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

Dario Righelli1*, Lukas M. Weber2*, Helena L. Crowell3,4*, Brenda Pardo5,6, Leonardo Collado-Torres6, Shila Ghazanfar7, Aaron T. L. Lun8, Stephanie C. Hicks2†, Davide Risso1†

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), 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


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


profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National
Academy of Sciences of the USA. 2016;113: 11046–11051.

RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy


cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nature Communications. 2018;9:
2419.


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


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

### Supplementary Tables

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

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

**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

(A) Visium human DLPFC

(B) seqFISH mouse embryogenesis: Sox2

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.