Package 'AbsFilterGSEA'

September 21, 2017

Type Package

Title Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering

Version 1.5.1

Author Sora Yoon <yoonsora@unist.ac.kr>

Maintainer Sora Yoon <yoonsora@unist.ac.kr>

Description Gene-set enrichment analysis (GSEA) is popularly used to assess the enrichment of differential signal in a pre-defined gene-set without using a cutoff threshold for differential expression. The significance of enrichment is evaluated through sample- or gene-permutation method. Although the sample-permutation approach is highly recommended due to its good false positive control, we must use gene-permuting method if the number of samples is small. However, such gene-permuting GSEA (or preranked GSEA) generates a lot of false positive gene-sets as the inter-gene correlation in each gene set increases. These false positives can be successfully reduced by filtering with the one-tailed absolute GSEA results. This package provides a function that performs gene-permuting GSEA calculation with or without the absolute filtering. Without filtering, users can perform (original) two-tailed or one-tailed absolute GSEA.

License GPL-2

LazyData TRUE

RoxygenNote 6.0.1

Depends

LinkingTo Rcpp, RcppArmadillo

Imports Rcpp, Biobase, stats, DESeq, limma

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-09-21 13:39:52 UTC

R topics documented:

`example` *Normalized RNA-seq count data*

Description

This is toy example of RNA-seq raw read count table. It containes 5000 genes and 6 samples (three for case and other three for control group).

Usage

```
data("example")
```

Format

A data frame with 5000 observations on the following 6 variables.

`groupA1` a numeric vector for RNA-seq counts for case samples 1.

`groupA2` a numeric vector for RNA-seq counts for case samples 2.

`groupA3` a numeric vector for RNA-seq counts for case samples 3.

`groupB1` a numeric vector for RNA-seq counts for control samples 1.

`groupB2` a numeric vector for RNA-seq counts for control samples 2.

`groupB3` a numeric vector for RNA-seq counts for control samples 3.

Details

This read count dataset was simulated based on the negative binomial distribution. Mean and dispersion parameters were assessed from TCGA KIRC RNA-seq dataset. Normalization was done by using edgeR package.Geneset_41~45 are up-regulated and Geneset_46~50 are down-regulated gene sets.

Source

Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499(7456):43-49.

References

Chen, Y., et al. edgeR: differential expression analysis of digital gene expression data User's Guide. 2015.

Examples

```
data(example)
```

GenePermGSEA *Gene permuting GSEA with or without filtering by absolute GSEA.*

Description

Gene-set enrichment analysis (GSEA) is popularly used to assess the enrichment of differential signal in a pre-defined gene-set without using a cutoff threshold for differential expression. The significance of enrichment is evaluated through sample- or gene-permutation method. Although the sample-permutation approach is highly recommended due to its good false positive control, we must use gene-permuting method if the number of samples is small. However, such gene-permuting GSEA (or preranked GSEA) generates a lot of false positive gene-sets as the inter-gene correlation in each gene set increases. These false positives can be successfully reduced by filtering with the one-tailed absolute GSEA results. This package provides a function that performs gene-permuting GSEA calculation with or without the absolute filtering. Without filtering, users can perform (original) two-tailed or one-tailed absolute GSEA.

Usage

```
GenePermGSEA(countMatrix, GeneScoreType, idxCase, idxControl, GenesetFile,
  normalization, minGenesetSize = 10, maxGenesetSize = 300, q = 1,
  nPerm = 1000, absoluteGeneScore = FALSE, GSEAtype = "absFilter",
  FDR = 0.05, FDRfilter = 0.05, minCount = 3)
```

Arguments

countMatrix	Normalized RNA-seq read count matrix.
GeneScoreType	Type of gene score. Possible gene score is "moderated_t","SNR", "FC" (log fold change score) or "RANKSUM" (zero centered).
idxCase	Indices for case samples in the count matrix. e.g., 1:3
idxControl	Indices for control samples in the count matrix. e.g., 4:6
GenesetFile	File path for gene set file. Typical GMT file or its similar 'tab-delimited' file is available. e.g., "C:/geneset.gmt"
normalization	Type 'DESeq' if the input matrix is composed of raw read counts. It will normalize the raw count data using DESeq method. Or type 'AlreadyNormalized' if the input matrix is already normalized.
minGenesetSize	Minimum size of gene set allowed. Gene-sets of which sizes are below this value are filtered out from the analysis. Default = 10
maxGenesetSize	Maximum size of gene set allowed. Gene-sets of which sizes are larger this value are filtered out from the analysis. Default = 300
q	Weight exponent for gene score. For example, if q=0, only rank of gene score is reflected in calculating gene set score (preranked GSEA). If q=1, the gene score itself is used. If q=2, square of the gene score is used.
nPerm	The number of gene permutation. Default = 1000.

absoluteGeneScore
: Boolean. Whether to take absolue to gene score (TRUE) or not (FALSE).

GSEAtype
: Type of GSEA. Possible value is "absolute", "original" or "absFilter". "absolute" for one-tailed absolute GSEA. "original" for the original two-tailed GSEA. "absFilter" for the original GSEA filtered by the results from the one-tailed absolute GSEA.

FDR
: FDR cutoff for the original or absolute GSEA. Default = 0.05

FDRfilter
: FDR cutoff for the one-tailed absolute GSEA for absolute filtering (only working when GSEAtype is "absFilter"). Default = 0.05

minCount
: Minimum median count of a gene to be included in the analysis. It is used for gene-filtering to avoid genes having small read counts. Default = 0

Details

Typical usages are GenePermGSEA(countMatrix = countMatrix, GeneScoreType = "moderated_t", idxCase = 1:3, idxControl = 4:6, GenesetFile = 'geneset.txt', GSEAtype = "absFilter")

Value

GSEA result table sorted by FDR Q-value.

Source

Nam, D. Effect of the absolute statistic on gene-sampling gene-set analysis methods. Stat Methods Med Res 2015. Subramanian, A., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA 2005;102(43):15545-15550. Li, J. and Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research 2013;22(5):519-536.

References

Nam, D. Effect of the absolute statistic on gene-sampling gene-set analysis methods. Stat Methods Med Res 2015. Subramanian, A., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA 2005;102(43):15545-15550. Li, J. and Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research 2013;22(5):519-536. Simon Anders and Wolfgang Huber (2010): Differential expression analysis for sequence count data. Genome Biology 11:R106

Examples

```
data(example)

# Create a gene set file and save it to your local directory.
# Note that you can use your local gene set file (tab-delimited like *.gmt file from mSigDB).
# But here, we will generate a toy gene set file to show the structure of this gene set file.
# It consists of 50 gene sets and each contains 100 genes.
```

```
for(Geneset in 1:50)
{
  GenesetName = paste("Geneset", Geneset, sep = "_")
  Genes = paste("Gene", (Geneset*100-99):(Geneset*100), sep="", collapse = '\t')
  Geneset = paste(GenesetName, Genes, sep = '\t')
  write(Geneset, file = "geneset.txt", append = TRUE, ncolumns = 1)
}

# Run Gene-permuting GSEA
RES = GenePermGSEA(countMatrix = example, GeneScoreType = "moderated_t", idxCase = 1:3,
                   idxControl = 4:6, GenesetFile = 'geneset.txt', normalization = 'DESeq',
                     GSEAtype = "absFilter")
RES
```

Index