# Package ‘GSEMA’

October 14, 2024

**Type** Package

**Title** Gene Set Enrichment Meta-Analysis

**Version** 0.99.3

**Description**
Performing the different steps of gene set enrichment meta-analysis. It provides different functions that allow the application of meta-analysis based on the combination of effect sizes from different pathways in different studies to obtain significant pathways that are common to all of them.

**License** GPL-2

**Depends** R (>= 4.4)

**Suggests** BiocStyle, qpdf, BiocGenerics, RUnit, knitr, rmarkdown

**Imports** Biobase, limma, Rdpack, singscore, GSVA, progress, pbapply,
doParallel, parallel, BiocParallel, metafor, RColorBrewer,
pheatmap, methods, scales, plyr, grDevices, stats, impute

**VignetteBuilder** knitr

**biocViews** StatisticalMethod, GeneSetEnrichment, Pathways

**RdMacros** Rdpack

**Encoding** UTF-8

**LazyData** false

**RoxygenNote** 7.3.2

**NeedsCompilation** no

**Author** Juan Antonio Villatoro-García [aut, cre],
Pablo Jurado-Bascón [aut],
Pedro Carmona-Sáez [aut]

**Maintainer** Juan Antonio Villatoro-García <juanantoniovillatorogarcia@gmail.com>

**Repository** CRAN

**Date/Publication** 2024-10-14 09:20:17 UTC

# Contents



---

`calculateESpath` *Calculation of Effects Sizes and their variance for the different Gene Sets and studies*

---

## Description

This function uses different estimators to calculate the different effects size and their variance for each gene and for each dataset

## Usage

```
calculateESpath(
  objectMApath,
  measure = c("limma", "SMD", "MD"),
  WithinVarCorrect = TRUE,
  missAllow = 0.3
)
```

## Arguments

`objectMApath` A list of list. Each list contains two elements. The first element is the Gene Set matrix (gene sets in rows and samples in columns) and the second element is a vector of zeros and ones that represents the state of the different samples of the Gene Sets matrix. 0 represents one group (controls) and 1 represents the other group (cases).

`measure` A character string that indicates the type of effect size to be calculated. The options are "limma", "SMD" and "MD". The default value is "limma". See details for more information.

`WithinVarCorrect` A logical value that indicates if the within variance correction should be applied. The default value is TRUE. See details for more information.

`missAllow` a number that indicates the maximum proportion of missing values allowed in a sample. If the sample has more proportion of missing values the sample will be eliminated. In the other case the missing values will be imputed using the K-NN algorithm.

### Details

The different estimator methods that can be applied are:

1. "limma"
2. "SMD"
3. "MD"

The **"SMD"** (Standardized mean different) method calculates the effect size using the Hedges'g estimator (Hedges, 1981).

The **"MD"** (raw mean different) calculates the effects size as the difference between the means of the two groups (Borenstein, 2009).

The **"limma"** method used the limma package to calculate the effect size and the variance of the effect size. The effect size is calculated from the moderated Student's t computed by limma. From it, the estimator of Hedges'g and its corresponding variance are obtained based on (Rosenthal, R., & Rosnow, R. L., 2008)) In this way, some of the false positives obtained by the "SMD" method are reduced.

The **WithinVarCorrect** parameter is a logical value that indicates if the within variance correction should be applied. In the case of applying the correction, the variance of the gene sets in each of the studies is calculated based on the mean of the estimators and not on the estimator of the study itself as described in formula (21) by (Lin L and Aloe AM 2021.)

### Value

A list formed by two elements:

- First element (ES) is a dataframe were columns are each of the studies (datasets) and rows are the genes sets. Each element of the dataframe represents the Effect Size.
- Second element (Var) is a dataframe were columns are each of the studies (datasets) and rows are the genes sets. Each element of the dataframe represents the variance of the Effect size.

