Package 'ssdGSA'

July 22, 2024

Type Package

Title Single Sample Directional Gene Set Analysis

Version 0.1.0

Description A method that inherits the standard gene set variation analysis (GSVA) method and also provides the option to use summary statistics from any analysis (disease vs healthy, lesional side vs nonlesional side, etc..) input to define the direction of gene sets used for directional gene set score calculation for a given disease. Hanzelmann, S., Castelo, R., and Guinney, J. (2013) <doi:10.1186/1471-2105-14-7>.

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Imports GSVA, dplyr, purrr, stringr, tibble, vctrs, clusterProfiler, stats, org.Hs.eg.db, tidyselect, utils

VignetteBuilder knitr

Depends R (>= 3.5.0)

NeedsCompilation no

Author Xingpeng Li [aut, cre], Qi Qian [aut]

Maintainer Xingpeng Li <xingpeng.li@pfizer.com>

Repository CRAN

Date/Publication 2024-07-22 17:30:02 UTC

Contents

`avg_expression` *Function to Calculate the Average of Gene Expressions*

Description

This function is to calculate the average of gene expressions for genes in the given gene sets.

Usage

```
avg_expression(Data, pathway.db)
```

Arguments

`Data`	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.
`pathway.db`	A list of gene sets.

Details

Within the ssdGSA function, when GSA_method = "avg.exprs", this function is used to calculate the average of gene expressions for genes in the given gene sets.

Value

Matrix of average gene expression in each gene set with rows corresponding to gene sets and columns corresponding to samples will be returned.

check_genes_missing *Function to Check if Genes in Gene Sets Have Missing Information in Data Matrix and Direction Matrix*

Description

This function is to check if genes in gene sets to be analyzed have missing information in data matrix and direction matrix.

Usage

```
check_genes_missing(Gene_sets, Data, Direction_matrix)
```

Arguments

Gene_sets	A list of gene sets to be analyzed, with gene set names as component names, and each component is a vector of gene entrez ID.
Data	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.
Direction_matrix	Matrix containing directionality information for each gene, such as effect size, p value of summary statistics. Each row of the matrix is for one gene, and there should be at least two columns (with the 1st column containing gene entrez ID, and 2nd column containing directionality information).

Details

Before single sample directional gene set analysis, it is necessary to check if genes in the gene sets have missing information in data matrix and direction matrix. If not, warning messages would be given such that users can double check whether the gene set analysis results are reliable.

Value

When at least one gene in the gene sets have information missing in data matrix or direction matrix, warning messages will be given, as well as the percentages (missing number/total number) of gene sets. If less than 10 gene sets have missing information, percentages (missing number/total number) of genes in each gene set that have missing information in data matrix and direction matrix will also be reported. However, if more than 10 gene sets have missing information, no detailed individual gene set missing information will be reported. Also note that if a gene set has 100% information missing in the data or direction matrix, the name of the gene set will be notated.

`check_genes_missing_noDir` *Function to Check if Genes in Gene Sets Have Missing Information in Data Matrix*

Description

This function is to check if genes in gene sets to be analyzed have missing information in data matrix.

Usage

```
check_genes_missing_noDir(Gene_sets, Data)
```

Arguments

`Gene_sets`	A list of gene sets to be analyzed, with gene set names as component names, and each component is a vector of gene entrez ID.
`Data`	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.

Details

Before single sample directional gene set analysis, it is necessary to check if genes in the gene sets have missing information in data matrix. If not, warning messages would be given such that users can double check whether the gene set analysis results are reliable.

Value

When at least one gene in the gene sets have information missing in data matrix, warning messages will be given, as well as the percentages (missing number/total number) of gene sets. If less than 10 gene sets have missing information, percentages (missing number/total number) of genes in each gene set that have missing information in data matrix and direction matrix will also be reported. However, if more than 10 gene sets have missing information, no detailed individual gene set missing information will be reported. Also note that if a gene set has 100% information missing in the data matrix, the name of the gene set will be notated.

`check_genes_missing_total` *Function to Check if 100% of Genes in Gene Sets Have Missing Information in Data Matrix*

Description

This function is to check if 100% of genes in gene sets to be analyzed have missing information in data matrix.

Usage

```
check_genes_missing_total(Gene_sets, Data)
```

Arguments

Gene_sets	A list of gene sets to be analyzed, with gene set names as component names, and each component is a vector of gene entrez ID.
Data	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.

Details

Before single sample directional gene set analysis, it is necessary to check if genes in the gene sets have missing information in data matrix. If a gene set has 100% information missing in the data matrix, the name of the gene set will be returned as a list named 'Total_missing_in_data_matrix'; If no such gene sets exist, nothing will be returned.

