Bioinformatics, 38(11), 2022, 3128–3131
https://doi.org/10.1093/bioinformatics/btac299
Advance Access Publication Date: 28 April 2022
Applications Note

Gene expression

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,*,†] and Davide Risso [1,*,†]

[1]Department of Statistical Sciences, University of Padova, 35121 Padova, Italy, [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, 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 76230, Mexico, [6]Lieber Institute for Brain Development, Baltimore, MD 21205, USA, [7]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom and [8]Genentech, South San Francisco, CA 94080, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors and last two authors should be regarded as Joint Authors.
Associate Editor: Valentina Boeva

Received on August 18, 2021; revised on February 2, 2022; editorial decision on April 19, 2022; accepted on April 25, 2022

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*, *TENxVisiumData* and *ggspavis* packages are available from Bioconductor. The package versions described in this manuscript are available in Bioconductor version 3.15 onwards.

Contact: davide.risso@unipd.it or shicks19@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Spatially-resolved transcriptomics (SRT) 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 SRT platforms include Spatial Transcriptomics (Ståhl *et al.*, 2016), 10x Genomics Visium (10x Genomics, 2021a), Slide-seq (Rodriques *et al.*, 2019), Slide-seqV2 (Stickels *et al.*, 2020), sci-Space (Srivatsan *et al.*, 2021), seqFISH (Lubeck *et al.*, 2014; Shah *et al.*, 2016), seqFISH+ (Eng *et al.*, 2019), osmFISH (Codeluppi *et al.*, 2018) and MERFISH (Chen *et al.*, 2015; Moffitt *et al.*, 2016; Xia *et al.*, 2019). 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 large sets of distinct individual messenger RNA (mRNA) molecules *in situ* at up to sub-cellular resolution (seqFISH, seqFISH+, osmFISH and MERFISH). SRT platforms have been applied to investigate spatial patterns of gene expression in a variety of biological systems, including the human brain (Maynard *et al.*, 2021), mouse brain (Ortiz *et al.*, 2020), cancer (Berglund *et al.*, 2018; Ji *et al.*, 2020) and mouse embryogenesis (Lohoff *et al.*, 2021; Srivatsan *et al.*, 2021). 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* (Amezquita *et al.*, 2020) class for single-cell RNA sequencing (scRNA-seq) data to the spatial context, with observations taking place at the level of spots or molecules instead

of cells. Several recent studies have reused or extended existing single-cell infrastructure to store additional spatial information (Lohoff *et al.*, 2021; Maynard *et al.*, 2021). In addition, several comprehensive analysis workflows have been developed using modified single-cell infrastructure adapted for spatial data, including *Seurat* (Hao *et al.*, 2021), *Giotto* (Dries *et al.*, 2021) and *Squidpy* (Palla *et al.*, 2022). However, while each of these workflows enables powerful analyses, it remains difficult for users to combine elements in a modular way, since each workflow relies on a separate infrastructure. There does not yet exist a common, standardized infrastructure for storing and accessing SRT data in R, which would allow users to easily build workflows combining methods and software developed by different groups. A well-designed independent data infrastructure simplifies 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 (Amezquita *et al.*, 2020; Huber *et al.*, 2015). 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.

2 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. barcode IDs or cell IDs); and (iv) `reducedDims`, reduced dimension representations (e.g. principal component analysis) of the measurements. *SpatialExperiment* objects also contain the following further components to store spatial information: (v) additional metadata stored in `colData` describing spatial characteristics of the spatial coordinates (spots) or cells (e.g. indicators for whether spots are located within the region overlapping with tissue); (vi) `spatialCoords`, spatial coordinates associated with each observation (e.g. x and y coordinates on the tissue slide); and, (vii) `imgData`, image files (e.g. histology images) and information related to the images (e.g. resolution in pixels) (Fig. 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 (Amezquita *et al.*, 2020) [e.g. preprocessing and normalization methods from *scater* (McCarthy *et al.*, 2017), downstream methods from *scran* (Lun *et al.*, 2016) and visualization tools from *iSEE* (Rue-Albrecht *et al.*, 2018)] 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 easily to downstream spatial analysis packages in R outside Bioconductor [e.g. packages from geostatistics such as *sp* (Pebesma & Bivand, 2005) and *sf* (Pebesma, 2018)], consistent with `reducedDims` in *SingleCellExperiment*. 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* (Lun, 2021)]. 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 (Pagès *et al.*, 2021). Image information is stored in `imgData` as a table containing sample IDs, image IDs, any other information such as scaling factors, and the underlying image data. The image data can be stored as either a fully realized in-memory object (for small images), a path to a local file that is loaded into memory on demand (for large images) or a URL to a remotely hosted image that is retrieved on demand. *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 (10x Genomics, 2020). For Visium data, `colData` includes the columns `in_tissue`, `array_row` and `array_col`. Measurements from multiple biological samples can be stored within a single object, and linked across the components by providing unique sample IDs. In addition, we provide the associated data packages *STexampleData* (example datasets from several platforms) and *TENxVisiumData* (publicly available Visium datasets provided by 10x Genomics), and the *ggspavis* package providing visualization functions designed for *SpatialExperiment* objects (Supplementary Fig. S1 and Supplementary Tables S1 and S2).

