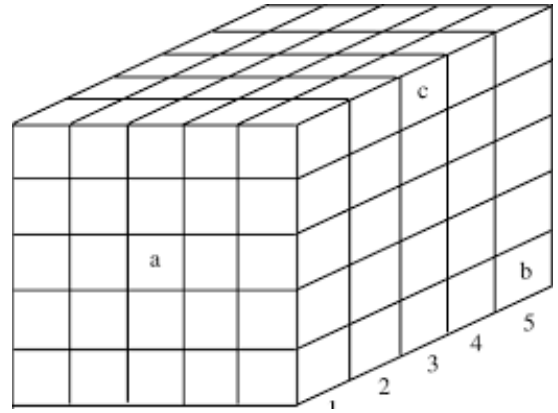
Hierarchical Data Formats for Large Storage

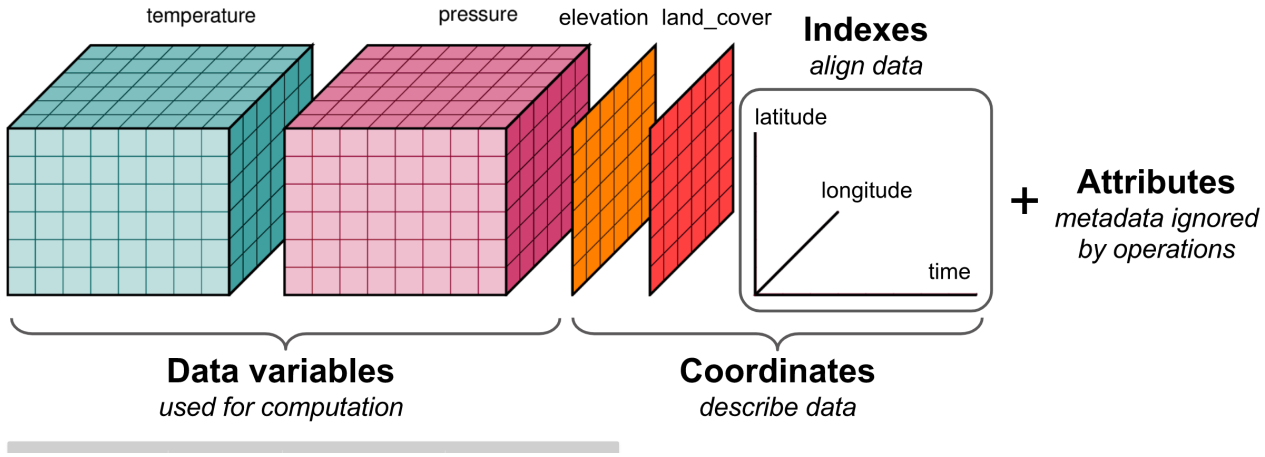
HDF

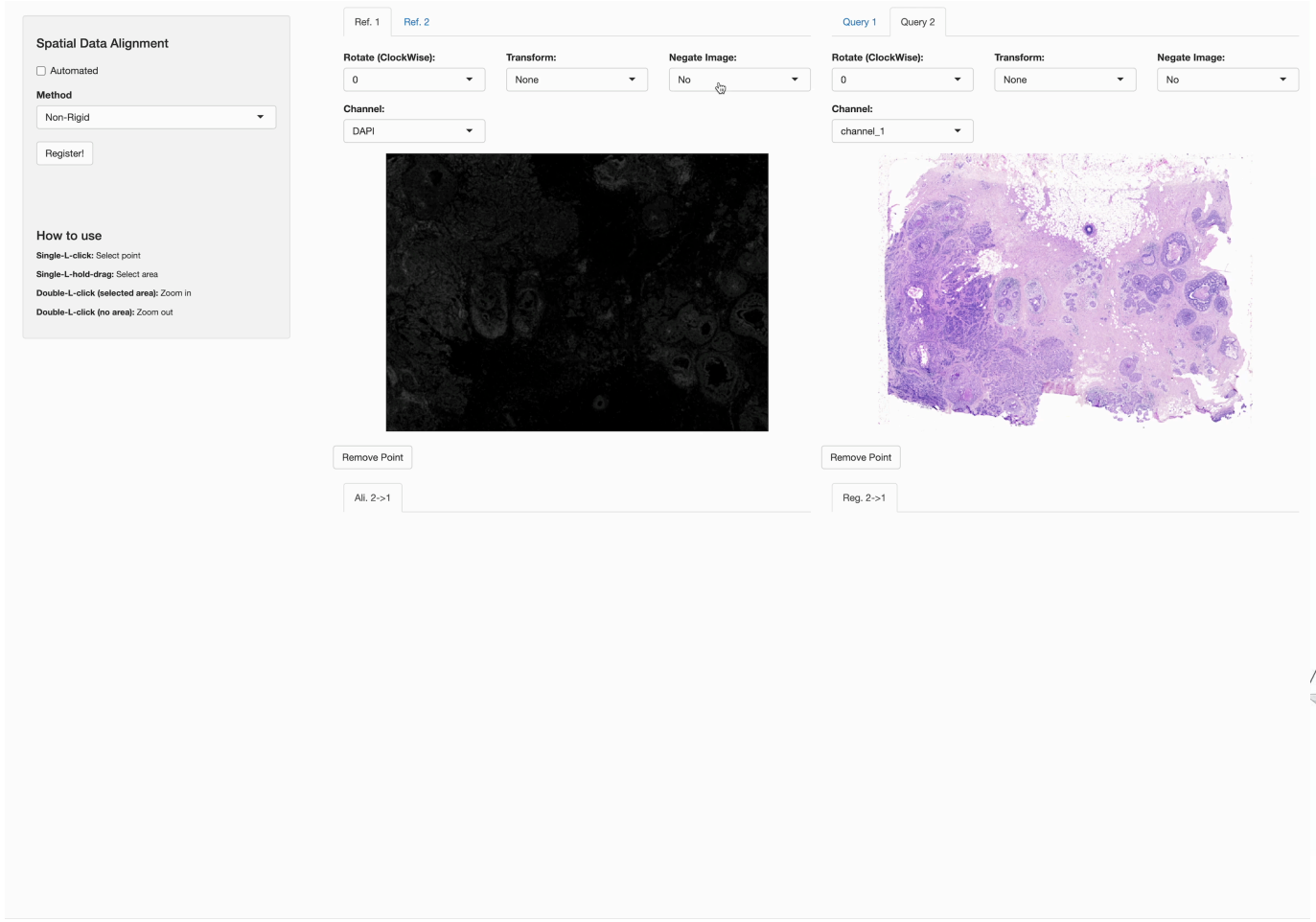
Zarr

HDF5 and Zarr

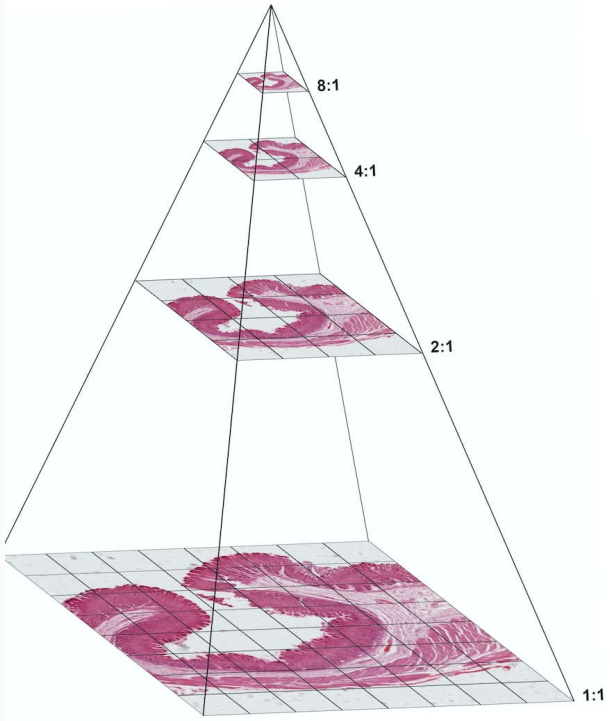


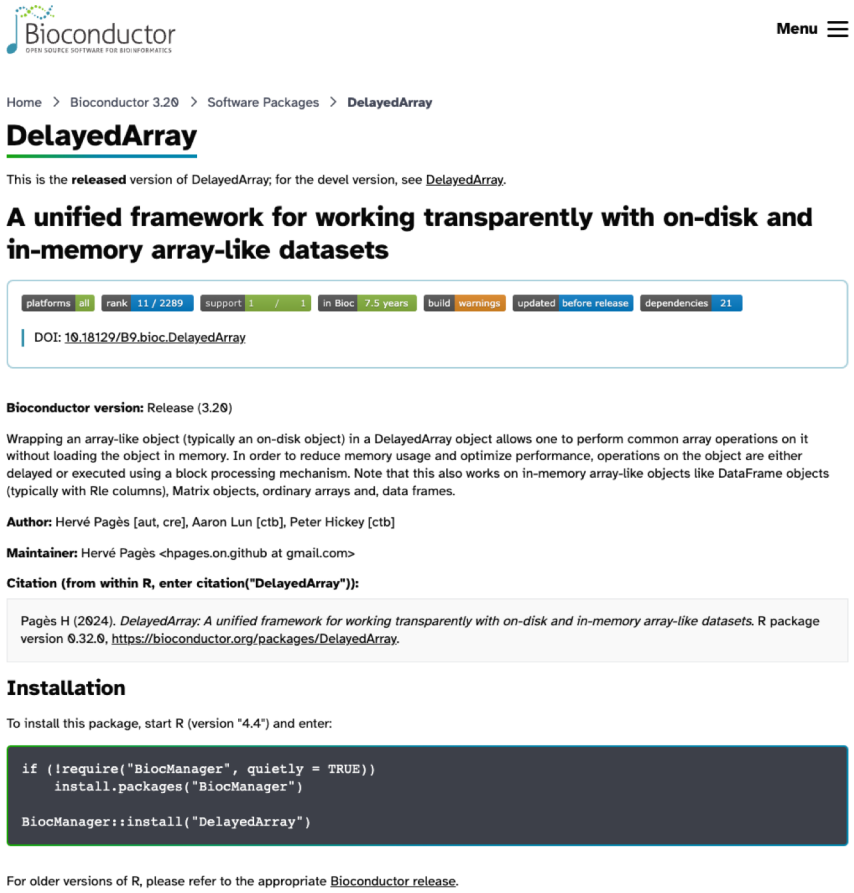






Multiscale Image Pyramids





Bioconductor

Menu

Home > Bioconductor 3.20 > Software Packages > DelayedArray

DelayedArray

This is the **released** version of DelayedArray; for the devel version, see DelayedArray.

A unified framework for working transparently with on-disk and in-memory array-like datasets

