Package 'scDEA'

August 16, 2021

Type Package

Title Performing differential expression analysis in single cell RNA sequencing data via ensemble learning

Version 0.1.0

Author Hui-sheng, Li

Maintainer Hui-sheng, Li<lihs@mails.ccnu.edu.cn>

Description scDEA is an ensemble learning method for differential expression analysis in single cell RNA sequencing data. It ensembles results from multiple individual differential expression analysis methods. The current implementation of scDEA integrates twelve state-of-the-art methods: BPSC, DEsingle, DESeq2, edgeR, MAST, monocle, scDD, T-test, Wilcoxon, limma, Seurat, zingeR.edgeR.

Depends R(>= 3.6.3)

Imports BPSC(>= 0.99.2), DEsingle(>= 1.6.0), DESeq2(>= 1.26.0), edgeR(>= 3.28.1), MAST(>= 1.12.0), monocle(>= 2.14.0), scDD(>= 1.10.0), limma(>= 3.42.2), Seurat(>= 3.2.2), zingeR(>= 0.1.0), SingleCellExperiment(>= 1.8.0), scater(>= 1.14.6), aggregation(>= 1.0.1), BiocGenerics(>= 0.32.0), S4Vectors(>= 0.24.4), VGAM(>= 1.1-4), methods(>= 3.6.3), parallel(>= 3.6.3), SummarizedExperiment(>= 1.16.1), stats(>= 3.6.3), Matrix(>= 1.2-18)

Suggests knitr, rmarkdown

biocViews

VignetteBuilder knitr

RoxygenNote 7.1.1

License GPL(>= 2)

Encoding UTF-8

LazyData true

NeedsCompilation no

R topics documented:

`data_process` *Data process*

Description

This function focus on dealing various single-cell RNA-seq input and unifying output format.

Usage

```
data_process(Data, group, norm.form = "CPM", is.normalized = FALSE)
```

Arguments

`Data`	single-cell RNA-seq matrix. The format could be raw-counts, FPKM/RPKM, TPM or UMI-counts. The matrix need include gene names and cell names.
`group`	group information. The cell need be divided into two category.
`norm.form`	character item. We provide several normalized method for raw-counts data. The method include "TMM","RLE", "CPM", "TPM". The default is "CPM".
`is.normalized`	logical. A logical flag to determin whether or not the input dataset normalizes. If TRUE, we will take the Data as normcounts and input for downstream analysis. If not, we provide method for the process.

Details

We take `relative2abs` transfering relative expression values into absolute transcript counts. However, the process maybe break the original dataset statistical properties. Hence, we advise user don't normalize firstly.

Value

- **sce** : A `SingleCellExperiment` item. The object include expression matrix, group information. The expression matrix contains counts and normcounts.

Author(s)

Huisheng, Li, <lihs@mails.ccnu.edu.cn>

Examples

```
data("Grun.counts.matrix")
data("Grun.group.information")
sce <- data_process(Data = Grun.counts.matrix, group = Grun.group.information)
```

Grun.counts.matrix *Gene expression count matrix of Grun*

Description

We obtain the preprocessing Grun data from Soneson. Two group cell types, WG and YPS, are selected to perform DE analysis. We use the function FindVariableFeatures in Seurat R package to select 20000 highly variable genes. In this study, the Grun data contains 20000 genes, 338 cells of WG and 378 cells of YPS.

Usage

```
data(Grun.counts.matrix)
```

Format

a large matrix

Examples

```
data(Grun.counts.matrix)
```

Grun.group.information *Cell type lebals of Grun data*

Description

Cell type lebals of Grun data

Usage

```
data(Grun.group.information)
```

Format

a vector

Examples

```
data(Grun.group.information)
```

lancaster.combination *Combining the p-values identified by DE ananlysis methods*

Description

This function is used to combine the results of differential expression anlaysis and obtain an uniform result.

Usage

```
lancaster.combination(Pvals, weight = TRUE, trimmed = 0.2)
```

Arguments

Pvals	a p-value matrix. The rows represent genes and the columns correspond to the individual DE analysis methods.
weight	a boolean variable that defines whether to use spearman correlation measure the similarity between different DE analysis methods. Default is "TRUE".
trimmed	a real number between 0 and 0.5 to trim p-values. Default value is 0.2.

Value

a vector represents the combined p-value for each gene.

Author(s)

Huisheng, Li, <lihs@mails.ccnu.edu.cn>

normalized *Normalized process*

Description

The function provide several normalized methods

Usage

```
normalized(counts.matrix, method = "CPM")
```

Arguments

counts.matrix	count expression matrix
method	character. "TMM", "RLE", "CPM", "TPM". The default value is "CPM".

`scDEA.p.adjust` *adjusting the p-values identified by individual DE ananlysis methods and scDEA*

Description

This function is used to adjust the p-values generated by individual DE analysis methods and scDEA.

