SpatialExperiment: infrastructure for spatially resolved transcriptomics data in R using Bioconductor

Dario Righelli[1]*, Lukas M. Weber[2]*, Helena L. Crowell[3,4]*, Brenda Pardo[5,6], Leonardo Collado-Torres[6], Shila Ghazanfar[7], Aaron T. L. Lun[8], Stephanie C. Hicks[2]†, Davide Risso[1]†

[1] Department of Statistical Sciences, University of Padova, Padova, Italy
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[3] Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[4] SIB Swiss Institute of Bioinformatics, Zurich, Switzerland
[5] Escuela Nacional de Estudios Superiores Unidad Juriquilla, Universidad Nacional Autónoma de México, Queretaro, Mexico
[6] Lieber Institute for Brain Development, Baltimore, MD, USA
[7] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom
[8] Genentech, South San Francisco, CA, USA

* Equal contributions (first authors)
† Equal contributions (senior authors)

16 August 2021

Abstract

Summary: *SpatialExperiment* is a new data infrastructure for storing and accessing spatially resolved transcriptomics data, implemented within the R/Bioconductor framework, which provides advantages of modularity, interoperability, standardized operations, and comprehensive documentation. Here, we demonstrate the structure and user interface with examples from the 10x Genomics Visium and seqFISH platforms, and provide access to example datasets and visualization tools in the *STexampleData*, *TENxVisiumData*, and *ggspavis* packages.
Availability and Implementation: The *SpatialExperiment*, *STexampleData*, and *TENxVisiumData* packages are available from Bioconductor. The package versions described in this manuscript are available in Bioconductor version 3.14 onwards. The *ggspavis* package is available from GitHub and has been submitted to Bioconductor.
Contact: risso.davide@gmail.com, shicks19@jhu.edu
Supplementary Information: Supplementary Tables and Figures are available online.

Introduction

Spatially resolved transcriptomics (ST) refers to a new set of high-throughput technologies, which measure up to transcriptome-wide gene expression along with the spatial coordinates of the measurements. Technological platforms differ in terms of the number of measured genes (from hundreds to full transcriptome) and spatial resolution (from multiple cells per coordinate to approximately single-cell to sub-cellular). Examples of ST platforms include Spatial Transcriptomics [1], 10x Genomics Visium [2], Slide-seq [3], Slide-seqV2 [4], sci-Space [5], seqFISH [6,7], seqFISH+ [8], and MERFISH [9–11]. These can be classified into spot-based and molecule-based platforms. Spot-based platforms measure transcriptome-wide gene expression at a series of spatial coordinates (spots) on a tissue slide (Spatial Transcriptomics, 10x Genomics Visium, Slide-seq, Slide-seqV2, and sci-Space), while molecule-based platforms detect up to thousands of distinct individual messenger RNA (mRNA) molecules *in situ* at up to sub-cellular resolution (seqFISH, seqFISH+, and MERFISH). ST platforms have been applied to investigate spatial patterns of gene expression in a variety of biological systems, including the human brain [12], mouse brain [13], cancer [14,15], and mouse embryogenesis [5,16]. By combining molecular and spatial information, these platforms promise to continue to generate new insights about biological processes that manifest with spatial specificity within tissues.

However, to effectively analyze these data, specialized and robust data infrastructures are required, to facilitate storage, retrieval, subsetting, and interfacing with downstream tools. Here, we describe *SpatialExperiment*, a new data infrastructure developed within the R/Bioconductor framework, which extends the popular *SingleCellExperiment* [17] class for single-cell RNA sequencing data to the spatial context, with observations taking place at the level of spots or molecules instead of cells. While several recent studies have reused or extended existing single-cell infrastructure to store additional spatial information [12,16], there does not yet exist a common, standardized infrastructure for storing and accessing ST data in R. A well-designed data infrastructure will simplify the work of various users, including developers of downstream analysis methods who can reuse the structure to store inputs and outputs, and analysts who can rely on the structure to connect packages from different developers into analysis pipelines. By working within the Bioconductor framework, we take advantage of long-standing Bioconductor principles of modularity, interoperability, continuous testing, and comprehensive documentation [17,18]. Furthermore, we can ensure compatibility with existing analysis packages designed for the *SingleCellExperiment* structure for single-cell data, providing a robust, flexible, and user-friendly resource for the research community. In addition to the *SpatialExperiment* package, we provide the *STexampleData* and *TENxVisiumData* packages (example datasets) and *ggspavis* package (visualization tools), for use in examples, tutorials, demonstrations, and teaching.

