Package 'SingleCellMultiModal'

July 3, 2025

Type Package

Title Integrating Multi-modal Single Cell Experiment datasets

Version 1.20.0

Description SingleCellMultiModal is an ExperimentHub package that serves multiple datasets obtained from GEO and other sources and represents them as MultiAssayExperiment objects. We provide several multi-modal datasets including scNMT, 10X Multiome, seqFISH, CITEseq, SCoPE2, and others. The scope of the package is is to provide data for benchmarking and analysis. To cite, use the 'citation' function and see <https://doi.org/10.1371/journal.pcbi.1011324>.

License Artistic-2.0

BugReports https://github.com/waldronlab/SingleCellMultiModal/issues

Depends R (>= 4.2.0), MultiAssayExperiment

Imports AnnotationHub, BiocBaseUtils, BiocFileCache, ExperimentHub, graphics, HDF5Array, S4Vectors, SingleCellExperiment, SpatialExperiment, SummarizedExperiment, Matrix, methods, utils

Suggests BiocStyle, ggplot2, knitr, RaggedExperiment, rmarkdown, scater, scran, UpSetR, uwot

VignetteBuilder knitr

biocViews ExperimentData, SingleCellData, ReproducibleResearch, ExperimentHub, GEO

Encoding UTF-8

RoxygenNote 7.3.2

Roxygen list(markdown = TRUE)

Date 2025-04-07

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

git_branch RELEASE_3_21

git_last_commit 32afbd3

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-07-03

Author Marcel Ramos [aut, cre] (ORCID:
<https://orcid.org/0000-0002-3242-0582>),
Ricard Argelaguet [aut],
Al Abadi [ctb],
Dario Righelli [aut],
Christophe Vanderaa [ctb],
Kelly Eckenrode [aut],
Ludwig Geistlinger [aut],
Levi Waldron [aut]

Maintainer Marcel Ramos <marcel.ramos@roswellpark.org>

Contents

`SingleCellMultiModal-package`
SingleCellMultiModal-package

Description

The SingleCellMultiModal package provides a convenient and user-friendly representation of multi-modal data from project such as `scNMT` for mouse gastrulation.

Author(s)

Maintainer: Marcel Ramos <marcel.ramos@roswellpark.org> (ORCID)

Authors:

- Ricard Argelaguet <ricard@ebi.ac.uk>
- Dario Righelli <dario.righelli@gmail.com>

- Kelly Eckenrode <kelly.eckenrode@sph.cuny.edu>
- Ludwig Geistlinger <ludwig_geistlinger@hms.harvard.edu>
- Levi Waldron <lwaldron.research@gmail.com>

Other contributors:

- Al Abadi [contributor]
- Christophe Vanderaa <christophe.vanderaa@uclouvain.be> [contributor]

See Also

Useful links:

- Report bugs at https://github.com/waldronlab/SingleCellMultiModal/issues

Examples

```
help(package = "SingleCellMultiModal")
```

.CITEseqMaeToSce *CITEseqMaeToSce*

Description

converts a `MultiAssayExperiment` object with CITEseq data into a `SingleCellExperiment` object to be used with already known methods and packages in literature.

Note that for creating a `SingleCellExperiment` object the following function subsets all the assays present in the `MultiAssayExperiment` with only the common cells across all the modalities. This could result in a not complete object.

Usage

```
.CITEseqMaeToSce(mae)
```

Arguments

`mae`	a MultiAssayExperiment object with scRNA and/or scADT and/or scHTO named experiments.