Value

A list 'Total_missing_in_data_matrix' with names of the gene sets that have 100% of genes that have information missing in the data matrix will be returned. If there are no such gene sets, NULL list will be returned.

---

check_gene_name_match *Function to Check if Gene IDs in the Input Gene Data Match Well*

Description

This function is to check if the gene IDs in the gene sets, data matrix and direction matrix match well.

Usage

```
check_gene_name_match(
  genes_in_Gene_sets,
  genes_in_Data,
  genes_in_Direction_matrix
)
```

Arguments

genes_in_Gene_sets	A list of gene names from the gene sets.
genes_in_Data	A list of gene names from the data matrix.
genes_in_Direction_matrix	A list of gene names from the direction matrix.

Details

Before single sample directional gene set analysis, it is necessary to check whether the gene ID types in the gene sets, data matrix and direction matrix have the same gene ID type. If not, the ssdGSA and ssdGSA_individual would stop, and users should double check to make gene ID types in different parts match one another.

Value

If there are more than 10\ the single sample directional gene set analysis would stop.

`check_gene_name_match_noDir` *Function to Check if Gene IDs in the Input Gene Data Match Well When Direction Matrix Is Missing*

Description

This function is to check if the gene IDs in the gene sets and data matrix match well.

Usage

```
check_gene_name_match_noDir(genes_in_Gene_sets, genes_in_Data)
```

Arguments

`genes_in_Gene_sets`	A list of gene names from the gene sets.
`genes_in_Data`	A list of gene names from the data matrix.

Details

Before single sample directional gene set analysis, it is necessary to check whether the gene ID types in the in the gene sets, data matrix and direction matrix have the same ID type. If not, users should double check to make gene ID types match one another.

Value

If there are more than 10\ the single sample directional gene set analysis would stop.

data_matrix *This is data to be included in package*

Description

This is data to be included in package

Usage

```
data_matrix
```

Format

An example data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.

S1 gene expression for this gene from the 1st sample

S2 gene expression for this gene from the 2nd sample

S3 gene expression for this gene from the 3rd sample

S4 gene expression for this gene from the 4th sample

S5 gene expression for this gene from the 5th sample

S6 gene expression for this gene from the 6th sample

S7 gene expression for this gene from the 7th sample

S8 gene expression for this gene from the 8th sample

S9 gene expression for this gene from the 9th sample

S10 gene expression for this gene from the 10th sample

data_matrix_entrezID *This is data to be included in package*

Description

This is data to be included in package

Usage

```
data_matrix_entrezID
```

Format

An example data matrix of gene expressions with gene entrez ID as row names and columns corresponding to different samples.

S1 gene expression for this gene from the 1st sample

S2 gene expression for this gene from the 2nd sample

S3 gene expression for this gene from the 3rd sample

S4 gene expression for this gene from the 4th sample

S5 gene expression for this gene from the 5th sample

S6 gene expression for this gene from the 6th sample

S7 gene expression for this gene from the 7th sample

S8 gene expression for this gene from the 8th sample

S9 gene expression for this gene from the 9th sample

S10 gene expression for this gene from the 10th sample

`direction_matrix` *This is data to be included in package*

Description

This is data to be included in package

Usage

```
direction_matrix
```

Format

An example direction matrix containing directionality information from summary statistics such as effect size (ES) or p value, with each row for one gene.

gene gene entrez ID

ES effect size (SE) for this gene from summary statistics

pval p value for this gene from summary statistics

gene_sets *This is data to be included in package*

Description

This is data to be included in package

Usage

```
gene_sets
```

Format

An example disease gene sets in the form of a list, with gene set names as list component names, and each component is a vector of gene entrez ID. In this sample gene sets list, there are 10 gene sets in total.

TCELL.KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY gene entrez ID related to this pathway

TLR.KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY gene entrez ID related to this pathway

BCELL.KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY gene entrez ID related to this pathway

NEUROTROPHIN.KEGG_NEUROTROPHIN_SIGNALING_PATHWAY gene entrez ID related to this pathway

ERBB.KEGG_ERBB_SIGNALING_PATHWAY gene entrez ID related to this pathway

CALCIUM.KEGG_CALCIUM_SIGNALING_PATHWAY gene entrez ID related to this pathway

CHEMOKINE.KEGG_CHEMOKINE_SIGNALING_PATHWAY gene entrez ID related to this pathway

GNRH.KEGG_GNRH_SIGNALING_PATHWAY gene entrez ID related to this pathway

VEGF.KEGG_VEGF_SIGNALING_PATHWAY gene entrez ID related to this pathway

median_expression *Function to Calculate Median of Gene Expressions*

Description

This function is to calculate the median of gene expressions for genes in the given gene sets.