Results

The *SpatialExperiment* package provides access to the core data infrastructure (referred to as a class), as well as functions to create, modify, and access instances of the class (objects). Objects contain the following components adapted from the existing *SingleCellExperiment* class: (i) `assays`, tables of measurement values such as raw and transformed transcript counts (note that within the Bioconductor framework, rows usually correspond to features, and columns to observations); (ii) `rowData`, additional information (metadata) describing the features (e.g. gene IDs and names); (iii) `colData`, metadata describing the observations (e.g. spatial barcode IDs or cell IDs); and (iv) `reducedDims`, reduced dimension representations (e.g. principal component analysis) of the measurements. In addition, *SpatialExperiment* objects contain the following components to store spatial information: (v) `spatialCoords`, spatial coordinates associated with each observation (e.g. x and y coordinates on the tissue slide); (vi) `spatialData`, metadata describing spatial characteristics of the spatial coordinates (spots) or cells (e.g. indicators for whether spots are located within the region overlapping with tissue); and (vii) `imgData`, image files (e.g. histology images) and information related to the images (e.g. resolution in pixels) (**Figure 1**).

Accessor and replacement functions allow each of these components to be extracted or modified. Since *SpatialExperiment* extends *SingleCellExperiment*, methods developed for single-cell analyses [17] (e.g. preprocessing and normalization methods from scater [19], downstream methods from scran [20], and visualization tools from iSEE [21]) can be applied to *SpatialExperiment* objects, treating spots as single cells. Spatial coordinates are stored in `spatialCoords` as a numeric matrix, allowing these to be provided to downstream spatial analysis packages in R outside Bioconductor (e.g. from geostatistics, such as sp [22] and sf [23]). For spot-based data, `assays` contains a table named `counts` containing the gene counts, while for molecule-based data, `assays` may contain two tables named `counts` and `molecules` containing total gene counts per cell as well as molecule-level information such as spatial coordinates per molecule (formatted as a `BumpyMatrix` [24]). For datasets that are too large to store in-memory, *SpatialExperiment* can reuse existing Bioconductor infrastructure for sparse matrices and on-disk data representations through the *DelayedArray* framework [25]. *SpatialExperiment* objects can be created with a general constructor function, `SpatialExperiment()`, or alternatively with a dedicated constructor function for the 10x Genomics Visium platform, `read10xVisium()`, which creates an object from the raw input files from the 10x Genomics Visium Space Ranger software [26]. Measurements from multiple biological samples can be stored within a single object, and linked across the components by providing unique sample IDs. Image files can be stored in-memory, as local files, or hosted remotely. In addition, we provide the associated packages *STexampleData* and *TENxVisiumData* containing example datasets formatted as *SpatialExperiment* objects, and the *ggspavis* package providing visualization functions designed for *SpatialExperiment* objects (**Supplementary Figure 1** and **Supplementary Table 1**).

