Package 'CTdata'

April 26, 2024

Title Data companion to CTexploreR

Version 1.3.0

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 7.2.3

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 b221d27

git_last_commit_date 2023-10-24

Repository Bioconductor 3.19

Date/Publication 2024-04-26

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

CCLE_correlation_matrix . . . 2
CCLE_data . . . 3
CTdata . . . 4
CT_genes . . . 4
CT_mean_methylation_in_tissues . . . 6
CT_methylation_in_tissues . . . 7
DAC_treated_cells . . . 8
DAC_treated_cells_multimapping . . . 8
GTEX_data . . . 9
makeTags . . . 10
normal_tissues_multimapping_data . . . 11
scRNAseq_HPA . . . 12
TCGA_CT_methylation . . . 13
TCGA_TPM . . . 13
testis_sce . . . 14

Index **16**

`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 298 rows and 24327 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 24327 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 at a highly level (TPM >= 10).
- A column percent_of_negative_CCLE_cell_lines that gives the percent of CCLE cell lines (all cell lines combined) in which genes are completely repressed (TPM < 0.1)
- 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 highly expressed (TPM >= 10) in at least one cell line and repressed (TPM <= 0.1) in at least 20% of cell lines. "not_activated" category corresponds to genes repressed (TPM <= 0.1) in at least 20% of cell lines and never expressed (TPM >= 10) in any cell line. "leaky" category corresponds to genes repressed (TPM <= 0.1) in less than 20% of cell lines.

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 298 rows and 36 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 `family` gives the gene family name.
- Columns `chr`, `strand` and `transcription_start_site` give the genegenomic location.
- Column `X_linked` indicates if the gene is on the chromosome X (TRUE) or not (FALSE).
- Column `TPM_testis` gives the gene expression level in testis (using GTEx database).
- Column `max_TPM_somatic` gives the maximum expression level found in a somatic tissue (using GTEx database).
- 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 `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 `testis_specificity` gives the testis-specificity of genes assigned to each gene using `GTEX_category` and `multimapping_analysis` ("testis_specific" or "testis_preferential").
- Column`testis_cell_type` specifies the testis cell-type showing the highest mean expression of each gene (based on testis scRNAseq data).
- Column `Higher_in_somatic_cell_type` specifies if a somatic cell type was found to express the gene at a higher level than any germ cell type (based on scRNAseq data of different tissues).
- 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 >= 10).
- Column `percent_of_negative_CCLE_cell_lines` gives the percentage of CCLE cancer cell lines in which genes are repressed (TPM <= 0.1).
- 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 one cell line (TPM >= 10) 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 >= 10).
- Column `percent_neg_tum` gives the percentage of TCGA cancer samples in which genes are repressed (TPM <= 0.1).
- Column `max_TPM_in_TCGA` gives the highest expression level of genes in TCGA cancer sample.
- Column `TCGA_category` gives the category assigned to each gene using TCGA data. "activated" category corresponds to genes expressed in at least one tumor (TPM >= 10) 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.

- 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 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).
- 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 levels in normal tissues (when available, has to be methylated in somatic and unmethylated in germline).
- 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).
- 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.
- Columns oncogene and tumor_suppressor informs if oncogenic and tumor-suppressor functions have been associated to genes (source: Cancermine).

Source

See scripts/make_CT_genes.R for details on how this list of curated CT genes was created.

CT_mean_methylation_in_tissues

CT genes' promoters mean methylation

Description

Mean methylation values of all CpGs located within Cancer-Testis (CT) promoters in a set of normal tissues

Format

A SummarizedExperiment object with 298 rows and 14 columns

- Rows correspond to CT genes
- 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_CT_mean_methylation_in_tissues.R` for details).

`CT_methylation_in_tissues`

Methylation of CpGs within CT promoters

Description

Methylation values of CpGs located within Cancer-Testis (CT) promoters in a set of normal tissues.

Format

A `RangedSummarizedExperiment` object with 51725 rows and 14 columns

- Rows correspond to CpGs (located within CT genes promoters)
- 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_CT_methylation_in_t`
for details).

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 24359 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 <= 0.05) in at least one cell line.

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 24359 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 <= 0.05) in at least one cell line.

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.

GTEX_data *Genes expression in GTEX*

Description

Gene expression data in normal tissues from GTEx database.

Format

A SummarizedExperiment object with 24359 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, quantile 75% of expression in somatic tissues < 0.5 TPM, 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 `max_TPM_somatic` giving the maximum expression level found 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).

`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
```

---

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 24359 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.

Source

RNAseq fastq files were downloaded from Encode database (see `scripts/make_normal_tissues_multimapping.R` 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_CT_methylation *Methylation of CT promoters in TCGA samples*

Description

Methylation values of probes located within Cancer-Testis (CT) promoters in samples from TCGA (tumor and peritumoral samples)

Format

A `RangedSummarizedExperiment` object with 666 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_CT_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 24350 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 >= 10), and the percentage of tumors that are negative (TPM < 0.1). In column `TCGA_category`, genes are labelled as "activated" when the percentage of positive tumors is > 0 and when at least 20% of tumors are negative. Genes are labelled as "not_activated" when the percentage of positive tumors is 0. Genes are labelled as "leaky" when less than 20% of tumors are negative.

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 19777 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).
- Column`testis_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: GSE11201). 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

CCLE_correlation_matrix, 2
CCLE_data, 3
CT_genes, 4
CT_mean_methylation_in_tissues, 6
CT_methylation_in_tissues, 7
CTdata, 4
CTdata(), *4*

DAC_treated_cells, 8
DAC_treated_cells_multimapping, 8

GTEX_data, 9

makeTags, 10

normal_tissues_multimapping_data, 11

scRNAseq_HPA, 12

TCGA_CT_methylation, 13
TCGA_TPM, 13
testis_sce, 14