Usage

```
median_expression(Data, pathway.db)
```

Arguments

Data	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.
pathway.db	A list of gene sets.

Details

Within the ssdGSA function, when GSA_method = "median.exprs", this function is used to calculate the average of gene expressions for genes in the given gene sets.

Value

Matrix of average gene expression in each gene set with rows corresponding to gene sets and columns corresponding to samples will be returned.

---

ssdGSA *Single Sample Directional Gene Set Analysis (ssdGSA)*

Description

This function is to calculate directional (disease weighted) gene set scores by incorporating each gene's correlation to a disease or pathway in the gene set.

Usage

```
ssdGSA(
  Data,
  Gene_sets,
  Direction_matrix = NULL,
  GSA_weight = "equal_weighted",
  GSA_weighted_by = "sum.ES",
  GSA_method = "gsva",
  min.sz = 1,
  max.sz = 2000,
  mx.diff = TRUE
)
```

Arguments

Data	Data matrix of gene expressions with gene ID as row names and columns corresponding to different samples.
Gene_sets	A list of gene sets with gene set names as component names, and each component is a vector of gene ID.

`Direction_matrix`
: Matrix containing directionality information for each gene, such as effect size, t statistics, p value of summary statistics. Each row of the direction matrix is for one gene, and there should be at least two columns (with the 1st column containing gene entrez ID, and 2nd column containing directionality information). Note that the default is "Direction_matrix = NULL", meaning that no direction matrix is inputted, then the classic single sample gene set scores without direction information would be calculated and returned.

`GSA_weight`
: Method to calculate weight in GSA. By default this is set to "group_weighted". Other option is "equal_weighted".

`GSA_weighted_by`
: When "group_weighted" is chosen to calculate GSA_weight, further specifications are needed to specify how group weights are calculated. By default, this is set to "avg.ES" (average of group ES). Other options are "sum.ES" (sum of group ES) and "median.ES" (median of group ES).

`GSA_method`
: Method to employ in the estimation of gene set enrichment scores per sample. By default this is set to "gsva" (Hanzelmann et al, 2013). Other options are "ssgsea" (Barbie et al, 2009), "zscore" (Lee et al, 2008), "avg.exprs" (average value of gene expressions in the gene set), and "median.exprs" (median of gene expressions in the gene set).

`min.sz`
: GSVA parameter to define the minimum size of the resulting gene sets. By default this is set to 1.

`max.sz`
: GSVA parameter to define the maximum size of the resulting gene sets. By default this is set to 2000.

`mx.diff`
: GSVA parameter to offer two approaches to calculate the enrichment statistic from the KS random walk statistic. mx.diff = FALSE: enrichment statistic is calculated as the maximum distance of the random walk from 0. mx.diff=TRUE (default): enrichment statistic is calculated as the magnitude difference between the largest positive and negative random walk deviations.

Details

Single sample directional gene set analysis inherits the standard gene set variation analysis(GSVA) method, but also provides the option to use summary statistics from any analysis (disease vs healthy, lesional side vs nonlesional side, etc..) input to define the direction of gene sets used for directional gene set score calculation for a given disease or directional function. This function is specific for using group weighted scores.

Value

Matrix of directional gene set scores with rows corresponding to gene sets and columns corresponding to different samples will be return.

References

Xingpeng Li, Qi Qian. ssdGSA - Single sample directional gene set analysis tool.

Barbie, D.A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature, 462(5):108-112, 2009.

Hanzelmann, S., Castelo, R. and Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 14:7, 2013.

Lee, E. et al. Inferring pathway activity toward precise disease classification. PLoS Comp Biol, 4(11):e1000217, 2008.

Tomfohr, J. et al. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics, 6:225, 2005.

See Also

ssdGSA_individual

Examples

```
ssdGSA(Data = data_matrix_entrezID,
       Gene_sets = gene_sets[c(1,2,4)],
       Direction_matrix = direction_matrix,
       GSA_weight = "group_weighted",
       GSA_weighted_by = "sum.ES",
       GSA_method = "gsva",
       min.sz = 1,
       max.sz = 2000,
       mx.diff = TRUE
       )
```

`ssdGSA_individual` *Single Sample Directional Gene Set Analysis Using Individual Weighted Scores*

Description

This function is to calculate single sample directional (disease weighted) gene set scores for a given disease using individual weighted scores.

Usage

```
ssdGSA_individual(Data, Gene_sets, Direction_matrix)
```

Arguments

`Data`	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.
`Gene_sets`	A list of gene sets with gene set names as component names, and each component is a vector of gene entrez ID.