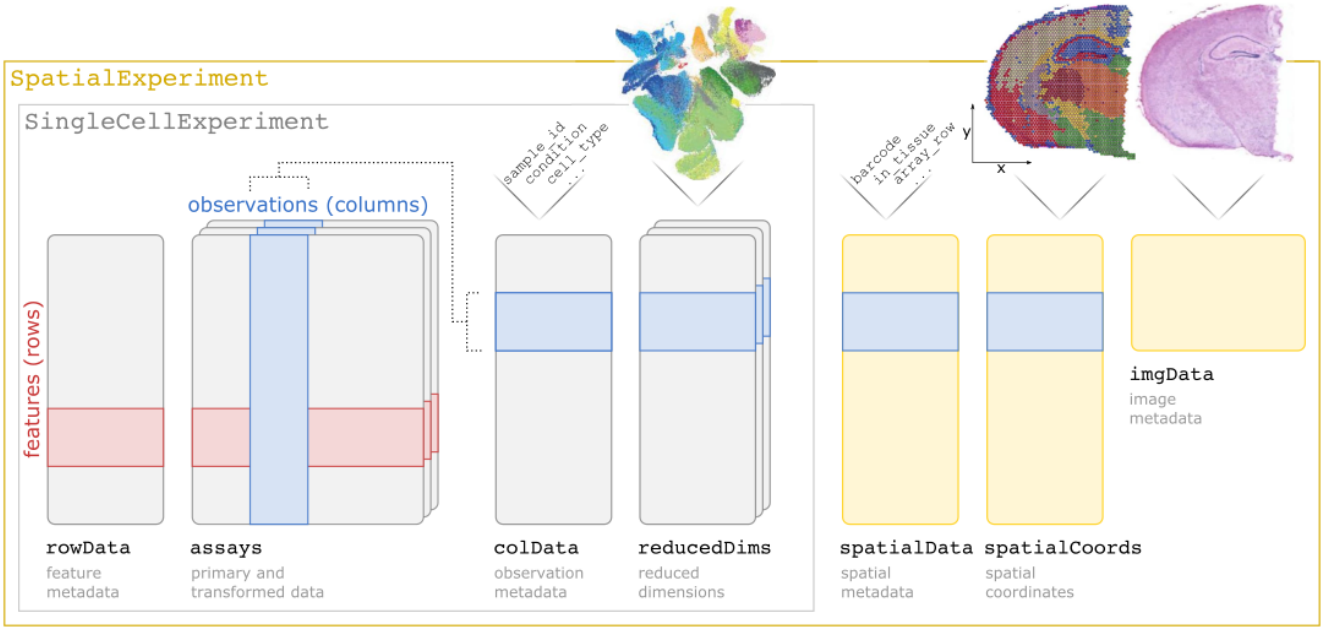


Figure 1. Overview of the *SpatialExperiment* class structure, including `assays` (tables of measurement values), `rowData` (metadata describing features), `colData` (metadata describing observations), `reducedDims` (reduced dimension representations), `spatialCoords` (spatial coordinates associated with the observations), `spatialData` (metadata describing spatial characteristics of the observations), and `imgData` (image files and information).

Discussion

Standardized data infrastructure for single-cell RNA sequencing data (e.g. *SingleCellExperiment* [17] and *Seurat* [27,28] in R, and *AnnData* [29] in Python) has greatly streamlined the work of downstream method developers and data analysts. For example, relying on common formats for inputs and outputs from individual packages allows users to connect packages into complete analysis pipelines, and operations such as subsetting by row (gene) or column (barcode or cell) across the entire object helps avoid errors. For single-cell data, this has enabled the development of comprehensive workflows and tutorials [17,30], which are an invaluable resource for new users. Here, we provide a new data infrastructure for ST data, extending the existing *SingleCellExperiment* class within the Bioconductor framework. In addition, we provide associated packages containing example datasets (*STexampleData* and *TENxVisiumData*) and visualization functions (*ggspavis*), for use in examples, tutorials, demonstrations, and teaching. ST technologies are still in their infancy, and the coming years are likely to see ongoing development of existing platforms as well as the emergence of novel experimental approaches. *SpatialExperiment* is ideally positioned to be extended to accommodate data from new platforms in the future, e.g. through extensions of the more general underlying *SummarizedExperiment* [31] or by integrating with *MultiAssayExperiment* [32] to store measurements from further assay types (e.g. transcriptomics, proteomics or spatial immunofluorescence, or epigenomics) or multiple assays from the same spatial coordinates. Similarly, three-dimensional spatial data [33] or data from multiple

timepoints could be accommodated within *SpatialExperiment* by storing additional spatial or temporal coordinates, and datasets that are too large to store in-memory can be stored using existing Bioconductor infrastructure for sparse matrices and on-disk data representations through the *DelayedArray* framework [25]. The ability to store image files within the objects (in-memory, locally, or remotely) will assist with correctly keeping track of images in datasets with large numbers of samples, e.g. from consortium efforts. Interoperability between *SpatialExperiment* and other data formats (e.g. *AnnData* [29] and *Loompy* [34] in Python) can also be ensured through the use of existing conversion packages [34,35]. *SpatialExperiment* provides the research community with a robust, flexible, and extendable core data infrastructure for ST data, assisting both method developers and analysts to generate reliable and reproducible biological insights from these platforms.

