# Package ‘scRNAseq’

April 14, 2020

**Title** Collection of Public Single-Cell RNA-Seq Datasets

**Version** 2.0.2

**Date** 2019-11-10

**Description** Gene-level counts for a collection of public scRNA-seq datasets, provided as SingleCellExperiment objects with cell- and gene-level metadata.

**License** CC0

**NeedsCompilation** no

**Depends** SingleCellExperiment

**Imports** SummarizedExperiment, S4Vectors, BiocGenerics, ExperimentHub

**Suggests** BiocStyle, AnnotationHub, AnnotationDbi, knitr, rmarkdown, BiocFileCache, testthat

**VignetteBuilder** knitr

**Encoding** UTF-8

**biocViews** ExperimentHub, ExperimentData, ExpressionData, SequencingData, RNASeqData, SingleCellData

**BuildResaveData** no

**RoxygenNote** 6.1.1

**git_url** https://git.bioconductor.org/packages/scRNAseq

**git_branch** RELEASE_3_10

**git_last_commit** c087c40

**git_last_commit_date** 2019-11-11

**Date/Publication** 2020-04-14

**Author** Davide Risso [aut, cre, cph],
Michael Cole [aut],
Aaron Lun [ctb]

**Maintainer** Davide Risso <risso.davide@gmail.com>

## R topics documented:

---

`scRNAseq-package` *Collection of Public Single-Cell RNA-Seq Datasets*

---

## Description

Gene-level counts for a collection of public scRNA-seq datasets, provided as SingleCellExperiment objects with cell- and gene-level metadata.

## Details

This package contains a collection of three publicly available single-cell RNA-seq datasets.

The dataset `fluidigm` contains 65 cells from Pollen et al. (2014), each sequenced at high and low coverage.

The dataset `th2` contains 96 T helper cells from Mahata et al. (2014).

The dataset `allen` contains 379 cells from the mouse visual cortex. This is a subset of the data published in Tasic et al. (2016).

See the package vignette for details on the pre-processing of the data.

## Author(s)

NA

Maintainer: NA

### References

Pollen, Nowakowski, Shuga, Wang, Leyrat, Lui, Li, Szpankowski, Fowler, Chen, Ramalingam, Sun, Thu, Norris, Lebofsky, Toppani, Kemp II, Wong, Clerkson, Jones, Wu, Knutsson, Alvarado, Wang, Weaver, May, Jones, Unger, Kriegstein, West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, 32, 1053-1058 (2014).

Mahata, Zhang, Kolodziejczyk, Proserpio, Haim-Vilmovsky, Taylor, Hebenstreit, Dingler, Moignard, Gottgens, Arlt, McKenzie, Teichmann. Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. Cell Reports, 7(4): 1130–1142 (2014).

Tasic, Menon, Nguyen, Kim, Jarsky, Yao, Levi, Gray, Sorensen, Dolbeare, Bertagnolli, Goldy, Shapovalova, Parry, Lee, Smith, Bernard, Madisen, Sunkin, Hawrylycz, Koch, Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience, 19, 335–346 (2016).

---

## `AztekinTailData` *Obtain the Aztekin tail data*

---

### Description

Obtain the Xenopus tail single-cell RNA-seq data from Aztekin et al. (2019).

### Usage

```
AztekinTailData()
```

### Details

Column metadata is provided in the same form as supplied in E-MTAB-7761. This contains information such as the treatment condition, batch, putative cell type, putative cell cycle phase.

The UMAP results are available as the `"UMAP"` entry in the `reducedDims`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/aztekin-tail`.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun

### References

Aztekin C et al. (2019). Identification of a regeneration-organizing cell in the Xenopus tail. *Science* 364(6441), 653-658

### Examples

```
sce <- AztekinTailData()
```

## BachMammaryData — *Obtain the Bach mammary data*

### Description

Obtain the mouse mammary gland single-cell RNA-seq data from Bach et al. (2017).

### Usage

```
BachMammaryData(samples = c("NP_1", "NP_2", "G_1", "G_2", "L_1", "L_2",
  "PI_1", "PI_2"))
```

### Arguments

| | |
|---|---|
| samples | A character vector with at least one element, specifying which samples(s) to retrieve. |

### Details

Column metadata is extracted from the sample annotation in GSE106273, and refers to the developmental stage of the mammary gland.

