Package 'MouseGastrulationData'

December 10, 2024

Title Single-Cell -omics Data across Mouse Gastrulation and Early Organogenesis

Version 1.20.0

Description Provides processed and raw count data for single-cell RNA sequencing, single-cell ATAC-seq, and seqFISH (spatial transcriptomic) experiments performed along a timecourse of mouse gastrulation and early organogenesis.

Depends R (>= 4.1), SingleCellExperiment, SummarizedExperiment, SpatialExperiment

Imports methods, ExperimentHub, BiocGenerics, S4Vectors, BumpyMatrix

Suggests BiocStyle, knitr, rmarkdown, testthat

VignetteBuilder knitr

License GPL-3

NeedsCompilation no

Encoding UTF-8

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

URL https://github.com/MarioniLab/MouseGastrulationData

BugReports https://github.com/MarioniLab/MouseGastrulationData/issues

RoxygenNote 7.3.0

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

git_branch RELEASE_3_20

git_last_commit a8d1ad2

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-12-10

Author Jonathan Griffiths [aut, cre], Aaron Lun [aut]

Maintainer Jonathan Griffiths <jonathan.griffiths.94@gmail.com>

Contents

MouseGastrulationData-package 2
AtlasSampleMetadata 3
BPSATACData 4
EmbryoAtlasData 6
EmbryoCelltypeColours 8
GuibentifExtraData 8
LohoffSeqFISHData 9
RAMultiomeData 11
RASampleMetadata 13
Tal1ChimeraData 13
TChimeraData 15
WTChimeraData 17

Index **20**

`MouseGastrulationData-package`
Single-Cell -omics Data across Mouse Gastrulation and Early Organogenesis

---

Description

Provides processed and raw count data for single-cell RNA sequencing, single-cell ATAC-seq, and seqFISH (spatial transcriptomic) experiments performed along a timecourse of mouse gastrulation and early organogenesis.

Details

This package contains the processed 10X Genomics data from Pijuan-Sala et al. (2019), Pijuan-Sala et al. (2020), Guibentif et al. (2020) and Lohoff et al. (2020.).

The data were processed as described in the methods that accompany the paper.

Author(s)

Jonathan Griffiths [aut, cre], Aaron Lun [aut]

Maintainer: Jonathan Griffiths <jonathan.griffiths.94@gmail.com>

References

Blanca Pijuan-Sala*, Jonathan A. Griffiths*, Carolina Guibentif*, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C.V. Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D. Simons, Jennifer Nichols, John C. Marioni, Berthold Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature, 566, pp490-495 (2019).

Blanca Pijuan-Sala, Nicola K. Wilson, Jun Xia, Xiaomeng Hou, Rebecca L. Hannah, Sarah Kinston, Fernando J. Calero-Nieto, Olivier Poirion, Sebastian Preissl, Feng Liu, Berthold Göttgens. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. Nat Cell Biol, 22, 487–497 (2020).

Carolina Guibentif*, Jonathan A. Griffiths*, Ivan Imaz-Rosshandler, Shila Ghazanfar, Jennifer Nichols, Valerie Wilson, Berthold Göttgens, John C. Marioni. Diverse Routes toward Early Somites in the Mouse Embryo. Dev Cell. 2020 Dec 1:S1534-5807(20)30889-3.

T. Lohoff*, S. Ghazanfar*, A. Missarova, N. Koulena, N. Pierson, J.A. Griffiths, E.S. Bardot, C.-H.L. Eng, R.C.V. Tyser, R. Argelaguet, C. Guibentif, S. Srinivas, J. Briscoe, B.D. Simons, A.-K. Hadjantonakis, B. Göttgens, W. Reik, J. Nichols, L. Cai, J.C. Marioni. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. bioRxiv 2020.11.20.391896.

`AtlasSampleMetadata` *Sample metadata from the Pijuan-Sala et al. embryo atlas*

Description

A data frame containing stage and embryo pool information for the atlas dataset.

Usage

```
AtlasSampleMetadata
```

Format

A data frame containing information for each 10x sample of the embryo atlas. This object contains:

`sample`**:** Integer, 10x sample index.

`stage`**:** Character, developmental stage from which sample was generated.

`pool_index`**:** Integer, index for pools of embryos; samples with the same values are from the same pool of dissociated cells.

`seq_batch`**:** Integer, seqencing batch index; samples with the same values were multiplexed for sequencing.