Acknowledgments

We thank the participants of the EuroBioc2020 “Birds of a Feather” session (14 December 2020) and workshop (16 December 2020) on the topic of infrastructure for ST data in Bioconductor, as well as the members of the *spatial* and *SpatialExperiment* channels of the Bioconductor community Slack workspace, for helpful feedback and suggestions.

Author contributions

D. Righelli, LMW, and HLC designed the *SpatialExperiment* class structure, with input from all other authors. D. Righelli led the implementation of the *SpatialExperiment* class, with significant code input from HLC. LMW developed the example data package *STexampleData* and the visualization package *ggspavis*. HLC developed the data package *TENxVisiumData* and provided functions for the *ggspavis* package. BP and LCT tested an earlier version of the *SpatialExperiment* class and provided input on design choices for the final class structure. SG provided input and examples for applying the *SpatialExperiment* class to molecule-based ST data. ATLL provided input on design choices for the *SpatialExperiment* class structure. SCH and D. Risso provided supervision and input on design choices for the *SpatialExperiment* class structure. LMW drafted the paper with input from all other authors. All authors approved the final version of the manuscript.

Code and data availability

The *SpatialExperiment* package is available from Bioconductor at https://bioconductor.org/packages/SpatialExperiment. The *STexampleData* and *TENxVisiumData* packages are available from Bioconductor at https://bioconductor.org/packages/STexampleData and https://bioconductor.org/packages/TENxVisiumData respectively. The *ggspavis* package is available from GitHub at https://github.com/lmweber/ggspavis and has been submitted to Bioconductor. The package versions described in this manuscript are available in Bioconductor version 3.14 onwards. Datasets from Supplementary Tables 1 and 2 and Supplementary Figure 1 are available as *SpatialExperiment* objects from the *STexampleData* and *TENxVisiumData* packages, and the full original datasets are available from the sources listed in Supplementary Tables 1 and 2 [12,16,36,37].

Conflicts of interest

The authors declare that they have no financial conflicts of interest.

Funding

This work was supported by CZF2019-002443 (LMW, D. Righelli, SCH, D. Risso) from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. LMW, SCH and LC-T were supported by NIH/NIMH U01MH122849 to SCH and LC-T. D. Risso was supported by “Programma per Giovani Ricercatori Rita Levi Montalcini” granted by the Italian Ministry of Education, University, and Research and by the National Cancer Institute of the National Institutes of Health (2U24CA180996). SG was supported by a Royal Society Newton International Fellowship (NIF\R1\181950).

References

1. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353: 78–82.

2. 10x Genomics. 10x Genomics Visium Spatial Gene Expression Solution. (Website). 2021.

3. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363: 1463–1467.

4. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature Biotechnology. 2020. Available: https://www.ncbi.nlm.nih.gov/pubmed/33288904

5. Srivatsan SR, Regier MC, Barkan E, Franks JM, Packer JS, Grosjean P, et al. Embryo-scale, single-cell spatial transcriptomics. Science. 2021;373: 111–117.

6. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nature Methods. 2014;11: 360–361.

7. Shah S, Lubeck E, Zhou W, Cai L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron. 2016;92: 342–357.

8. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568: 235–239.

9. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348: aaa6090.

10. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences of the USA. 2016;113: 11046–11051.

11. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences of the USA. 2019;116: 19490–19499.

12. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience (in press). 2021.

13. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, Meletis K. Molecular atlas of the adult mouse brain. Science Advances. 2020;6: eabb3446.

14. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020;182: 1661–1662.

15. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nature Communications. 2018;9: 2419.

16. Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, et al. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. bioRxiv (preprint). 2020.

17. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating Single-Cell Analysis with Bioconductor. Nature Methods. 2019;17: 137–145.

18. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015;12: 115–121.

19. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and

visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33: 1179–1186.

20. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5: 2122.

21. Rue-Albrecht K, Marini F, Soneson C, Lun ATL. iSEE: Interactive SummarizedExperiment Explorer. F1000Research. 2018;7: 741.