If multiple samples are specified in `samples`, the count matrices will be `cbind`ed together. Cells originating from different samples are identifiable by the `"Sample"` field in the column metadata.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/bach-mammary`.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun

### References

Bach K et al. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun.* 8(1), 2128

### Examples

```
sce <- BachMammaryData()
```

## BaronPancreasData — *Obtain the Baron pancreas data*

### Description

Obtain the human/mouse pancreas single-cell RNA-seq data from Baron et al. (2017).

### Usage

```
BaronPancreasData(which = c("human", "mouse"), ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| `which` | String specifying the species to get data for. |
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Column metadata is provided in the same form as supplied in GSE84133. This contains information such as the cell type labels and donor ID (for humans) or strain (for mouse).

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/baron-pancreas`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Baron M et al. (2017). Single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* 3(4), 346-360.

### Examples

```
sce.human <- BaronPancreasData()

sce.mouse <- BaronPancreasData("mouse")
```

`BuettnerESCData` *Obtain the Buettner ESC data*

## Description

Obtain the mouse embryonic stem cell single-cell RNA-seq data from Buettner et al. (2015).

## Usage

```
BuettnerESCData(remove.htseq = TRUE)
```

## Arguments

| | |
|---|---|
| `remove.htseq` | Logical scalar indicating whether HT-seq alignment statistics should be removed. |

## Details

Rows corresponding to HT-seq's alignment statistics are removed by default. These can be retained by setting `remove.htseq=FALSE`.

Column metadata contains the experimentally determined cell cycle phase for each cell.

Counts for ERCC spike-ins are stored in the `"ERCC"` entry in the `altExps`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/buettner-esc`.

## Value

A SingleCellExperiment object with a single matrix of read counts.

## Author(s)

Aaron Lun

## References

Buettner F et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33(2), 155-160.

## Examples

```
sce <- BuettnerESCData()
```

`CampbellBrainData` *Obtain the Campbell brain data*

## Description

Obtain the mouse brain single-cell RNA-seq data from Campbell et al. (2017).

## Usage

```
CampbellBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the row names of the returned object should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in GSE93374. This contains information such as the diet of the mice, sex and proposed cell type for each cell.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/campbell-brain`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

Campbell R et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20, 484-496.

## Examples

```
sce <- CampbellBrainData()
```

ChenBrainData *Obtain the Chen brain data*

## Description

Obtain the mouse brain single-cell RNA-seq data from Chen et al. (2017).

## Usage

```
ChenBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in GSE87544. This contains the putative cell type assigned by the original authors.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/chen-brain`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

Chen R et al. (2017). Single-Cell RNA-Seq reveals hypothalamic cell diversity. *Cell Rep.* 18, 3227-3241.

## Examples

```
sce <- ChenBrainData()
```

## GrunHSCData — *Obtain the Grun HSC data*

### Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq data from Grun et al. (2016).

### Usage

```
GrunHSCData(ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Row metadata contains the symbol and chromosomal location for each gene. Column metadata contains the extraction protocol used for each sample, as described in GSE76983.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/grun-hsc`.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun

### References

Grun D et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19(2), 266-277.

### Examples

```
sce <- GrunHSCData()
```

GrunPancreasData *Obtain the Grun pancreas data*

## Description

Obtain the human pancreas single-cell RNA-seq data from Grun et al. (2016).

## Usage

```
GrunPancreasData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| ensembl | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Row metadata contains fields for the symbol and chromosomal location of each gene, as derived from the row names.

Column metadata is derived from the column names of the count matrix with the sample annotations in GSE81076. This includes the donor identity for each cell and the type of sample.

The "ERCC" entry in the altExps contains count data for the ERCC spike-in transcripts.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/grun-pancreas.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

## References

Grun D et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19(2), 266-277.

## Examples

```
sce <- GrunPancreasData()
```

`KolodziejczykESCData` *Obtain the Kolodziejcyzk ESC data*

### Description

Obtain the mouse embryonic stem cell single-cell RNA-seq data from Kolodziejczyk et al. (2015).

### Usage

```
KolodziejczykESCData(remove.htseq = TRUE)
```

### Arguments

| | |
|---|---|
| `remove.htseq` | Logical scalar indicating whether HT-seq alignment statistics should be removed. |

### Details

Column metadata is generated from the column names, and contains the culture conditions and the plate of origin for each cell.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry in the `altExps`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/kolodziejczyk-esc`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Messmer T et al. (2019). Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep* 26(4), 815-824.e4

### Examples