platforms all | rank 11 / 2289 | support 1 / 1 | in Bioc 7.5 years | build warnings | updated before release | dependencies 21

DOI: 10.18129/B9.bioc.DelayedArray

Bioconductor version: Release (3.20)

Wrapping an array-like object (typically an on-disk object) in a DelayedArray object allows one to perform common array operations on it without loading the object in memory. In order to reduce memory usage and optimize performance, operations on the object are either delayed or executed using a block processing mechanism. Note that this also works on in-memory array-like objects like DataFrame objects (typically with Rle columns), Matrix objects, ordinary arrays and, data frames.

Author: Hervé Pagès [aut, cre], Aaron Lun [ctb], Peter Hickey [ctb]

Maintainer: Hervé Pagès <hpages.on.github at gmail.com>

Citation (from within R, enter citation("DelayedArray")):

Pagès H (2024). *DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets*. R package version 0.32.0, https://bioconductor.org/packages/DelayedArray.

Installation

To install this package, start R (version "4.4") and enter:

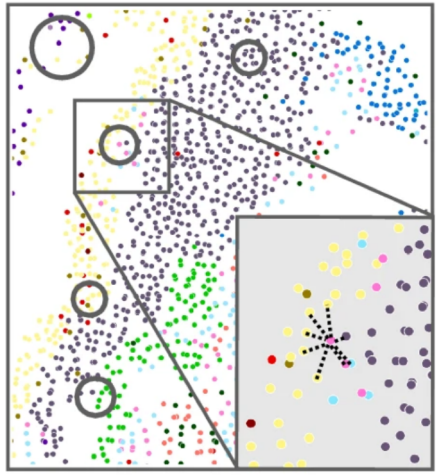
```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DelayedArray")
```

For older versions of R, please refer to the appropriate Bioconductor release.