### Author(s)

Juan Antonio Villatoro Garcia, <juanantoniovillatorogarcia@gmail.com>

### References

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 221–235). New York: Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. doi:10.2307/1164588

Lin L, Aloe AM (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med. 2021 Jan 30;40(2):403-426. doi:10.1002/sim.8781

Rosenthal, R., & Rosnow, R. L. (2008). Essentials of behavioral research: Methods and data analysis. McGraw-Hill.

## See Also

createObjectMApath

## Examples

```
data("simulatedData")
calculateESpath(objectMApath = objectMApathSim, measure = "limma")
```

---

createObjectMApath *Creation of the object to use in gene set enrichment meta-analysis*

---

## Description

It allows the creation of an object to perform gene set enrichment meta-analysis.

## Usage

```
createObjectMApath(
  listEX,
  listPheno = NULL,
  namePheno = c(rep(1, length(listEX))),
  expGroups = c(rep(1, length(listEX))),
  refGroups = c(rep(2, length(listEX))),
  geneSets,
  pathMethod = c("GSVA", "Zscore", "ssGSEA", "Singscore"),
  minSize = 7,
  kcdf = "Gaussian",
  normalize = TRUE,
  n.cores = 1,
  internal.n.cores = 1
)
```

## Arguments

| | |
|---|---|
| listEX | A list of dataframes or matrix (genes in rows and sample in columns). A list of ExpressionSets can be used too |
| listPheno | A list of phenodatas (dataframes or matrix). If the object listEX is a list of ExpressionSets this element can be null. |
| namePheno | A list or vector of the different colunm names or positions from the phenodatas where the experimental and reference groups are identified. Each element of namePheno correspont to its equivalent element in the listPheno (default a vector of 1, all the first columns of each elements of listPheno are selected). |
| expGroups | A list of vectors or a vector containing the names or the positions with which we identify the elements of the experiment groups (cases) of the namePheno element (default a vector of 1, all the first groups are selected) |

- `refGroups` A list of vectors or a vector containing the names or the positions with which we identify the elements of the reference groups (control) of the namePheno elements (default a vector of 1, all the first groups are selected)
- `geneSets` List of gene sets to check. Object similar to the one used in the fgsea package
- `pathMethod` The single sample enrichment method used to obtain the enrichment score of each sample and gene set. See details for more information
- `minSize` Minimum size of the resulting gene sets after gene identifier mapping. By default, the minimum size is 7.
- `kcdf` Only neccesary for the GSVA method. Character vector of length 1 denoting the kernel to use during the non-parametric estimation of the cumulative distribution function of expression levels across samples. By default, kcdf="Gaussian" which is suitable when input expression values are continuous, such as microarray fluorescent units in logarithmic scale, RNA-seq log-CPMs, log-RPKMs or log-TPMs. When input expression values are integer counts, such as those derived from RNA-seq experiments, then this argument should be set to kcdf="Poisson".
- `normalize` boolean specifying if the gen set matrices should be normalized. Default value "TRUE".
- `n.cores` Number of cores to use in the parallelization of the datsets. By default, n.cores=1.
- `internal.n.cores` Number of cores to use in the parallelization of the single sample enrichment methods. By default internal.n.cores= 1.

## Details

The single sample scoring methods that can be used to obtain the enrichment score of each sample and gene set are:

1. "GSVA": Gene Set Variation method (Hänzelmann S, 2013)
2. "Zscore": Z-score method (Lee E, 2008)
3. "ssGSEA": Single Sample Gene Set Enrichment Analysis method (Barbie DA, 2009)
4. "Singscore": Single sample scoring of molecular phenotypes (Foroutan M, 2018)

In parallelization, several aspects must be considered. n.cores refers to the parallelization of studies or datasets. Therefore, if we have 3 studies, the maximum number for n.cores will be 3. internal.n.cores refers to the parallelization of single sample enrichment methods. This is especially recommended for the ssGSEA method. For Singscore and GSVA, it may also be advisable. The process is parallelized based on the samples in each study. Therefore, the larger the number of samples, the slower the process will be. The number of cores that the computer will use is the multiplication of both parameters n.cores * internal.n.cores = total cores.