Value

a SingleCellExperiment object as widely with scRNA data as counts and scADT, scHTO data as altExps. If only one modality is present, it has returned as main assay of the SCE.

addCTLabels *addCTLabels*

Description

addCTLabels

Usage

```
addCTLabels(
  cd,
  out,
  outname,
  ct,
  mkrcol = "markers",
  ctcol = "celltype",
  overwrite = FALSE,
  verbose = TRUE
)
```

Arguments

cd	the colData DataFrame
out	list data structure returned by getCellGroups
outname	character indicating the name of the out data structure
ct	character indicating the celltype to assign in the ctcol
mkrcol	character indicating the cd column to store the markers indicated by outname (default is markers)
ctcol	character indicating the column in cd to store the cell type indicated by ct (default is celltype)
overwrite	logical indicating if the cell types have to be overwritten without checking if detected barcodes were already assigned to other celltypes
verbose	logical for having informative messages during the execution

Value

an updated version of the cd DataFrame

CITEseq *CITEseq*

Description

function assembles data on-the-fly from ExperimentHub to provide a MultiAssayExperiment container. Actually the dataType argument provides access to the available datasets associated to the package.

Usage

```
CITEseq(
  DataType = c("cord_blood", "peripheral_blood"),
  modes = "*",
  version = "1.0.0",
  dry.run = TRUE,
  filtered = FALSE,
  verbose = TRUE,
  DataClass = c("MultiAssayExperiment", "SingleCellExperiment"),
  ...
)
```

Arguments

DataType	character(1) indicating the identifier of the dataset to retrieve. (default "cord_blood")
modes	character() The assay types or modes of data to obtain these include scADT and scRNA-seq data by default.
version	character(1) Either version '1.0.0' depending on data version required.
dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
filtered	logical(1) indicating if the returned dataset needs to have filtered cells. See Details for additional information about the filtering process.
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
DataClass	either MultiAssayExperiment or SingleCellExperiment data classes can be returned (default MultiAssayExperiment)
...	Additional arguments passed on to the ExperimentHub-class constructor

Details

CITEseq data are a combination of single cell transcriptomics and about a hundread of cell surface proteins. Available datasets are:

- cord_blood: a dataset of single cells of cord blood as provided in Stoeckius et al. (2017).
  - scRNA_Counts - Stoeckius scRNA-seq gene count matrix

  - scADT - Stoeckius antibody-derived tags (ADT) data

- peripheral_blood: a dataset of single cells of peripheral blood as provided in Mimitou et al. (2019). We provide two different conditions controls (CTRL) and Cutaneous T-cell Limphoma (CTCL). Just build appropriate `modes` regex for subselecting the dataset modes.
  - scRNA - Mimitou scRNA-seq gene count matrix
  - scADT - Mimitou antibody-derived tags (ADT) data
  - scHTO - Mimitou Hashtag Oligo (HTO) data
  - TCRab - Mimitou T-cell Receptors (TCR) alpha and beta available through the object metadata.
  - TCRgd - Mimitou T-cell Receptors (TCR) gamma and delta available through the object metadata.

If `filtered` parameter is `FALSE` (default), the `colData` of the returned object contains multiple columns of `logicals` indicating the cells to be discarded. In case `filtered` is `TRUE`, the `discard` column is used to filer the cells. Column `adt.discard` indicates the cells to be discarded computed on the ADT assay. Column `mito.discard` indicates the cells to be discarded computed on the RNA assay and mitocondrial genes. Column `discard` combines the previous columns with an `OR` operator. Note that for the `peripheral_blood` dataset these three columns are computed and returned separately for the `CTCL` and `CTRL` conditions. In this case the additional `discard` column combines the `discard.CTCL` and `discard.CTRL` columns with an `OR` operator. Cell filtering has been computed for `cord_blood` and `peripheral_blood` datasets following section 12.3 of the Advanced Single-Cell Analysis with Bioconductor book. Executed code can be retrieved in the CITEseq_filtering.R script of this package.

Value

A single cell multi-modal `MultiAssayExperiment` or informative `data.frame` when `dry.run` is `TRUE`. When `DataClass` is `SingleCellExperiment` an object of this class is returned with an RNA assay as main experiment and other assay(s) as `AltExp(s)`.

