Package 'CTdata'

September 18, 2024

Title Data companion to CTexploreR

Version 1.5.2

Description Data from publicly available databases (GTEx, CCLE, TCGA and ENCODE) that go with CTexploreR in order to re-define a comprehensive and thoroughly curated list of CT genes and their main characteristics.

License Artistic-2.0

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 4.2)

biocViews Transcriptomics, Epigenetics, GeneExpression, DataImport, ExperimentHubSoftware

Imports ExperimentHub, utils

Suggests testthat (>= 3.0.0), DT, BiocStyle, knitr, rmarkdown, SummarizedExperiment, SingleCellExperiment

VignetteBuilder knitr

BugReports https://github.com/UCLouvain-CBIO/CTdata/issues

Config/testthat/edition 3

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

git_branch devel

git_last_commit 100a61f

git_last_commit_date 2024-09-18

Repository Bioconductor 3.20

Date/Publication 2024-09-18

Author Axelle Loriot [aut] (<https://orcid.org/0000-0002-5288-8561>), Julie Devis [aut] (<https://orcid.org/0000-0001-5525-5666>), Anna Diacofotaki [ctb], Charles De Smet [ths], Laurent Gatto [aut, ths, cre] (<https://orcid.org/0000-0002-1520-2268>)

Maintainer Laurent Gatto <laurent.gatto@uclouvain.be>

Contents

`all_genes` *All genes genes description table*

Description

All genes description, according to the analysis done for CT genes

Format

A `tibble` object with 24488 rows and 47 columns.

- Rows correspond to genes
- Columns give genes characteristics

Details

When the promoter is mentionned, it has been determined as 1000 nt upstream TSS and 200 nt downstream TSS.

CT_genes characteristics column:

- Column `CT_gene_type` indicates if the gene is a CT specific gene ("CT_gene" : testis_specific in `testis_specificity`) and activated in "TCGA_category" and "CCLE_category) or CT preferential gene ("CTP_gene" : testis_preferential in `testis_specificity`) and activated in"TCGA_category" and "CCLE_category").
- Column `testis_specificity` gives the testis-specificity of genes assigned to each gene using `GTEX_category` and `multimapping_analysis` ("testis_specific" or "testis_preferential"). Genes were assigned "testis-preferential" if testis-specific in these categories but not testis specific in `HPA_category` or leaky in `CCLE_category` or `TCGA_category`.
- Column `regulated_by_methylation` indicates if the gene is regulated by methylation (`TRUE`) based on DAC induction (has to be TRUE) and on promoter methylation level in normal somatic tissues (when available, has to be methylated in somatic tissues).
- Column `X_linked` indicates if the gene is on the chromosome X (TRUE) or not (FALSE).
- Columns `chr`, `strand` and `transcription_start_site` give the genomic location.
- Column `GTEX_category` gives the category ("testis_specific", "testis_preferential" or "lowly_expressed") assigned to each gene using GTEx database (see `?GTEX_data` for details).
- Column `q75_TPM_somatic` gives the q75 expression level found in a somatic tissue (using GTEx database).
- Column `max_TPM_somatic` gives the maximum expression level found in a somatic tissue (using GTEx database).
- Column `ratio_testis_somatic` gives the ratio between expression in testis and the highest expression found in a somatic tissue (using GTEx database).
- Column `TPM_testis` gives the gene expression level in testis (using GTEx database).
- Column `lowly_expressed_in_GTEX` indicates if the gene is lowly expressed in GTEX database and thus needed to be analysed with multimapping allowed.
- Column `multimapping_analysis` informs if the gene (flagged as "lowly_expressed" in GTEX_data) was found to be testis-specific when multi-mapped reads were counted for gene expression in normal tissues ("not_analysed" or "testis_specific") (see `?normal_tissues_multimapping_data` for details).
- Column `HPA_RNA_single_cell_type_specific_nTPM` specifies the cell types in which genes were detected in the HPA single cell data (see `?HPA_cell_type_specificity` for details).
- Column `max_HPA_germcell` specifies if the maximum expression value in a germ cell type. (see `?HPA_cell_type_specificity` for details).
- Column `max_HPA_somatic` specifies if the maximum expression value in a somatic cell type. (see `?HPA_cell_type_specificity` for details).
- Column `not_detected_in_somatic_HPA` specifies if the gene is detected or not in a somatic cell type. (see `?HPA_cell_type_specificity` for details).
- Column `HPA_ratio_germ_som` gives the ratio between `max_HPA_germcell` and `max_HPA_somatic` columns.
- Column `percent_of_positive_CCLE_cell_lines` gives the percentage of CCLE cancer cell lines in which genes are expressed (genes were considered as expressed if TPM >= 1).
- Column `percent_of_negative_CCLE_cell_lines` gives the percentage of CCLE cancer cell lines in which genes are repressed (TPM <= 0.5).