3 Discussion

Standardized data infrastructure for scRNA-seq data [e.g. *SingleCellExperiment* (Amezquita *et al.*, 2020) within the R/Bioconductor framework] has greatly streamlined the work of data analysts and downstream method developers. 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 (Amezquita *et al.*, 2020; Lun *et al.*, 2021), which are an invaluable resource for new users. Here, we provide a new data infrastructure for SRT 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.

Existing alternative infrastructure for SRT data includes object classes provided in the *Seurat* (Hao *et al.*, 2021) and *Giotto* (Dries *et al.*, 2021) packages in R, and *Squidpy*/*AnnData* (Palla *et al.*, 2022; Virshup *et al.*, 2021) in Python, which provide similar underlying functionality such as storing annotated tables of measurement values and related spatial and image information. Compared to these alternatives, a key advantage of *SpatialExperiment* is that it has been developed independently of any individual analysis workflow and is compatible with any downstream analysis packages that use the *SpatialExperiment* or *SingleCellExperiment* class within Bioconductor. This allows analysts to easily build customized, modular workflows consisting of packages developed by various research groups, including the latest methods (which may not yet have been integrated into published workflows) as well as any of the wide variety of methods for single-cell data that have been released through Bioconductor.

SRT 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* (Morgan *et al.*, 2021) or by integrating with *MultiAssayExperiment* (Ramos *et al.*, 2017) to store measurements from further assay types (transcriptomics, proteomics or spatial immunofluorescence, or epigenomics) or multiple assays from the same spatial coordinates. For example, the *SingleCellMultiModal* package (Eckenrode *et al.*, 2021) stores *MultiAssayExperiment* objects containing scRNA-seq and SRT