`ncells`**:** Integer, number of cells (post-QC) per sample.

Note that sample 11 is missing by design due to experimental failure: it is not available for download.

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
head(AtlasSampleMetadata)
```

BPSATACData *E8.25 snATAC-seq data*

Description

Obtain the processed or raw counts for the Pijuan-Sala et al. (2020) E8.25 single-nucleus ATAC-seq dataset.

Usage

```
BPSATACData(type = c("processed", "raw"), Csparse.assays = TRUE)
```

Arguments

type	String specifying the type of data to obtain, see Details. Default behaviour is to return processed data.
Csparse.assays	Logical indicating whether to convert assay matrices into the column major format that is more performant with contemporary software packages. Default behaviour is to perform the conversion.

Details

This function downloads the data for the E8.25 single-nucleus ATAC-seq data from Pijuan-Sala et al. (2020). The dataset is provided as a single sample.

In the processed data, QC-passing libraries have already been identified in each sample. The count matrix contains the number of counts for each identified peak for each cell. Note that you may want to binarise this matrix for downstream analyses. Full details of the methods used in analyses can be found in the paper (see References, below).

The column metadata for cells contains:

`sample`: Integer, sample index (for consistency across MGD datasets).

`stage`: Character, collection timepoint (for consistency across MGD datasets).

`barcode`: Character, unique cell identifier.

`nuclei_type`: Character, whether cells were selected using flow gates. Note that these are probably not doublets, but cells in different cell cycle phases.

`num_of_reads`: Integer, number of reads.

`promoter_coverage`: Numeric, fraction of promoters accessible "in the majority of datasets based on ENCODE DNase Hypersensitive Sites and ATAC-seq data".

`read_in_promoter`: Integer, number of reads in promoters.

`doublet_scores`: Numeric, doublet scores (calculated with scrublet v0.4).

`read_in_peak`: Integer, Number of reads in across-cell-calculated peaks.

`ratio_peaks`: Numeric, fraction of reads in across-cell peaks.

`final_clusters`: Integer, final cluster indices.

`celltype`: Character, celltype label.

`al_haem_endo_clusters`: Character, clusters from the focused blood, allantois, endothelium celltypes (or NA, for other celltypes).

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object. These are `topics` and `umap`. Please see the methods of the manuscript (see References, below) for more details on the topic modelling approach.

For both raw and processed data, the row metadata is relatively complex. It contains:

`peakID`: Character, unique peak identifier.

`peak_chr`: Character, chromosome ID for each peak.

`peak_start`: Integer, start position for each peak. As this is from a bed file (I think), this is 0-indexed, and the peak is inclusive of this position.

`peak_end`: Integer, end position for each peak. As this is from a bed file (I think), this is 0-indexed, and the peak is exclusive of this position.

`Annotation.General`: Character, general peak annotation (TSS (-1kb to +100bp), TTS (-100bp to +1kb), intron, exon, intergenic).

`distance_from_TSS`: Integer, distance from the TSS that peaks been annotated to if the region is intergenic. Note: the authors have annotated peaks to multiple genes; distances for different genes are comma-separated in this column.

`geneName`: Character, gene name (MGI). Note: the authors have annotated peaks to multiple genes; names for different genes are comma-separated in this column.

`geneID`: Character, gene ID (Ensembl gene ID, v92). Note: the authors have annotated peaks to multiple genes; IDs for different genes are comma-separated in this column.

`strand`: Character, strand for linked genes. Note: the authors have annotated peaks to multiple genes; strands for different genes are comma-separated in this column.

`celltype_specificity`: Character, celltype specificity of the peak. For multiple celltypes, authors have semicolon-separated celltype names.

`topic`: Character, topic membership of the peak. For multiple topics, authors have semicolon-separated topic names.

`topic_stringent`: Character, topic membership of the peak if it contributes to only a single topic; else "Nonspecific".

`accessibility`: Integer, number of nuclei with where peak is accessible.

`accessibility_log`: Numeric, log-transformed number of nuclei with where peak is accessible (base e, with an added 1 to the count).

`accessibility_ratio`: Numeric, fraction of nuclei where peak is accessible.

`umap_X`: Numeric, umap x-coordinate of peak.

`umap_Y`: Numeric, umap y-coordinate of peak.

`Pattern_endothelium`: Integer, index for dynamic pattern during endothelial establishment (else NA).

Value

If `type="processed"`, a SingleCellExperiment is returned containing the processed data.

If `type="raw"`, a SingleCellExperiment is returned containing the raw data.

Author(s)

Aaron Lun, with modification by Jonathan Griffiths

References

Pijuan-Sala B et al. (2020). Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nature Cell Biology* 22, 4:487–97.

Examples

