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] Institute 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)

27 January 2021

Abstract

Motivation: Spatially resolved transcriptomics is a new set of technologies to measure gene expression for up to thousands of genes at near-single-cell, single-cell, or sub-cellular resolution, together with the spatial positions of the measurements. Analyzing combined molecular and spatial information has generated new insights about biological processes that manifest in a spatial manner within tissues. However, to efficiently analyze these data, specialized data infrastructure is required, which facilitates storage, retrieval, subsetting, and interfacing with downstream tools.
Results: Here, we describe *SpatialExperiment*, a new data infrastructure for storing and accessing spatially resolved transcriptomics data, implemented within the Bioconductor framework in the R programming language. *SpatialExperiment* extends the existing *SingleCellExperiment* for single-cell data from the Bioconductor framework, which brings with it advantages of modularity, interoperability, standardized operations, and comprehensive documentation. We demonstrate the structure and user interface with examples from the 10x Genomics Visium and seqFISH platforms. *SpatialExperiment* is extendable to alternative technological platforms measuring expression and to new types of data modalities, such as spatial immunofluorescence or proteomics, in the future. We also provide access to example datasets and visualization tools in the *STexampleData*, *TENxVisiumData*, and *ggspavis* packages.
Availability and Implementation: *SpatialExperiment* is freely available from Bioconductor at https://bioconductor.org/packages/SpatialExperiment. The *STexampleData*, *TENxVisiumData*, and *ggspavis* packages are available from GitHub and will be submitted to Bioconductor.

Introduction

Recent advances in high-throughput technologies have now led to simultaneously measuring transcriptome-wide gene expression (or near-transcriptome-wide up to thousands of genes) at near-single-cell, single-cell, or sub-cellular resolution, along with the spatial coordinates of each measurement. These technologies are referred to as spatially resolved transcriptomics (ST). Examples of ST platform technologies include 10x Genomics Visium [1] and its earlier iteration Spatial Transcriptomics [2], Slide-seq [3] and Slide-seqV2 [4], seqFISH [5,6] and seqFISH+ [7], MERFISH [8–10], CARTANA [11], and others. These platforms can be divided into two major groups depending on the resolution -- spot-based and molecule-based. Current spot-based technologies measure transcriptome-wide expression at a series of spatial coordinates (spots) on a two-dimensional tissue slide (Visium, Spatial Transcriptomics, Slide-seqV2, Slide-seq), and potentially in three dimensions and across time in the future. Molecule-based technologies detect up to thousands of distinct individual messenger RNA (mRNA) molecules at single-cell or sub-cellular resolution (seqFISH, seqFISH+, MERFISH). These ST technologies have been successfully applied to investigate spatial patterns of gene expression in applications including the human brain [12], mouse brain [13], cancer [14,15], and mouse embryogenesis [16].

To efficiently analyze data generated with these technologies, we require robust data infrastructure, which facilitates reliable storage, retrieval, operations such as subsetting, and interfacing with downstream analysis and visualization tools. There are a number of similarities between ST data and single-cell RNA sequencing (scRNA-seq) data, with observations taking place at the level of spots or molecules instead of cells. For scRNA-seq, several well-maintained existing data infrastructures, or classes, are available, including *SingleCellExperiment* [17] in the Bioconductor [18] framework in the R programming language, *Seurat* [19,20] in R, and *AnnData* [21] in Python. Each of these provides a core data class, which stores measurements (e.g. transcript counts) together with additional information (referred to as metadata) describing the rows, columns, and overall experiment. These classes provide consistent storage and access to the data, including operations such as subsetting. Due to the similarity with scRNA-seq, recent studies have successfully reused or extended these classes with some customizations to store the spatial information [12,16]. However, currently there does not exist a standardized data infrastructure for storing and accessing ST data, which is crucial for downstream analyses and visualization functions that use these data. A well-designed infrastructure that standardizes these choices will simplify user actions such as generating plots and building analysis pipelines.

Here, we describe *SpatialExperiment*, a new data infrastructure for ST, implemented within the Bioconductor framework in R. *SpatialExperiment* extends the existing *SingleCellExperiment*, with added functionality for storing and accessing spatial information corresponding to the units of measurement. In extending *SingleCellExperiment*, we enable building upon existing analysis methods implemented for single-cell data [17,22]. Furthermore, the Bioconductor framework provides advantages of modularity, standardized infrastructure, and interoperability between packages from different developers [17]. In addition to the *SpatialExperiment* package, we have also developed R packages to provide example ST datasets (*STexampleData* and *TENxVisiumData*) and visualization tools (*ggspavis*), to use in examples, tutorials, demonstrations, and teaching. In the following sections, we describe in detail the structure and usage of the *SpatialExperiment* class, and show examples of functionality from the data and visualization packages.