Author(s)

Dario Righelli

References

Stoeckius et al. (2017), Mimitou et al. (2019)

Examples

```
mae <- CITEseq(DataType="cord_blood", dry.run=FALSE)
experiments(mae)
```

getCellGroups *getCellGroups*

Description

Shows the cells/barcodes in two different plots (scatter and density) divinding the space in four quadrant indicated by the two thresholds given as input parameters. The x/y-axis represent respectively the two ADTs given as input. It returns a list of one element for each quadrant, each with barcodes and percentage (see Value section for details).

Usage

```
getCellGroups(mat, adt1 = "CD19", adt2 = "CD3", th1 = 0.2, th2 = 0)
```

Arguments

mat	matrix of counts or clr transformed counts for ADT data in CITEseq
adt1	character indicating the name of the marker to plot on the x-axis (default is CD19).
adt2	character indicating the name of the marker to plot on the y-axis (default is CD3).
th1	numeric indicating the threshold for the marker on the x-axis (default is 0.2).
th2	numeric indicating the threshold for the marker on the y-axis (default is 0).

Details

helps to do manual gating for cell type indentification with CITEseq or similar data, providing cell markers. Once identified two interesting markers for a cell type, the user has to play with the thresholds to identify the cell populations specified by an uptake (+) o downtake (-) of the couple of markers (ADTs) previously selected.

Value

a list of four different element, each one indicating the quarter where the thresholds divide the plotting space, in eucledian order I, II, III, IV quadrant, indicating respectively +/+, +/-, -/+, -/- combinations for the couples of selected ADTs. Each element of the list contains two objects, one with the list of detected barcodes and one indicating the percentage of barcodes falling into that quadrant. .

GTseq *Parallel sequencing data of single-cell genomes and transcriptomes*

Description

GTseq assembles data on-the-fly from ExperimentHub to provide a MultiAssayExperiment container. The DataType argument provides access to the mouse_embryo_8_cell dataset as obtained from Macaulay et al. (2015). Protocol information for this dataset is available from Macaulay et al. (2016). See references.

Usage

```
GTseq(
  DataType = "mouse_embryo_8_cell",
  modes = "*",
  version = "1.0.0",
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

DataType	character(1) Indicates study that produces this type of data (default: 'mouse_embryo_8_cell')
modes	character() A wildcard / glob pattern of modes, such as "*omic". A wildcard of "*" will return all modes including copy numbers ("genomic") and RNA-seq read counts ("transcriptomic"), which is the default.
version	character(1) Currently, only version '1.0.0'.
dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
...	Additional arguments passed on to the ExperimentHub constructor

Details

G&T-seq is a combination of Picoplex amplified gDNA sequencing (genome) and SMARTSeq2 amplified cDNA sequencing (transcriptome) of the same cell. For more information, see Macaulay et al. (2015). * mouse_embryo_8_cell: this dataset was filtered for bad cells as specified in Macaulay et al. (2015). * genomic - integer copy numbers as detected from scDNA-seq * transcriptomic - raw read counts as quantified from scRNA-seq

Value

A single cell multi-modal MultiAssayExperiment or informative data.frame when dry.run is TRUE

metadata

The `MultiAssayExperiment` metadata includes the original function call that saves the function call and the data version requested.

Source

https://www.ebi.ac.uk/ena/browser/view/PRJEB9051

References

Macaulay et al. (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods, 12:519–22.

Macaulay et al. (2016) Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc, 11:2081–103.

See Also

SingleCellMultiModal-package

Examples

```
GTseq()
```

`ontomap` *Obtain a map of cell types for each dataset*

Description

The `ontomap` function provides a mapping of all the cell names across the all the data sets or for a specified data set.

Usage

```
ontomap(dataset = c("scNMT", "scMultiome", "SCoPE2", "CITEseq", "seqFISH"))
```

Arguments

`dataset`	`character()` One of the existing functions within the package. If missing, a map of all cell types in each function will be provided.

Details

Note that `CITEseq` does not have any cell annotations; therefore, no entries are present in the `ontomap`.

Value

A `data.frame` of metadata with cell types and ontologies

Examples

