

Package: SubcellularSpatialData (via r-universe)

October 10, 2026

Title Annotated spatial transcriptomics datasets from 10x Xenium, NanoString CosMx and BGI STOmics.

Version 1.8.0

Description This is a data package that hosts annotated sub-cellular localised datasets from the STOmics, Xenium and CosMx platforms. Specifically, it hosts datasets analysed in the publication Bhuvu et. al, 2024 titled ``Library size confounds biology in spatial transcriptomics data". Raw transcript detections are hosted and functions to convert them to SpatialExperiment objects have been implemented.

biocViews ExperimentHub, SpatialData, Homo_sapiens_Data, Mus_musculus_Data, LungCancerData

License GPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

URL <https://davislaboratory.github.io/SubcellularSpatialData>

BugReports <https://github.com/DavisLaboratory/SubcellularSpatialData/issues>

Imports dplyr, hexbin, Matrix, SpatialExperiment, stats

NeedsCompilation no

VignetteBuilder knitr

Suggests BiocStyle, ExperimentHub, ggplot2, knitr, prettydoc, rmarkdown, testthat (>= 3.0.0)

Depends R (>= 4.4)

LazyData false

Config/testthat/edition 3

Config/pak/sysreqs libmagick++-dev gsfonts libicu-dev libssl-dev zlib1g-dev

Repository Bioconductor 3.23

Date/Publication 2026-04-28 12:37:17 UTC
RemoteUrl <https://github.com/bioc/SubcellularSpatialData>
RemoteRef RELEASE_3_23
RemoteSha 7ec55d259b50cc903d346b773bf1108b9c7c877d

Contents

tx_small	2
tx2spe	2
Index	5

tx_small	<i>Sample annotated sub-cellular localised spatial 'omics dataset</i>
----------	---

Description

This object stores a sample transcript table to demonstrate the format of data stored within this package. The sample dataset contains 100,000 transcript detections from 2 breast cancer samples profiled using the 10x Xenium platform.

Usage

tx_small

Format

A data.frame where each row represents a transcript detection. The full description of these transcript tables is available in the vignettes or in the documentation for the tx2spe() function.

tx2spe	<i>Compute cell/bin-level summaries and save to a SpatialExperiment object</i>
--------	--

Description

This function summarises transcript detection (or DNA nanoball spot counts) into cells or hexbins and returns a SpatialExperiment object. Annotations of cells or hexbins are summarised such that the most frequent annotation is used.

Usage

tx2spe(x, bin = c("cell", "hex", "square", "region"), nbins = 100)

Arguments

<code>x</code>	a data.frame containing a sub-cellular localised dataset from the SubcellularSpatialData package.
<code>bin</code>	a character, stating whether transcripts should be binned into cells or hexbins. The default is 'cell', however, it is advisable to use 'hexbin' when cell identification is not accurate (e.g., when cell boundaries are inferred only from nuclei and not directly measured using cytoplasmic/cell membrane stains).
<code>nbins</code>	a numeric, stating the number of bins to create in the x and y axes

Details

Data description:

This function works on all data associated with this package. For new datasets, the transcript table MUST have the following columns:

- `sample_id` Unique identifier for the sample the transcript belongs to.
- `cell` Unique identifier for the cell the transcript belongs to (NA if it is not allocated to any cell).
- `gene` Gene name or identity of the transcript.
- `genotype` Type of target (e.g., gene or control probe). True genes should be annotated as "Gene". Other target types can be named as desired.
- `x` x-coordinate of the transcript.
- `y` y-coordinate of the transcript.
- `counts` Number of transcripts detected at this location (always 1 for NanoString CosMx and 10x Xenium as each row is an individual detection).
- `technology` Name of the platform the data was sequenced on. "STOmics" is treated differently as the number of spots as well as the number of nuclei/cells are counted. No specific format required for other technologies.

The "region" column is optional and only required if binning by regions. If present, it must be a character or factor. For datasets within this package, this column stored the independently annotated histological region.

Summarisation:

When transcript counts are aggregated (using any approach), aggregation of numeric columns is performed using the `mean(..., na.rm = TRUE)` function and aggregation of character/factor columns is performed such that the most frequent class becomes the representative class. As such, for a hex bin, the highest frequency region for the transcripts allocated to the bin becomes the bin's region annotation. New coordinates for cells, bins, or regions are computed using the mean function as well, therefore represent the center of mass for the object. For hex and square bins, the average coordinate is computed by default, however, x and y indices for the bin are stored in the `colData` under the `bin_x` and `bin_y` columns.

When aggregation is performed using bins or regions, an additional column, `ncells`, is computed that indicates how many unique cells are present within the bin/region. Do note that if a cell overlaps multiple bins/regions, it will be counted in each bin/region.

Please note that BGI only performs nuclear binning of counts therefore cellular counts obtained will represent nuclear counts only. We recommended that for fair evaluations during benchmarking, alternative binning algorithms that are fair are explored, or analysis be performed on square or hex binned counts.

Value

a SpatialExperiment object containing cell/bin/region-level expression quantification.

Examples

```
data(tx_small)
head(tx_small)
tx2spe(tx_small, "region")
```

Index

* **datasets**

tx_small, [2](#)

tx2spe, [2](#)

tx_small, [2](#)