Package 'CDSeq'

February 10, 2021

Type Package

Title A Complete Deconvolution Method using Sequencing Data

Depends R (>= 3.6.0)

biocViews

Version 1.0.8

Date 2021-01-29

Maintainer Kai Kang <kangkai0714@gmail.com>

Description Estimate cell-type-specific gene expression profiles and sample-specific cell-type proportions simultaneously using bulk sequencing data. Kang et al. (2019) <doi:10.1371/journal.pcbi.1007510>.

License GPL-3

Imports Rcpp (>= 1.0.3), MASS, foreach, doParallel, dirmult, RcppThread, iterators, parallel, qlcMatrix, gplots, grDevices, clue, Biobase, Seurat, ggplot2, magrittr, dplyr, rlang, Matrix, matrixStats, ggpubr

LinkingTo Rcpp, RcppArmadillo,

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

URL https://github.com/kkang7/CDSeq_R_Package, https://doi.org/10.1371/journal.pcbi.1007510

BugReports https://github.com/kkang7/CDSeq_R_Package/issues

Config/testthat/edition 3

NeedsCompilation yes

Author Kai Kang [aut, cre, cph],
Caizhi Huang [aut],
Qian Meng [ctb],
Igor Shats [ctb],
Melissa Li [ctb],
David Umbach [ctb],
Leping Li [aut, cph],
Yuanyuan Li [ctb],
Xiaoling Li [ctb]

Repository CRAN

Date/Publication 2021-02-10 16:10:02 UTC

R topics documented:

CDSeq-R-package *CDSeq: A package for complete deconvolution using sequencing data*

Description

CDSeq-R-package takes bulk RNA-seq data as input and simultaneously returns estimates of both cell-type-specific gene expression profiles and sample-specific cell-type proportions.

Reduce-Recover

CDSeq uses reduce-recovery strategy and CPU parallel computing to speed up the deconvolution.

Hyperparameter estimation

Estimate hyperparameter for cell-type-specific GEPs (i.e. beta) using reference profile when cell_type_number is scalar.

Estimating number of cell type

Estimate number of cell types when cell_type_number is a vector of integers.

Partition on input bulk RNA-seq data

When block_number (number of partition on the bulk RNASeq data) is 1, whole bulk_data will be used. GEP is not from reduce-recovery.

Author(s)

Kai Kang, David Huang, <kangkai0714@gmail.com>

References

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007510

CDSeq *Complete deconvolution using sequencing data.*

Description

CDSeq takes bulk RNA-seq data as input and simultaneously returns estimates of both cell-type-specific gene expression profiles and sample-specific cell-type proportions.

Usage

```
CDSeq(
  bulk_data,
  beta = 0.5,
  alpha = 5,
  cell_type_number = NULL,
  mcmc_iterations = 700,
  dilution_factor = 1,
  gene_subset_size = NULL,
  block_number = 1,
  cpu_number = NULL,
  gene_length = NULL,
  reference_gep = NULL,
  verbose = FALSE,
  print_progress_msg_to_file = 0
)
```

Arguments

bulk_data	RNA-Seq read counts matrix. Columns represent samples and rows represent genes.
beta	beta is a scalar or a vector of length G where G is the number of genes; default value for beta is 0.5; When beta=Null, CDSeq uses reference_gep to estimate beta.
alpha	alpha is a scalar or a vector of length cell_type_number where cell_type_number is the number of cell type; default value for alpha is 5.
cell_type_number	number of cell types. cell_type_number can be an integer or a vector of different integers. To estimate the number of cell types, please provide a vector for cell_type_number, e.g. cell_type_number <- 2:30, then CDSeq will estimate the number of cell types.
mcmc_iterations	number of iterations for the Gibbs sampler; default value is 700.
dilution_factor	a scalar to dilute the read counts for speeding up; default value is 1. CDSeq will use bulk_data/dilution_factor.

`gene_subset_size`
: number of genes randomly sampled for each block. Default is NULL.

`block_number`
: number of blocks. Each block contains gene_subset_size genes. Default is 1.

`cpu_number`
: number of cpu cores that can be used for parallel computing; Default is NULL and CDSeq will detect the available number of cores on the device and use number of all cores - 1 for parallel computing.

`gene_length`
: a vector of the effective length (gene length - read length + 1) of each gene; Default is NULL.