```
## Not run:
# dataset large enough to cause bioc build issues
atac.data <- BPSATACData()
atac.data <- BPSATACData(type="processed")

## End(Not run)
```

`EmbryoAtlasData` *Mouse gastrulation timecourse data*

Description

Obtain the processed or raw counts for the mouse gastrulation scRNAseq dataset.

Usage

```
EmbryoAtlasData(
  type = c("processed", "raw"),
  samples = NULL,
  get.spliced = FALSE,
  Csparse.assays = TRUE
)
```

Arguments

`type`	String specifying the type of data to obtain, see Details. Default behaviour is to return processed data.
`samples`	Integer or character vector specifying the samples for which data (processed or raw) should be obtained. If `NULL` (default), data are returned for all (36) samples.
`get.spliced`	Logical indicating whether to also download the spliced/unspliced/ambiguously spliced count matrices.
`Csparse.assays`	Logical indicating whether to convert assay matrices into the column major format that is more performant with contemporary software packages. Default behaviour is to perform the conversion.

Details

This function downloads the data for the embryo atlas from Pijuan-Sala et al. (2019). The dataset contains 36 10X Genomics samples; sample 11 is absent due to QC failure. The `AtlasSampleMetadata` variable contains information about each of these samples.

In the processed data, cell-containing libraries have already been identified in each sample using the `emptyDrops` function from **DropletUtils**. The count matrix contains the raw count vectors for the cells called from all samples in this manner. Size factors were computed using the `computeSumFactors` function from **scran**. The column metadata for called cells contains:

`cell`: Character, unique cell identifier across all samples.

`barcode`: Character, cell barcode from the 10X Genomics experiment.

`sample`: Integer, index of the sample from which the cell was taken.

`pool`: Integer, index of the embryo pool from which the sample derived. Samples with the same value are technical, not biological, replicates

`stage`: Character, stage of the mouse embryo at which the sample was taken.

`sequencing.batch`: Integer, sequencing run in which sample was multiplexed.

`theiler`: Character, Theiler stage from which the sample was taken; alternative scheme to `stage`.

`doub.density`: Numeric, output of (a now-outdated run of) `scran::doubletCells`, performed on each sample separately.

`doublet`: Logical, whether a cell was called as a doublet.

`cluster`: Integer, top-level cluster to which cell was assigned across all samples.

`cluster.sub`: Integer, cluster to which cell was assigned when clustered within each `cluster`.

`cluster.stage`: Integer, top-level cluster to which cell was assigned within individual timepoints.

`cluster.theiler`: Integer, top-level cluster to which cell was assigned within individual Theiler stages.

`stripped`: Logical, whether a cell was called as a cytoplasm-stripped nucleus.

`celltype`: Character, cell type to which the cell was assigned.

`colour`: Integer, cell type colour (hex) as in Pijuan-Sala et al. (2019).

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object. These are `pca.corrected` and `umap`.

If spliced counts were requested, these will be in the assays slot of the SingleCellExperiment object. Spliced count matrices were collated using *velocyto* version 0.17.17. Spliced count matrices will not have had swapped molecules removed, as *velocyto* and `DropletUtils::swappedDrops` are not compatible. However, these should still be effective for calculating RNA velocity estimates using various different tools.

The raw data contains the unfiltered count matrix for each sample, as generated directly from the CellRanger software. Swapped molecules have been removed using `DropletUtils::swappedDrops`. No filtering has been performed to identify cells. This may be useful if performing analyses that need to account for the ambient RNA pool.

For both raw and processed data, the row metadata contains the Ensembl ID and MGI symbol for each gene.

Value

If `type="processed"`, a SingleCellExperiment is returned containing processed data from selected samples.

If `type="raw"`, a List of SingleCellExperiments is returned, each containing the raw counts for a single sample. List elements are named after the corresponding sample.

Author(s)

Aaron Lun, with modification by Jonathan Griffiths

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
atlas.data <- EmbryoAtlasData(samples = 1:2)

atlas.data <- EmbryoAtlasData(type="processed", samples = 1:2)
```

`EmbryoCelltypeColours` *Celltype colours from Pijuan-Sala et al.*

Description

A vector containing the colour hexcodes that were used in Pijuan-Sala et al.