- Column max_TPM_in_CCLE gives the highest expression level of genes in CCLE cell lines.
- Column CCLE_category gives the category assigned to each gene using CCLE data. "Activated" category corresponds to genes expressed in at least 1% of cell lines (TPM >= 1) and repressed in at least 20% of cell lines.
- Column percent_pos_tum gives the percentage of TCGA cancer samples in which genes are expressed (genes were considered as expressed if TPM >= 1).
- Column percent_neg_tum gives the percentage of TCGA cancer samples in which genes are repressed (TPM <= 0.5).
- Column max_TPM_in_TCGA gives the highest expression level of genes in TCGA cancer sample.
- Column max_q75_in_NT gives the maximum q75 expression in normal peritumoral tissues from TCGA.
- Column TCGA_category gives the category assigned to each gene using TCGA data. "activated" category corresponds to genes expressed in at least 1% of tumors (TPM >= 1) and repressed in at least 20% of samples. "multimapping_issue" corresponds to genes that need multi-mapping to be allowed in order to be analysed properly.
- Columns external_transcript_name, ensembl_transcript_id, and transcript_biotype give the references and informations about the most biologically relevant transcript associated to each gene.
- Column IGV_backbone indicates if a gene has been removed from CT genes as RNA-Seq reads were not properly aligned on exons, but were instead spread across a wide genomic region spanning the genes.
- Column family gives the gene family name.
- Column DAC_induced summarises the results (TRUE or FALSE) of a differential expression evaluating gene induction upon DAC treatment in a series of cell lines.
- Column named CpG_density, gives the density of CpG within each promoter (number of CpG / promoter length * 100).
- Column CpG_promoter classifies the promoters according to their CpG densities: "low" (CpG_density < 2), "intermediate" (CpG_density >= 2 & CpG_density < 4), and "high" (CpG_density >= 4).
- Column somatic_met_level that gives the mean methylation level of each promoter in somatic tissues.
- Column sperm_met_level that gives the methylation level of each promoter in sperm.
- Column somatic_methylation indicates if the promoter's mean methylation level in somatic tissues is higher than 50%.
- Column germline_methylationindicates if the promoter is methylated in germline, based on the ratio with somatic tissues (FALSE if somatic_met_level is at least twice higher than germline_met_level).
- Columns oncogene and tumor_suppressor informs if oncogenic and tumor-suppressor functions have been associated to genes (source: Cancermine).

Source

See scripts/make_all_genes_prelim.R and scripts/make_all_genes_and_CT_genes.R for details on how this list of genes was created.

CCLE_correlation_matrix *Gene correlations in CCLE cancer cell lines*

Description

Correlation coefficients between Cancer-Testis genes and all genes found on the CCLE database.

Format

A `matrix` object with 238 rows and 24483 columns.

- Rows correspond to CT genes
- Columns correspond to all genes from CCLE database

Details

Correlation coefficients (Pearson) between CT genes and all other genes are given in the matrix. These correlation coefficients were calculated using log transformed expression values from `CCLE_data` (all cell lines).

Source

See `scripts/make_CCLE_correlation_matrix.R` for details.

CCLE_data *Genes expression data in CCLE*

Description

Gene expression data in cancer cell lines from CCLE

Format

A `SummarizedExperiment` object with 24473 rows and 1229 columns

- Rows correspond to genes (ensembl_gene_id)
- Columns correspond to CCLE cell lines
- Expression data from the assay are TPM values
- Cell lines metadata are stored in colData

Details

The rowData contains

- A column `percent_of_positive_CCLE_cell_lines` that gives the percentage of CCLE cell lines (all cell lines combined) expressing the gene (TPM >= 1).
- A column `percent_of_negative_CCLE_cell_lines` that gives the percent of CCLE cell lines (all cell lines combined) in which genes are repressed (TPM < 0.5)
- A column `max_TPM_in_CCLE` that gives the maximal expression (in TPM) found in all cell lines.
- A column `CCLE_category` gives the category ("activated", "not_activated", "leaky") assigned to each gene. "activated" category corresponds to genes expressed (TPM >= 1) in at least 1% of cell lines, repressed (TPM <= 0.5) in at least 20% of cell lines with a maximal expression higher than 5 TPM. "not_activated" category corresponds to genes repressed (TPM <= 0.5) in at least 20% of cell lines but expressed (TPM >= 1) less than 1%. "leaky" category corresponds to genes repressed (TPM <= 0.5) in less than 20% of cell lines. "lowly_expressed" corresponds to genes repressed (TPM <= 0.5) in at least 20%, expressed (TPM >= 1) in more than 1 % of cell lines, with a maximum expression lower than 5 TPM.

Source

TPM values downloaded using depmap bioconductor package (see `scripts/make_CCLE_data.R` for details).

`CTdata` *All CTdata datasets*

Description

This is the companion Package for `CTexploreR` containing omics data to select and characterise CT genes.

Data come from public databases and include expression and methylation values of genes in normal and tumor samples as well as in tumor cell lines, and expression in cells treated with a demethylating agent is also available.