System and Methods

The *SpatialExperiment* package provides access to the core data structure, referred to as a class. In the following, we use *italics* to refer to packages and class names, and `code font` for function names and parts of the class accessible to users via functions known as accessor functions. In addition, we provide the *STexampleData*, *TENxVisiumData*, and *ggspavis* packages, which provide access to example datasets in the *SpatialExperiment* format and visualization functions. The *SpatialExperiment* package is available from Bioconductor at https://bioconductor.org/packages/SpatialExperiment, while the *STexampleData*, *TENxVisiumData*, and *ggspavis* packages are available from GitHub at https://github.com/lmweber/STexampleData, https://github.com/HelenaLC/TENxVisiumData, and https://github.com/lmweber/ggspavis, and will be submitted to Bioconductor. The *STexampleData* package contains several small datasets from different platforms, which can easily be loaded and used for examples and demonstrations. The *TENxVisiumData* contains a set of 13 datasets from the 10x Genomics Visium platform. For the examples in the next section, we use example datasets *STexampleData* (**Table 1**). The names of the datasets contain the platform name and dataset name, separated by an underscore.

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 [1]	Spot-based	Human brain	1	3,639	33,538	Yes	Yes	[12,23]
Visium_mouseCoronal	10x Genomics Visium [1]	Spot-based	Mouse brain	1	2,702	32,285	Yes	Yes	[24]
seqFISH_mouseEmbryo	seqFISH [5,6]	Molecule-based	Mouse embryo	1	11,026	351	No	No	[16]

Table 1. Summary of example datasets from the *STexampleData* package used to demonstrate the *SpatialExperiment* structure.

Implementation

The *SpatialExperiment* class extends Bioconductor's widely-used *SingleCellExperiment* class (**Figure 1A**). Instances of the class, known as objects, contain the following components that are reused from *SingleCellExperiment*: (i) `assays`, which contain tables of measurement values such as raw and transformed transcript counts, (ii) `rowData`, which contains additional information (metadata) describing the features (e.g. gene IDs and gene names), (iii) `colData`, which contains metadata describing the spatial coordinates or cells (e.g. barcode IDs for spot-based ST or cell IDs for molecule-based ST), and optionally (iv) `reducedDims`, which stores reduced dimension representations of the measurements (e.g. from principal component analysis) of the features. Note that in the Bioconductor framework, features are mostly commonly stored in rows, and observations in columns. In addition, to store the spatial information in ST data, we extend *SingleCellExperiment* to include the following components: (v) `spatialData`, which contains spatial coordinates (e.g. x and y

coordinates) and other spatial metadata describing the barcodes or cells, and optionally (vi) `imgData`, which stores image files (e.g. histology images, if available) and information describing these (e.g. resolution in pixels). An additional accessor function `spatialCoords()` can be used to extract the spatial coordinates as a numeric matrix, which enables easier input to some downstream methods such as visualization functions. For spot-based data, `assays` contains a single table of measurements named `counts`, consisting of the messenger RNA (mRNA) counts per gene per spot. For molecule-based data, `assays` contains two tables named `counts` and `molecules`, with `counts` containing summarized mRNA counts per gene per cell, and `molecules` storing the spatial coordinates and intensities for each individual mRNA molecule (formatted as a `BumpyMatrix` [25] object). In addition, for molecule-based data, the spatial coordinates of the cell centroids are stored in `spatialData`, and (optionally) segmentation vertices outlining the spatial coordinates of each cell are stored in `colData` (as a `SplitDataFrameList` object).

By building *SpatialExperiment* as an extension of *SingleCellExperiment* within the Bioconductor framework, we ensure that users can rely on standard operations, such as subsetting objects by column (barcode or cell) or row (gene), which simplifies the user interface and helps to avoid errors by consistently subsetting across all components of the object. The modularity and interoperability of the Bioconductor framework [17] ensures that users can continue to apply methods that were originally developed for *SingleCellExperiment* objects (e.g. preprocessing methods [26]). For datasets that are too large to store in memory (e.g. millions of spots), *SpatialExperiment* can easily take advantage of existing Bioconductor infrastructure for sparse matrices and on-disk data representations, such as *DelayedArray* [27]. *SpatialExperiment* objects can be created with a function known as a constructor function (`SpatialExperiment()`), which accepts the individual components via input arguments and returns a *SpatialExperiment* object. Alternatively, for the 10x Genomics Visium platform, we also provide a dedicated constructor function (`read10xVisium()`), which creates objects directly from the raw files generated by the 10x Genomics Visium processing software (Space Ranger [28]).