Usage

```
EmbryoCelltypeColours
```

Format

A vector of hexcodes named according to the appropriate celltype; celltypes match those in the metadata.

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
head(EmbryoCelltypeColours)
```

`GuibentifExtraData` *Guibentif et al. accessory data*

Description

Obtain the trajectory and NMP ordering data used in Guibentif et al.

Usage

```
GuibentifExtraData()
```

Details

This function downloads the data used in some of the analyses from Guibentif et al. (2020). Specifically, it contains the NMP cell orderings, and the atlas somitogenesis trajectory data.

This data is stored in a list. The first element of the list is named `atlas_somite_trajectories`, and is itelf a list that contains:

- `masses`:A data.frame containing the mass allocated to each cell from each trajectory (note: excluding extraembryonic, mixed_gastrulation timepoint, and doublet or stripped nuclei cells).
- `membership`:A data.frame containing the somite trajectory labels used in the paper, calculated from `masses`.

The second element is named `nmp_orderings`, and is also a list, which contains:

- `atlas`:A data.frame containing the position for each cell in the NMP ordering from the embryo atlas (see `EmbryoAtlasData`).
- `wt_chimera`:A data.frame containing the position for each cell in the NMP ordering from the WT chimera data (see `WTChimeraData`).
- `t_chimera`:A data.frame containing the position for each cell in the NMP ordering from the T chimera data (see `TChimeraData`).

Value

A `list` of the relevant somitogenesis trajectory and NMP ordering data will be returned. Details of the list structure are described in Details, below.

Author(s)

Jonathan Griffiths

References

Guibentif C, Griffiths JA et al. (2020). Title. *Journal* 566, 7745:490-495.

Examples

```
data <- GuibentifExtraData()
```

`LohoffSeqFISHData` *seqFISH data of E8.5 mouse embryos*

Description

Obtain the observed or imputed counts for the Lohoff et al. E8.5 mouse embryo seqFISH dataset.

Usage

```
LohoffSeqFISHData(
  type = c("observed", "imputed"),
  samples = NULL,
  get.molecules = FALSE
)
```

Arguments

type	String specifying the type of data to obtain, see Details. Default behaviour is to return the observed data.
samples	Integer or character vector specifying the samples for which data (observed or imputed) should be obtained. If NULL (default), data are returned for all (6) samples.
get.molecules	Logical indicating whether to also download the positions of each observed mRNA molecule in each cell.

Details

This function downloads the seqFISH data from Lohoff et al. (2020). The dataset contains 6 seqFISH samples; consecutive samples (1 and 2, etc.) are from different sections taken from the same embryo.

In the observed data, mRNA counts and molecule locations are available for the 351 genes in the seqFISH panel. The count matrix contains the raw count vectors for the cells called from all samples in this manner. Size factors were computed from the total observed counts for each cell, excluding the sex-specific gene Xist. For both observed and imputed data, the row metadata contains the Ensembl ID and MGI symbol for each gene. The column metadata for cells contains:

cell: Character, unique cell identifier across all samples.

embryo: Character, embryo ID for each cell.

pos: Character, name of the imaging region (i.e., encodes batch effects within a sample).

embryo_pos: Character, concatenated embryo name and imaging region.

embryo_pos_z: Character, concatenated embryo name, imaging region, and z position. Represents groups of (in principle) batch-effect-free cells.

area: Integer, number of pixels enclosed by the segmentation mask for each cell.

celltype: Character, celltype label for each cell. Note that it does not match exactly with the scRNAseq atlas.

sample: Integer, represents groups of cells from a single embryo at a single z-position. See above for details.

segmentation_vertices: DataFrameList, contains a DataFrame for each cell with x and y segmentation vertices. These were calculated from cadherin (i.e. cell membrane) staining.

sizeFactor: Numeric, size factor for normalisation.

A UMAP representation of the data is also available in the reducedDims slot of the SingleCellExperiment object.

If molecule positions were requested, these will be in the assays slot of the SingleCellExperiment object. These are represented as a BumpyMatrix object, with each cell of the matrix containing a DataFrame of positions for each mRNA molecule. Each entry contains the positions of each mRNA molecule for the corresponding gene (row) and cell (column) of the SpatialExperiment object.

The "observed" data the observed molecule counts from the experiment for the 351 genes assayed. The "imputed" data contains transcriptome-wide data imputed from the scRNAseq atlas. These data are in the assay slot imputed_logcounts. Note that these are much less sparse than scRNA-seq matrices, and are large in memory. The imputed SpatialExperiment object is identical to that obtained for the observed data, except for the difference in assay slots, rowData, and the absence of sizeFactors (as the data was imputed from the normalised atlas).

Value

If `type="observed"`, a SpatialExperiment is returned containing the observed seqFISH data from selected samples.

If `type="imputed"`, a SpatialExperiment is returned containing the transcriptome-wide logcounts imputed from the scRNA-seq data from selected samples.

Author(s)

Jonathan Griffiths

References

Lohoff T, Ghazanfar S et al. (2020). Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. *bioRxiv* 2020.11.20.391896.

Examples