The `CTdata()` function returns a `data.frame` with all the annotated datasets provided in the package. For details on these individual datasets, refer to their respective manual pages.

See the vignette and the respective manuals pages for more details about the package and the data themselves.

Usage

```
CTdata()
```

Value

A `data.frame` describing the data available in `CTdata`.

Author(s)

Laurent Gatto

Examples

```
CTdata()
```

`CT_genes` *CT genes description table*

Description

Cancer-Testis (CT) genes description

Format

A `tibble` object with 280 rows and 47 columns.

- Rows correspond to CT genes
- Columns give CT genes characteristics

Details

When the promoter is mentionned, it has been determined as 1000 nt upstream TSS and 200 nt downstream TSS.

CT_genes characteristics column:

- Column `CT_gene_type` indicates if the gene is a CT specific gene ("CT_gene" : testis_specific in `testis_specificity`) and activated in "TCGA_category" and "CCLE_category) or CT preferential gene ("CTP_gene" : testis_preferential in `testis_specificity`) and activated in"TCGA_category" and "CCLE_category").
- Column `testis_specificity` gives the testis-specificity of genes assigned to each gene using `GTEX_category` and `multimapping_analysis` ("testis_specific" or "testis_preferential"). Genes were assigned "testis-preferential" if testis-specific in these categories but not testis specific in `HPA_category` or leaky in `CCLE_category` or `TCGA_category`.
- Column `regulated_by_methylation` indicates if the gene is regulated by methylation (`TRUE`) based on DAC induction (has to be TRUE) and on promoter methylation level in normal somatic tissues (when available, has to be methylated in somatic tissues).
- Column `X_linked` indicates if the gene is on the chromosome X (TRUE) or not (FALSE).
- Columns `chr`, `strand` and `transcription_start_site` give the genomic location.
- Column `GTEX_category` gives the category ("testis_specific", "testis_preferential" or "lowly_expressed") assigned to each gene using GTEx database (see `?GTEX_data` for details).
- Column `q75_TPM_somatic` gives the q75 expression level found in a somatic tissue (using GTEx database).
- Column `max_TPM_somatic` gives the maximum expression level found in a somatic tissue (using GTEx database).
- Column `ratio_testis_somatic` gives the ratio between expression in testis and the highest expression found in a somatic tissue (using GTEx database).
- Column `TPM_testis` gives the gene expression level in testis (using GTEx database).
- Column `lowly_expressed_in_GTEX` indicates if the gene is lowly expressed in GTEX database and thus needed to be analysed with multimapping allowed.

- Column multimapping_analysis informs if the gene (flagged as "lowly_expressed" in GTEX_data) was found to be testis-specific when multi-mapped reads were counted for gene expression in normal tissues ("not_analysed" or "testis_specific") (see ?normal_tissues_multimapping_data for details).
- Column HPA_RNA_single_cell_type_specific_nTPM specifies the cell types in which genes were detected in the HPA single cell data (see ?HPA_cell_type_specificity for details).
- Column max_HPA_germcell specifies if the maximum expression value in a germ cell type. (see ?HPA_cell_type_specificity for details).
- Column max_HPA_somatic specifies if the maximum expression value in a somatic cell type. (see ?HPA_cell_type_specificity for details).
- Column not_detected_in_somatic_HPA specifies if the gene is detected or not in a somatic cell type. (see ?HPA_cell_type_specificity for details).
- Column HPA_ratio_germ_som gives the ratio between max_HPA_germcell and max_HPA_somatic columns.
- Column percent_of_positive_CCLE_cell_lines gives the percentage of CCLE cancer cell lines in which genes are expressed (genes were considered as expressed if TPM >= 1).
- Column percent_of_negative_CCLE_cell_lines gives the percentage of CCLE cancer cell lines in which genes are repressed (TPM <= 0.5).
- Column max_TPM_in_CCLE gives the highest expression level of genes in CCLE cell lines.
- Column CCLE_category gives the category assigned to each gene using CCLE data. "Activated" category corresponds to genes expressed in at least 1% of cell lines (TPM >= 1) and repressed in at least 20% of cell lines.
- Column percent_pos_tum gives the percentage of TCGA cancer samples in which genes are expressed (genes were considered as expressed if TPM >= 1).
- Column percent_neg_tum gives the percentage of TCGA cancer samples in which genes are repressed (TPM <= 0.5).
- Column max_TPM_in_TCGA gives the highest expression level of genes in TCGA cancer sample.
- Column max_q75_in_NT gives the maximum q75 expression in normal peritumoral tissues from TCGA.
- Column TCGA_category gives the category assigned to each gene using TCGA data. "activated" category corresponds to genes expressed in at least 1% of tumors (TPM >= 1) and repressed in at least 20% of samples. "multimapping_issue" corresponds to genes that need multi-mapping to be allowed in order to be analysed properly.
- Columns external_transcript_name, ensembl_transcript_id, and transcript_biotype give the references and informations about the most biologically relevant transcript associated to each gene.
- Column IGV_backbone indicates if a gene has been removed from CT genes as RNA-Seq reads were not properly aligned on exons, but were instead spread across a wide genomic region spanning the genes.
- Column family gives the gene family name.
- Column DAC_induced summarises the results (TRUE or FALSE) of a differential expression evaluating gene induction upon DAC treatment in a series of cell lines.
- Column named CpG_density, gives the density of CpG within each promoter (number of CpG / promoter length * 100).