`reference_gep`
: a reference gene expression profile can be used to determine the cell type and/or estimate beta; Default is NULL.

`verbose`
: if TRUE, then print progress message to the console. Default is FALSE.

`print_progress_msg_to_file`
: print progress message to a text file. Set 1 if need to print progress msg to a file and set 0 if no printing. Default is 0;

Value

CDSeq returns estimates of both cell-type-specific gene expression profiles and sample-specific cell-type proportions. CDSeq will also return estimated number of cell types. and the log posterior values for different number of cell types.

Examples

```
result1<-CDSeq(bulk_data =  mixtureGEP, cell_type_number = 6, mcmc_iterations = 5,
       dilution_factor = 50, block_number = 1, gene_length = as.vector(gene_length),
       reference_gep = refGEP, cpu_number=1, print_progress_msg_to_file=0)
```

---

cdseq.result *Output of synthetic mixtures of PBMC scRNAseq data*

Description

Output of synthetic mixtures of PBMC scRNAseq data

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

This is the CDSeq output of synthetic PBMC mixtures

`Cell2RNA` *Cell proportion to RNA proportion* `Cell2RNA` *converts Cell proportion to RNA proportion*

Description

Cell proportion to RNA proportion `Cell2RNA` converts Cell proportion to RNA proportion

Usage

```
Cell2RNA(eta, cellprop)
```

Arguments

`eta`	numeric vector represents the different amounts of RNA produced by different cell types
`cellprop`	sample-specific cell-type proportion

Value

Cell2RNA returns sample-specific cell-type RNA proportion

`cellTypeAssign` *Assign cell types using correlation matrix computed using cell-type-specific GEPs and reference GEPs.* `cellTypeAssign` *assigns CDSeq-identified cell types to reference profile.*

Description

Assign cell types using correlation matrix computed using cell-type-specific GEPs and reference GEPs. `cellTypeAssign` assigns CDSeq-identified cell types to reference profile.

Usage

```
cellTypeAssign(corMat, threshold = 0.8)
```

Arguments

`corMat`	correlation matrix between CDSeq-estimated GEPs and reference GEPs.
`threshold`	only the correlations that are above threshold will be considered.

Value

cellTypeAssign returns a vector of cell type assignment to the reference profile.

cellTypeAssignMarkerGenes

cellTypeAssignMarkerGenes *assigns CDSeq-identified cell types using user-provided marker gene list and plots heatmap.*

Description

cellTypeAssignMarkerGenes assigns CDSeq-identified cell types using user-provided marker gene list and plots heatmap.

Usage

```
cellTypeAssignMarkerGenes(
  cell_gep = NULL,
  marker_gene_list = NULL,
  threshold = 2,
  fig_path = getwd(),
  rowlabelsize = 1,
  collabelsize = 1,
  margins = c(3, 0),
  fig_width = 100,
  fig_height = 100,
  keysize = 1,
  srtcol = 45,
  keypar = c(3.5, 0, 3, 0),
  heatmap_name = "cellTypeAssign_heatmap.pdf",
  heatmap_name_fuzzy_assign = "cellTypeAssign_heatmap_fuzzy.pdf",
  verbose = FALSE
)
```

Arguments

cell_gep	gene expression profile matrix with G rows (genes) and M columns (cell types).
marker_gene_list	a G (genes) by C (cell types with known identities) matrix or dataframe that contains the marker genes for each cell type. Column names must be CellType and GeneName.
threshold	a numeric value that provides the threshold of whether a known cell type in the marker gene list can be identified.
fig_path	the location where the heatmap figure is saved.
rowlabelsize	row label size
collabelsize	column label size
margins	a vector of length 2 indicates row and column label margins
fig_width	figure width for pdf figure
fig_height	figure height for pdf figure

keysize	color key size for heatmap
srtcol	column label angle
keypar	color key layout
heatmap_name	the name of heatmap figure of one-to-one assignment.
heatmap_name_fuzzy_assign	the name of heatmap figure of fuzzy assignment.
verbose	if TRUE, some information will be printed.

