Package 'scDesign'

September 28, 2018

Title A statistical simulator for rational scRNA-seq experimental design

Version 0.0.1

Date 2018-08-30

Author Wei Vivian Li, Jingyi Jessica Li

Maintainer Wei Vivian Li <liw@ucla.edu>

Description A flexible and robust simulator for researchers to quantitatively assess practical scRNA-seq experimental design in the context of differential gene expression analysis. In addition to experimental design, scDesign also assists computational method development by generating high-quality synthetic scRNA-seq datasets under customized experimental settings.

Depends R (>= 3.4.2), parallel, ggplot2

Imports MAST, tidyr, dplyr, gridExtra, data.table, stats, GenomicRanges, SummarizedExperiment

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

`design_data` *use scDesign to simulate scRNA-seq data*

Description

use scDesign to simulate scRNA-seq data

Usage

```
design_data(realcount, S = 1e+08, ncell, ngroup = 1, pUp = 0.05,
  pDown = 0.05, fU = 5, fL = 1.5, ncores = 1)
```

Arguments

realcount	A numeric matrix with rows representing genes and columns representing cells. Gene names are given as row names.
S	A number specifying the total number of RNA-seq reads. Default to 1e8.
ncell	An integer specifying the number of cells. When ngroup > 1, ncell is the number of cells in each cell state.
ngroup	An integer giving the number of cell states to simulate. Defaults to 1.
pUp	A value between 0 and 1 specifying the proportion of up regulated genes between two adjacent cell states. Defaults to 0.05 and only used when ngroup > 1.
pDown	A value between 0 and 1 specifying the proportion of down regulated genes between two adjacent cell states. Defaults to 0.05 and only used when ngroup > 1.
fU	A value specifying the upper bound of fold changes of differentially expressed genes. Deaults to 5.
fL	A value specifying the lower bound of fold changes of differentially expressed genes. Deaults to 1.5.
ncores	An integer specifying the number of cores used for parallel computation. Defaults to 1.

Value

When ngroup = 1, design_data returns a simulated count matrix with rows representing genes and columns representing cells. When ngroup > 1, design_data returns a list of ngroup elements. The g-th element corresponds to the g-th cell state, and is a list containing three elements:

count: a count matrix with rows representing genes and columns representing cells;

genesUp: a character vector giving the names of up-regulated genes from state g-1 to g;

genesDown: a character vector giving the names of down-regulated genes from state g-1 to g.

Author(s)

Wei Vivian Li, <liw@ucla.edu>

Jingyi Jessica Li, <jli@stat.ucla.edu>

---

design_joint *use scDesign to make experimental design assuming two cell states are sequenced together*

Description

use scDesign to make experimental design assuming two cell states are sequenced together

Usage

```
design_joint(realcount1, realcount2, prop1, prop2, S = 1e+08,
  ncell = round(2^seq(6, 13, 1)), B = 100, de_method = "ttest",
  p_thre = 10^seq(-2, -6, -1), plot_dir = "./", ncores = 1, rank = 1000)
```

Arguments

realcount1	A numeric matrix with rows representing genes and columns representing cells (cell state 1). Gene names are given as row names.
realcount2	A numeric matrix with rows representing genes and columns representing cells (cell state 2). Gene names are given as row names.
prop1	A number giving the proportion of state 1 cells in the cell population.
prop2	A number giving the proportion of state 2 cells in the cell population.
S	A number specifying the total number of RNA-seq reads for the cell population. Default to 1e8.
ncell	An integer vector specifying the total number of cells to sequence. Defaults to `2^seq(6,13,1)`.
B	An integer giving the number of experiments to repeat in order the calculate the average DE analysis accuracy. Defaults to 100.
de_method	A character specifying the differential expression analysis method to use. Currently supports "ttest" (default) or "mast".
p_thre	A numeric vector specifying the FDR thresholds used to identify differentially expressed genes. Defaults to `10^seq(-2,-6,-1)`.
plot_dir	A character giving the directory to save experimental design results Defaults to "./".
ncores	An integer specifying the number of cores used for parallel computation. Defaults to 1.
rank	An integer specifying the number of top DE genes to identify from scImpute's results as the standard in DE analysis. Defaults to 1000.

Value

A list of five elments:

precision: a matrix of precision.

recall: a matrix of recall (true positive rate).

TN: a matrix of TN (true negative rate).

F1: a matrix of F1 (precision vs. recall).

F2: a matrix of F2 (TN vs. recall).

In all the matrices, rows correspond to different FDR thresholds and columns correspond to the cell numbers specified in `ncell`. `design_joint` also writes the list to design_summary.txt and saves it to `plot_dir`. The corresponding plots are also saved to `plot_dir`.

Author(s)

Wei Vivian Li, <liw@ucla.edu>

Jingyi Jessica Li, <jli@stat.ucla.edu>

`design_sep` *use scDesign to make experimental design assuming two cell states are sequenced independently*

Description

use scDesign to make experimental design assuming two cell states are sequenced independently

Usage

```
design_sep(realcount1, realcount2, S1 = 1e+08, S2 = 1e+08, ncell = NULL,
  B = 100, de_method = "ttest", p_thre = 10^seq(-2, -6, -1),
  plot_dir = "./", ncores = 1, rank = 1000)
```

Arguments

`realcount1`	A numeric matrix with rows representing genes and columns representing cells (cell state 1). Gene names are given as row names.
`realcount2`	A numeric matrix with rows representing genes and columns representing cells (cell state 2). Gene names are given as row names.
`S1`	A number specifying the total number of RNA-seq reads for cell state 1. Default to 1e8.
`S2`	A number specifying the total number of RNA-seq reads for cell state 2. Default to 1e8.
`ncell`	A two-column matrix specifying the numbers of cells. Column 1 is for cell state 1 and column 2 is for cell state 2. By default, the following cell number matrix is used:

64	64
128	128
256	256
512	512
1024	1024
2048	2048
4096	4096

`B`	An integer giving the number of experiments to repeat in order the calculate the average DE analysis accuracy. Defaults to 100.
`de_method`	A character specifying the differential expression analysis method to use. Currently supports "ttest" (default) or "mast".
`p_thre`	A numeric vector specifying the FDR thresholds used to identify differentially expressed genes. Defaults to `10^seq(-2,-6,-1)`.
`plot_dir`	A character giving the directory to save experimental design results Defaults to "./".
`ncores`	An integer specifying the number of cores used for parallel computation. Defaults to 1.
`rank`	An integer specifying the number of top DE genes to identify from scImpute's results as the standard in DE analysis. Defaults to 1000.

Value

A list of five elments:

precision: a matrix of precision.

recall: a matrix of recall (true positive rate).

TN: a matrix of TN (true negative rate).

F1: a matrix of F1 (precision vs. recall).

F2: a matrix of F2 (TN vs. recall).

In all the matrices, rows correspond to different FDR thresholds and columns correspond to the cell numbers specified in `ncell`. `design_sep` also writes the list to design_summary.txt and saves it to `plot_dir`. The corresponding plots are also saved to `plot_dir`.

Author(s)

Wei Vivian Li, <liw@ucla.edu>

Jingyi Jessica Li, <jli@stat.ucla.edu>

Index