```
seqfish.data <- LohoffSeqFISHData(samples = 1:2)
```

RAMultiomeData *Mouse gastrulation joint ATAC/RNA data*

Description

Obtain the processed counts for the mouse gastrulation "multi-omics" dataset.

Usage

```
RAMultiomeData(type = c("all", "rna", "peaks", "tss"), samples = NULL)
```

Arguments

`type`	String specifying the type of data to obtain, see Details. Default behaviour is to return all three data types.
`samples`	Integer or character vector specifying the samples for which data (processed or raw) should be obtained. If `NULL` (default), data are returned for all (11) samples.

Details

This function downloads the data for the embryo atlas from Argelaguet et al. (2022). The dataset contains 11 10X Genomics multiome samples.

The column metadata contains columns from the following set, depending on modality:

`barcode`: Character: cell barcode from the 10X Genomics experiment (with appended "-1" from Cellranger).

`sample`: Integer: index of the sample from which the cell was taken.

`sample_name`: Character: descriptive name of the sample from which the cell was taken.

`stage`: Character: stage of the mouse embryo at which the sample was taken.

`genotype`: Character: cell genotype, wild type (WT) or Brachyury KO (T_KO)

`celltype`: Character: cell type to which the cell was assigned by mapping to RNA atlas.

`nFeature_RNA`: Integer: number of genes detected in RNAseq data for the cell.

`nCount_RNA`: Integer: number of RNA molecules detected in RNAseq data for the cell.

`mitochondrial_percent_RNA`: Numeric: percent of RNA molecules detected from mitochondrial genome for the cell.

`ribosomal_percent_RNA`: Numeric: percent of RNA molecules detected from ribosomal genes for the cell.

`nFrags_atac`: Numeric: number of ATAC fragments detected per cell.

`TSSEnrichment_atac`: Numeric: Quality control metric that represents the ratio of ATAC peaks near the transcription start site relative to the flanking regions. Derived from the ArchR package.

`doublet_score`: Numeric: doublet score for each cell calculated using the `cxds_bcds_hybrid` function from the `scds` package.

`doublet_call`: Logical: doublet call for each cell calculated from the "doublet_score" column. Cells with a doublet score larger than 1.25 are assumed to be doublets and thus were removed from downstream analysis.

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object. These are UMAPs calculated either across all the data, or per stage (`perstage`). Those labelled either `rna` or `atac` alone were calculated from the processed count matrices of these modalities; `rna_atac`-labelled UMAPs were calculated from the MOFA factors calculated cross-modality.

For the RNA and TSS gene score data, the row metadata contains the Ensembl ID and MGI symbol for each gene. The ATAC peak row metadata contains information for each of those peaks Unlike other datasets in MouseGastrulationData, the rownames for these objects are gene symbols.

Value

If `type="all"`, a SingleCellExperiment object is returned containing processed data from selected samples for all data types. RNA-seq data is in the primary assay slot, while the other data types are in the altExp slot. The default `counts` slot on the first level of the SingleCellExperiment object will be occupied by the RNA data. The other modalities can be accessed using `SingleCellExperiment::altExp`, where the counts slot will again be occupied by the data for each modality for compatability with many function defaults.

If `type="rna"`, `type="peaks"`, or `type="tss"`, a SingleCellExperiment object is returned containing information for a single data type. Each assay will be in the primary `counts` slot. RNA data corresponds to RNA-seq read counts. Peak data corresponds to read counts from ATAC-seq quantified over peaks defined using ArchR's peak calling strategy. TSS data corresponds to read counts from ATAC-seq quantified over transcriptions start sites using ArchR's Gene Scores model.

Author(s)

Jonathan Griffiths

References

Argelaguet R et al. (2022). Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv* 2022.06.15.496239

Examples