Value

cellTypeAssignMarkerGenes returns a list containing: GEP_markerSum (a A by B matrix where A is nrow(marker_gene_list), B is ncol(cell_gep)),

GEP_markerSum_zscore (row-wise z score of GEP_markerSum),

GEP_matched is cell_gep(,cell_type_idx),

cell_type_idx (column index of cell_gep that are considered matching with cell types in marker_gene_list),

cell_type_matched stores the cell types in marker_gene_list that are considered to be matched with cell_gep,

GEP_markerSum_zscore_matched contains only the rows of GEP_markerSum_zscore that are considered to be matched with some cell types in cell_gep. GEP_markerSum_zscore_matched and GEP_markerSum_zscore have same columns.

cell_type_matched_fuzzy is a zero-one matrix that has the same size as GEP_markerSum_zscore_matched. If (i,j) element is one, means ith cell type in marker_gene_list is assigned to jth element in cell_gep.

---

cellTypeAssignSCRNA *cellTypeAssignSCRNA assigns CDSeq-identified cell types using single cell RNAseq data.*

Description

cellTypeAssignSCRNA assigns CDSeq-identified cell types using single cell RNAseq data.

Usage

```
cellTypeAssignSCRNA(
  cdseq_gep = NULL,
  cdseq_prop = NULL,
  cdseq_gep_sample_specific = NULL,
  sc_gep = NULL,
  sc_annotation = NULL,
  nb_size = NULL,
  nb_mu = NULL,
  seurat_count_threshold = 100,
  seurat_scale_factor = 10000,
  seurat_norm_method = "LogNormalize",
```

```
  seurat_select_method = "vst",
  seurat_nfeatures = 100,
  seurat_npcs = 50,
  seurat_dims = 1:10,
  seurat_reduction = "pca",
  seurat_resolution = 0.8,
  seurat_find_marker = FALSE,
  seurat_DE_test = "wilcox",
  seurat_DE_logfc = 0.25,
  seurat_top_n_markers = 10,
  sc_pt_size = 1,
  cdseq_pt_size = 3,
  plot_umap = 1,
  plot_tsne = 1,
  plot_per_sample = 0,
  fig_save = 0,
  fig_path = getwd(),
  fig_name = "cellTypeAssignSCRNA",
  fig_format = "pdf",
  fig_dpi = 300,
  verbose = FALSE
)
```

Arguments

cdseq_gep	CDSeq-estimated gene expression profile matrix with G rows (genes) and T columns (cell types).
cdseq_prop	CDSeq-estimated sample-specific cell-type proportion, a matrix with T rows (cell type) and M (sample size).
cdseq_gep_sample_specific	CDSeq-estimated sample-specific cell type gene expression, in the form of read counts. It is a 3 dimension array, i.e. gene by sample by cell type. The element cdseq_gep_sample_specific[i,j,k] represents the reads mapped to gene i from cell type k in sample j.
sc_gep	a G (genes) by N (cell) matrix or dataframe that contains the gene expression profile for N single cells.
sc_annotation	a dataframe contains two columns "cell_id" and "cell_type". cell_id needs to match with the cell_id in sc_gep but not required to have the same size. cell_type is the cell type annotation for the single cells.
nb_size	size parameter for negative binomial distribution, check rnbinom for details.
nb_mu	mu parameter for negative binomial distribution, check rnbinom for details.
seurat_count_threshold	this parameter will be passed to Seurat subset function (subset = nCount_RNA > seurat_count_threshold) for filtering out single cells whose total counts is less this threshold.
seurat_scale_factor	this parameter will be passed to scale.factor in Seurat function NormalizeData.

`seurat_norm_method`
: this parameter will be passed to normalization.method in Seurat function NormalizeData.

`seurat_select_method`
: this parameter will be passed to selection.method in Seurat function FindVariableFeatures

`seurat_nfeatures`
: this parameter will be passed to nfeatures in Seurat function FindVariableFeatures.

`seurat_npcs`
: this parameter will be passed to npcs in Seurat function RunPCA.

`seurat_dims`
: this parameter will be passed to dims in Seurat function FindNeighbors.

`seurat_reduction`
: this parameter will be passed to reduction in Seurat function FindNeighbors.

`seurat_resolution`
: this parameter will be passed to resolution in Seurat function FindClusters.

`seurat_find_marker`
: this parameter controls if run seurat FindMarker function, default is FALSE.

`seurat_DE_test`
: this parameter will be passed to test.use in Seurat function FindAllMarkers.

`seurat_DE_logfc`
: this parameter will be passed to logfc.threshold in Seurat function FindAllMarkers.