```
sce <- KolodziejczykESCData()
```

`LaMannoBrainData` *Obtain the La Manno brain data*

## Description

Obtain the mouse/human brain scRNA-seq data from La Manno et al. (2016).

## Usage

```
LaMannoBrainData(which = c("human-es", "human-embryo", "human-ips",
  "mouse-adult", "mouse-embryo"), ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `which` | A string specifying which dataset should be obtained. |
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in the supplementary tables in GSE71585. This contains information such as the time point and cell type.

The various settings of `which` will obtain different data sets.

- `"human-es"`, human embryonic stem cells.
- `"human-embryo"`, human embryo midbrain.
- `"human-ips"`, human induced pluripotent stem cells.
- `"mouse-adult"`, mouse adult dopaminergic neurons.
- `"mouse-embryo"`, mouse embryo midbrain.

Unfortunately, each of these datasets uses a different set of features. If multiple datasets are to be used simultaneously, users will have to decide how to merge them, e.g., by taking the intersection of common features across all datasets.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lamanno-brain`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

La Manno A et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 167(2), 566-580.

### Examples

```
sce.h.es <- LaMannoBrainData()

sce.h.em <- LaMannoBrainData("human-embryo")

sce.h.ip <- LaMannoBrainData("human-ips")

sce.m.ad <- LaMannoBrainData("mouse-adult")

sce.m.em <- LaMannoBrainData("mouse-embryo")
```

---

## LawlorPancreasData *Obtain the Lawlor pancreas data*

---

### Description

Provides the human pancreas single-cell RNA-seq data from Lawlor et al. (2017).

### Usage

```
LawlorPancreasData()
```

### Details

Column metadata is provided in the same form as supplied in GSE86469. This contains information such as the cell type labels and patient status.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lawlor-pancreas`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Lawlor N et al. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* 27(2), 208-222.

### Examples

```
sce <- LawlorPancreasData()
```

`LengESCData` *Obtain the Leng ESC data*

## Description

Obtain the human embryonic stem cell single-cell RNA-seq data from Leng et al. (2015).

## Usage

```
LengESCData()
```

## Details

Column metadata contains the cell line, experiment number and experimentally determined cell cycle phase for each cell,

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/leng-esc`.

## Value

A SingleCellExperiment object with a single matrix of normalized expected read counts.

## Author(s)

Aaron Lun

## References

Leng F et al. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* 12(10), 947-950.

## Examples

```
sce <- LengESCData()
```

`LunSpikeInData` *Obtain the Lun spike-in data*

## Description

Obtain the spike-in single-cell RNA-seq data from Lun et al. (2017).

## Usage

```
LunSpikeInData(which = c("416b", "tropho"))
```

## Arguments

| | |
|---|---|
| `which` | String specifying whether the 416B or trophoblast data should be obtained. |

## Details

Row data contains a single "Length" field describing the total exonic length of each feature.

Column metadata is provided in the same form as supplied in E-MTAB-5522. This contains information such as the cell type, plate of origin, spike-in addition order and oncogene induction.

Two sets of spike-ins were added to each cell in each dataset. These are available as the "SIRV" and "ERCC" entries in the `altExps`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/lun-spikein`.

## Value

A SingleCellExperiment object with a single matrix of read counts.

## Author(s)

Aaron Lun

## References

Lun ATL et al. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* 27(11), 1795-1806.

## Examples

```
sce <- LunSpikeInData()

sce <- LunSpikeInData("tropho")
```

---

# MacoskoRetinaData *Obtain the Macosko retina data*

---

## Description

Obtain the mouse retina single-cell RNA-seq data from Macosko et al. (2016).

## Usage

```
MacoskoRetinaData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata contains the cluster identity as reported in the paper. Note that some cells will have `NA` identities as they are present in the count matrix but not in the metadata file. These are presumably low-quality cells that were discarded prior to clustering.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/macosko-retina`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

Macosko E et al. (2016). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161(5), 1202-1214.

## Examples

```
sce <- MacoskoRetinaData()
```

---

# `MarquesBrainData` *Obtain the Marques brain data*

---

## Description

Obtain the mouse brain single-cell RNA-seq data from Marques et al. (2016).

## Usage

```
MarquesBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in GSE75330. This contains information such as the cell type and age/sex of the mouse of origin for each cell.

