# Package ‘mGSZ’

February 20, 2015

**Type**  Package  

**Title**  Gene set analysis based on GSZ-scoring function and asymptotic p-value  

**Version**  1.0  

**Date**  2014-02-19  

**Author**  Pashupati Mishra, Petri Toronen  

**Maintainer**  Pashupati Mishra &lt;pashupati.mishra@helsinki.fi&gt;  

**Depends**  R(>= 3.0.0), Biobase,GSA,limma,MASS,ismev  

**Description**  Performs gene set analysis based on GSZ-scoring function and asymptotic p-value. It is different from GSZ in that it implements asymptotic p-values instead of empirical p-values. Asymptotic p-values are calculated by fitting suitable distribution model to the null distribution. Unlike empirical p-values, resolution of asymptotic p-values are independent of the number of permutations and hence requires considerably fewer permutations. In addition, this package allows gene set analysis with seven other popular gene set analysis methods.

**License**  GPL (>= 2)  

**NeedsCompilation**  no  

**Repository**  CRAN  

**Date/Publication**  2014-02-19 12:24:13  

## R topics documented:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>calc_z_var</td>
<td>2</td>
</tr>
<tr>
<td>count.prob.sum</td>
<td>3</td>
</tr>
<tr>
<td>count_hyge_var_mean</td>
<td>3</td>
</tr>
<tr>
<td>diffFCscore</td>
<td>3</td>
</tr>
<tr>
<td>diffScore</td>
<td>3</td>
</tr>
<tr>
<td>emp</td>
<td>4</td>
</tr>
<tr>
<td>emp.wrs</td>
<td>4</td>
</tr>
<tr>
<td>FC</td>
<td>4</td>
</tr>
<tr>
<td>flipListStruct</td>
<td>4</td>
</tr>
<tr>
<td>geneSetsList</td>
<td>5</td>
</tr>
<tr>
<td>KS.p.values</td>
<td>5</td>
</tr>
</tbody>
</table>
calc_z_var

Internal mGSZ function

Description

Internal mGSZ function not to be called by the users

Author(s)

Pashupati Mishra, Petri Toronen
<table>
<thead>
<tr>
<th>Function</th>
<th>Internal mGSZ function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>count.prob.sum</strong></td>
<td>Internal mGSZ function</td>
</tr>
<tr>
<td><strong>count_hyge_var_mean</strong></td>
<td>Internal mGSZ function</td>
</tr>
<tr>
<td><strong>diffFCscore</strong></td>
<td>Internal mGSZ function</td>
</tr>
<tr>
<td><strong>diffScore</strong></td>
<td>Internal mGSZ function</td>
</tr>
</tbody>
</table>

**Description**

Internal mGSZ functions not to be called by the users

**Author(s)**

Pashupati Mishra, Petri Toronen
<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>Internal mGSZ function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>emp.wrs</td>
<td>Internal WRS function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>FC</td>
<td>Internal mGSZ function</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>flipListStruct</td>
<td>Internal mGSZ function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
</tbody>
</table>
**geneSetsList**

Convert gene set data in gmt file to R list

**Description**

Converts gene set data in gmt file to R list readable by mGSZ program

**Usage**

```r
geneSetsList(data)
```

**Arguments**

- `data` Gene set data in gmt file format

**Value**

Gene set data in list format with gene set name as list names

**Author(s)**

Pashupati Mishra, Petri Toronen

**Examples**

```r
### gene.sets <- geneSetsList(filename.gmt) ###
```

---

**KS.p.values**

*Internal KS function*

**Description**

Internal KS function not to be called by the users

**Author(s)**

Pashupati Mishra, Petri Toronen
<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS.score</td>
<td>Internal KS function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>listToClMatrix</td>
<td>Internal mGSZ function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>logEvcdf</td>
<td>Internal mGSZ function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
<tr>
<td>logNORMcdf</td>
<td>Internal mGSZ function not to be called by the users</td>
<td>Pashupati Mishra, Petri Toronen</td>
</tr>
</tbody>
</table>
mAllez.p.values  

Internal mGSZ function

Description

Internal mGSZ function not to be called by the users

Author(s)

Pashupati Mishra, Petri Toronen

mGSA.p.values  

Internal mGSZ function

Description

Internal mGSZ functions

Author(s)