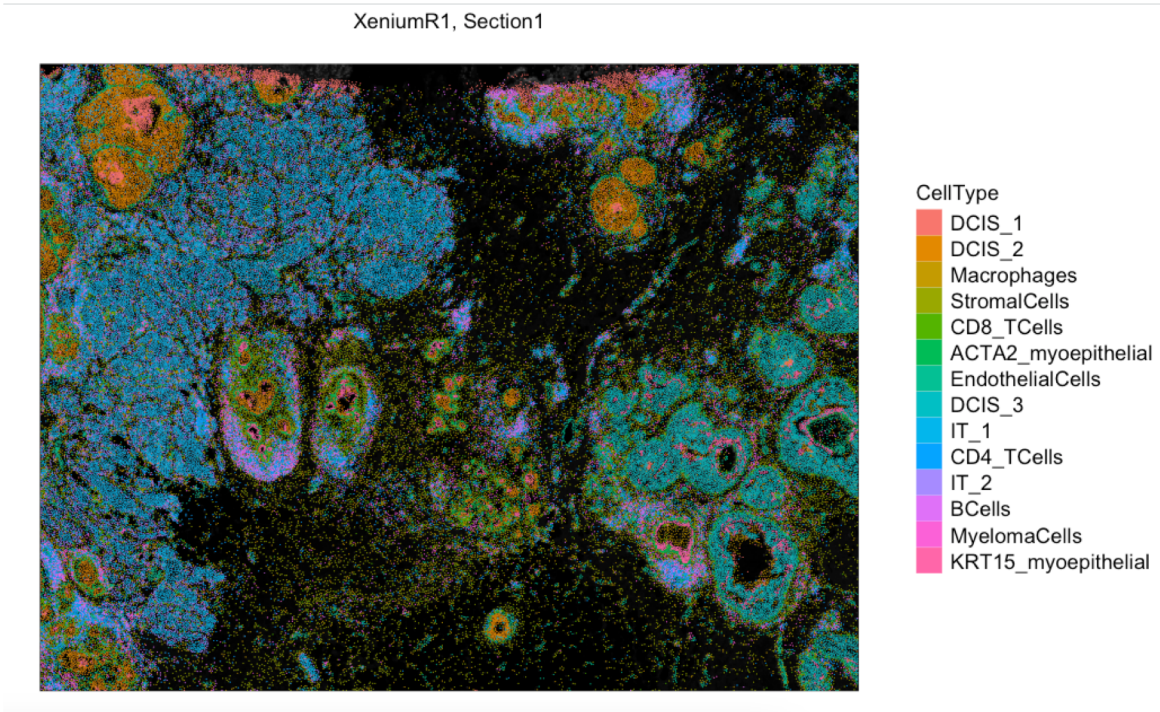
- In Memory (regular array):
  1. Load data (from disk)
  2. apply operation
  3. Show results

- On Disk (DelayedArray or BPCells):
  1. Find where data is (on disk)
  2. Save operation
  3. Load only needed data
  4. apply operation
  5. Show results