Direction_matrix
: Matrix containing directionality information for each gene, such as effect size, t statistics, p value of summary statistics. Each row of the direction matrix is for one gene, and there should be at least two columns (with the 1st column containing gene entrez ID, and 2nd column containing directionality information).

Details

Single sample directional gene set analysis using individual weighted scores inherits the standard gene set variation analysis(GSVA) method, but also provides the option to use summary statistics from any analysis (disease vs healthy, lesional side vs nonlesional side, etc..) input to define the direction of gene sets used for directional gene set score calculation for a given disease. This function is specific for using individual weighted scores.

Value

Matrix of directional gene set scores with rows corresponding to gene sets and columns corresponding to different samples will be return.

See Also

ssdGSA

Examples

```
ssdGSA_individual(Data = data_matrix_entrezID,
                  Gene_sets = gene_sets[c(1,2,4)],
                  Direction_matrix = direction_matrix
                  )
```

---

ssGSA *Function to Calculate Single Sample Gene Set Scores without Direction Matrix*

Description

This function is to calculate traditional single sample gene set scores without considering the direction of each gene.

Usage

```
ssGSA(
  Data,
  Gene_sets,
  GSA_weight = "equal_weighted",
  GSA_weighted_by = "sum.ES",
  GSA_method = "gsva",
```

```
  min.sz = 1,
  max.sz = 2000,
  mx.diff = TRUE
)
```

Arguments

Data	Data matrix of gene expressions with gene ID as row names and columns corresponding to different samples.
Gene_sets	A list of gene sets with gene set names as component names, and each component is a vector of gene ID.
GSA_weight	Method to calculate weight in GSA. By default this is set to "group_weighted". Other option is "equal_weighted".
GSA_weighted_by	When "group_weighted" is chosen to calculate GSA_weight, further specifications are need to specify how group weights are calculated. By default this is set to "avg.ES" (average of group ES). Other options are "sum.ES" (sum of group ES) and "median.ES" (median of group ES).
GSA_method	Method to employ in the estimation of gene-set enrichment scores per sample. By default this is set to "gsva" (Hanzelmann et al, 2013). Other options are "ssgsea" (Barbie et al, 2009), "zscore" (Lee et al, 2008), "avg.exprs" (average value of gene expressions in the gene set), and "median.exprs" (median of gene expressions in the gene set).
min.sz	GSVA parameter to define the minimum size of the resulting gene sets. By default this is set to 1.
max.sz	GSVA parameter to define the maximum size of the resulting gene sets. By default this is set to 2000.
mx.diff	GSVA parameter to offer two approaches to calculate the enrichment statistic from the KS random walk statistic. mx.diff = FALSE: enrichment statistic is calculated as the maximum distance of the random walk from 0. mx.diff=TRUE (default): enrichment statistic is calculated as the magnitude difference between the largest positive and negative random walk deviations.

Details

Single sample directional gene set analysis inherits the standard gene set variation analysis(GSVA) method, but also provides the option to use summary statistics from any analysis (disease vs healthy, LS vs NL, etc..) input to define the direction of gene sets used for directional gene set score calculation for a given disease or directional function. However, when the directionality information is missing for genes, gene set scores from traditional single sample gene set analysis will be returned.

Value

Matrix of gene set scores (without considering directionality information of each gene) with rows corresponding to gene sets and columns corresponding to different samples will be return.

References

Xingpeng Li, Qi Qian. ssdGSA - Single sample direction gene set analysis tool.

Barbie, D.A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature, 462(5):108-112, 2009.

Hanzelmann, S., Castelo, R. and Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 14:7, 2013.

Lee, E. et al. Inferring pathway activity toward precise disease classification. PLoS Comp Biol, 4(11):e1000217, 2008.

Tomfohr, J. et al. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics, 6:225, 2005.

See Also

ssdGSA, ssdGSA_individual

`transform_ensembl_2_entrez`

Function to Transform ENSEMBL ID to ENTREZ ID

Description

This function is to uniform gene ID types in data matrices, i.e., from ENSEMBL ID to ENTREZ ID.

Usage

```
transform_ensembl_2_entrez(Data)
```

Arguments

`Data`	Data matrix of gene expressions with gene ensembl ID as row names and columns corresponding to different samples.

Details

Since gene IDs in data matrices from different sources may be in different formats (ensembl ID or entrez ID), this function is to transform the gene IDs in the data matrix from ensembl ID to entrez ID, to assist the following single sample directional gene set analysis.

Value

Data matrix of gene expressions with ENSEMBL ID as row names and columns corresponding to samples will be return.

Examples

```
transform_ensembl_2_entrez(Data = data_matrix)
```

Index