Package 'SFEData'

November 12, 2024

Title Example SpatialFeatureExperiment datasets

Version 1.8.0

Description Example spatial transcriptomics datasets with Simple Feature annotations as SpatialFeatureExperiment objects. Technologies include Visium, slide-seq, Nanostring CoxMX, Vizgen MERFISH, and 10X Xenium. Tissues include mouse skeletal muscle, human melanoma metastasis, human lung, breast cancer, and mouse liver.

Imports BiocFileCache, AnnotationHub, ExperimentHub, utils

Suggests BiocStyle, knitr, rmarkdown, SeuratObject, SpatialFeatureExperiment, testthat (>= 3.0.0)

License Artistic-2.0

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.1

biocViews ExperimentHub, ExpressionData, Mus_musculus_Data, Homo_sapiens_Data, SpatialData, Tissue, SingleCellData

VignetteBuilder knitr

Config/testthat/edition 3

URL https://github.com/pachterlab/SFEData

BugReports https://github.com/pachterlab/SFEData/issues

git_url https://git.bioconductor.org/packages/SFEData

git_branch RELEASE_3_20

git_last_commit da5376a

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-12

Author Lambda Moses [aut, cre] (<https://orcid.org/0000-0002-7092-9427>), Alik Huseynov [aut] (<https://orcid.org/0000-0002-1438-4389>), Kayla Jackson [aut] (<https://orcid.org/0000-0001-6483-0108>), Lior Pachter [aut, ths] (<https://orcid.org/0000-0002-9164-6231>)

Maintainer Lambda Moses <dlu2@caltech.edu>

Contents

`BiermannMelaMetasData` *Melanoma metastasis slide-seq2 data*

Description

This function can download one of the human melanoma brain metastasis (MBM) samples and one of the melanoma extracranial metastasis (ECM) samples from the paper Dissecting the treatment-naive ecosystem of human melanoma brain metastasis, Biermann et al. The datasets are `GSM6025935_MBM05_rep1` and `GSM6025946_ECM01_rep1`. The raw counts and cell metadata were downloaded from GEO. The raw counts, QC metrics such as number of UMIs and genes detected per barcode, and centroid coordinates as `sf` POINT geometry, are included in the SFE object.

Usage

```
BiermannMelaMetasData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

`dataset`	Which dataset to use, must be one of "MBM05_rep1" and "ECM01_rep1".
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Value

A `SpatialFeatureExperiment` object.

Examples

```
sfe <- BiermannMelaMetasData()
```

CosMXOutput *CosMX output from mouse brain*

Description

This is a small subset of the CosMX output from the mouse quarter brain downloaded from the Nanostring website. When downloaded, the files are in the output format of CosMX, not an SFE object. The purpose of this dataset is to demonstrate and test `readCosMX` in SFE.

Usage

```
CosMXOutput(dataset = datasets, file_path = ".", force = FALSE, verbose = TRUE)
```

Arguments

`dataset`	Can only be "subset1".
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Value

Path to the tarball containing the output directory.

HeNSCLCData *Nanostring FFPE CosMX human NSCLC data*

Description

One of the CosMX example datasets for human non small cell lung cancer (NSCLC, Lung5_Rep1) from the Nanostring website was downloaded and formatted into an SFE object. The dataset is described in the paper High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging, He et al. Since there's no easy way to get the cell segmentation polygon coordinates from the Nanostring website, the polygon coordinates were downloaded from Seurat's vignette. The raw count matrix, QC metrics, cell segmentation in one z-plane, and other cell attributes such as area, aspect ratio, mean DAPI level, mean immunofluorescence signal, and etc. are included in the SFE object.

Usage

```
HeNSCLCData(dataset = datasets, file_path = ".", force = FALSE, verbose = TRUE)
```

Arguments

dataset	Which dataset to use, for now can only be "Lung5_Rep1".
file_path	Path to save downloaded files for the *Output functions which don't return an SFE object.
force	Logical, whether to force redownload if the files are already present. Defaults to FALSE.
verbose	Whether to display progress of download.

Value

A SpatialFeatureExperiment object.

JanesickBreastData *Xenium FFPE human breast cancer data*

Description

This dataset was downloaded from the 10X website, and described in the paper High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue, Janesick et al. The dataset might not be representative of later Xenium data. There are two samples, which can both be downloaded with this package. For each sample, the raw gene counts, QC metrics, cell and nuclei segmentation polygons in one z-plane, and cell centroids are included in the SFE object. The two samples are in separate SFE objects. A small number of nuclei polygons are invalid due to self-intersection; these cases were resolved by making a buffer of distance 0 and then removing the holes. Additional cell metadata provided by 10X, such as cell area, are also included.

Usage

```
JanesickBreastData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

dataset	Which dataset to use, must be one of "rep1" and "rep2".
file_path	Path to save downloaded files for the *Output functions which don't return an SFE object.
force	Logical, whether to force redownload if the files are already present. Defaults to FALSE.
verbose	Whether to display progress of download.