```
ontomap(dataset = "scNMT")
```

`scmmCache` *Manage cache / download directories for study data*

Description

Managing data downloads is important to save disk space and re-downloading data files. This can be done effortlessly via the integrated `BiocFileCache` system.

Usage

```
scmmCache(...)

setCache(
  directory = tools::R_user_dir("SingleCellMultiModal", "cache"),
  verbose = TRUE,
  ask = interactive()
)

removeCache(accession)
```

Arguments

`...`	For `scmmCache`, arguments passed to `setCache`
`directory`	`character(1)` The file location where the cache is located. Once set, future downloads will go to this folder. See `setCache` section for details.
`verbose`	Whether to print descriptive messages
`ask`	`logical(1)` (default TRUE when `interactive()`) Confirm the file location of the cache directory
`accession`	`character(1)` A single string indicating the accession number of the study

Value

The directory / option of the cache location

scmmCache

Get the directory location of the cache. It will prompt the user to create a cache if not already created. A specific directory can be used via `setCache`.

setCache

Specify the directory location of the data cache. By default, it will go into the user's home and package name directory as given by R_user_dir (default: varies by system e.g., for Linux: '$HOME/.cache/R/SingleCellMultiMod

removeCache

Some files may become corrupt when downloading, this function allows the user to delete the tarball associated with a study number in the cache.

Examples

```
getOption("scmmCache")
scmmCache()
```

---

scMultiome *Single-cell Multiome ATAC + Gene Expression*

Description

10x Genomics Multiome technology enables simultaneous profiling of the transcriptome (using 3' gene expression) and epigenome (using ATAC-seq) from single cells to deepen our understanding of how genes are expressed and regulated across different cell types. Data prepared by Ricard Argelaguet.

Usage

```
scMultiome(
  DataType = "pbmc_10x",
  modes = "*",
  version = "1.0.0",
  format = c("MTX", "HDF5"),
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

DataType	character(1) Indicates study that produces this type of data (default: 'mouse_gastrulation')
modes	character() A wildcard / glob pattern of modes, such as "acc*". A wildcard of "*" will return all modes including Chromatin Accessibilty ("acc"), Methylation ("met"), RNA-seq ("rna") which is the default.
version	character(1) Either version '1.0.0' or '2.0.0' depending on data version required (default '1.0.0'). See version section.
format	character(1) Either MTX or HDF5 data format (default MTX)

dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
...	Additional arguments passed on to the ExperimentHub-class constructor

Details

Users are able to choose from either an MTX or HDF5 file format as the internal data representation. The MTX (Matrix Market) format allows users to load a sparse dgCMatrix representation. Choosing HDF5 gives users a sparse HDF5Array class object. * pbmc_10x: 10K Peripheral Blood Mononuclear Cells provided by 10x Genomics website Cell quality control filters are available in the object colData together with the celltype annotation labels.

Value

A 10X PBMC MultiAssayExperiment object

Examples

```
scMultiome(DataType = "pbmc_10x", modes = "*", dry.run = TRUE)
```

---

scNMT *Single-cell Nucleosome, Methylation and Transcription sequencing*

Description

scNMT assembles data on-the-fly from ExperimentHub to provide a MultiAssayExperiment container. The DataType argument provides access to the mouse_gastrulation dataset as obtained from Argelaguet et al. (2019; DOI: 10.1038/s41586-019-1825-8). Pre-processing code can be seen at https://github.com/rargelaguet/scnmt_gastrulation. Protocol information for this dataset is available at Clark et al. (2018). See the vignette for the full citation.

