Package 'PathwaySplice'

April 12, 2018

Type Package

Title An R Package for Unbiased Splicing Pathway Analysis

Version 1.0.0

Author Aimin Yan, Xi Chen, Lily Wang

Maintainer Aimin Yan <aimin.at.work@gmail.com>

Depends R (>= 3.4.0)

Imports goseq, Biobase, DOSE, reshape2, igraph, org.Hs.eg.db, org.Mm.eg.db, BiocGenerics, AnnotationDbi, JunctionSeq, BiasedUrn, GO.db,gdata, geneLenDataBase, grDevices, graphics, stats, utils, VennDiagram, RColorBrewer, ensembldb, AnnotationHub, S4Vectors, dplyr, plotly, webshot, htmlwidgets , mgcv ,gridExtra, grid ,gplots, tibble

Suggests testthat, knitr, rmarkdown

VignetteBuilder knitr

biocViews AlternativeSplicing, DifferentialSplicing, GeneSetEnrichment, GO, RNASeq, Sequencing, Software, Visualization, NetworkEnrichment, Network, Pathways, GraphAndNetwork, Regression

Description

Pathway analysis of alternative splicing would be biased without accounting for the different number of exons associated with each gene, because genes with higher number of exons are more likely to be included in the 'significant' gene list in alternative splicing.
PathwaySplice is an R package that:
(1) performs pathway analysis that explicitly adjusts for the number of exons associated with each gene
(2) visualizes selection bias due to different number of exons for each gene
(3) formally tests for presence of bias using logistic regression
(4) supports gene sets based on the Gene Ontology terms, as well as more broadly defined gene sets (e.g. MSigDB) or user defined gene sets
(5) identifies the significant genes driving pathway significance
(6) organizes significant pathways with an enrichment map, where pathways with large number of overlapping genes are grouped together in a network graph

License LGPL(>=2)

LazyData TRUE

RoxygenNote 6.0.1

NeedsCompilation no

R topics documented:

`compareResults` *compareResults*

Description

This function helps with visualizing the effects of bias adjustment in pathway analysis, by comparing the distributions of bias factors (e.g. number of exon bins) in genes associated with the most significant gene sets, before and after adjusting for bias factors in splicing pathway analysis.

Usage

```
compareResults(n.go, adjusted, unadjusted, gene.based.table,
  output.dir = tempdir(), type.boxplot = c("All", "Only3"))
```

Arguments

`n.go`	Distributions of bias factor in genes associated with the most significant `n.go` gene sets will be compared
`adjusted`	An object returned by `runPathwaySplice`, should correspond to gene set anlaysis results adjusting for biases in splicing analysis
`unadjusted`	An object returned by `runPathwaySplice`, should correspond to gene set analysis results NOT adjusting for biases
`gene.based.table`	An object returned by `makeGeneTable`, should correspond to a table with one p-value for each gene
`output.dir`	Directory for output files
`type.boxplot`	Options are 'All' and 'Only3', corresponding to drawing 5 boxplots or 3 boxplots. 5 boxplots: `all genesets`, `sig.adjusted` (sig gene sets in adjusted analysis), `sig.unadjusted` (sig gene sets in unadjusted analysis), `sig.adjusted.only` (sig gene sets in adjusted analysis only), `sig.unadjusted.only` (sig gene sets in unadjusted analysis only) 3 boxplots: `all genesets`, `adjusted.sig`, `unadjusted.sig`

Value

The output include 3 files in `output.dir`: (1) a venn diagram comparing significant gene sets before and after adjusting for bias factors (2) a .csv file with gene set names belonging to different sections of the venn diagram (3) a box plot showing the distributions of number of features within all genes in significant gene sets, with and without adjusting for bias factors

Examples

```
dir.name <- system.file('extdata', package='PathwaySplice')
hallmark.pathway.file <- file.path(dir.name,'h.all.v6.0.symbols.gmt.txt')

hallmark <- gmtGene2Cat(hallmark.pathway.file,genomeID='hg19')

gene.based.table <- makeGeneTable(featureBasedData)

res.adj <- runPathwaySplice(gene.based.table,genome='hg19',
                        id='ensGene',gene2cat=hallmark,
                        go.size.limit = c(5, 200),
                        method='Wallenius', output.file=tempfile())

res.unadj <- runPathwaySplice(gene.based.table,genome='hg19',
                        id='ensGene',gene2cat=hallmark,go.size.limit = c(5, 200),
                        method='Hypergeometric',output.file=tempfile())

compareResults(20, res.adj, res.unadj, gene.based.table, type.boxplot='Only3')

## Not run:
# illustrate specification of output directory
compareResults(20, res.adj, res.unadj, gene.based.table, type.boxplot='Only3',output.dir='C:/Temp')
output.file.dir <- '~/OutputTestPathwaySplice'
compareResults(20,res.adj, res.unadj,gene.based.table,output.file.dir,type.boxplot='Only3')

## End(Not run)
```

`enrichmentMap` *enrichmentMap*

Description

This function draws an enrichment map based on the overlap of gene sets as measured by the Jaccard Coefficient(JC)

Usage

```
enrichmentMap(pathway.res, n = 50, fixed = TRUE, node.label.font = 1,
  similarity.threshold, scaling.factor = 1, output.file.dir = tempdir(),
  label.node.by.index = FALSE, add.numSIGInCat = FALSE, ...)
```

Arguments

`pathway.res`	Pathway analysis results, an object returned by `runPathwaySplice`
`n`	The top *n* most significant gene sets are shown on enrichment map
`fixed`	If set to FALSE, will invoke tkplot (an interactive graphing facility in R) that allows one to draw an interactive enrichment map. Users can then manually adjust the layout of the enrichment map. Note: on OS X system, users need to have XQuartz installed to run this function . tcltk R package is also required, but in most distributions of R tcltk is already included

`node.label.font`
: Font size of node label

`similarity.threshold`
: Gene sets with Jaccard Coefficient > `similarity.threshold` will be connected on the enrichment map

`scaling.factor`
: Scaling factor that users can use to adjust the edge thickness of the network, which is based on value of sqrt(JC coefficient * 5) * scaling.factor

`output.file.dir`
: Output files directory, see `Details` section below.