```
RA_rna <- RAMultiomeData(samples=1, type = "rna")
```

RASampleMetadata *Sample metadata from the Argelaguet et al. multiome atlas*

Description

A data frame containing stage and genotype information for the multiome atlas dataset.

Usage

```
RASampleMetadata
```

Format

A data frame containing information for each 10x sample of the embryo atlas. This object contains:

`sample`: Integer, 10x sample index.

`sample_name`: Character, sample name provided by authors.

`stage`: Character, developmental stage from which sample was generated.

`ncells`: Integer, number of cells (post-QC) per sample.

`genotype`: Character, T_KO if brachyury knockout sample, otherwise WT.

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
head(RASampleMetadata)
```

Tal1ChimeraData *Tal1 chimera data*

Description

Obtain the processed or raw counts for the Tal1 chimeric mouse embryo dataset.

Usage

```
Tal1ChimeraData(
  type = c("processed", "raw"),
  samples = NULL,
  Csparse.assays = TRUE
)
```

Arguments

type	String specifying the type of data to obtain, see Details. Default behaviour is to return processed data.
samples	Integer or character vector specifying the samples for which data (processed or raw) should be obtained. If NULL (default), data are returned for all (four) samples.
Csparse.assays	Logical indicating whether to convert assay matrices into the column major format that is more performant with contemporary software packages. Default behaviour is to perform the conversion.

Details

This function downloads the data for the E8.5 Tal1 chimera experiment from Pijuan-Sala et al. (2019). The dataset contains four 10X Genomics samples:

- Sample 1: _Tal1_ knock-out cells (tomato positive)
- Sample 2: _Tal1_ knock-out cells (tomato positive)
- Sample 3: wild-type cells (tomato negative)
- Sample 4: wild-type cells (tomato negative)

All samples are from E8.5, from the same pool of chimeric embryos. Different samples with the same Tomato status are therefore technical replicates of each other.

In the processed data, cell-containing libraries have already been identified in each sample using the `emptyDrops` function from **DropletUtils**. The count matrix contains the raw count vectors for the cells called from all samples in this manner. Size factors were computed using the `computeSumFactors` function from **scran**. The column metadata for called cells contains:

`cell`: Character, unique cell identifier across all samples.

`barcode`: Character, cell barcode from the 10X Genomics experiment.

`sample`: Integer, number of the sample from which the cell was taken.

`stage`: Character, stage of the mouse embryo at which the sample was taken.

`tomato`: Logical, whether this cell expressed td-Tomato during FACS.

`stage.mapped`: Character, stage of the mouse embryo atlas to which the cell was mapped.

`celltype.mapped`: Character, cell type of the mouse embryo atlas to which the cell was mapped.

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object.

The raw data contains the unfiltered count matrix for each sample, as generated directly from the CellRanger software. Swapped molecules have been removed using `DropletUtils::swappedDrops`. No filtering has been performed to identify cells. This may be useful if performing analyses that need to account for the ambient RNA pool.

For both raw and processed data, the row metadata contains the Ensembl ID and MGI symbol for each gene.

Value

If `type="processed"`, a SingleCellExperiment is returned containing processed data from selected samples.

If `type="raw"`, a List of SingleCellExperiments is returned, each containing the raw counts for a single sample. List elements are named after the corresponding sample.