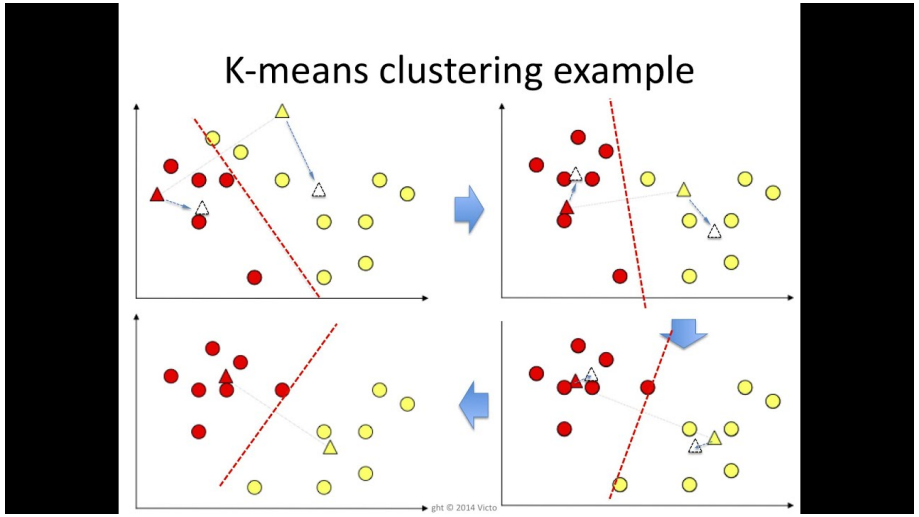
Niche Clustering

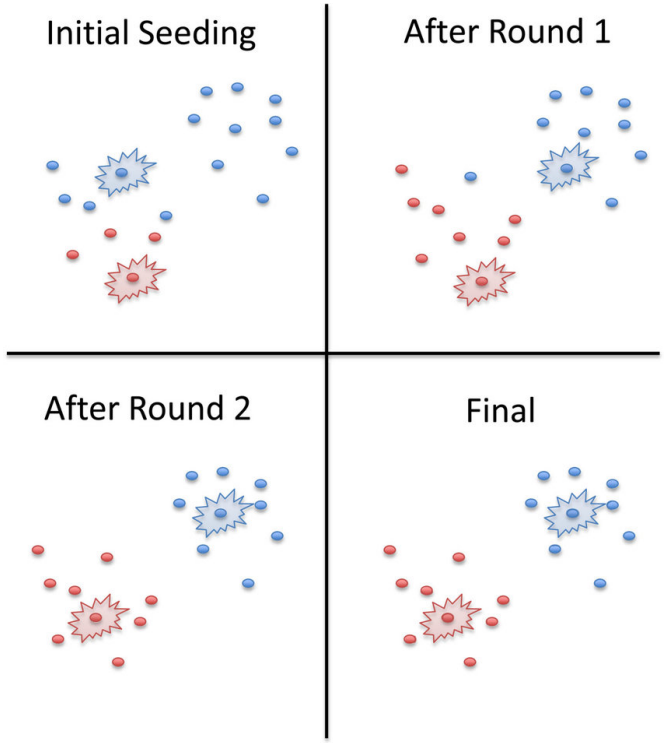


Define cellular niches
(*k* spatial nearest neighbors)

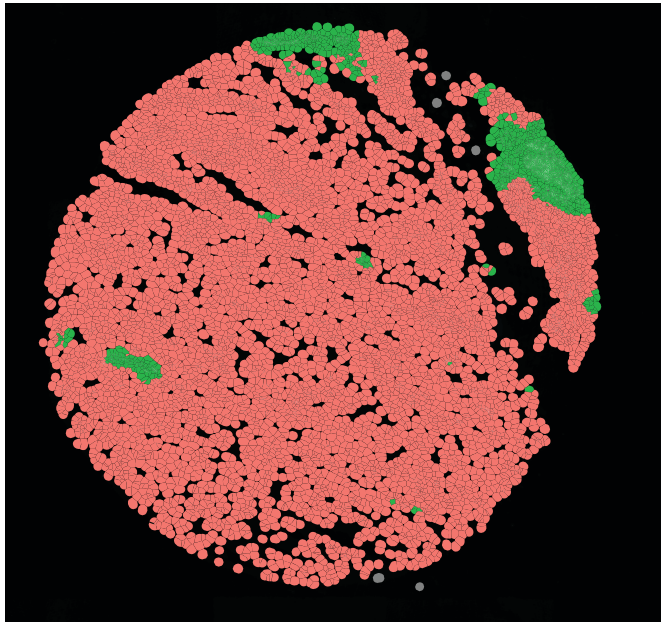


KMEANS CLUSTERING





Hot Spot Analysis



Hot Spot Analysis

Getis–Ord statistics

Local statistics [edit]

There are two different versions of the statistic, depending on whether the data point at the target location i is included or not[6]

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j}$$

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

Here x_i is the value observed at the i^{th} spatial site and w_{ij} is the spatial weight matrix which constrains which sites are connected to one another. For G_i^* the denominator is constant across all observations.