Note that some genes may be present in multiple rows corresponding to different genomic locations. These additional rows are identified by a `_loc[2-9]` suffix in their row names. Users may wish to consider either removing them or merging them, e.g., with `scater::sumCountsAcrossFeatures`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained. All searching is performed after removing the `_loc[2-9]` suffix.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/marques-brain`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

Marques A et al. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* 352(6291), 1326-1329.

## Examples

```
sce <- MarquesBrainData()
```

---

`MessmerESCData` *Obtain the Messmer ESC data*

---

## Description

Obtain the human embryonic stem cell single-cell RNA-seq data from Messmer et al. (2019).

## Usage

```
MessmerESCData()
```

## Details

Row data contains a single `"Length"` field describing the total exonic length of each feature.

Column metadata is provided in the same form as supplied in E-MTAB-6819. This contains information such as the cell phenotype (naive or primed) and the batch of origin. Note that counts for technical replicates have already been summed together.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/messmer-esc`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Messmer T et al. (2019). Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep* 26(4), 815-824.e4

### Examples

```
sce <- MessmerESCData()
```

---

## `MuraroPancreasData` *Obtain the Muraro pancreas data*

---

### Description

Obtain the human pancreas single-cell RNA-seq data from Muraro et al. (2016).

### Usage

```
MuraroPancreasData(ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Row data contains fields for the symbol and chromosomal location of each gene.

Column metadata is derived from the columns of the count matrix provided in GSE85241, with additional cell type labels obtained from the authors (indirectly, via the Hemberg group). Some cells have `NA` labels and were presumably removed prior to downstream analyses.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/muraro-pancreas`.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

### References

Muraro MJ et al. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* 3(4), 385-394.

### Examples

```
sce <- MuraroPancreasData()
```

---

## NestorowaHSCData *Obtain the Nestorowa HSC data*

---

### Description

Obtain the mouse haematopoietic stem cell single-cell RNA-seq data from Nestorowa et al. (2015).

### Usage

```
NestorowaHSCData(remove.htseq = TRUE)
```

### Arguments

| | |
|---|---|
| `remove.htseq` | Logical scalar indicating whether HT-seq alignment statistics should be removed. |

### Details

Rows corresponding to HT-seq's alignment statistics are removed by default. These can be retained by setting `remove.htseq=FALSE`.

Column metadata includes the cell type mapping, as described on the website (see References), and the FACS expression levels of selected markers. Note that these are stored as nested matrices within the `colData`.

Diffusion map components are provided as the `"diffusion"` entry in the `reducedDims`.

Counts for ERCC spike-ins are stored in the `"ERCC"` entry in the `altExps`.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/nestorowa-hsc`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

## References

Nestorowa S et al. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation *Blood* 128, e20-e31.

Gene and protein expression in adult haematopoiesis: Data. http://blood.stemcells.cam.ac.uk/single_cell_atlas.html#data.

## Examples

```
sce <- NestorowaHSCData()
```

---

ReprocessedAllenData *Reprocessed single-cell data sets*

---

## Description

Obtain the legacy count matrices for three publicly available single-cell RNA-seq datasets. Raw sequencing data were downloaded from NCBI's SRA or from EBI's ArrayExpress, aligned to the relevant genome build and used to quantify gene expression.

## Usage

```
ReprocessedAllenData(assays = NULL, ensembl = FALSE)

ReprocessedTh2Data(assays = NULL, ensembl = FALSE)

ReprocessedFluidigmData(assays = NULL, ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| assays | Character vector specifying one or more assays to return. Choices are "tophat_counts", "cufflinks_fpkm", "rsem_counts" and "rsem_tpm". If NULL, all assays are returned. |
| ensembl | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

ReprocessedFluidigmData returns a dataset of 65 human neural cells from Pollen et al. (2014), each sequenced at high and low coverage (SRA accession SRP041736).

ReprocessedTh2Data returns a dataset of 96 mouse T helper cells from Mahata et al. (2014), obtained from ArrayExpress accession E-MTAB-2512. Spike-in counts are stored in the "ERCC" entry of the altExps.