Pashupati Mishra, Petri Toronen

mGSZ  

Gene set analysis based on Gene Set Z-scoring function and asymptotic p-value

Description

Gene set analysis based on Gene Set Z-scoring function and asymptotic p-value

Usage

mGSZ(x,y,l,f=FALSE,s="T",log=TRUE,g=FALSE,min.sz=5,o=FALSE,pv=0,w1=0.2,w2=0.5,vc=10,p=200)
Arguments

x  Gene expression data matrix (rows as genes and columns as samples)
y  Gene set data (dataframe/table/matrix/list)
l  Vector of response values (example:1,2)
f  TRUE if gene set data is list with genes as list names
s  Gene level statistics (example: T-score/FC/P-value)
log TRUE for log fold change as gene level statistics
g  TRUE for analysis with both gene and sample permutation data as the null distributions
min.sz Minimum size of gene sets (number of genes in a gene set) to be included in the analysis
o  TRUE for gene set analysis with other methods (see the manuscript for details)
pv  Estimate of the variance associated with each observation
w1  Weight 1, parameter used to calculate the prior variance obtained with class size var.constant. This penalizes especially small classes and small subsets. Default is 0.2. Values around 0.1 - 0.5 are expected to be reasonable.
w2  Weight 2, parameter used to calculate the prior variance obtained with the same class size as that of the analyzed class. This penalizes small subsets from the gene list. Default is 0.5. Values around 0.3 and 0.5 are expected to be reasonable.
vc  Size of the reference class used with wgt1. Default is 10
p  Number of permutations for p-value calculation

details

A function for Gene set analysis based on Gene Set Z-scoring function and asymptotic p-value. It differs from GSZ (Toronen et al 2009) in that it implements asymptotic p-values instead of empirical p-values. Asymptotic p-values are based on fitting suitable distribution model to the permutation data. Unlike empirical p-values, the resolution of asymptotic p-values are independent of the number of permutations and hence requires considerably fewer permutations. In addition to GSZ, this function allows the users to carry out analysis with seven other scoring functions (visit http://ekhidna.biocenter.helsinki.fi/downloads/pashupati/mGSZ.html for a more detailed description) and compare the results.

Value

mGSZ  Dataframe with gene sets (in decreasing order based on the significance) reported by mGSZ method and their sizes, scores, p-values and gene set expression summary
mGSA  Dataframe with gene sets (in decreasing order based on the significance) reported by mGSA method and their sizes, scores, p-values and gene set expression summary
mAllez  Dataframe with gene sets (in decreasing order based on the significance) reported by mAllez method and their sizes, scores, p-values and gene set expression summary
WRS
Dataframe with gene sets (in decreasing order based on the significance) reported by WRS method and their sizes, scores, p-values and gene set expression summary

SUM
Dataframe with gene sets (in decreasing order based on the significance) reported by SUM method and their sizes, scores, p-values and gene set expression summary

SS
Dataframe with gene sets (in decreasing order based on the significance) reported by SS method and their sizes, scores, p-values and gene set expression summary

KS
Dataframe with gene sets (in decreasing order based on the significance) reported by KS method and their sizes, scores, p-values and gene set expression summary

wKS
Dataframe with gene sets (in decreasing order based on the significance) reported by wKS method and their sizes, scores, p-values and gene set expression summary

sample.labels
Vector of response values used

perm.number
Number of permutations used for p-value calculation

expr.data
For internal use

gene.sets
For internal use

flip.gene.sets
For internal use

min.cl.sz
For internal use

other.methods
For internal use

pre.var
For internal use

wgt1
For internal use

wgt2
For internal use

var.constant
For internal use

start.val
For internal use

select
For internal use

is.log
For internal use

gene.perm.log
For internal use

Author(s)
Pashupati Mishra, Petri Toronen

References

Examples

gene.names <- paste("g", 1:100, sep = ")

# create random gene expression data matrix

set.seed(100)
x <- matrix(rnorm(100*10), ncol=10)
rownames(x) <- gene.names
b <- matrix(2*rnorm(50), ncol=5)
ind <- sample(1:10, replace=FALSE)
x[ind, 6:10] <- x[ind, 6:10] + b

l <- rep(1:2, c(5, 5))

# create random gene sets

y <- vector("list", 20)
for(i in 1:length(y)){
y[i] <- sample(gene.names, size = 10)
}
names(y) <- paste("set", as.character(1:20), sep="")

mGSZ.obj <- mGSZ(x, y, l, p = 100)
top.mGSZ.sets <- toTable(mGSZ.obj, n = 10)

# scoring function profile data across the ordered gene list for top 2 gene sets
data4plot <- StabPlotData(mGSZ.obj, rank.vector=c(1, 2))

# profile plot for the top gene set
plotProfile(data4plot, 1)

# gene sets in a gmt format can be converted to mGSZ readable format as follows:
# gene.sets <- geneSetsList("gene.sets.gmt")

mGSZ.adj.mean.std Internal mGSZ function

Description

Internal mGSZ function not to be called by the users

Author(s)

Pashupati Mishra, Petri Toronen
mGSZ.p.values

Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

mGSZ.test.score

Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

mGSZ.test.score.stb2

Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

mGSZ.test.score.stb3

Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen
**mGSZ.test.score.stb4**  *Internal mGSZ function*