`seurat_top_n_markers`
: the number of top DE markers saved from Seurat output.

`sc_pt_size`
: point size of single cell data in umap and tsne plots

`cdseq_pt_size`
: point size of CDSeq-estimated cell types in umap and tsne plots

`plot_umap`
: set 1 to plot umap figure of scRNAseq and CDSeq-estimated cell types, 0 otherwise.

`plot_tsne`
: set 1 to plot tsne figure of scRNAseq and CDSeq-estimated cell types, 0 otherwise.

`plot_per_sample`
: currently disabled for debugging

`fig_save`
: 1 or 0. 1 means save figures to local and 0 means do not save figures to local.

`fig_path`
: the location where the heatmap figure is saved.

`fig_name`
: the name of umap and tsne figures. Umap figure will have the name of fig_name_umap_date and tsne figure will be named fig_name_tsne_date.

`fig_format`
: "pdf", "jpeg", or "png".

`fig_dpi`
: figure dpi

`verbose`
: if TRUE, some calculation information will be print.

Value

cellTypeAssignSCRNA returns a list containing following fields: fig_path: same as the input fig_path

fig_name: same as the input fig_name

cdseq_synth_scRNA: synthetic scRNAseq data generated using CDSeq-estiamted GEPs

cdseq_scRNA_umap: ggplot figure of the umap outcome

cdseq_scRNA_tsne: ggplot figure of the tsne outcome

cdseq_synth_scRNA_seurat: Seurat object containing the scRNAseq combined with CDSeq-estimated cell types. Cell id for CDSeq-estimated cell types start with "CDSeq".

seurat_cluster_purity: for all cells in a Seurat cluster i, the ith value in seurat_cluster_purity is the proportion of the mostly repeated cell annotation from sc_annotation. For example, after Seurat clustering, suppose there are 100 cells in cluster 1, out of these 100 cells, 90 cells' annotation in sc_annotation is cell type A, then the fist value in seurat_cluster_purity is 0.9. This output can be used to assess the agreement between Seurat clustering and the given sc_annotation.

seurat_unique_clusters: Unique Seurat cluster numbering. This can be used together with seurat_cluster_gold_label to match the Seurat clusters with given annotations.

seurat_cluster_gold_label: The cell type annotations for each unique Seurat cluster based on sc_annotation.

seurat_markers: DE genes for each Seurat cluster.

seurat_top_markers: Top seurat_top_n_markers DE genes for each Seurat cluster.

CDSeq_cell_type_assignment_df: cell type assignment for CDSeq-estimated cell types.

cdseq_prop_merged: CDSeq-estimated cell type proportions with cell type annotations.

cdseq_gep_sample_specific_merged: sample-specific cell-type read counts. It is a 3d array with dimensions: gene, sample, cell type.

input_list: values for input parameters

cdseq_sc_comb_umap_df: dataframe for umap plot

cdseq_sc_comb_tsne_df: dataframe for tsne plot

`gene2rpkm` *gene2rpkm outputs the rpkm normalizations of the CDSeq-estimated GEPs.* `gene2rpkm` *outputs the rpkm normalizations of the CDSeq-estimated GEPs.*

Description

gene2rpkm outputs the rpkm normalizations of the CDSeq-estimated GEPs. `gene2rpkm` outputs the rpkm normalizations of the CDSeq-estimated GEPs.

Usage

```
gene2rpkm(gene_rate, gene_effective_length, cell_line_counts)
```

Arguments

`gene_rate`	CDSeq-estimated GEP normalized by gene length.
`gene_effective_length`	gene effective length which is the gene length minus the read length plus one.
`cell_line_counts`	RNA-Seq read counts data of cell lines

Value

gene2rpkm returns rpkm normalization of the CDSeq-estimated GEPs.

`gene_length` *Gene length*

Description

Gene length

Usage

```
data(SyntheticMixtureData)
```

Format

numeric vector

Author(s)

Kai Kang

`gibbsSampler` *This is the Gibbs sampler for CDSeq.* `GibbsSampler` *returns estimated GEPs and cell type proportions.*

Description

This is the Gibbs sampler for CDSeq. `GibbsSampler` returns estimated GEPs and cell type proportions.