ReprocessedAllenData return a dataset of 379 mouse brain cells from Tasic et al. (2016). This is a re-processed subset of the data from TasicBrainData, and contains spike-in information stored as in the altExps.

In each dataset, the first columns of the colData are sample quality metrics from FastQC and Picard. The remaining fields were obtained from the original study in their GEO/SRA submission and/or as Supplementary files in the associated publication. These two categories of colData are distinguished by a `which_qc` element in the `metadata`, which contains the names of the quality-related columns in each object.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/legacy-allen`, `scRNAseq/legacy-fluidigm` or `scRNAseq/legacy-th2`.

### Value

A SingleCellExperiment object containing one or more expression matrices of counts and/or TPMs, depending on `assays`.

### Pre-processing details

FASTQ files were either obtained directly from ArrayExpress, or converted from SRA files (downloaded from the Sequence Read Archive) using the SRA Toolkit.

Reads were aligned with TopHat (v. 2.0.11) to the appropriate reference genome (GRCh38 for human samples, GRCm38 for mouse). RefSeq mouse gene annotation (GCF_000001635.23_GRCm38.p3) was downloaded from NCBI on Dec. 28, 2014. RefSeq human gene annotation (GCF_000001405.28) was downloaded from NCBI on Jun. 22, 2015.

featureCounts (v. 1.4.6-p3) was used to compute gene-level read counts. Cufflinks (v. 2.2.0) was used to compute gene-leve FPKMs. Reads were also mapped to the transcriptome using RSEM (v. 1.2.19) to compute read counts and TPM's.

FastQC (v. 0.10.1) and Picard (v. 1.128) were used to compute sample quality control (QC) metrics. However, no filtering on the QC metrics has been performed for any dataset.

### References

Pollen AA et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32(10), 1053-8.

Mahata B et al. (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep*, 7(4), 1130-42.

Tasic A et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19(2), 335-46.

### Examples

```
sce <- ReprocessedAllenData()
```

---

## `RichardTCellData` *Obtain the Richard T cell data*

---

### Description

Obtain the mouse CD8+ T cell single-cell RNA-seq data from Richard et al. (2018).

### Usage

```
RichardTCellData()
```

### Details

Column metadata is provided in the same form as supplied in E-MTAB-6051. This contains information such as the stimulus, time after stimulation, age of the mice and sequencing batch.

Count data for ERCC spike-ins are stored in the "ERCC" entry of the altExps.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/richard-tcell.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Richard AC et al. (2018). T cell cytolytic capacity is independent of initial stimulation strength. *Nat. Immunol.* 19(8), 849-858.

### Examples

```
sce <- RichardTCellData()
```

---

## RomanovBrainData    *Obtain the Romanov brain data*

---

### Description

Obtain the mouse brain single-cell RNA-seq dataset from Romanov et al. (2017).

### Usage

```
RomanovBrainData(ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| ensembl | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Column metadata is provided in the same form as supplied in GSE74672. This contains information such as the reporter gene expressed in each cell, the mouse line, dissection type and so on.

Counts for ERCC spike-ins are stored in the "ERCC" entry of the altExps. Note that some of the spike-in rows have NA observations for some (but not all) cells.

If ensembl=TRUE, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for scRNAseq/romanov-brain.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun, based on code by Vladimir Kiselev and Tallulah Andrews.

### References

Romanov RA et al. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* 20, 176-188.

### Examples

```
sce <- RomanovBrainData()
```

---

## SegerstolpePancreasData — *Obtain the Segerstolpe pancreas data*

---

### Description

Download the human pancreas single-cell RNA-seq (scRNA-seq) dataset from Segerstolpe et al. (2016)

### Usage

```
SegerstolpePancreasData(ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Row data contains fields for the gene symbol and RefSeq transcript IDs corresponding to each gene. The rows of the output object are named with the symbol, but note that these are not unique.

Column metadata were extracted from the `Characteristics` fields of the SDRF file for ArrayExpress E-MTAB-5061. This contains information such as the cell type labels and patient status.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/segerstolpe-pancreas`.

### Value

A SingleCellExperiment object with a single matrix of read counts.

### Author(s)

Aaron Lun

### References

Segerstolpe A et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24(4), 593-607.