Figure 1B displays an example of a spot-based ST dataset in *SpatialExperiment* format (*Visium_humanDLPFC* from the *STexampleData* package). This is a single biological sample (sample 151673 [12]) from the human brain dorsolateral prefrontal cortex (DLPFC) region, measured with the 10x Genomics Visium platform. The full dataset containing 12 biological samples was previously published in *SingleCellExperiment* format in the *spatialLIBD* package [12,23]. **Figure 1C** shows an example of a molecule-based ST dataset (*seqFISH_mouseEmbryo* dataset in *STexampleData*). This is a subset of cells (embryo 1, z-slice 2 [16]) from a published dataset investigating mouse embryogenesis [16], generated using the seqFISH platform. Additional details on both datasets are provided in **Table 1**. Code to download these example datasets reproduce these figures is available in the *STexampleData* and *ggspavis* package vignettes.

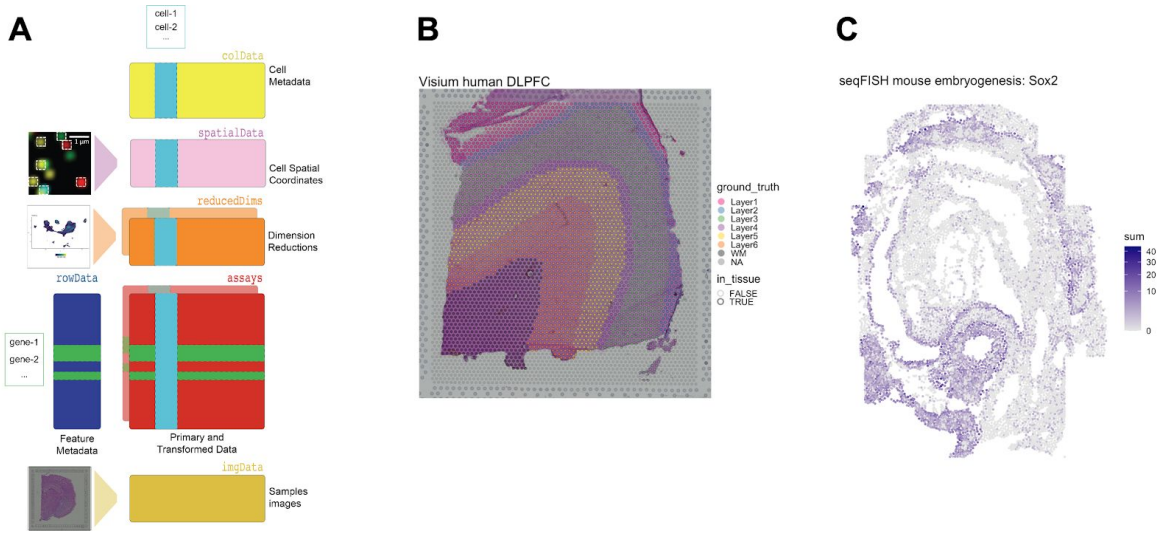

Figure 1: (A) Illustration of the structure of the *SpatialExperiment* class, including `assays`, `rowData`, `colData`, `reducedDims`, `spatialData`, and `imgData` components, as described in the text. **(B)** Example of spot-based dataset from the 10x Genomics Visium platform [1] (available as *Visium_humanDLPFC* [12,23] in the *STexampleData* package). Image shows histology image as background, grid of spots, highlighting for spots overlapping with tissue, and colors for ground truth cluster labels available in this dataset. Original dataset is also available from *spatialLIBD* package [23]. **(C)** Example of molecule-based dataset from the seqFISH platform [5,6] (available as *seqFISH_mouseEmbryo* [16] in the *STexampleData* package). Color scale shows total mRNA counts per cell for the *Sox2* gene. Additional details on both datasets are provided in **Table 1**. Figures generated using plotting functions from the *ggspavis* package.

Discussion

Providing robust, standardized data infrastructure and containers for single-cell data (such as *SingleCellExperiment* [17], *Seurat* [19,20], and *AnnData* [21]) has greatly streamlined the work of users, including data analysts and method developers. In particular, a consistent infrastructure brings several benefits in terms of interoperability and modularity -- method developers can implement methods that accept and generate standard objects as inputs and outputs, which enables data analysts to build pipelines consisting of methods from different developers. For single-cell data, this has led to the development of comprehensive tutorials and pipelines [17,22], which are an invaluable resource for new users to learn how to analyze these data types. ST is a new technology, which yields data with some similarities to single-cell data, but does not fit neatly into existing data infrastructure. By developing *SpatialExperiment*, we aim to create similar benefits of standardization, interoperability, and modularity for ST data. We anticipate that our contribution will simplify the work of method developers, spur the development of resources such as tutorials and example pipelines, and make it easier for users and experimental researchers to analyze these data. Since *SpatialExperiment* extends *SingleCellExperiment*, we also ensure that existing tools, originally designed for single-cell data, can easily be applied to ST data, and that existing pipelines can be adapted. Our associated packages (*STexampleData*, *TENxVisiumData*, and *ggspavis*) provide several example datasets and visualization functions, for use in method development, as well as examples, tutorials, demonstrations, and teaching.