Usage

```
scNMT(
  DataType = "mouse_gastrulation",
  modes = "*",
  version = "1.0.0",
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

DataType	character(1) Indicates study that produces this type of data (default: 'mouse_gastrulation')
modes	character() A wildcard / glob pattern of modes, such as "acc*". A wildcard of "*" will return all modes including Chromatin Accessibilty ("acc"), Methylation ("met"), RNA-seq ("rna") which is the default.
version	character(1) Either version '1.0.0' or '2.0.0' depending on data version required (default '1.0.0'). See version section.
dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
...	Additional arguments passed on to the ExperimentHub-class constructor

Details

scNMT is a combination of RNA-seq (transcriptome) and an adaptation of Nucleosome Occupancy and Methylation sequencing (NOMe-seq, the methylome and chromatin accessibility) technologies. For more information, see Reik et al. (2018) DOI: 10.1038/s41467-018-03149-4

- mouse_gastrulation - this dataset provides cell quality control filters in the object colData starting from version 2.0.0. Additionally, cell types annotations are provided through the lineage colData column.
  - rna - RNA-seq
  - acc_* - chromatin accessibility
  - met_* - DNA methylation
    - cgi - CpG islands
    - CTCF - footprints of CTCF binding
    - DHS - DNase Hypersensitive Sites
    - genebody - gene bodies
    - p300 - p300 binding sites
    - promoter - gene promoters

Special thanks to Al J Abadi for preparing the published data in time for the 2020 BIRS Workshop, see the link here: https://github.com/BIRSBiointegration/Hackathon/tree/master/scNMT-seq

Value

A single cell multi-modal MultiAssayExperiment or informative data.frame when dry.run is TRUE

versions

Version '1.0.0' of the scNMT mouse_gastrulation dataset includes all of the above mentioned assay technologies with filtering of cells based on quality control metrics. Version '2.0.0' contains all of the cells without the QC filter and does not contain CTCF binding footprints or p300 binding sites.

metadata

The MultiAssayExperiment metadata includes the original function call that saves the function call and the data version requested.

Source

http://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation/

References

Argelaguet et al. (2019)

See Also

SingleCellMultiModal-package

Examples

```
scNMT(DataType = "mouse_gastrulation", modes = "*",
    version = "1.0.0", dry.run = TRUE)
```

SCoPE2 *Single-cell RNA sequencing and proteomics*

Description

SCoPE2 assembles data on-the-fly from ExperimentHub to provide a MultiAssayExperiment container. The DataType argument provides access to the SCoPE2 dataset as provided by Specht et al. (2020; DOI: http://dx.doi.org/10.1101/665307). The article provides more information about the data acquisition and pre-processing.

Usage

```
SCoPE2(
  DataType = "macrophage_differentiation",
  modes = "*",
  version = "1.0.0",
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

DataType	character(1) Indicates study that produces this type of data (default: 'macrophage_differentiation')
modes	character() A wildcard / glob pattern of modes, such as "rna". A wildcard of "*" will return all modes, that are transcriptome ("rna") or proteome ("protein") which is the default.
version	character(1), currently only version '1.0.0' is available
dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
...	Additional arguments passed on to the ExperimentHub-class constructor

Details

The SCoPE2 study combines scRNA-seq (transcriptome) and single-cell proteomics.

- macrophage_differentiation: the cells are monocytes that undergo macrophage differentiation. No annotation is available for the transcriptome data, but batch and cell type annotations are available for the proteomics data in the celltype colData column. The transcriptomics and proteomics data were not measured from the same cells but from a distinct set of cell cultures. This dataset provides already filtered bad quality cells.
  - scRNAseq1 - single-cell transcriptome (batch 1)
  - scRNAseq2 - single-cell transcriptome (batch 2)
  - scp - single-cell proteomics

Value

A single cell multi-modal MultiAssayExperiment or informative data.frame when dry.run is TRUE

Source

All files are linked from the slavovlab website https://scope2.slavovlab.net/docs/data

References

Specht, Harrison, Edward Emmott, Aleksandra A. Petelski, R. Gray Huffman, David H. Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. 2020. "Single-Cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity." bioRxiv. https://doi.org/10.1101/665307.