### Examples

```
sce <- SegerstolpePancreasData()
```

---

## `ShekharRetinaData` *Obtain the Shekhar retina data*

---

### Description

Obtain the mouse retina single-cell RNA-seq dataset from Shekhar et al. (2016).

### Usage

```
ShekharRetinaData(ensembl = FALSE)
```

### Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

### Details

Column metadata contains the cluster identities as reported in the paper. Note that some cells will have `NA` identities as they are present in the count matrix but not in the metadata file. These are presumably low-quality cells that were discarded prior to clustering.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/shekhar-retina`.

### Value

A SingleCellExperiment object with a single matrix of UMI counts.

### Author(s)

Aaron Lun

### References

Shekhar K et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166(5), 1308-1323.

## Examples

```
sce <- ShekharRetinaData()
```

---

`TasicBrainData` *Obtain the Tasic brain data*

---

## Description

Obtain the mouse brain single-cell RNA-seq data from Tasic et al. (2015).

## Usage

```
TasicBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in GSE71585. This contains information such as the reporter gene expressed in each cell, the mouse line, dissection type and so on.

Count data for ERCC spike-ins are stored in the `"ERCC"` entry of the `altExps`. Note that some of the spike-in rows have `NA` observations for some (but not all) cells.

The last 9 columns (containing `_CTX_` in their names) correspond to no-cell control libraries.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/tasic-brain`.

## Value

A SingleCellExperiment object with a single matrix of read counts.

## Author(s)

Aaron Lun

## References

Tasic A et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19(2), 335-46.

## Examples

```
sce <- TasicBrainData()
```

`UsoskinBrainData` *Obtain the Usoskin brain data*

## Description

Obtain the mouse brain single-cell RNA-seq dataset from Usoskin et al. (2015).

## Usage

```
UsoskinBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Column metadata is provided in the same form as supplied in External Table 2 of `http://linnarssonlab.org/drg/`. This contains information such as the library of origin and the cell type.

The count matrix contains information for repeats, marked with `r_` prefixes in the row names; as well as mitochondrial transcripts, marked with `mt-` prefixes.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/usoskin-brain`.

## Value

A SingleCellExperiment object with a single matrix of RPMs.

## Author(s)

Aaron Lun

## References

Usoskin A et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18(1), 145-53.

## Examples

```
sce <- UsoskinBrainData()
```

---

`XinPancreasData` *Obtain the Xin pancreas data*

---

## Description

Obtain the human pancreas single-cell RNA-seq dataset from Xin et al. (2016).

## Usage

```
XinPancreasData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Row data contains fields for the Entrez ID and symbol for each gene. Column metadata was obtained from the authors (indirectly, via the Hemberg group) and contains information such as the cell type labels and donor status.

If `ensembl=TRUE`, the Entrez IDs are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/xin-pancreas`.

## Value

A SingleCellExperiment object with a single matrix of RPKMs.

## Author(s)

Aaron Lun, using additional metadata obtained by Vladimir Kiselev.

## References

Xin A et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24(4), 608-615.

## Examples

```
sce <- XinPancreasData()
```

ZeiselBrainData *Obtain the Zeisel brain data*

## Description

Obtain the mouse brain single-cell RNA-seq dataset from Zeisel et al. (2015).

## Usage

```
ZeiselBrainData(ensembl = FALSE)
```

## Arguments

| | |
|---|---|
| `ensembl` | Logical scalar indicating whether the output row names should contain Ensembl identifiers. |

## Details

Row data contains a single `"featureType"` field describing the type of each feature (endogenous genes, mitochondrial genes, spike-in transcripts and repeats). Spike-ins and repeats are stored as separate entries in the `altExps`.

Column metadata is provided in the same form as supplied in `http://linnarssonlab.org/cortex/`. This contains information such as the cell diameter and the published cell type annotations.

If `ensembl=TRUE`, the gene symbols are converted to Ensembl IDs in the row names of the output object. Rows with missing Ensembl IDs are discarded, and only the first occurrence of duplicated IDs is retained.

All data are downloaded from ExperimentHub and cached for local re-use. Specific resources can be retrieved by searching for `scRNAseq/zeisel-brain`.

## Value

A SingleCellExperiment object with a single matrix of UMI counts.

## Author(s)

Aaron Lun

## References

Zeisel A et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226), 1138-42.

## Examples

```
sce <- ZeiselBrainData()
```

# Index