Package ‘ebGSEA’

June 3, 2020

Type Package

Title Empirical Bayes Gene Set Enrichment Analysis

Version 0.1.0

Date 2020-6-3

Author Andrew E. Teschendorff, Tianyu Zhu

Maintainer Andrew E. Teschendorff <andrew@picb.ac.cn>, Tianyu Zhu <zhutianyu@picb.ac.cn>

Description Gene Set Enrichment Analysis is one of the most common tasks in the analysis of omic data, and is critical for biological interpretation. In the context of Epigenome Wide Association Studies, which typically rank individual cytosines according to the level of differential methylation, enrichment analysis of biological pathways is challenging due to differences in CpG/probe density between genes. ebGSEA implements an empirical Bayes Gene Set Enrichment Analysis algorithm, which does not rank CpGs but genes according to the overall level of differential methylation of its CpGs/probes, allowing unbiased and sensitive detection of enriched pathways. ebGSEA is a GSEA tool for EWAS that use Illumina HM450k and EPIC beadarrays.

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Depends R (>= 3.6)

Imports globaltest, kpmt, parallel, org.Hs.eg.db, AnnotationDbi, stats

VignetteBuilder knitr

Suggests roxygen2, BiocStyle, knitr, rmarkdown

NeedsCompilation no

R topics documented:

ebGSEA-package 2
convertIDs 2
doGSEAft 3

doGSEAwt 5
doGT 6
dualmap450kEID 7
dualmap850kEID 8
ebgseaDATA 8
MSigDB-28Feb14-data 9
SampleCpG 10
selEIDfromSelCpG 10

`ebGSEA-package` *Empirical Bayes Gene Set Enrichment Analysis*

Description

ebGSEA (Empirical Bayes Gene Set Enrichment Analysis) implements a GSEA designed for Illumina Infinium Methylation beadchips, based on an empirical Bayes method to rank genes based on their level of differential methylation, subsequently assessing enrichment of biological terms using this ranked list.

Details

ebGSEA leverages the evidence of differential methylation from all CpGs/probes mapping to a given gene, to rank genes according to their overall level of differential methylation. A key property of ebGSEA is that it does not favour genes with high or low CpG/probe representation, thus avoiding bias, whilst also rendering the method sensitive enough to detect true biological enrichment. With genes ranked by this empirical Bayes regression model, GSEA can subsequently be performed using a non parametric Wilcoxon rank sum test or the known population median test (KPMT), thus allowing GSEA to be performed in a threshold independent manner.

Author(s)

Andrew E Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[1].

Examples

```
### see example in tutorial
```

[1] https://doi.org/10.1093/bioinformatics/btz073

convertIDs *Convert between different gene IDs*

Description

Convert gene IDs based on ‘select‘ function in AnnotationDbi package

Usage

```
convertIDs(ids, from, to, db, ifMultiple=c("putNA", "useFirst"))
```

Arguments

`ids`	A vector of the gene IDs to convert from.
`from`	The name of the gene identifier to convert from. All possible keys are returned by using the ‘keys‘ method.
`to`	The name of the gene identifier to convert to. All possible keys are returned by using the ‘keys‘ method.
`db`	The AnnotationDb object with the annoation of gene IDs to convert from and to.
`ifMultiple`	If there are multiple hits for an input gene ID, whether to return 'NA' (ifMultiple = "putNA") or the first hit (ifMultiple = "useFirst").

Value

A vector of converted gene IDs.

Author(s)

Andrew E. Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[2].

Examples

```
#data("sgtm")
#rankEID.v <- rownames(sgt.m)
#sym.v <- convertIDs(rankEID.v, 'ENTREZID', 'SYMBOL', org.Hs.eg.db, ifMultiple="useFirst"
```

[2] `https://doi.org/10.1093/bioinformatics/btz073`

doGSEAft *GSEA with Fisher's Exact Test*

Description

Perform GSEA with Fisher's Exact Test to a group of selected genes

Usage

```
doGSEAft(selEID.v, ptw.ls, allEID.v, ncores = 4, minN = 5, adjPVth = 0.05)
```

Arguments

`selEID.v`	A vector of selected Entrez Gene ID.
`ptw.ls`	Lists of Gene EntrezID in each pathway of interest. You can get the 8567 biological terms from Molecular Signatures Database by 'data("MSigDB-28Feb14-data")'.
`allEID.v`	A vector of the universal set of Entrez Gene ID which you select genes from.
`ncores`	Number of cores used for parallel running. (default = 4)
`minN`	For each pathway, the minium number of genes(i.e. available in the ranked gene list) to conduct GSEA. If less than this value, the p value of this pathway would be set 1. (default = 5)
`adjPVth`	Adjusted p value threshold to infer a pathway to be significantly enriched or not. P value was derived from Wilcoxon rank sum test and adjusted with BH method. (default = 0.05)

Details

GSEA with Fisher's Exact Test is an extended enrichement test for user specified genes that doesn't require for ranks.