ST technologies are still in their infancy and the next years will see a continuing development of existing platforms as well as the emergence of novel experimental approaches. For instance, new

platforms are being developed to measure multiple assay types (e.g. transcriptomics, proteomics, and epigenomics) at single-cell or near-single-cell level in a spatial context [29]. These and other developments will necessitate further extensions to data infrastructure. As part of the Bioconductor framework, *SpatialExperiment* is ideally positioned to be extended in this manner. For example, multiple assay types can be stored within multiple tables of measurements in `assays` using the *MultiAssayExperiment* [30] framework (e.g. as demonstrated in the *SingleCellMultiModal* package [31]). New platforms and datasets will also extend ST data into three dimensions [32] or across multiple timepoints -- these types of measurements could easily be stored in *SpatialExperiment* by including additional columns of spatial or temporal coordinates. We also expect that datasets will soon become too large to store in memory, with up to millions of spots or cells measured per sample: *SpatialExperiment* already has the capabilities to deal with such cases through its integration with the existing Bioconductor infrastructure for sparse matrices and on-disk data representations, such as *DelayedArray* [27]. Finally, interoperability between *SpatialExperiment* and other data formats (such as *AnnData* [21] and *Loompy* [33] in Python) will be ensured through the use of existing conversion packages [33,34].

By providing data infrastructure and examples, and working with the Bioconductor community to further develop downstream analysis methods and visualizations (e.g. *dittoSeq* [35] and *iSEE* [36]), we aim to streamline the work of method developers and data analysts, offer an appropriate standard for data sharing, and ultimately make it easier for experimental researchers to obtain robust and reproducible biological insights from these platforms.

Acknowledgments

The authors thank members 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 and 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 code input 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 freely available from Bioconductor at https://bioconductor.org/packages/SpatialExperiment. This manuscript was prepared using version 1.1.42 of the *SpatialExperiment* package. The data packages *STexampleData* and *TENxVisiumData*, and the visualization package *ggspavis*, are available from GitHub at http://github.com/lmweber/STexampleData, https://github.com/HelenaLC/TENxVisiumData, and https://github.com/lmweber/ggspavis, and will be submitted to Bioconductor. All datasets described in Table 1 and Figure 1 are available in the *STexampleData* package. Datasets are also freely available from the original sources listed in Table 1 [12,16,23,24].

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. 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. 10x Genomics. 10x Genomics Visium Spatial Gene Expression Solution. Website: https://www10xgenomics.com/spatial-transcriptomics/. 2020.

2. 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.

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.

5. 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.

6. 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.

7. 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.

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

9. 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.

10. 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.

11. CARTANA: Part of 10x Genomics. CARTANA: Spatial Single Cell Analysis of Tissue Samples. Website: https://www.cartana.se/. 2020.

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. bioRxiv. 2020.

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 J. A., Bardot E. S., Eng C.-H. L., Tyser R. C. V., Argelaguet R., Guibentif C., Srinivas S., Briscoe J., Simons B. D., Hadjantonakis A.-K., Göttgens B., Reik W., Nichols J., Cai L., Marioni J. C. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. bioRxiv. 2020;391896.

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. 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.

20. 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.

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

22. Lun A, Amezquita R, Hicks S, Gottardo R. Orchestrating Single-Cell Analysis with Bioconductor. Online book: https://osca.bioconductor.org/. 2020.

23. Collado-Torres L, Maynard KR, Jaffe AE. spatialLIBD: LIBD Visium spatial transcriptomics human pilot data inspector; R package version 1.2.1. http://www.bioconductor.org/packages/spatialLIBD. 2020.

24. 10x Genomics. Mouse Brain Section (Coronal). Website: https://support10xgenomics.com/spatial-gene-expression/datasets/110/V1_Adult_Mouse_Brain. 2020.

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

26. 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.

27. Pagès H, Hickey P, Lun A. DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets; R package, version 0.16.0. http://bioconductor.org/packages/DelayedArray. 2020.

28. 10x Genomics. Space Ranger: Spatial Gene Expression. Website: https://support10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger. 2020.

29. 10x Genomics. Visium Spatial Proteomics. Website: https://www10xgenomics.com/products/spatial-proteomics. 2020.

30. 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.

31. Ramos M, Eckenrode K, Waldron L. SingleCellMultiModal: Integrating Multi-modal Single Cell Experiment datasets; R package, version 1.2.2. https://bioconductor.org/packages/SingleCellMultiModal. 2020.

32. 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.

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

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

35. Bunis D, Andrews J. dittoSeq: User Friendly Single-Cell and Bulk RNA Sequencing Visualization; R package, version 1.2.2. https://www.bioconductor.org/packages/dittoSeq. 2021.

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