Details

As the SFE and Voyager packages are in the experimental stage and they were originally developed and tested on relatively small Visium datasets, they are not yet very scalable to larger smFISH datasets. While 10X provided transcript spot locations, these are not included in the SFE objects for now as we are not sure if `spatstat` can work with such a large dataset for spatial point process analyses, nor does SFE integrate with `spatstat`. In a future version of this package, the transcript locations might be added as a separate dataset, but this is not guaranteed.

Value

A `SpatialFeatureExperiment` object.

`LohoffGastrulationData` *seqFISH mouse gastrulation dataset*

Description

This dataset was downloaded from the companion website titled, Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis Lohoff et al. There are three biological replicates available in this dataset, each representing a different embryo. For each dataset, the raw gene counts, metadata, and cell segmentation in one z-plane are provided in the SFE object. Segmentation data were not provided for provided for all cells in the count matrix for embryos 1 and 2. In these cases, the segmentation data are represented by empty polygons.

Usage

```
LohoffGastrulationData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

`dataset`	Which dataset to use, must be one of "rep1", "rep2", and "rep3".
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Details

While the authors provided the spot location for each mRNA molecule, these are not included in the SFE objects.

Value

A `SpatialFeatureExperiment` object.

`McKellarMuscleData` *Download McKellar et al. mouse skeletal muscle data*

Description

In the first version of this package, only the first time point, 2 days after notexin injury, is available. We may add the later time points in later versions of this package.

Usage

```
McKellarMuscleData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

`dataset`	Which dataset to use. Whether the full dataset ("full"), the first small subset ("small"), or the second small subset ("small2"). The second small subset has a different `sample_id`.
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Details

All datasets are `SpatialFeatureExperiment` (SFE) objects, with a `counts` assay for the raw gene counts. Column metadata includes total UMI counts (`nCounts`) and number of genes (`nGenes`) detected per spot. Row metadata includes means, variances, and CV2 of each gene in the full dataset. Column geometry includes Visium spot polygons (`spotPoly`). Annotation geometry includes tissue boundary (`tissueBoundary`), myofiber segmentation (full resolution `myofiber_full` and simplified `myofiber_simplified`), nuclei segmentation (`nuclei`), and nuclei centroids (`nuclei_centroid`).

Myofibers were segmented manually with the LabKit ImageJ plugin on a 4x downsized H&E image, downsized so the image can be loaded into LabKit, and the `terra` R package was used to convert the TIFF segmentation masks into polygons. Coordinates in the SFE objects are in pixels in the full resolution H&E image. Hence the coordinates of the myofiber segmentations were scaled up to match the other coordinates. The full resolution myofiber segmentation looks pixelated; the `mapshaper` R package was used to simplify polygons while conserving contiguity. Morphological (area, perimeter, eccentricity, angle) and Haralick (see `EBImage::computeFeatures.haralick`) metrics were computed for the myofibers with the `EBImage` R package.

Nuclei were segmentated with StarDist. About 3000 nuclei from randomly selected regions in the H&E image from this and later time points were manually annotated with LabKit to train the StarDist model, which was then used to segment all nuclei. OpenCV was used to convert segmentation masks into polygons and compute morphological metrics.

Tissue boundary was obtained by first thresholding the H&E image by grascale intensity and then converting the mask into polygons with OpenCV. Small pieces which are debris were removed.

Value

A SpatialFeatureExperiment object.

Examples

```
sfe <- McKellarMuscleData("small")
```

SeuratTestData *Seurat objects to test coercion function*

Description

Some of the other test datasets in this package are made as Seurat object to unit test function to convert Seurat objects to SFE.