Usage

```
gibbsSampler(
  ALPHA,
  BETA,
  mixtureSamples,
  T,
  NN,
  OUTPUT,
  processID,
  data_block_idx,
  CDSeq_tmp_log,
  write_2_file,
  verbose
)
```

Arguments

ALPHA	hyperparameter for cell type proportion.
BETA	hyperparameter for cell-type-specific GEPs.
mixtureSamples	bulk RNA-seq data in form of read counts.
T	number of cell types.
NN	number of MCMC iteration.
OUTPUT	MCMC progress output control.
processID	worker process ID when using parallel computing.
data_block_idx	index for data blocks from bulk RNA-seq input.
CDSeq_tmp_log	temporary log file recording the workers' jobs.
write_2_file	print to progress msg to CDSeq_tmp_log if it is 1, not printing otherwise.
verbose	if greater than or euqal to 1, then print working progress in console, otherwise do not print in console.

Value

random integers uniformly distributed in 0..(2^32 - 1).

---

hungarian_Rcpp *This is the Hungarian algorithm wrapper for cell type assignment* hungarian_Rcpp *returns cell type assignment given reference GEPs*

Description

This is the Hungarian algorithm wrapper for cell type assignment hungarian_Rcpp returns cell type assignment given reference GEPs

Usage

```
hungarian_Rcpp(costMat)
```

Arguments

costMat	correlation matrix

Value

cost for the assignment and cell type assignment

`intersection` `intersection` *take intersection of multiple lists and return the common set and index*

Description

`intersection` take intersection of multiple lists and return the common set and index

Usage

```
intersection(list.vector, order = "sort")
```

Arguments

`list.vector`	this is a list of list contain all the data.
`order`	this is either sort or stable. If choose sort, the output common value will be sorted. If choose stable, the output common value will be in the same order as appear in the first element in list.vector.

Value

The common values among lists and their indices. intersection function: input: list.vector is a list of list contain all the data for example, if we need to find the common elements of a, b, c, then input should be list(a,b,c)

`logpost` *logpost computes the log posterior of the CDSeq model.* `logpost` *outputs the value of log posterior.*

Description

logpost computes the log posterior of the CDSeq model. `logpost` outputs the value of log posterior.

Usage

```
logpost(estProp, estGEP, mydata, alpha, beta)
```

Arguments

`estProp`	CDSeq-estimated cell type proportions.
`estGEP`	CDSeq-estimated cell-type-specific GEPs.
`mydata`	input bulk RNA-seq data.
`alpha`	hyperparameter for cell type proportion estimation.
`beta`	hyperparameter for cell-type-specific GEP estimation.

Value

logpost returns log posterior values.

`max_rep` `max_rep` *Find the element that repeats the most in a given vector and calculate its proportion.*

Description

`max_rep` Find the element that repeats the most in a given vector and calculate its proportion.

Usage

```
max_rep(v)
```

Arguments

v	a vector

Value

max_rep_value contains two elements: max_element and max_element_proportion. max_element is the element that repeats the most in v, and max_element_proportion is its proportion.

`merge_df` *Data frame for keeping the CDSeq-estimated cell type proportions for PBMC mixtures*

Description

Data frame for keeping the CDSeq-estimated cell type proportions for PBMC mixtures

Usage

```
data(SyntheticMixtureData)
```

Format

data frame

Author(s)

Kai Kang

Source

CDSeq estimated cell type proportions for cell type number 3, 6, 9 and 12

mixtureGEP — *Synthetic bulk RNA-seq read counts data of six cell types*

Description

Synthetic bulk RNA-seq read counts data of six cell types

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

we downloaded the pure cell line RNA-seq data and generated the mixing proportions randomly and produced the mixtures

pbmc_ggplot — *ggplot figures of comparison between CDSeq-estimated cell type proportion and ground truth*

Description

ggplot figures of comparison between CDSeq-estimated cell type proportion and ground truth

Usage

```
data(SyntheticMixtureData)
```

Format

ggplot object

Author(s)

Kai Kang

Source

CDseq-estimated cell type proportion and ground truth

pbmc_mix *Synthetic bulk RNA-seq read counts data of PBMC single cell data*

Description

Synthetic bulk RNA-seq read counts data of PBMC single cell data

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

we downloaded the PBMC scRNAseq and generated the mixing proportions randomly and produced the mixtures

read2gene *read2gene outputs the GEP normalized by gene length of the CDSeq-estimated GEPs.* read2gene *outputs the gene length normalized CDSeq-estimated GEP.*

Description

read2gene outputs the GEP normalized by gene length of the CDSeq-estimated GEPs. read2gene outputs the gene length normalized CDSeq-estimated GEP.