- Column CpG_promoter classifies the promoters according to their CpG densities: "low" (CpG_density < 2), "intermediate" (CpG_density >= 2 & CpG_density < 4), and "high" (CpG_density >= 4).
- Column somatic_met_level that gives the mean methylation level of each promoter in somatic tissues.
- Column sperm_met_level that gives the methylation level of each promoter in sperm.
- Column somatic_methylation indicates if the promoter's mean methylation level in somatic tissues is higher than 50%.
- Column germline_methylationindicates if the promoter is methylated in germline, based on the ratio with somatic tissues (FALSE if somatic_met_level is at least twice higher than germline_met_level).
- Columns oncogene and tumor_suppressor informs if oncogenic and tumor-suppressor functions have been associated to genes (source: Cancermine).

Source

See scripts/make_all_genes_prelim.R and scripts/make_all_genes_and_CT_genes.Rfor details on how this list of curated CT genes was created.

DAC_treated_cells *DE genes with/without demethylating agent*

Description

Gene expression values in a set of cell lines treated or not with 5-Aza-2'-Deoxycytidine (DAC), a demethylating agent.

Format

A SummarizedExperiment object with 24516 rows and 32 columns

- Rows correspond to genes (ensembl_gene_id).
- Columns correspond to samples.
- Expression data correspond to counts that have been normalised (by DESeq2 method) and log-transformed (log1p).
- The colData contains the SRA references of the fastq files that were downloaded, and informations about the cell lines and the DAC treatment.
- The rowData contains the results of a differential expression evaluating the DAC treatment effect. For each each cell line, the log2FC between treated and control cells is given, as well as the p-adjusted value. The column induced flags genes significantly induced (log2FoldChange >= 2 and padj <= 0.1) in at least one cell line. The threshold is not too stringent as DAC is expected to induce low expression levels (demethylation doesn't necessarily occurs in all treated cells...). When all cells lines already express the gene before DAC treatment, no assessment of induction was done.

Details

Differential expression analysis was done using DESeq2_1.36.0, using as design = ~ treatment (see scripts/make_DAC_treated_cells.R for details).

Source

RNAseq

fastq files were downloaded from Encode database. SRA reference of samples are stored in the colData.

`DAC_treated_cells_multimapping`
DE genes treated or not with a demethylating agent

Description

Gene expression values in a set of cell lines treated or not with 5-Aza-2'-Deoxycytidine (DAC), a demethylating agent. Many CT genes belong to gene families from which members have identical or nearly identical sequences. Some CT can only be detected in RNAseq data in which multimapping reads are not discarded.

Format

A `SummarizedExperiment` object with 24516 rows and 32 columns

- Rows correspond to genes (ensembl_gene_id).
- Columns correspond to samples.
- Expression data correspond to counts that have been normalised (by DESeq2 method) and log-transformed (log1p).
- The colData contains the SRA references of the fastq files that were downloaded, and informations about the cell lines and the DAC treatment.
- The rowData contains the results of a differential expression evaluating the DAC treatment effect. For each each cell line, the log2FC between treated and control cells is given, as well as the p-adjusted value. The column `induced` flags genes significantly induced (log2FoldChange >= 2 and padj <= 0.1) in at least one cell line. The threshold is not too stringent as DAC is expected to induce low expression levels (demethylation doesn't necessarily occurs in all treated cells...). When all cells lines already express the gene before DAC treatment, no assessment of induction was done.

Details

Differential expression analysis was done using DESeq2_1.36.0, using as `design = ~ treatment` (see `scripts/make_DAC_treated_cells_multimapping.R` for details).

Source

RNAseq fastq files were downloaded from Encode database. SRA reference of samples are stored in the colData.

embryo_sce_Petropoulos *Embryo scRNAseq*

Description

Human embryo single cell RNAseq data in RPKM from 'Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in

Format

A `SingleCellExperiment` object with 26178 rows and 1481 columns

- Rows correspond to genes (gene names as rownames)
- Columns correspond to cells

Details

Description of the colData:

- Column `individual` gives the sample the cell is coming from.
- Column `stage` specifies the stage of the early embryo.
- Column `sex` is the sex inference made using the expression of 11 Y-linked genes, made for each day individually.
- Column `ambigous` indicates if the inference of the embryo's sex was ambigous due to some cells expression of the Y-linked genes.