## Value

The object needed to perform gene set enrichment meta-analysis. Each list contains two elements: The first element is the gene set matrix (gene sets in rows and samples in columns) The second element is a vector of zeros and ones that represents the state of the diffenrent samples of the gene sets matrix. 0 represents reference group (controls) and 1 represents experimental group (cases).

## References

Hänzelmann S, Castelo R, Guinney J. (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14: 7. doi:10.1186/1471-2105-14-7

Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. (2008) Inferring Pathway Activity toward Precise Disease Classification. PLOS Computational Biology. 2008;4: e1000217. doi:10.1371/journal.pcbi.1000217

Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462: 108–112. doi:10.1038/nature08460

Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. (2018) Single sample scoring of molecular phenotypes. BMC Bioinformatics. 2018;19: 404. doi:10.1186/s12859-018-2435-4

Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. (2021) Fast gene set enrichment analysis. bioRxiv; 2021. p. 060012. doi:10.1101/060012

## Examples

```
data("simulatedData")
listMatrices <- list(study1Ex, study2Ex)
listPhenodata <- list(study1Pheno, study2Pheno)
phenoGroups <- c("Condition","Condition")
phenoCases <- list("Case", "Case")
phenoControls <- list("Healthy", "Healthy")
objectMApathSim <- createObjectMApath(
   listEX = listMatrices,
   listPheno = listPhenodata, namePheno = phenoGroups,
   expGroups = phenoCases, refGroups = phenoControls,
   geneSets = GeneSets,
   pathMethod = "Zscore")
```

---

# filteringPaths *Fuction for filtering gene sets with low expression*

---

## Description

This function eliminates gene sets with low expression in both groups in a study

## Usage

```
filteringPaths(objectMApath, threshold = 0.65, n_cores = 1)
```

### Arguments

objectMApath — A list of list. Each list contains two elements. The first element is the Gene Set matrix (gene sets in rows and samples in columns) and the second element is a vector of zeros and ones that represents the state of the different samples of the Gene Sets matrix. 0 represents one group (controls) and 1 represents the other group (cases).

threshold — A number that indicates the threshold to eliminate a gene set. For a eliminate a gene set is necessary that the median for both groups are less than the threshold If threshold = "sd" the threshold will be the standard deviation of the gene set. The default value is 0.65.

n_cores — A number that indicates the number of cores to use in the parallelization. The default value is 1.

### Value

The same objectMApath list but with the gene sets that do not meet the threshold eliminated.

### Author(s)

Juan Antonio Villatoro Garcia, `<juanantoniovillatorogarcia@gmail.com>`

### See Also

`createObjectMApath`

### Examples

```
data("simulatedData")
newObject <- filteringPaths(objectMApathSim, threshold = "sd")
```

---

## heatmapPaths — *Visualization of the gene set enrichment meta-analysis results*

---

### Description

It allows to see how the different significant gene sets are expressed in the different samples

### Usage

```
heatmapPaths(
  objectMApath,
  resMA,
  scaling = c("zscor", "rscale", "swr", "none"),
  regulation = c("all", "up", "down"),
  breaks = c(-2, 2),
  fdrSig = 0.05,
```

```
  comES_Sig = 0.5,
  numSig = "all",
  color = colorRampPalette(rev(brewer.pal(n = 7, name = "RdYlBu")))(100),
  na_col = "darkgrey",
  legend = TRUE,
  cluster_cols = FALSE,
  cluster_rows = FALSE,
  show_rownames = TRUE,
  show_colnames = FALSE,
  fontsize = 10,
  fontsize_row = fontsize,
  fontsize_col = fontsize
)
```