Author(s)

Aaron Lun

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
tal1.data <- Tal1ChimeraData(samples = 1)

tal1.data <- Tal1ChimeraData(type="processed", samples = 1)
```

TChimeraData *T chimera data*

Description

Obtain the processed or raw counts for the T chimeric mouse embryo dataset.

Usage

```
TChimeraData(
  type = c("processed", "raw"),
  samples = c(1:2, 5:16),
  Csparse.assays = TRUE
)
```

Arguments

`type`	String specifying the type of data to obtain, see Details. Default behaviour is to return processed data.
`samples`	Integer or character vector specifying the samples for which data (processed or raw) should be obtained. If `NULL` (default), data are returned for all QC-passing (fourteen) samples.
`Csparse.assays`	Logical indicating whether to convert assay matrices into the column major format that is more performant with contemporary software packages. Default behaviour is to perform the conversion.

Details

This function downloads the data for the T chimera experiment from Guibentif et al. (2020). The dataset contains sixteen 10X Genomics samples from sets of embryo pools:

- Sample 1: E8.5 injected cells (tomato positive), pool 1
- Sample 2: E8.5 host cells (tomato negative), pool 1
- Sample 3: E7.5 injected cells (tomato positive), pool 2
- Sample 4: E7.5 host cells (tomato negative), pool 2

- Sample 5: E8.5 injected cells (tomato positive), pool 3
- Sample 6: E8.5 host cells (tomato negative), pool 3
- Sample 7: E8.5 injected cells (tomato positive), pool 4
- Sample 8: E8.5 host cells (tomato negative), pool 4
- Sample 9: E8.5 injected cells (tomato positive), pool 5
- Sample 10: E8.5 host cells (tomato negative), pool 5
- Sample 11: E7.5 injected cells (tomato positive), pool 6
- Sample 12: E7.5 host cells (tomato negative), pool 6
- Sample 13: E7.5 injected cells (tomato positive), pool 7
- Sample 14: E7.5 host cells (tomato negative), pool 7
- Sample 15: E7.5 injected cells (tomato positive), pool 8
- Sample 16: E7.5 host cells (tomato negative), pool 8

Samples from the same pool are paired in the experimental design. Each pool is a biological replicate. Samples 3 and 4 were excluded from analyses, as in these chimeras host cells seemed to form only ExE ectoderm. The data is available to download if you like, but will not be fetched by default.

In the processed data, cell-containing libraries have already been identified in each sample using the `emptyDrops` function from **DropletUtils**. The count matrix contains the raw count vectors for the cells called from all samples in this manner. Size factors were computed using the `computeSumFactors` function from **scran**. The column metadata for called cells contains:

`cell`**:** Character, unique cell identifier across all samples.

`barcode`**:** Character, cell barcode from the 10X Genomics experiment.

`sample`**:** Integer, number of the sample from which the cell was taken.

`stage`**:** Character, stage of the mouse embryo at which the sample was taken.

`tomato`**:** Logical, whether this cell expressed td-Tomato during FACS.

`pool`**:** Integer, embryo pool from which cell derived; samples with same value are matched.

`stage.mapped`**:** Character, stage of the mouse embryo atlas to which the cell was mapped.

`celltype.mapped`**:** Character, cell type of the mouse embryo atlas to which the cell was mapped.

`closest.cell`**:** Character, closest cell in the atlas dataset (see `EmbryoAtlasData`) after MNN mapping.

`doub.density`**:** Numeric, output of (a now-outdated run of) `scran::doubletCells`, performed on each sample separately.

`trajectory.mapped`**:** Character, trajectory membership for somite/NMP formation.

`somite.subct.mapped`**:** Character, somite subcluster to which cells mapped.

`sizeFactor`**:** Numeric, cell sizefactor.

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object.

The raw data contains the unfiltered count matrix for each sample, as generated directly from the CellRanger software. Swapped molecules have been removed using `DropletUtils::swappedDrops`. No filtering has been performed to identify cells. This may be useful if performing analyses that need to account for the ambient RNA pool.

For both raw and processed data, the row metadata contains the Ensembl ID and MGI symbol for each gene.

Value

If `type="processed"`, a SingleCellExperiment is returned containing processed data from selected samples

If `type="raw"`, a List of SingleCellExperiments is returned, each containing the raw counts for a single sample. List elements are named after the corresponding sample.

Author(s)

Aaron Lun, with modification by Jonathan Griffiths

References

Guibentif C, Griffiths JA et al. (2020). Diverse Routes towards Early Somites in the Mouse Embryo *Developmental Cell* In press.

Examples

```
t.data <- TChimeraData(samples = 1)

t.data <- TChimeraData(type="processed", samples = 1)
```

`WTChimeraData` *WT chimera data*

Description

Obtain the processed or raw counts for the WT chimeric mouse embryo dataset.

Usage