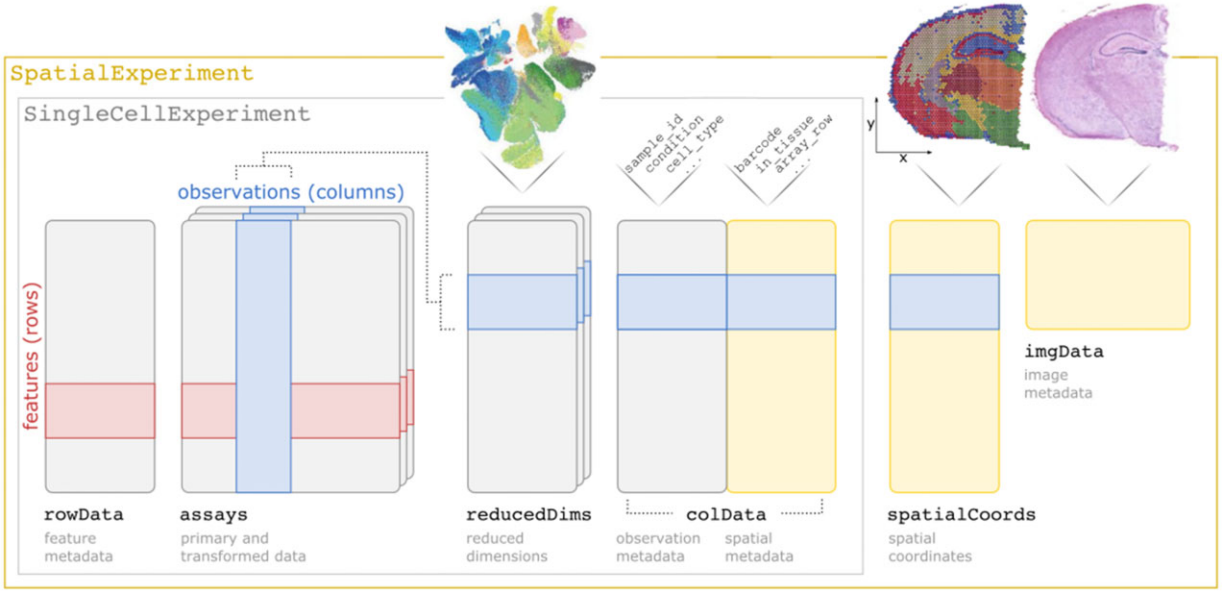

Fig. 1. Overview of the *SpatialExperiment* class structure, including `assays` (tables of measurement values), `rowData` (metadata describing features), `reducedDims` (reduced dimension representations), `colData` (non-spatial and spatial metadata describing observations), `spatialCoords` (spatial coordinates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

data as *SingleCellExperiment* and *SpatialExperiment* objects, respectively. Three-dimensional spatial data (Wang *et al.*, 2018) or data from multiple timepoints could be accommodated within *SpatialExperiment* by storing additional spatial or temporal coordinates. 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 (Pagès *et al.*, 2021). 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. in Python) can also be ensured through the use of existing conversion packages such as *zellkonverter* (Zappia & Lun, 2021) and *LoomExperiment* (Morgan & Van Twisk, 2021). *SpatialExperiment* provides the research community with a robust, flexible and extendable core data infrastructure for SRT data, assisting both method developers and analysts to generate reliable and reproducible biological insights from these platforms.

Acknowledgements

The authors 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 SRT 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.Rig., L.M.W. and H.L.C. designed the *SpatialExperiment* class structure, with input from all other authors. D.Rig. led the implementation of the *SpatialExperiment* class, with significant code input from H.L.C. L.M.W. developed the example data package *STexampleData* and the visualization package *ggspavis*. H.L.C. developed the data package *TENxVisiumData* and provided functions for the *ggspavis* package. B.P. and L.C.-T. tested an earlier version of the *SpatialExperiment* class and provided input on design choices for the final class structure. S.G. provided input and examples for applying the *SpatialExperiment* class to molecule-based SRT data. A.T.L.L. provided input on design choices for the *SpatialExperiment* class structure. S.C.H. and D.Ris. provided supervision and input on design choices for the *SpatialExperiment* class structure. L.M.W. drafted the article with input from all other authors. All authors approved the final version of the manuscript.

Funding

This work was supported by CZF2019-002443 [to L.M.W., D.Rig., S.C.H., D.Ris.] from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. L.M.W., S.C.H. and L.C.-T. were supported by National Institutes of Health/NIMH U01MH122849 to S.C.H. and L.C.-T. D.Ris. 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]. S.G. was supported by a Royal Society Newton International Fellowship [NIF\R1\181950].

Conflict of Interest: none declared.

Data availability

The *SpatialExperiment* package is available from Bioconductor at https://bioconductor.org/packages/SpatialExperiment. The *STexampleData*, *TENxVisiumData* and *ggspavis* packages are available from Bioconductor at https://bioconductor.org/packages/STexampleData, https://bioconductor.org/packages/TENxVisiumData and https://bioconductor.org/packages/ggspavis, respectively. The package versions described in this manuscript are available in Bioconductor version 3.15 onwards. Datasets from Supplementary Tables S1 and S2 and Supplementary Figure S1 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 S1 and S2 (10x Genomics, 2021b; Lohoff *et al.*, 2021; Maynard *et al.*, 2021; Pardo *et al.*, 2022).

References

10x Genomics. (2020) *Space Ranger: Spatial Gene Expression*. https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger (2 May 2022, date last accessed).

10x Genomics. (2021a) *10x Genomics Visium Spatial Gene Expression Solution* (Website). https://www.10xgenomics.com/products/spatial-gene-expression (2 May 2022, date last accessed).

10x Genomics. (2021b) *Mouse Brain Section Coronal* (Website). https://www.10xgenomics.com/resources/datasets/mouse-brain-section-coronal-1-standard-1-1-0 (2 May 2022, date last accessed).