Value

`Rank(P)`	A matrix showing enriched pathways ranked by adjusted Fisher's Exact Test p values. "nREP" is the number of genes in the pathway, "nOVL" is the number of selected genes in the pathway, "OR" is the odds ratio of Fisher's Exact Test, "P" is the p value of Fisher's Exact Test, "adjP" is the adjusted p value of Fisher's Exact Test (method='BH'), "Genes" is all the selected genes in the pathway.
`Rank(OR)`	A matrix showing enriched pathways ranked by odds ratio. The columns are samely defined as in Rank(P).

Author(s)

Andrew E. Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[3].

Examples

```
# topGSEAft.lm <- doGSEAft(selEID.v = sigEID.ls$selEID, ptw.ls = listEZ.lv, allEID.v = na

# Details can be found in tutorial
```

---

doGSEAwt *GSEA with Wilcoxon Rank Sum Test and the Known-Population Median Test*

Description

Perform GSEA with wilcoxon rank sum test and known-population test using the ranked gene list from global test.

Usage

```
doGSEAwt(rankEID.m, ptw.ls, ncores = 4, minN = 5, adjPVth = 0.05)
```

Arguments

`rankEID.m` The resulted matrix from doGT function, with genes by row and ranked by statistics from global test. Rownames of the matrix should be gene EntrezID.

`ptw.ls` Lists of Gene EntrezID in each pathway of interest. You can get the 8567 biological terms from Molecular Signatures Database by ‘data("MSigDB-28Feb14-data")‘.

`ncores` Number of cores used for parallel running. (default = 4)

`minN` For each pathway, the minium number of genes(i.e. available in the ranked gene list) to conduct GSEA. If less than this value, the p value of this pathway would be set 1. (default = 5)

`adjPVth` Adjusted p value threshold to infer a pathway to be significantly enriched or not. P value was derived from Wilcoxon rank sum test and adjusted with BH method. (default = 0.05)

Details

GSEA with Wilcoxon Rank Sum Test and the Known-Population Median Test is the second step of ebGSEA algorithm. Once the ranks of genes were derived from *doGT*, enrichment of biological terms can be performed using either a standard one-tailed Wilcoxon rank sum test (WT), or a recently introduced more powerful version called Known-Population Median Test (KPMT). A group of enriched pathways can then be defined over adjusted pvalue from Wilcoxon rank sum test.

[3] `https://doi.org/10.1093/bioinformatics/btz073`

Value

`Rank(P)` A matrix showing enriched pathways ranked by adjusted Wilcox test p values. "nREP" is the number of mapped genes in the pathway, "AUC" is the Area under curve of wilcox test, "P(WT)" is the p-value of wilcox test, "P(KPMT)" is the p-value of the Known-Population Median Test, "adjP" is the adjusted p-value of wilcox test.

`Rank(AUC)` A matrix showing enriched pathways ranked by AUC. The columns are defined samely as Rank(P).

`Genestat` Lists of gene symbols in each enriched pathway. Each object contains the statistic and p-value from global test of each gene.

Author(s)

Andrew E. Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[4].

Examples

```
# data("MSigDB-28Feb14-data")
# data("sgtm")
# topGSEA.lm <- doGSEAwt(rankEID.m = sgt.m, ptw.ls = listEZ.lv, ncores = 10, minN = 5, ac
```

`doGT` *Empirical Bayes Global Test*

Description

Function to assess the overall level of differential methylation using all the probes mapping to a gene.

Usage

```
doGT(pheno.v, data.m, model = c("linear"), array = c("450k", "850k"), ncores = 4
```

[4] https://doi.org/10.1093/bioinformatics/btz073

Arguments

`pheno.v` A vector of phenotype information, must be matched to columns of the input beta matrix.

`data.m` A matrix of beta values with probes by row and samples by column. Missing values shoud be excluded.

`model` The regression model for global test. Default is "linear".

`array` Array type for the input data. "450k" for Illumina HumanMethylation450 data and "850k" for Illumina MethylationEPIC data.

`ncores` Number of cores used for parallel running. (default = 4)

Details

Global test is the first step of ebGSEA algorithm. ebGSEA ranks genes according to their overall level of differential methylation by adapting the global test from *Geoman et al(2006)*, which can be interpreted as an empirical Bayes generalized regression model. The global test evaluates whether DNA methylation patterns of CpGs mapping to a given gene *g* differ significantly between two phenotypes.

Value

A matrix with genes in row ranked by statistic from global test.

Author(s)

Andrew E. Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[5].