```
WTChimeraData(
  type = c("processed", "raw"),
  samples = NULL,
  Csparse.assays = TRUE
)
```

Arguments

`type`	String specifying the type of data to obtain, see Details. Default behaviour is to return processed data.
`samples`	Integer or character vector specifying the samples for which data (processed or raw) should be obtained. If `NULL` (default), data are returned for all (ten) samples.
`Csparse.assays`	Logical indicating whether to convert assay matrices into the column major format that is more performant with contemporary software packages. Default behaviour is to perform the conversion.

Details

This function downloads the data for the WT chimera experiment from Pijuan-Sala et al. (2019). The dataset contains ten 10X Genomics samples from sets of embryo pools:

- Sample 1: E7.5 injected cells (tomato positive), pool 1
- Sample 2: E7.5 host cells (tomato negative), pool 1
- Sample 3: E7.5 injected cells (tomato positive), pool 2
- Sample 4: E7.5 host cells (tomato negative), pool 2
- Sample 5: E8.5 injected cells (tomato positive), pool 3
- Sample 6: E8.5 host cells (tomato negative), pool 3
- Sample 7: E8.5 injected cells (tomato positive), pool 4
- Sample 8: E8.5 host cells (tomato negative), pool 4
- Sample 9: E8.5 injected cells (tomato positive), pool 5
- Sample 10: E8.5 host cells (tomato negative), pool 5

Samples from the same pool are paired in the experimental design. Each pool is a biological replicate. Only samples 5 and 6 were used in the analyses of Pijuan-Sala et al. (2019).

In the processed data, cell-containing libraries have already been identified in each sample using the `emptyDrops` function from **DropletUtils**. The count matrix contains the raw count vectors for the cells called from all samples in this manner. Size factors were computed using the `computeSumFactors` function from **scran**. The column metadata for called cells contains:

`cell`: Character, unique cell identifier across all samples.

`barcode`: Character, cell barcode from the 10X Genomics experiment.

`sample`: Integer, number of the sample from which the cell was taken.

`stage`: Character, stage of the mouse embryo at which the sample was taken.

`tomato`: Logical, whether this cell expressed td-Tomato during FACS.

`pool`: Integer, embryo pool from which cell derived; samples with same value are matched.

`stage.mapped`: Character, stage of the mouse embryo atlas to which the cell was mapped.

`celltype.mapped`: Character, cell type of the mouse embryo atlas to which the cell was mapped.

`closest.cell`: Character, closest cell in the atlas dataset (see `EmbryoAtlasData`) after MNN mapping.

`doub.density`: Numeric, output of (a now-outdated run of) `scran::doubletCells`, performed on each sample separately.

Reduced dimension representations of the data are also available in the `reducedDims` slot of the SingleCellExperiment object.

The raw data contains the unfiltered count matrix for each sample, as generated directly from the CellRanger software. Swapped molecules have been removed using `DropletUtils::swappedDrops`. No filtering has been performed to identify cells. This may be useful if performing analyses that need to account for the ambient RNA pool.

For both raw and processed data, the row metadata contains the Ensembl ID and MGI symbol for each gene.

Value

If `type="processed"`, a SingleCellExperiment is returned containing processed data from selected samples

If `type="raw"`, a List of SingleCellExperiments is returned, each containing the raw counts for a single sample. List elements are named after the corresponding sample.

Author(s)

Aaron Lun, with modification by Jonathan Griffiths

References

Pijuan-Sala B, Griffiths JA, Guibentif C et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 7745:490-495.

Examples

```
wt.data <- WTChimeraData(samples = 1)

wt.data <- WTChimeraData(type="processed", samples = 1)
```

Index

∗ **datasets**
AtlasSampleMetadata, 3
EmbryoCelltypeColours, 8
RASampleMetadata, 13

AtlasSampleMetadata, 3

BPSATACData, 4

EmbryoAtlasData, 6, *9*, *16*, *18*
EmbryoCelltypeColours, 8

GuibentifExtraData, 8

List, *7*, *14*, *17*, *19*
LohoffSeqFISHData, 9

MouseGastrulationData *(*MouseGastrulationData-package*)*, 2
MouseGastrulationData-package, 2

RAMultiomeData, 11
RASampleMetadata, 13

SingleCellExperiment, *5*, *7*, *12*, *14*, *17*, *19*
SpatialExperiment, *11*

Tal1ChimeraData, 13
TChimeraData, *9*, 15

WTChimeraData, *9*, 17