## Arguments

| | |
|---|---|
| `objectMApath` | A list of list. Each list contains two elements. The first element is the Gene Set matrix (gene sets in rows and samples in columns) and the second element is a vector of zeros and ones that represents the state of the different samples of the Gene Sets matrix. 0 represents one group (controls) and 1 represents the other group (cases). |
| `resMA` | Output generated by the function that performs meta-analysis (metaAnalysisES-path). |
| `scaling` | Character variable to choose between different scaling approaches. See "Details" for more information. |
| `regulation` | Character variable that indicates whether we want the heatmap to show all significant paths ("all"), only the up-regulated paths ("up") or only the down-regulated paths("down") |
| `breaks` | Numeric vector of length 2 that contains the extreme values (minimum and maximum) of the range of values in which the heatmap color scale will be distributed. Default a vector By default a vector of -2 and 2 as extreme values. |
| `fdrSig` | Adjusted p-value from which a gene set is considered significant. Default 0.05 |
| `comES_Sig` | In absolute value. Combine effect size threshold from which gene sets are considered. Default 0.5 |
| `numSig` | The number of most significant paths to be represented. If numSig = "all", all significant paths that meet the selected parameters will be represented. |
| `color` | vector of colors used in heatmap. |
| `na_col` | color of the NA cell in the heatmap. |
| `legend` | logical to determine if legend should be drawn or not. |
| `cluster_cols` | boolean values determining if columns should be clustered. |
| `cluster_rows` | boolean values determining if rows should be clustered. |
| `show_rownames` | boolean specifying if row names are be shown. |
| `show_colnames` | boolean specifying if column names are be shown. |
| `fontsize` | base fontsize for the plot |
| `fontsize_row` | fontsize for rownames (Default: fontsize) |
| `fontsize_col` | fontsize for colnames (Default: fontsize) |

## Details

Scaling approaches that can be used are:

- "rscale": it applies rescale function of *scales* package. Values will be between -1 and 1)
- "zscor": It calculates a z-score value for each gene, that is, the mean gene expression from each gene is subtracted from each gene expression value and then it is divided by the standard deviation
- "swr": it applys scaling relative to reference dataset approach
- "none": any scaling approach it is applied.

## Value

The matrix represented in the heatmap

## Author(s)

Juan Antonio Villatoro Garcia, <juanantoniovillatorogarcia@gmail.com>

## References

Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

Lazar, C, Meganck, S, Taminau, J, and et al. 2013. “Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey,” 469–90.

Raivo Kolde 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

## See Also

createObjectMApath, metaAnalysisESpath

## Examples

```
data("simulatedData")
resultsMA <- metaAnalysisESpath(objectMApathSim, typeMethod="REM")
heatmapPaths(objectMA=objectMApathSim, resultsMA,
    scaling = "zscor", regulation = "all",breaks=c(-2,2),
    fdrSig = 0.05, comES_Sig = 1.5, numSig=20)
```

metaAnalysisESpath *Performing Gene Set Enrichment Meta-analysis*

## Description

It performs Gene Sets Enrichment meta-analysis by applying Effects size combination methods

## Usage

```
metaAnalysisESpath(
  objectMApath = NULL,
  effectS = NULL,
  measure = c("limma", "SMD", "MD"),
  WithinVarCorrect = TRUE,
  typeMethod = c("REM", "FEM"),
  missAllow = 0.3,
  numData = length(objectMApath)
)
```