10x Genomics. (2021c) *Spatial Gene Expression Datasets*. https://support.10xgenomics.com/spatial-Gene-Expression/datasets (2 May 2022, date last accessed).

Amezquita,R.A. *et al.* (2020) Orchestrating single-cell analysis with Bioconductor. *Nat. Methods*, **17**, 137–145.

Berglund,E. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.

Cable,D.M. *et al.* (2021) Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.*, **1**, 1.

Chen,K.H. *et al.* (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.

Codeluppi,S. *et al.* (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, **15**, 932–935.

Dries,R. *et al.* (2021) Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.*, **22**, 78.

Eckenrode,K.B. *et al.* (2021) Curated Single Cell Multimodal Landmark Datasets for R/Bioconductor. bioRxiv (preprint).

Eng,C.-H.L. *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, **568**, 235–239.

Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587. e29.

Huber,W. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.

Ji,A.L. *et al.* (2020) Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, **182**, 1661–1662.

Lohoff,T. *et al.* (2021) Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.*, **1**, 1.

Lubeck,E. *et al.* (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, **11**, 360–361.

Lun,A.T.L. (2021) *BumpyMatrix; Version 1.2.0. R/Bioconductor Package*. https://doi.org/10.18129/B9.bioc.BumpyMatrix.

Lun,A.T.L. *et al.* (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, **5**, 2122.

Lun,A.T.L. *et al.* (2021) *Orchestrating Single-Cell Analysis with Bioconductor* (Online Book). https://bioconductor.org/books/release/OSCA/

Maynard,K.R. *et al.* (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.*, **24**, 425–436.

McCarthy,D.J. *et al.* (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.

Moffitt,J.R. *et al.* (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. USA*, **113**, 11046–11051.

Morgan,M. and Van Twisk,D. (2021) *LoomExperiment; Version 1.12.0. R/Bioconductor Package*. https://doi.org/10.18129/B9.bioc.LoomExperiment.

Morgan,M. *et al.* (2021) *SummarizedExperiment: SummarizedExperiment Container; R Package Version 1.24.0. R/Bioconductor Package*. https://doi.org/10.18129/B9.bioc.SummarizedExperiment.

Ortiz,C. *et al.* (2020) Molecular atlas of the adult mouse brain. *Sci. Adv.*, **6**, eabb3446.

Pagès,H. *et al.* (2021) *DelayedArray: A Unified Framework for Working Transparently with On-disk and In-memory Array-like Datasets; Version 0.20.0. R/Bioconductor Package*. https://doi.org/10.18129/B9.bioc.DelayedArray.

Palla,G. *et al.* (2022) Squidpy: a scalable framework for spatial single cell analysis. *Nat. Meth*. https://www.nature.com/articles/s41592-021-01358-2.

Pardo,B. *et al.* (2022) spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. bioRxiv (in press).

Pebesma,E. (2018) Simple features for R: standardized support for spatial vector data. *R J.*, **10**, 439–446.

Pebesma,E.J. and Bivand,R.S. (2005) Classes and methods for spatial data in R. *R News*, **5**, 9–13.

Ramos,M. *et al.* (2017) Software for the integration of multi-omics experiments in Bioconductor. *Cancer Res.*, **77**, e39–e42.

Rodriques,S.G. *et al.* (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.

Rue-Albrecht,K. *et al.* (2018) iSEE: interactive SummarizedExperiment Explorer. *F1000Research*, **7**, 741.

Shah,S. *et al.* (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, **92**, 342–357.

Srivatsan,S.R. *et al.* (2021) Embryo-scale, single-cell spatial transcriptomics. *Science*, **373**, 111–117.

Ståhl,P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.

Stickels,R.R. *et al.* (2021) Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqV2. *Nat. Biotechnol.*, **39**, 313–319.

Virshup,I. *et al.* (2021) anndata: Annotated data. bioRxiv (preprint).

Wang,X. *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, **361**, 6400.

Xia,C. *et al.* (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. USA*, **116**, 19490–19499.

Zappia,L. and Lun,A. (2021) *zellkonverter; Version 1.4.0. R/Bioconductor Package*. https://doi.org/10.18129/B9.bioc.zellkonverter.