See Also

SingleCellMultiModal-package

Examples

```
SCoPE2(DataType = "macrophage_differentiation",
       modes = "*",
       version = "1.0.0",
       dry.run = TRUE)
```

`seqFISH` *Single-cell spatial + Gene Expression*

Description

seqFISH function assembles data on-the-fly from `ExperimentHub` to provide a `MultiAssayExperiment` container. Actually the `DataType` argument provides access to the available datasets associated to the package.

Usage

```
seqFISH(
  DataType = "mouse_visual_cortex",
  modes = "*",
  version,
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

`DataType`	`character(1)` indicating the identifier of the dataset to retrieve. (default "mouse_visual_cortex")
`modes`	`character()` The assay types or modes of data to obtain these include seq-FISH and scRNA-seq data by default.
`version`	`character(1)` Either version '1.0.0' or '2.0.0' depending on data version required (default '1.0.0'). See version section.
`dry.run`	`logical(1)` Whether to return the dataset names before actual download (default `TRUE`)
`verbose`	`logical(1)` Whether to show the dataset currently being (down)loaded (default `TRUE`)
`...`	Additional arguments passed on to the ExperimentHub-class constructor

Details

seq FISH data are a combination of single cell spatial coordinates and transcriptomics for a few hundreds of genes. seq-FISH data can be combined for example with scRNA-seq data to unveil multiple aspects of cellular behaviour based on their spatial organization and transcription.

Available datasets are:

- mouse_visual_cortex: combination of seq-FISH data as obtained from Zhu et al. (2018) and scRNA-seq data as obtained from Tasic et al. (2016), Version 1.0.0 returns the full scRNA-seq data matrix, while version 2.0.0 returns the processed and subsetted scRNA-seq data matrix (produced for the Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Types 2020 Workshop) The returned seqFISH data are always the processed ones for the same workshop. Additionally, cell types annotations are available in the `colData` through the `class` column in the seqFISH `assay`.
  - scRNA_Counts - Tasic scRNA-seq gene count matrix
  - scRNA_Labels - Tasic scRNA-seq cell labels
  - seqFISH_Coordinates - Zhu seq-FISH spatial coordinates
  - seqFISH_Counts - Zhu seq-FISH gene counts matrix
  - seqFISH_Labels - Zhu seq-FISH cell labels

Value

A `MultiAssayExperiment` of seq-FISH data

Author(s)

Dario Righelli <dario.righelli gmail.com>

Examples

```
seqFISH(DataType = "mouse_visual_cortex", modes = "*", version = "2.0.0",
    dry.run = TRUE)
```

SingleCellMultiModal *Combining Modalities into one MultiAssayExperiment*

Description

Combine multiple single cell modalities into one using the input of the individual functions.

Usage

```
SingleCellMultiModal(
  DataTypes,
  modes = "*",
  versions = "1.0.0",
  dry.run = TRUE,
  verbose = TRUE,
  ...
)
```

Arguments

DataTypes	character() A vector of data types as indicated in each individual function by the DataType parameter. These can be any of the following: "mouse_gastrulation", "pbmc_10x", "macrophage_differentiation", "cord_blood", "peripheral_blood", "mouse_visual_cortex", "mouse_embryo_8_cell"
modes	list() A list or CharacterList of modes for each data type where each element corresponds to one data type.
versions	character() A vector of versions for each DataType. By default, version 1.0.0 is obtained for all data types.
dry.run	logical(1) Whether to return the dataset names before actual download (default TRUE)
verbose	logical(1) Whether to show the dataset currently being (down)loaded (default TRUE)
...	Additional arguments passed on to the ExperimentHub-class constructor

Value

A multi-modality MultiAssayExperiment

metadata

The metadata in the MultiAssayExperiment contains the original function call used to generate the object (labeled as call), a call_map which provides traceability of technology functions to DataType prefixes, and lastly, R version information as version.

Examples

```
SingleCellMultiModal(c("mouse_gastrulation", "pbmc_10x"),
    modes = list(c("acc*", "met*"), "rna"),
    version = c("2.0.0", "1.0.0"), dry.run = TRUE, verbose = TRUE
)
```

Index