Source

RPKM and metadata files were downloaded from https://www.ebi.ac.uk/biostudies/files/E-MTAB-3929/ The data were converted in a SingleCellExperiment (see `scripts/make_embryo_sce_Petropoulos.R` for details).

embryo_sce_Zhu *Embryo scRNAseq*

Description

Human embryo single cell RNAseq data in FPKM from 'Single Cell DNA Methylome Sequencing of Human Preimplantation

Format

A `SingleCellExperiment` object with 26255 rows and 50 columns

- Rows correspond to genes (gene names as rownames)
- Columns correspond to cells

Details

Description of the colData:

- Column `embryo` gives the embryo the cell is coming from.
- Column `stage` specifies the stage of the early embryo.
- Column `sex` is the sex inference made using the expression of RPS4Y1. If mean expression of RPS4Y1 is higher than 50 FPKM, the sample is male.

Source

50 FPKM files were downloaded from GEO (accession: GSE81233). The data were converted in a SingleCellExperiment (see `scripts/make_embryo_sce_Zhu.R` for details).

`FGC_sce` *Fetal gonad scRNAseq*

Description

Human fetal gonad single cell RNAseq data from `Single-cell roadmap of human gonadal development` (Garcia-Alonso, Nature 2022)

Format

A `SingleCellExperiment` object with 22489 rows and 10850 columns

- Rows correspond to genes (gene names as rownames)
- Columns correspond to cells

Details

Description of the colData:

- Column `type` gives the gender and the cell type.
- Column `stage` specifies if the cell type is "pre-meiotic" or "meiotic".
- Column `germcell` is set to TRUE when the cell type is a germ cell.

Source

ee58527e-e1e4-465d-8dc8-800ee40f14f2.rds file dowloaded from https://cellxgene.cziscience.com/collections/661a402a
2a5a-4c71-9b05-b346c57bc451Data. The data were converted in a SingleCellExperiment (see `scripts/make_FGC_sce.R` for details).

GTEX_data *Genes expression in GTEX*

Description

Gene expression data in normal tissues from GTEx database.

Format

A SummarizedExperiment object with 24504 rows and 32 columns

- Rows correspond to genes (ensembl_gene_id as rownames)
- Columns correspond to tissues
- Expression data from the assay are TPM values

Details

The rowData contains

- A column named GTEX_category, specifying the tissue specificity category ("testis_specific", "testis-preferential", "lowly_expressed" or "other") assigned to each gene using expression values in testis and in somatic tissues, has been added to the rowData. "testis_specific" genes are expressed exclusively in testis (expression in testis >= 1 TPM, highest expression in somatic tissues < 0.5 TPM, and expressed at least 10x more in testis than in any somatic tissue). "testis-preferential" genes are genes expressed in testis but also in a few somatic tissues (expression in testis >= 1 TPM, and allowed in a minority of somatic tissues (q75_TPM_somatic < 0.5) and expressed at least 10x more in testis than in any somatic tissue). "lowly_expressed" genes are genes undetectable in GTEX database probably due to multi-mapping issues (expression in all GTEX tissues < 1 TPM).
- A column named q75_TPM_somatic giving the quantile 75% of TPM in a somatic tissue.
- A column named max_TPM_somatic giving the maximum expression level found in a somatic tissue.
- A column named ratio_testis_somatic giving the ratio between the TPM in testis and the max TPM in a somatic tissue

Source

Downloaded from https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz. Some categories of tissues were pooled (mean expression values are given in pooled tissues) (see scripts/make_GTEX_data.R for details).

hESC_data *Genes expression in hESC*

Description

Gene expression data in human embryonic stem cells

Format

A `SummarizedExperiment` object with 24488 rows and 4 columns

- Rows correspond to genes (ensembl_gene_id as rownames)
- Columns correspond to hESC types
- Expression data from the assay are TPM values

Details

The colData contains

- Column `genotype` gives the sexual genotype of the cells

Source

RNAseq fastq files were downloaded from Encode databas (see `scripts/make_hESC_data.R` for details).

HPA_cell_type_specificities *Cell type specificities (from HPA)*

Description

Cell type specificities based on scRNAseq data from the Human Protein Atlas (https://www.proteinatlas.org)

Format

A `tibble` object with 24504 rows and 7 columns.

- Rows correspond to genes (ensembl_gene_id)
- Columns give genes cell type specificities

Details

- Column HPA_scRNAseq_celltype_specific_nTPM gives the cell types in which genes were detected (corresponds to column RNA single cell type specific nTPM of proteinatlas.tsv file.
- Column max_HPA_germcell specifies if the maximum expression value in a germ cell type
- Column max_HPA_somatic specifies if the maximum expression value in a somatic cell type
- Column not_detected_in_somatic_HPA specifies if the gene is detected or not in a somatic cell type. Genes are flagged as TRUE if the max_HPA_somatic value is equal to 0, and FALSE if max_HPA_somatic value is > 0. NA is set when the original table from HPA had no values for that gene.
- Column HPA_ratio_germ_som gives the ratio between max_HPA_germcell and max_HPA_somatic columns.