**Description**
Internal mGSZ function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen

---

**mGSZ.test.score2**  *Internal mGSZ function*

**Description**
Internal mGSZ function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen

---

**mGSZ.test.score4**  *Internal mGSZ function*

**Description**
Internal mGSZ function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen

---

**pick.data.cols**  *Internal mGSZ function*

**Description**
Internal mGSZ function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen
### Description

Plot GSZ scoring function profile

### Usage

```r
plotProfile(data, rank)
```

### Arguments

- **data**  
  GSZ profile data
- **rank**  
  Rank of the gene set for the plot

### Details

Once significant gene sets are reported, it is useful to evaluate a gene set in more detail to see the behavior of the gene set. This can be done by visualizing the scoring function profile across the gene list as shown in the GSEA article (Subramanian et al., 2005). It is even more relevant to compare gene set score profile from positive and permuted data. Positive data corresponds to differential gene expression test scores calculated from gene expression data with correct sample labels and permuted data corresponds to differential gene expression test scores calculated from gene expression data with permuted sample labels. This function outputs the visualization that shows the gene set score profile of the analyzed gene set from positive data and a summary of the gene set score profile of the analyzed gene set from permuted data. The plot uses seven percentiles of the gene set score profile of the analyzed gene set from permuted data as a summary.

### Author(s)

Pashupati Mishra, Petri Toronen

### References


See Also

StabPlotData

Examples

gene.names <- paste("g", 1:100, sep = "")

# create create random gene expression data matrix

set.seed(100)
x <- matrix(rnorm(100*10), ncol=10)
rownames(x) <- gene.names
b <- matrix(2*rnorm(50), ncol=5)
ind <- sample(1:10, replace=FALSE)
x[ind, 6:10] <- x[ind, 6:10] + b

l <- rep(1:2, c(5,5))

# create random gene sets

y <- vector("list", 20)
for(i in 1:length(y)){
y[[i]] <- sample(gene.names, size=10)
} names(y) <- paste("set", as.character(1:20), sep="")

mGSZ.obj <- mGSZ(x, y, l, p = 100)
top.mGSZ.sets <- toTable(mGSZ.obj, n = 10)

# scoring function profile data across the ordered gene list for top 2 gene sets

data4plot <- StabPlotData(mGSZ.obj, rank.vector = c(1,2))

# profile plot for the top gene set

plotProfile(data4plot, 1)

---

rm.rows.with.noSetMembers

**Internal mGSZ function**

Description

Internal mGSZ function not to be called by the users

Author(s)

Pashupati Mishra, Petri Toronen
**rm.small.genesets**

**Description**

Internal mGSZ function not to be called by the users

**Author(s)**

Pashupati Mishra, Petri Toronen

---

**ss.p.values**

**Description**

Internal mGSZ function not to be called by the users

**Author(s)**

Pashupati Mishra, Petri Toronen

---

**ss.test.score**

**Description**

Internal mGSZ function not to be called by the users

**Author(s)**

Pashupati Mishra, Petri Toronen
StabPlotData  

**GSZ scoring function profile data**

**Description**

GSZ scoring function profile data

**Usage**

```r
StabPlotData(mGSZobj, rank.vector, sample.perm.data=FALSE)
```

**Arguments**

- `mGSZobj` : mGSZ object
- `rank.vector` : A vector of ranks for gene sets for which GSZ scoring function profile data is required.
- `sample.perm.data` : Profile data for sample permutation data when both gene and sample permutation are used.

**Details**

Once significant gene sets are reported, it is useful to evaluate a gene set in more detail to see the behavior of the gene set. This can be done by visualizing the scoring function profile across the gene list as shown in the GSEA article (Subramanian et al., 2005). It is even more relevant to compare signals from positive and permuted data. Positive data corresponds to differential gene expression test scores calculated from gene expression data with correct sample labels and permuted data corresponds to differential gene expression test scores calculated from gene expression data with permuted sample labels. This function outputs scoring function profile data for both positive and permuted data to be used as input for the visualization that shows the signal from positive data and a summary of the signal from permuted data.

**Value**

An R object with running GSZ scores for positive and permuted data to be used as input for profile plot.

**Author(s)**

Pashupati Mishra, Petri Toronen

**References**


genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550.