Usage

```
scDEA.p.adjust(combination.Pvals, adjusted.method = "bonferroni")
```

Arguments

`combination.Pvals`
: a vector of combined p-value for each gene.

`adjusted.method`
: a string variable specifying the type of p.adjust method used in t-test method. Possible values: "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none". Default is "bonferroni".

Value

adjusted p-value ,

`scDEA_individual_methods` *Run individual DE analysis methods to perform DE analysis on scRNA-seq datasets.*

Description

This function is implemented to perform individual DE analysis methods. The current implementation of scDEA integrates tweleve state-of-the-art methods: Beta-poisson mixture model (BPSC), DEsingle, DESeq2, edgeR, Model-based analysis of single-cell transcriptomics (MAST), monocle, scDD, T-test, Wilcoxon rank sum test (Wilcoxon test), limma, Seurat and zingeR.edgeR. This function depends on the follwing R package: BPSC, DEsingle, DESeq2, edgeR, MAST, monocle, scDD, limma, Seurat, zingeR, SingleCellExperiment, scater, dplyr. These packages will be automatically installed along with scDEA.

Usage

```
scDEA_individual_methods(
  raw.count,
  cell.label,
  is.normalized = FALSE,
  verbose = TRUE,
  BPSC = TRUE,
```

```
  DEsingle = TRUE,
  DESeq2 = TRUE,
  edgeR = TRUE,
  MAST = TRUE,
  monocle = TRUE,
  scDD = TRUE,
  Ttest = TRUE,
  Wilcoxon = TRUE,
  limma = TRUE,
  Seurat = TRUE,
  zingeR.edgeR = TRUE,
  BPSC.coef = 2,
  BPSC.normalize = "CPM",
  BPSC.parallel = TRUE,
  DEsingle.parallel = TRUE,
  DEsingle.normalize = "CPM",
  DESeq2.test = "LRT",
  DESeq2.parallel = TRUE,
  DESeq2.beta.prior = TRUE,
  DESeq2.fitType = "parametric",
  DESeq2.normalize = "CPM",
  edgeR.Test = "QLFT",
  edgeR.normalize = "TMM",
  limma.method.fit = "ls",
  limma.trend = TRUE,
  limma.robust = TRUE,
  limma.normalize = "CPM",
  Seurat.normalize = "CPM",
  Seurat.method = "bimod",
  MAST.method = "bayesglm",
  MAST.normalize = "CPM",
  MAST.parallel = TRUE,
  monocle.cores = 1,
  monocle.normalize = "CPM",
  scDD.alpha1 = 0.01,
  scDD.mu0 = 0,
  scDD.s0 = 0.01,
  scDD.a0 = 0.01,
  scDD.b0 = 0.01,
  scDD.normalize = "CPM",
  scDD.permutation = 0,
  Ttest.normalize = "CPM",
  Wilcoxon.normalize = "CPM",
  zingeR.edgeR.normalize = "CPM",
  zingeR.edgeR.maxit.EM = 100
)
```

Arguments

raw.count	single-cell RNA-seq matrix. The format could be raw read count or normalized matrix. The rows correspond to genes and the columns.
cell.label	cell labels information. The cells need be divided into two categories.

`is.normalized` a boolean variable that defines whether the input raw.count has been normalized? Default is FALSE.

`verbose` a boolean variable that defines whether to save the DE analysis results and name "Results_DE_individual.RData" in the current working directory.

`BPSC` a boolean variable that defines whether to perform DE analysis using the BPSC method. Default is TRUE.

`DEsingle` a boolean variable that defines whether to perform DE analysis using the DEsingle method. Default is TRUE.

`DESeq2` a boolean variable that defines whether to perform DE analysis using the DESeq2 method. Default is TRUE.

`edgeR` a boolean variable that defines whether to perform DE analysis using the edgeR method. Default is TRUE.

`MAST` a boolean variable that defines whether to perform DE analysis using the MAST method. Default is TRUE.

`monocle` a boolean variable that defines whether to perform DE analysis using the MONOCLE method. Default is TRUE.

`scDD` a boolean variable that defines whether to perform DE analysis using the scDD method. Default is TRUE.

`Ttest` a boolean variable that defines whether to perform DE analysis using the T-test method. Default is TRUE.

`Wilcoxon` a boolean variable that defines whether to perform DE analysis using the Wilcoxon method. Default is TRUE.

`limma` a boolean variable that defines whether to perform DE analysis using the limma method. Default is TRUE.

`Seurat` a boolean variable that defines whether to perform DE analysis using the Seurat method. Default is TRUE.

`zingeR.edgeR` a boolean variable that defines whether to perform DE analysis using the zingeR.edgeR method. Default is TRUE.

`BPSC.coef` an integer to point out the column index corresponding to the coefficient for the generalized linear mode (GLM) testing in BPSC. Default value is 2.

`BPSC.normalize` a string variable specifying the type of size factor estimation in BPSC method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`BPSC.parallel` a boolean variable that defines whether to execute parallel computing for BPSC method. Default is TRUE.