Source

proteinatlas.tsv was downloaded from the Human Protein Atlas (https://www.proteinatlas.org) See scripts/make_HPA_cell_type_specificities.R for details.

---

makeTags *A short function that returns the default CTdata tags and, if provided, additional data-specific tags.*

Description

A short function that returns the default CTdata tags and, if provided, additional data-specific tags.

Usage

```
makeTags(x)
```

Arguments

x	An optional character() containing specific tags.

Value

A character containing the default tags and optional data-specific tags. If x is missing or is of length 0, the default tags are returned. Otherwise, a vector of length equal to length(x) is returned.

Examples

```
CTdata:::makeTags() ## only default tags

CTdata:::makeTags(character()) ## only default tags

CTdata:::makeTags("myTag") ## one additional tag

CTdata:::makeTags(c("myTag", "myOtherTag")) ## two additional tag
```

mean_methylation_in_embryo

All genes' promoters mean methylation in embryos

Description

Mean methylation values of all CpGs located within all genes promoters in early embryos. Data is based on hg19 reference genome ! From Single Cell DNA Methylome Sequencing of Human Preimplantation Embry (Zhu et al. 2018)

Format

A RangedSummarizedExperiment object with 24441 rows and 492 columns

- Rows correspond to all genes (gene names as rownames)
- Mean methylation levels in embryos types are stored in columns
- rowRanges correspond to the hg19 promoter positions

Details

The rowData contains:

- A column named ensembl_gene_id containing gene ids.

Source

WGBS methylation data was downloaded from GEO. Mean methylation levels are evaluated using methylation values of CpGs located in promoter region (defined as 1000 nt upstream TSS and 500 nt downstream TSS) (see scripts/make_mean_methylation_in_embryos.R for details).

mean_methylation_in_FGC

All genes' promoters mean methylation in FGC

Description

Mean methylation values of all CpGs located within all genes promoters in fetal germ cells. Data is based on hg19 reference genome ! From Dissecting the epigenomic dynamics of human fetal germ cell develop

Format

A RangedSummarizedExperiment object with 24441 rows and 337 columns

- Rows correspond to all genes (gene names as rownames)
- Mean methylation levels in FGC types are stored in columns
- rowRanges correspond to the hg19 promoter positions

Details

The rowData contains:

- A column named `ensembl_gene_id` containing gene ids.

Source

WGBS methylation data was downloaded from GEO. Mean methylation levels are evaluated using methylation values of CpGs located in promoter region (defined as 1000 nt upstream TSS and 500 nt downstream TSS) (see `scripts/make_mean_methylation_in_FGC.R` for details).

`mean_methylation_in_hESC` *All genes' promoters mean methylation in hESC*

Description

Mean methylation values of all CpGs located within all genes promoters in human embryonic stem cells

Format

A `SummarizedExperiment` object with 24488 rows and 3 columns

- Rows correspond to all genes (gene names as rownames)
- Mean methylation levels in hESC types are stored in columns

Details

The rowData contains:

- A column named `ensembl_gene_id` containing gene ids.

The colData contains

- Column `genotype` gives the sexual genotype of the cells

Source

WGBS methylation data was downloaded from Encode. Mean methylation levels are evaluated using methylation values of CpGs located in promoter region (defined as 1000 nt upstream TSS and 200 nt downstream TSS) (see `scripts/make_mean_methylation_in_hESC.R` for details).

`mean_methylation_in_tissues` *All genes' promoters mean methylation*

Description

Mean methylation values of all CpGs located within all genes promoters in a set of normal tissues

Format

A `SummarizedExperiment` object with 24502 rows and 14 columns

- Rows correspond to all genes (gene names as rownames)
- Mean methylation levels in normal tissues are stored in columns
- CpG densities and results of methylation analysis are stored in rowData

Details

The rowData contains:

- A column named `CpG_density`, gives the density of CpG within each promoter (number of CpG / promoter length * 100).
- A column `CpG_promoter` that classifies the promoters according to their CpG densities: "low" (CpG_density < 2), "intermediate" (CpG_density >= 2 & CpG_density < 4), and "high" (CpG_density >= 4).
- A column `somatic_met_level` that gives the mean methylation level of each promoter in somatic tissues.
- A column `sperm_met_level` that gives the methylation level of each promoter in sperm.
- A column `somatic_methylation` indicates if the promoter's mean methylation level in somatic tissues is higher than 50%.
- A column `germline_methylation`indicates if the promoter is methylated in germline, based on the ratio with somatic tissues (FALSE if somatic_met_level is at least twice higher than germline_met_level).