Usage

```
read2gene(read_rate, gene_effective_length)
```

Arguments

read_rate	CDSeq-estimated GEP before normalized by gene length.
gene_effective_length	gene effective length which is the gene length minus the read length plus one.

Value

read2gene returns gene length normalized CDSeq-estimated GEPs.

refGEP *GEPs of six component pure cell lines*

Description

GEPs of six component pure cell lines

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

result1 *CDSeq result of synthetic bulk RNA-seq read counts data of six cell types*

Description

CDSeq result of synthetic bulk RNA-seq read counts data of six cell types

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

CDSeq-estimates of mixtureGEP

result2 *CDSeq result of synthetic bulk RNA-seq read counts data of six cell types*

Description

CDSeq result of synthetic bulk RNA-seq read counts data of six cell types

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

CDSeq estimates of mixtureGEP

result3 *CDSeq result of synthetic bulk RNA-seq read counts data of six cell types*

Description

CDSeq result of synthetic bulk RNA-seq read counts data of six cell types

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

CDSeq estimates of mixtureGEP

RNA2Cell *RNA proportion to cell proportion* RNA2Cell *converts RNA proportion to cell proportion*

Description

RNA proportion to cell proportion RNA2Cell converts RNA proportion to cell proportion

Usage

```
RNA2Cell(eta, rnaprop)
```

Arguments

eta	numeric vector represents the different amounts of RNA produced by different cell types
rnaprop	sample-specific cell-type RNA proportion

Value

RNA2Cell returns sample-specific cell-type proportion

sc_annotation *Cell type annotation of the PBMC single cell data*

Description

Cell type annotation of the PBMC single cell data

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

We used the annotation provided by the single cell data

sc_gep *PBMC single cell RNAseq read counts that used for creating synthetic PBMC mixtures*

Description

PBMC single cell RNAseq read counts that used for creating synthetic PBMC mixtures

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

we downloaded the PBMC RNA-seq data and generated the mixing proportions randomly and produced the mixtures

seedMT *This is the Mersenne Twister random number generator.* cokus *generates pseudorandom integers uniformly distributed in 0..(2^32 - 1).*

Description

This is the Mersenne Twister random number generator. cokus generates pseudorandom integers uniformly distributed in 0..(2^32 - 1).

Usage

```
seedMT(seed)
```

Arguments

seed	odd number for seeding

Value

random integers uniformly distributed in 0..(2^32 - 1).

Author(s)

Shawn Cokus (Cokus@math.washington.edu)

`SyntheticMixtureData` *Synthetic bulk RNA-seq read counts data of six cell types, PBMC mixtures using scRNASeq and some preliminary results*

Description

Synthetic bulk RNA-seq read counts data of six cell types, PBMC mixtures using scRNASeq and some preliminary results

Usage

```
data(SyntheticMixtureData)
```

Format

A matrix of read counts data containing 40 synthetic mixtures with 500 genes

Author(s)

Kai Kang

Source

we downloaded the pure cell line RNA-seq data and generated the mixing proportions randomly and produced the mixtures

Examples

```
data(SyntheticMixtureData)
```

`true_GEP_gene` *True GEPs of the six component cell types normalized by gene length*

Description

True GEPs of the six component cell types normalized by gene length

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

`true_GEP_read` *True GEPs of the six component cell types unnormalized by gene length*

Description

True GEPs of the six component cell types unnormalized by gene length

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

`true_GEP_rpkm` *True GEPs of the six component cell types RPKM normalization*

Description

True GEPs of the six component cell types RPKM normalization

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

true_prop — *True cell type proportion in the PBMC synthetic mixtures*

Description

True cell type proportion in the PBMC synthetic mixtures

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

randomly generated

true_prop_cell — *True cell proportions of the mixtures*

Description

True cell proportions of the mixtures

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

cell type proportions are randomly generated

`true_prop_RNA` *True cell type RNA proportions*

Description

True cell type RNA proportions

Usage

```
data(SyntheticMixtureData)
```

Format

numeric matrix

Author(s)

Kai Kang

Source

randomly generated

Index