Examples

```
# sgt.m <- doGT(pheno.v, data.m, array = c("450k"), ncores = 10)
# topGSEA.lm <- doGSEAwt(rankEID.m = sgt.m, ptw.ls = listEZ.lv, ncores = 10, minN = 5, ac
```

[5] `https://doi.org/10.1093/bioinformatics/btz073`

`dualmap450kEID` *Illumina HM450k probes annotation to Entrez Gene ID*

Description

This annotation file is derived from Illumina HM450k annotation

Usage

```
data("dualmap450kEID")
```

Format

A list with 485577 items and a list with 18649 items

Details

- map450ktoEID.lv : A list mapping 450k probes to Entrez Gene ID
- mapEIDto450k.lv : A list mapping Entrez Gene ID to 450k probes

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[6].

`dualmap850kEID` *Illumina EPIC probes annotation to Entrez Gene ID*

Description

This annotation file is derived from Illumina annotation file

Usage

```
data("dualmap850kEID")
```

Format

A list with 867531 items and a list with 24357 items

Details

- map850ktoEID.lv : A list mapping 850k probes to Entrez Gene ID
- mapEIDto850k.lv : A list mapping Entrez Gene ID to 850k probes

[6] https://doi.org/10.1093/bioinformatics/btz073

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[7].

`ebgseaDATA` *Sample DNAm data from a buccal swab study*

Description

A DNAm data matrix over 7933 CpGs and 325 samples, and a corresponding vector with 325 entries.

Usage

```
data("ebgseaDATA")
```

Format

Two objects: a matrix containing the DNAm data and a matched phenotype vector containing the smoking-pack-year information

Details

- dataSMK.m : The DNAm data matrix defined over 7933 CpGs and 325 samples.
- phenoSMK.v : The vector containing the smoking information.

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[8].

[7] `https://doi.org/10.1093/bioinformatics/btz073`

[8] `https://doi.org/10.1093/bioinformatics/btz073`

MSigDB-28Feb14-data — *Biological terms from Molecular Signatures Database 28Feb14*

Description

The 8567 biological terms from Molecular Signatures Database[9] 28Feb14

Usage

```
data("MSigDB-28Feb14-data")
```

Format

Two lists with 8567 items and a vector with 8567 items

Details

- listEZ.lv : Gene sets in NCBI (Entrez) Gene IDs for each biological term
- listG.lv : Gene stes in gene symbols for each biological term
- listclassALL.lv : The type of biological term defined by MSiDB

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[10].

Subramanian A, Tamayo P, Mootha VK, et al. *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A (2005) 102(43):15545-15550. doi: 10.1073/pnas.0506580102[11].

SampleCpG — *Sample CpGs for fisher's exact test in Tutorial*

Description

The differentialy methylated cytosines associated with smoking pack-years identified from buccal swab dataset

Usage

```
data("SampleCpG")
```

[9] https://www.gsea-msigdb.org/gsea/msigdb/
[10] https://doi.org/10.1093/bioinformatics/btz073
[11] https://doi.org/10.1073/pnas.0506580102

Format

A vector with length 40626 and a vector with length 484272

Details

- sampleCpG.v : The pre-selected differentially methylated cytosines
- allCpG.v : All the CpGs from buccal swab dataset

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[12].

`selEIDfromSelCpG` *Select significant genes from a selected set of CpGs*

Description

Select significant genes from a selected set of CpGs with binomial test.

Usage

```
selEIDfromSelCpG(selCpG.v, allCpG.v, pvth = 0.3/length(selCpG.v), array = c("450
```

Arguments

`selCpG.v` A vector of user selected CpGs.

`allCpG.v` A vector of all CpGs the user select the CpGs from.

`pvth` P-value threshold to infer the number of selected CpGs mapped to a gene is significant or not in a binomial test. (default = 0.3/length(selCpG.v))

`array` Array type for the input CpGs. "450k" for Illumina HumanMethylation450 data and "850k" for Illumina MethylationEPIC data.

Details

This function maps user specified CpGs to genes and return genes whose number of mapped CpGs is significant in binomial test.

Value

`selEID` A vector of significant genes in Entrez Gene ID.

`selPV` P-values of significant genes.

`allPV` P-values of all mapped genes.

[12] `https://doi.org/10.1093/bioinformatics/btz073`

Author(s)

Andrew E. Teschendorff, Tianyu Zhu

References

Dong D, Tian Y, Zheng SC, Teschendorff AE. *ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies.* BMC Bioinformatics (2019) 35(18):3514-3516. doi: 10.1093/bioinformatics/btz073[13].

Examples

```
# data("SampleCpG")
# sigEID.ls <- selEIDfromSelCpG(selCpG.v = sampleCpG.v, allCpG.v = allCpG.v, array = "450
```

[13] https://doi.org/10.1093/bioinformatics/btz073