Source

WGBS methylation data was downloaded from Encode and from GEO databases. Mean methylation levels are evaluated using methylation values of CpGs located in promoter region (defined as 1000 nt upstream TSS and 200 nt downstream TSS) (see `scripts/make_mean_methylation_in_tissues.R` for details).

methylation_in_embryo *Methylation of CpGs within all genes promoters in embryo*

Description

Methylation values of CpGs located within all genes promoters in embryo. Data is based on hg19 reference genome ! From Single Cell DNA Methylome Sequencing of Human Preimplantation Embryos (Zhu et al. 2018)

Format

A RangedSummarizedExperiment object with 1915545 rows and 492 columns

- Rows correspond to CpGs (located within all genes promoters (TSS +- 1000 nt))
- Columns correspond to cells
- Methylation values from scWGBS data
- rowRanges correspond to CpG positions

Details

Description of the colData:

- Column cell_type indicates the embryo type.
- Column bulk_or_single_cell specifies if the sample was indeed only a single cell or a bulk of several cells.
- Other information about the sequencing of each sample are clearly labelled

Source

scWGBS methylation data was downloaded from GEO database (see scripts/make_methylation_in_embryo.R for details).

methylation_in_FGC *Methylation of CpGs within all genes promoters in FGC*

Description

Methylation values of CpGs located within all genes promoters in fetal germ cells. Data is based on hg19 reference genome ! From Dissecting the epigenomic dynamics of human fetal germ cell development at

Format

A RangedSummarizedExperiment object with 1915545 rows and 337 columns

- Rows correspond to CpGs (located within all genes promoters (TSS +- 1000 nt))
- Columns correspond to cells
- Methylation values from scWGBS data
- rowRanges correspond to CpG positions

Details

Description of the colData:

- Column `type` indicates if the cell type is somatic or FGC
- Column `time_week` specifies the time of the embryo when cells were removed.
- Column `sex` indicates the sex of the cells.
- Other information about the sequencing of each sample are clearly labelled

Source

scWGBS methylation data was downloaded from GEO database (see `scripts/make_methylation_in_FGC.R` for details).

`methylation_in_hESC` *Methylation of CpGs within all genes promoters in hESC*

Description

Methylation values of CpGs located within all genes promoters in human embryonic stem cells.

Format

A `RangedSummarizedExperiment` object with 4280098 rows and 3 columns

- Rows correspond to CpGs (located within all genes promoters (TSS +- 5000 nt))
- Columns correspond to hESC
- Methylation values from WGBS data
- rowRanges correspond to CpG positions

Source

WGBS methylation data was downloaded from Encode (see `scripts/make_methylation_in_hESC.R` for details).

`methylation_in_tissues` *Methylation of CpGs within all genes promoters*

Description

Methylation values of CpGs located within all genes promoters in a set of normal tissues.

Format

A `RangedSummarizedExperiment` object with 4280327 rows and 14 columns

- Rows correspond to CpGs (located within all genes promoters (TSS +- 5000 nt))
- Columns correspond to normal tissues
- Methylation values from WGBS data
- rowRanges correspond to CpG positions

Source

WGBS methylation data was downloaded from Encode and from GEO databases (see `scripts/make_methylation_in_` for details).

`normal_tissues_multimapping_data`

Gene expression values in normal tissues

Description

Gene expression values (TPM) in a set of normal tissues obtained by counting or not multi-mapped reads. Many CT genes belong to gene families from which members have identical or nearly identical sequences. Some CT can only be detected in RNAseq data in which multimapping reads are not discarded.

Format

A `SummarizedExperiment` object with 24504 rows and 18 columns

- Rows correspond to genes (ensembl_gene_id)
- Columns correspond to normal tissues.
- First assay, `TPM_no_multimapping`, gives TPM expression values obtained when discarding multimapped reads.
- Second assay, `TPM_with_multimapping`, gives TPM expression values obtained by counting multimapped reads.

Details

A column named `multimapping_analysis` has been added to the rowData. It summarizes the testis specificity analysis of genes flagged as "lowly_expressed" in GTEX_data. Genes are considered "testis_specific" when, with multimapping allowed, they are detectable in testis (TPM >= 1), their TPM value has increased compared to without multimapping (ratio > 5), and their TPM value is at least 10 times higher in testis than in any other somatic tissue (where the maximum expression always has to be below 1 TPM). Genes are considered "testis_preferential" when, with multimapping allowed, they are detectable in testis (TPM >= 1), their TPM value has increased compared to without multimapping (ratio > 5), and their TPM value is at least 10 times higher in testis than in any other somatic tissue (where the maximum expression is above 1 TPM).

Source

RNAseq fastq files were downloaded from Encode database (see `scripts/make_normal_tissues_multimapping.R` for details).

oocytes_sce *Oocytes scRNAseq*

Description

Human oocytes single cell RNAseq data from Decoding dynamic epigenetic landscapes in human oocytes using
(Yan et al. Cell Stem Cell 2021)

Format

A SingleCellExperiment object with 26500 rows and 899 columns

- Rows correspond to genes(gene names as rownames)
- Columns correspond to cells

Details

Description of the colData:

- Column type gives the cell type.
- Column stage specifies if the cell type is "pre-meiotic" or "meiotic".
- Column germcell is set to TRUE when the cell type is a germ cell.