## Arguments

| | |
|---|---|
| objectMApath | A list of list. Each list contains two elements. The first element is the Gene Set matrix (gene sets in rows and samples in columns) and the second element is a vector of zeros and ones that represents the state of the different samples of the Gene Sets matrix. 0 represents one group (controls) and 1 represents the other group (cases). |
| effectS | A list of two elements. The first element is a dataframe with gene sets in rows and studies in columns. Each component of the dataframe is the effect of a gene set in a study. The second element of the list is also a dataframe with the same structure, but in this case each component of the dataframe represent the variance of the effect of a gene set in a study. This argument should be only used in the case that objectMApath argument is null. |
| measure | A character string that indicates the type of effect size to be calculated. The options are "limma", "SMD" and "MD". The default value is "limma". See details for more information. |
| WithinVarCorrect | A logical value that indicates if the within variance correction should be applied. The default value is TRUE. See details for more information. |
| typeMethod | A character that indicates the method to be performed. See "Details"for more information |
| missAllow | a number that indicates the maximum proportion of missing values allowed in a sample. If the sample has more proportion of missing values the sample will be eliminated. In the other case the missing values will be imputed using the K-NN algorithm. |

`numData` The minimum number of datasets in which a gene must be contained to be included in the emta-analysis. By default, the gene must be contained in all the datasets. If the number entered exceeds the number of studies, the total number of studies will be considered."

### Details

There are different ways to calculate the effect size of a gene set:

1. "MD": Raw Mean Difference (Borenstein, 2009)
2. "SMD": Standardized Mean Difference (Hedges, 1981)
3. "limma": Standardized Mean Difference calculated from the t-statistics and degrees of freedom obtained by the limma package by applying the transformation of Rosenthal and Rosnow, 2008). Its calculation is similar to the one proposed by (Marot et al., 2009) but considering the transformation of (Rosenthal and Rosnow, 2008).

The correction of the variance of the effect size is based on Lin L, Aloe AM (2021) in which the variance is calculated from the different estimators.

The meta-analysis methods that can be applied are:

1. "FEM": Fixed Effects model
2. "REM": Random Effects model (Default).

### Value

A dataframe with the meta-analysis results. For more information see the package vignette.

### Author(s)

Juan Antonio Villatoro Garcia, <juanantoniovillatorogarcia@gmail.com>

### References

Toro-Domínguez D., Villatoro-García J.A., Martorell-Marugán J., Román-Montoya Y., Alarcón-Riquelme M.E., Carmona-Sáez P. (2020). A survey of gene expression meta-analysis: methods and applications, Briefings in Bioinformatics, bbaa019, doi:10.1093/bib/bbaa019

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 221–235). New York: Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. doi:10.2307/1164588

Lin L, Aloe AM (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med. 2021 Jan 30;40(2):403-426. doi:10.1002/sim.8781

Marot, G., Foulley, J. L., Mayer, C. D., & Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics. 2692-2699. doi:10.1093/bioinformatics/btp444

Rosenthal, R., & Rosnow, R. L. (2008). Essentials of behavioral research: Methods and data analysis. McGraw-Hill.

### See Also

calculateESpath

### Examples

```
data("simulatedData")
results <- metaAnalysisESpath(objectMApath = objectMApathSim,
    measure = "limma", typeMethod = "REM")
```

---

## simulatedData — *GSEMA synthetic data*

---

### Description

- study1Ex, study2Ex: two expression matrices.
- study1Pheno, study2Pheno: two phenodata objects.
- GeneSets: a list of gene sets with each element are the genes that belong to a pathway.
- objectMApathSim: the meta-analysis object created from the different expression matrices and phenodatas.

### Usage

```
data(simulatedData)
```

### Format

arrays (`study1Ex`, `study2Ex`), data.frames (`study1Pheno`, `study2Pheno`), list vectors (`GeneSets`) and list of nested lists (`objectMApathSim`).

### Source

study1Ex, study2Ex, study1Pheno and study2Pheno are synthetic. GeneSets was created from the information of MsigDB database by adding a simulated pathway with simulated genes

objectMApathSim was created after using createObjectMA GSEMA function to the different studies objects and with the information of GeneSets object

maObjectDif was created after using createObjectMA DExMA function to the listExpressionSets object. maObject was obtained after using allSameID function to maObjectDif function.

### References

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1: 417–425

# Index