22. Pebesma EJ, Bivand RS. Classes and methods for spatial data in R. R News. 2005;5: 9–13.

23. Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. {The R Journal}. 2018;10: 439–446.

24. Lun A. BumpyMatrix; R package, version 0.99.6. http://bioconductor.org/packages/BumpyMatrix. 2021.

25. Pagès H, Hickey P, Lun A. DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets; version 0.16.1. R/Bioconductor package. 2021.

26. 10x Genomics. Space Ranger: Spatial Gene Expression. Website: https://support10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger. 2020.

27. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018;36: 411–420.

28. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell. 2019;177: 1–15.

29. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. Genome Biology. 2018;19: 15.

30. Lun A, Amezquita R, Hicks S, Gottardo R. Orchestrating Single-Cell Analysis with Bioconductor. (Online Book). 2021.

31. Morgan M, Obenchain V, Hester J, Pagès H. SummarizedExperiment: SummarizedExperiment container; R package version 1.22.0. R/Bioconductor package. 2021.

32. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Cabrera CR, et al. Software for the integration of multi-omics experiments in Bioconductor. Cancer Research. 2017;77: e39–42.

33. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361: 6400.

34. Morgan M, Van Twisk D. LoomExperiment; R package, version 1.8.0. http://bioconductor.org/packages/LoomExperiment/. 2021.

35. Zappia L, Lun A. zellkonverter; R package, version 1.0.0. http://bioconductor.org/packages/zellkonverter/. 2021.

36. Collado-Torres L, Maynard KR, Jaffe AE. spatialLIBD: LIBD Visium spatial transcriptomics human pilot data inspector; version 1.2.1. R/Bioconductor package. 2021.

37. 10x Genomics. Mouse Brain Section Coronal. (Website). 2021.

38. 10x Genomics. Spatial Gene Expression Datasets. Website: https://support10xgenomics.com/spatial-gene-expression/datasets. 2021.

Supplementary Tables

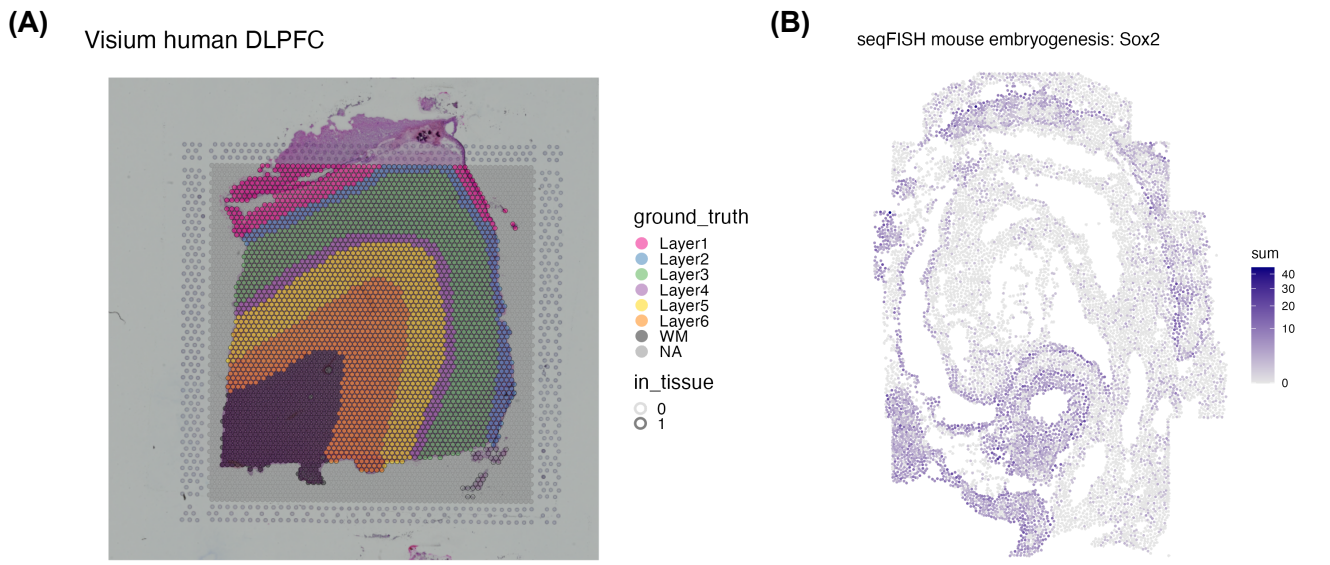
Dataset name	Platform	Type	Tissue	Number of samples	Number of spots or cells	Number of features (genes)	Contains ground truth labels?	Contains image data?	Source
Visium_humanDLPFC	10x Genomics Visium [2]	Spot-based	Human brain	1	3,639	33,538	Yes	Yes	[12,36]
Visium_mouseCoronal	10x Genomics Visium [2]	Spot-based	Mouse brain	1	2,702	32,285	Yes	Yes	[37]
seqFISH_mouseEmbryo	seqFISH [6,7]	Molecule-based	Mouse embryo	1	11,026	351	No	No	[16]