Usage

```
SeuratTestData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

dataset	Which dataset to use. Description of the datasets: **Visium** From SeuratData (ie stxBrain.SeuratData), subsetted to keep first 50 genes **VisiumHD8** Visium HD mouse brain data from the 10X website, with the first 50 genes in 8 um bins **VisiumHDmulti** Visium HD mouse brain data from the 10X website, with the first 50 genes in 8 um and 16 um bins **Vizgen** Same dataset in VizgenOutput with dataset = "hdf5" **VizgenMulti** Same as in Vizgen but with a subset of the data used as if it's another sample to test the coercion function for multiple samples **Xenium** Same as in XeniumOutput **XeniumMulti** Same as in Xenium but with a subset of the data used as if it's another sample to test the coercion function for multiple samples
file_path	Path to save downloaded files for the *Output functions which don't return an SFE object.
force	Logical, whether to force redownload if the files are already present. Defaults to FALSE.
verbose	Whether to display progress of download.

Value

A Seurat object

SFEData *Example SpatialFeatureExperiment datasets*

Description

Example spatial transcriptomics datasets with Simple Features annotations as `SpatialFeatureExperiment` objects.

Datasets

Below are datasets that are already SFE objects, used in examples and vignettes:

Full Visium dataset of the first time point, including spots outside tissue. (`McKellarMuscleData`)

Small subset of the full Visium dataset for function examples. (`McKellarMuscleData`)

A second small subset of the full Visium dataset with a different `sample_id` used for function examples involving multiple samples. (`McKellarMuscleData`)

Slide-seq2 human melanoma brain metastasis dataset (`BiermannMelaMetasData`)

Slide-seq2 human melanoma extracranial metastasis dataset (`BiermannMelaMetasData`)

10X Xenium formalin fixed paraffin embedded (FFPE) Xenium dataset for human breast cancer (2 biological replica, `JanesickBreastData`)

Nanostring CosMX FFPE human non small cell lung cancer data (`HeNSCLCData`)

Vizgen MERFISH mouse liver data (`VizgenLiverData`)

Below are small subsets of datasets in the original output format used to demonstrate and test data reading functions in the SFE package:

Unpublished Vizgen MERFISH human brain cancer data (`VizgenOutput`)

Subset of CosMX mouse brain data (`CosMXOutput`)

Subset of Xenium output from the mouse brain, generated with Xenium Onboarding Analysis v1 (`XeniumOutput`)

Subset of Xenium output from the human pancreas, generated with Xenium Onboarding Analysis v2 (`XeniumOutput`)

Seurat object for unit testing function to convert Seurat objects to SFE (`SeuratTestData`)

Author(s)

Maintainer: Lambda Moses <dlu2@caltech.edu> (ORCID)

Authors:

- Alik Huseynov (ORCID)
- Kayla Jackson <kaylajac@caltech.edu> (ORCID)
- Lior Pachter <lpachter@caltech.edu> (ORCID) [thesis advisor]

See Also

Useful links:

- https://github.com/pachterlab/SFEData
- Report bugs at https://github.com/pachterlab/SFEData/issues

`VizgenLiverData` *Vizgen MERFISH mouse liver data*

Description

This is one of the example datasets from Vizgen's website, downloaded from here. The gene count matrix, cell metadata provided by Vizgen, QC metrics, and cell segmentation in one z-plane are included in the SFE object. While it appears that 7 z-planes are in the cell boundaries in the hdf5 files on the website, all 7 z-planes are the same, so only one was used here.

Usage

```
VizgenLiverData(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

`dataset`	Which dataset to use, for now can only be "Liver1Slice1".
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Details

This is the largest SFE example dataset thus far. Whlie the SFE object can fit into memory due to the relatively small number of genes, we are considering making a `HDF5Array` version of this example dataset. Furthermore, the geometries of the large number of cells can also consume a lot of memory. We are considering using Apache Sedona and possibly SQLDataFrame for on disk geometries and geometric operations in a future version of `SpatialFeatureExperiment` and `Voyager`.

While Vizgen provides transcript spot locations, we don't yet know what to do with the huge dataset, so this is not included in the SFE object.

Value

A `SpatialFeatureExperiment` object.

VizgenOutput *Vizgen MERFISH output from human brain cancer*

Description

This dataset is unpublished, and is a small subset of the original dataset. When downloaded, the files are in the output format of Vizgen, not an SFE object. The purpose of this dataset is to demonstrate and test `readVizgen` in SFE.

Usage

```
VizgenOutput(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

`dataset`	Which dataset to use, can be "hdf5" where cell segmentations are stored in hdf5 files or "cellpose" where cell segmentations are stored in a parquet file.
`file_path`	Path to save downloaded files for the `*Output` functions which don't return an SFE object.
`force`	Logical, whether to force redownload if the files are already present. Defaults to `FALSE`.
`verbose`	Whether to display progress of download.

Value

Path to the tarball containing the output directory.

XeniumOutput *Xenium output*

Description

These are small subsets of Xenium output from data downloaded from the 10X website used to test `readXenium()` in SFE. The first subset comes from the mouse brain, generated with Xenium Onboarding Analysis (XOA) v1. The second subset comes from the human pancreas, generated with XOA v2. XOA v1 and v2 have different output formats, hence the separate subsets used for testing. To reduce download time, the zarr files in the original output were removed since they are not used by the SFE package. Also, only lower resolution images are kept.

Usage

```
XeniumOutput(
  dataset = datasets,
  file_path = ".",
  force = FALSE,
  verbose = TRUE
)
```

Arguments

dataset	Either "v1" or "v2", as described above.
file_path	Path to save downloaded files for the *Output functions which don't return an SFE object.
force	Logical, whether to force redownload if the files are already present. Defaults to FALSE.
verbose	Whether to display progress of download.

Details

For the v1 output, the cell and nucleus boundary parquet files have two versions, one with the arrow raw bytes and the other without. There are no longer raw bytes since XOA v1.4.

To make the test data smaller, the transcript spots have been down sampled in the v1 data but not in v2.

Value

Path to the tarball containing the output directory.

Index