Source

GSE154762_hO_scChaRM_count_matix.txt.gzwas downloaded from GEO (accession: GSE154762). The data were
for details).

scRNAseq_HPA *Gene expression in human cell types*

Description

Gene expression profiles in different human cell types based on scRNAseq data obtained from the Human Protein Atlas (https://www.proteinatlas.org)

Format

A SingleCellExperiment object with 20082 rows and 66 columns

- Rows correspond to genes (ensembl gene id as rownames)
- Columns correspond to cell types
- Expression values correspond to transcripts per million protein coding genes (pTPM)

Details

Description of the colData:

- Column `Cell_type` gives cell type.
- Column `group` gives the cell type group (defined in the Human Protein Atlas).

Description of the rowData:

- Column `max_TPM_in_a_somatic_cell_type` gives the maximum expression value found in a somatic cell type
- Column `max_in_germcells_group` gives the maximum expression value found in a germ cell type
- Column `Higher_in_somatic_cell_type` specifies if a somatic cell type

Source

Gene expression values in cell types, based on multiple scRNAseq datasets obtained from the Human Protein Atlas (https://www.proteinatlas.org/about/download) The data were converted in a SummarizedExperiment (see `scripts/14_make_scRNAseq_HPA.R` for details).

`TCGA_methylation` *Methylation of all genes promoters in TCGA samples*

Description

Methylation values of probes located within all genes promoters in samples from TCGA (tumor and peritumoral samples)

Format

A `RangedSummarizedExperiment` object with 79445 rows and 3423 columns

- Rows correspond to Infinium 450k probes
- Columns correspond to samples
- Methylation data from the assay are Beta values
- Clinical information are stored in colData
- Probe information (hg38 coordinates) are stored in rowRanges

Source

SKCM, LUAD, LUSC, COAD, ESCA, BRCA and HNSC methylation data were downloaded with TCGAbiolinks and subsetted to select probes located in CT genes promoter regions (see `scripts/make_TCGA_methylation.R` for details).

TCGA_TPM *Gene expression in TCGA samples*

Description

Gene expression data in TCGA samples (tumor and peritumoral samples).

Format

A `SummarizedExperiment` object with 24497 rows and 4141 columns

- Rows correspond to genes (ensembl_gene_id)
- Columns correspond to samples
- Expression data from the assay are TPM values
- Clinical information are stored in colData
- Genes information are stored in rowData

Details

- The colData contains clinical data from TCGA as well as global hypomethylation levels obtained from paper *DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load* from Jang et al., Nature Commun 2019 that were added (see `inst/scripts/make_TCGA_TPM.R` for details).
- The rowData contains genes information and, for each gene, the percentage of tumors that are positive (TPM >= 1), and the percentage of tumors that are negative (TPM < 0.5). In column `TCGA_category`, genes are labelled as "activated" when the percentage of positive tumors is > 1, with a maximal expression higher than 5 TPM, and when at least 20% of tumors are negative. Genes are labelled as "not_activated" when the percentage of positive tumors is lower than 1. Genes are labelled as "leaky" when less than 20% of tumors are negative. Genes are labelled as "lowly_expressed" when repressed (TPM <= 0.5) in at least 20%, expressed (TPM >= 1) in more than 1 % of cell lines, with a maximum expression lower than 5 TPM.

Source

SKCM, LUAD, LUSC, COAD, ESCA, BRCA and HNSC expression data were downloaded with TCGAbiolinks (see `scripts/make_TCGA_TPM.R` for details).

testis_sce *Testis scRNAseq data*

Description

Testis single cell RNAseq data from `The adult human testis transcriptional cell atlas` (Guo et al. 2018)

Format

A SingleCellExperiment object with 20891 rows and 6490 columns

- Rows correspond to genes (gene names as rownames)
- Columns correspond to testis cells

Details

Description of the colData:

- Column nGene gives the number of distinct genes detected per cell.
- Column nUMI gives the total UMI number per cell.
- Column clusters gives cluster number defined in the Guo's paper.
- Column type gives the testis cell type associated to the cluster number.
- Column Donor gives the Donor origin of the cells.

Description of the rowData:

- Column percent_pos_testis_germcells gives the percent of testis germ cells in wich the genes are detected (count > 0) (based on testis scRNAseq data).
- Column percent_pos_testis_somatic gives the percent of testis somatic cells in wich the genes are detected (count > 0) (based on testis scRNAseq data).
- Columntestis_cell_type specifies the testis cell-type showing the highest mean expression of each gene (based on testis scRNAseq data).

The rowData contains the testis_cell_type column, specifying the testis cell-type showing the highest mean expression of each gene.

Source

The count matrix GSE112013_Combined_UMI_table.txt.gz was downloaded from GEO (accession: GSE112013). Metadata correspond to TableS1 from the paper's supplemental data. The data were converted in a SingleCellExperiment (see scripts/13_make_testis_sce.R for details).

Index