`DEsingle.parallel`
a boolean variable that defines whether to execute parallel computing for DEsingle method. Default is TRUE.

`DEsingle.normalize`
a string variable specifying the type of size factor estimation in DEsingle method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`DESeq2.test` a string variable specifying the type of test the difference in deviance between a full and reduced model formula in DESeq2 method. Possible values: "Wald" or "LRT". The values represent Wald tests or likelihood ratio test. Default is "Wald".

`DESeq2.parallel`
a boolean variable that defines whether to execute parallel computing for DESeq2 method. Default is TRUE. The parallel computing may fail on Windows system. Default is TRUE.

`DESeq2.beta.prior`
: a boolean variable that defines whether or not to put a zero-mean normal prior on the non-intercept coefficient in DESeq2 method. Default is TRUE.

`DESeq2.fitType`
: a string variable specifying the type of fitting of dispersions to the mean intensity in DESeq2 method. Possible values: "parametric", "local", "mean". Default is "parametric".

`DESeq2.normalize`
: a string variable specifying the type of size factor estimation in DESeq2 method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`edgeR.Test`
: a string variable specifying the type of fitting distribution to count data for each gene. Possible values: "LRT", "QLFT". The values represent negative binomial generalized log-linear model and quasi-likelihood negative binomial generalized log-linear model. Default is "QLFT".

`edgeR.normalize`
: a string variable specifying the type of size factor estimation in edgeR method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`limma.method.fit`
: a string variable specifying the type of fitting method in limma method. Possible values: "ls", "robust". The values represent least squares and robust regression. Default is "ls".

`limma.trend`
: a boolean variable that defines whether or not to allow an intensity-trend for the prior variance in limma method. Default is TRUE.

`limma.robust`
: a boolean variable that defines whether or not to estimate defined prior information and variance prior against outlier sample variances in limma method. Default is TRUE.

`limma.normalize`
: a string variable specifying the type of size factor estimation in limma method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`Seurat.normalize`
: a string variable specifying the type of size factor estimation in Seurat method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`Seurat.method`
: a string variable specifying the type of test method in Seurat method. Possible values: "LR", "bimod", "roc". The values represent likelihood-ratio test, negative binomial generalized linear model, ROC analysis. Default is "bimod".

`MAST.method`
: a string variable specifying the type of test method in MAST method. Possible values: "glm", "glmer", "bayesglm". Default is "bayesglm".

`MAST.normalize`
: a string variable specifying the type of size factor estimation in Seurat method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`MAST.parallel`
: a boolean variable that defines whether to execute parallel computing for MAST method. Default is TRUE.

`monocle.cores`
: the number of cores to be used while testing each gene for differential expression.. Default is 1.

`monocle.normalize`
: a string variable specifying the type of size factor estimation in monocle method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`scDD.alpha1`
: prior parameter value to be used to model each gene as a mixture of DP normals in scDD method. Default is 0.01.

`scDD.mu0`
: prior parameter values to be used to model each gene as a mixture of DP normals in scDD method. Default is 0.

`scDD.s0` prior parameter values to be used to model each gene as a mixture of DP normals in scDD method. Default is 0.01.

`scDD.a0` prior parameter values to be used to model each gene as a mixture of DP normals in scDD method. Default is 0.01.

`scDD.b0` prior parameter values to be used to model each gene as a mixture of DP normals in scDD method. Default is 0.01.

`scDD.normalize` a string variable specifying the type of size factor estimation in scDD method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`scDD.permutation`
the number of permutations to be used in calculating empirical p-values in scDD method. If the parameter value is set to 0, the full Bayes Factor will not be performed. Else, scDD method takes the nonparametric Kolmogorove-Smirnov test to identify DGEs. Default is 0.

`Ttest.normalize`
a string variable specifying the type of size factor estimation in t-test method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`Wilcoxon.normalize`
a string variable specifying the type of size factor estimation in Wilcoxon method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`zingeR.edgeR.normalize`
a string variable specifying the type of size factor estimation in zingeR.edgeR method. Possible values: "TMM", "RLE", "CPM", "TPM". Default is "CPM".

`zingeR.edgeR.maxit.EM`
The number of iterations for EM-algorithm in zingeR.edgeR method. If the EM-algorithm does not stop automatically, then, the algorithm may not be convergence. The user need set a larger value. Default is 100.

Value

a p-values matrix contains the p-values of each differential expression anlysis methods.

Author(s)

Huisheng, Li, <lihs@mails.ccnu.edu.cn>

Examples

```
data("Grun.counts.matrix")
data("Grun.group.information")
# scDD is very slow
Pvals <- scDEA_individual_methods(raw.count = Grun.counts.matrix,
cell.label = Grun.group.information,  verbose = FALSE)
combination.Pvals <- lancaster.combination(Pvals, weight = TRUE, trimmed = 0.2)
adjusted.Pvals <- scDEA.p.adjust(combination.Pvals, adjusted.method = "bonferroni")
```

Index