See Also
plotProfile

Examples

geneNnames <- paste("g",1:100, sep = "")

# create random gene expression data matrix
set.seed(100)
x <- matrix(rnorm(100*10), ncol=10)
rownames(x) <- geneNnames
b <- matrix(2*rnorm(50), ncol=5)
ind <- sample(1:10, replace=FALSE)
x[ind,6:10] <- x[ind,6:10] + b

l <- rep(1:2,c(5,5))

# create random gene sets
y <- vector("list", 20)
for(i in 1:length(y)){
y[[i]] <- sample(geneNnames, size = 10)
}
names(y) <- paste("set", as.character(1:20), sep="")

mGSZ.obj <- mGSZ(x, y, 1, p = 100)
top.mGSZ.sets <- toTable(mGSZ.obj, n = 10)

# scoring function profile data across the ordered gene list for top 2 gene sets
data4plot <- StabPlotData(mGSZ.obj, rank.vector=c(1,2))

# profile plot for the top gene set
plotProfile(data4plot,1)
Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

sumVarMean_calc
Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

toMatrix
Internal mGSZ function

Description
Internal mGSZ function not to be called by the users

Author(s)
Pashupati Mishra, Petri Toronen

toTable
Table with top gene sets

Description
Table with top gene sets

Usage
toTable(mGSZobj, sample=FALSE, m=c("mGSZ","mGSA","mAllez","WRS","SS","SUM","KS","wKS"), n=5)
Arguments

\begin{itemize}
  \item \texttt{mGSZobj} \hspace{1cm} \text{mGSZ object}
  \item \texttt{sample} \hspace{1cm} TRUE for table of top gene sets based on sample permutation when both gene and sample permutations were used.
  \item \texttt{n} \hspace{1cm} Method for which table for top gene sets is required (Required only when other methods were used for the gene set analysis)
  \item \texttt{n} \hspace{1cm} Number of top gene sets in the table
\end{itemize}

Value

A table with top gene sets

Author(s)

Pashupati Mishra, Petri Toronen

References


See Also

\texttt{mGSZ}

Examples

```r
gene.names <- paste("g",1:100, sep = ")

# create random gene expression data matrix
set.seed(100)
x <- matrix(rnorm(100*10),ncol=10)ownames(x) <- gene.names
b <- matrix(2*rnorm(50),ncol=5)
ind <- sample(1:10,replace=FALSE)
x[ind,6:10] <- x[ind,6:10] + b
l <- rep(1:2,c(5,5))

# create random gene sets
y <- vector("list", 20)
for(i in 1:length(y)){
y[[i]] <- sample(gene.names, size = 10)
}
names(y) <- paste("set", as.character(1:20), sep="")

mGSZ.obj <- mGSZ(x, y, l, p = 100)
top.mGSZ.sets <- toTable(mGSZ.obj, n = 10)
```
WRS.p.values

**Description**
Internal WRS function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen

---

WRS.test.score

**Description**
Internal WRS function not to be called by the users

**Author(s)**
Pashupati Mishra, Petri Toronen
Index

*Topic \textasciitilde kwd1
  calc_z_var, 2
  count.prob.sum, 3
  count_hyge_var_mean, 3
  diffFscore, 3
  diffScore, 3
  emp, 4
  emp.wrs, 4
  FC, 4
  flipListStruct, 4
  KS.p.values, 5
  KS.score, 6
  listTOclMatrix, 6
  logEVcdf, 6
  logNORMcdf, 6
  mAllez.p.values, 7
  mGSA.p.values, 7
  mGSZ.adj.mean.std, 10
  mGSZ.p.values, 11
  mGSZ.test.score, 11
  mGSZ.test.score.stb2, 11
  mGSZ.test.score.stb3, 11
  mGSZ.test.score.stb4, 12
  mGSZ.test.score2, 12
  mGSZ.test.score4, 12
  pick.data.cols, 12
  rm.rows.with.noSetMembers, 14
  rm.small.genesets, 15
  SS.p.values, 15
  SS.test.score, 15
  sumTestscore, 17
  sumVarMean_calc, 18
  WRS.p.values, 20
  WRS.test.score, 20
*Topic \textasciitilde kwd2
  calc_z_var, 2
  count.prob.sum, 3
  count_hyge_var_mean, 3
  diffFscore, 3
  diffScore, 3
  emp, 4
  emp.wrs, 4
  FC, 4
flipListStruct, 4
geneSetsList, 5
KS.p.values, 5
KS.score, 6
listTOclMatrix, 6
logEVCdf, 6
logNORMcdf, 6
mAllez.p.values, 7
mGSA.p.values, 7
mGSZ, 7, 19
mGSZ.adj.mean.std, 10
mGSZ.p.values, 11
mGSZ.test.score, 11
mGSZ.test.score.stb2, 11
mGSZ.test.score.stb3, 11
mGSZ.test.score.stb4, 12
mGSZ.test.score2, 12
mGSZ.test.score4, 12
pick.data.cols, 12
plotProfile, 13, 17
rm.rows.with.noSetMembers, 14
rm.small.genesets, 15
SS.p.values, 15
SS.test.score, 15
StabPlotData, 14, 16
sumTestscore, 17
sumVarMean_calc, 18
toMatrix, 18
toTable, 18
WRS.p.values, 20
WRS.test.score, 20