Supplementary Table 1. Summary of example datasets provided in *SpatialExperiment* format in the *STexampleData* package. Table columns describe characteristics for each dataset, and provide the original references. For the *Visium_humanDLPFC* and *seqFISH_mouseEmbryo* datasets, the objects in the *STexampleData* package contain small subsets of the full original datasets, allowing users to easily download and load these datasets for examples and tutorials. The full datasets can be obtained from the original references.

Dataset name	Tissue	Number of samples	Targeted panel(s)	Number of spots	Number of genes
HumanBreastCancerIDC	Human invasive ductal carcinoma breast	2	–	7,785	36,601
HumanBreastCancerILC	Human invasive lobular carcinoma breast	1	– *Immunology*	4,325	36,601 *1,056*
HumanCerebellum	Human cerebellum	1	– *Neuroscience*	4,992	36,601 *1,186*
HumanColorectalCancer	Human invasive adenocarcinoma of the large intestine	1	– *Gene signature*	3,138	36,601 *1,142*
HumanGlioblastoma	Human glioblastoma multiforme	1	– *Pan-cancer*	3,468	36,601 *1,253*
HumanHeart	Human heart	1	–	4,247	36,601
HumanLymphNode	Human lymph node	1	–	4,035	36,601
HumanOvarianCancer	Human ovarian endometrial adenocarcinoma	1	– *Immunology* *Pan-cancer*	3,493	36,601 *1,056* *1,253*
HumanSpinalCord	Human spinal cord	1	– *Neuroscience*	2,812	36,601 *1,186*
MouseBrainCoronal	Mouse brain (coronal plane)	1	–	2,702	32,285
MouseBrainSagittalAnterior	Mouse brain (sagittal slice of the posterior)	2	–	5,520	32,285
MouseBrainSagittalPosterior	Mouse brain (sagittal slice of the anterior)	2	–	6,644	32,285
MouseKidneyCoronal	Mouse kidney	1	–	1,438	32,285

Supplementary Table 2. Summary of example datasets provided in *SpatialExperiment* format in the *TENxVisiumData* package. All data are spot-based, and were obtained using the 10x Genomics Visium platform [2]. Table columns describe characteristics for each dataset. For some datasets, targeted expression panels were measured in addition to whole-transcriptome analysis; these are indicated with the name of the panel and corresponding number of genes in italics. The original datasets can be obtained from [38].

Supplementary Figures



Supplementary Figure 1. (A) Example of visualization of spot-based ST data (*Visium_humanDLPFC* object from the *STexampleData* package). Image shows a histology image as background, grid of spatial coordinates (spots), highlighting for spots that overlap with tissue, and colors for ground truth cluster labels. The dataset represents a single biological sample (sample 151673) from the human brain dorsolateral prefrontal cortex (DLPFC) region [12,36], measured with the 10x Genomics Visium platform. The full dataset contains 12 biological samples, and is available in *SpatialExperiment* format in the *spatialLIBD* Bioconductor package [12,36]. **(B)** Example of visualization of molecule-based ST data (*seqFISH_mouseEmbryo* object from the *STexampleData* package). Color scale shows total mRNA counts per cell for the *Sox2* gene. The dataset represents a subset of cells (embryo 1, z-slice 2) from a published dataset investigating mouse embryogenesis [16], generated using the seqFISH platform. Additional details on the datasets are provided in Supplementary Table 1. Figures were generated using plotting functions from the *ggspavis* package.