`label.node.by.index`
: Options for labeling nodes on network.
FALSE indicates to label nodes by gene set names
TRUE indicates to label nodes by the index of gene sets

`add.numSIGInCat`
: Option for users to decide whether to add number of signficant genes of each gene set to the nodes in enrichment map or not

`...`
: Additional parameter

Details

In the enrichment map,

- the *node colors* are controlled by gene set p-values, where smaller p-values correspond to dark red color.
- *node sizes* are controlled by the number of significant genes in gene set.
- *thickness of the edges* correspond to Jaccard similarity coefficient between two gene sets.
- the numbers after ':' indicates the nubmer of significant genes in the gene set.

The Jaccard similarity coefficient ranges from 0 to 1. JC=0 indicates there are no overlapping genes between two gene sets, JC=1 indicates two gene sets are identical.

The output directory will include the following files:

(1) a network file (in GML format) that can be used as an input for Cytoscape software (2) when label.node.by.index=TRUE, also a gene set information file that includes full names of the gene sets and the gene set indices shown on the network.

Value

A list with edge and node information used to plot enrichment map

Author(s)

Aimin created this funciton based on enrichMap function in G Yu's DOSE R package

Examples

```
gene.based.table <- makeGeneTable(featureBasedData)

res <- runPathwaySplice(gene.based.table,genome='hg19',
                        id='ensGene',test.cats=c('GO:BP'),
                        go.size.limit=c(5,30),method='Wallenius')
```

```
# labeling each node by gene set name
enmap <- enrichmentMap(res,n=10,similarity.threshold=0.3,
label.node.by.index = FALSE)

# labeling each node by gene set index
enmap <- enrichmentMap(res,n=10,similarity.threshold=0.3,
label.node.by.index = TRUE)

## Not run:
# illustrates specification of output file directory
# Enable interactive map and label each node by gene set index
enmap <- enrichmentMap(res,n=10,fixed=FALSE, similarity.threshold=0.3,
                      label.node.by.index = TRUE, output.file.dir='C:/temp')

enmap <- enrichmentMap(res,n=10,similarity.threshold=0.3,
                      label.node.by.index = FALSE, output.file.dir='C:/temp')
## End(Not run)
```

`featureBasedData` *featureBasedData*

Description

This dataset includes analysis results of RNA-seq data in Dolatshad et al. (2015), which compared transcriptome of CD34+ cells from myelodysplastic syndrome (MDS) patients with SF3B1 mutations vs. healthy controls using RNA sequencing. The `JunctionSeq` package was used to assess differential usage of counting bins, which are non-overlapping segments of the exons or splicing junctions (see Fig 1 in Anders et al. (2012)). Because of the size limit, only counting bins associated with a subset of genes were included here for demonstration.

Usage

```
data(featureBasedData)
```

Format

A data frame with variables for gene identifier (`geneID`), gene feature identifier (`countbinID`), and p-value for gene feature (`pvalue`). Here we used "gene feature" and "counting bin" interchangeably

References

H Dolatshad, A Pellagatti, M Fernandez-Mercado1, B H Yip, L Malcovati, M Attwood, B Przychodzen N Sahgal, A A Kanapin, H Lockstone, L Scifo, P Vandenberghe, E Papaemmanuil, C W J Smith, P J Campbell, S Ogawa1, J P Maciejewski, M Cazzola, K I Savage1 and J Boultwood1 (2015) *Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells.* Leukemia (2015) 29, 1092-1103

Anders S, Reyes A, Huber W (2012) *Dececting differential usage of exons from RNA-seq data.* Genome Research 22(10): 2008-2017

gmtGene2Cat *gmtGene2Cat*

Description

Obtains all pathways associated with a set of genes

Usage

```
gmtGene2Cat(pathway.file, gene.anno.file = NULL, genomeID = c("mm10",
  "hg19", "hg38"))
```

Arguments

pathway.file	Input file for the gene sets in GMT format, must be in *gene symbols*
gene.anno.file	Gene annotation file that facilitate gene id conversions when gene ids in RNA-Seq data and pathway.file differ. If not specified, gmtGene2cat relies on gene annotations provided by R package AnnotationHub.
genomeID	Genome to be used. Options are 'mm10','hg19' or 'hg38'.

Details

This function reads a gene set file in https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#GMT:_Gene_Matrix_Transposed_file_format_.28.2A.gmt.29, and returns a list with its name being a gene id, and each element of the list being the pathways associated with the gene. When gene ids in RNA-Seq data differ from those in pathway database, gene.anno.file facilitate gene id conversions. Users can prepare this file based on the format of the example gene annotation file at https://raw.githubusercontent.com/aiminy/GOSJ/master/data/gene_annotation.txt

Value

A list where each entry is named by a gene and contains a vector of all the pathways associated with the gene

Examples

```
#using local file for pathways database
dir.name <- system.file('extdata', package='PathwaySplice')
hallmark.local.pathways <- file.path(dir.name,'h.all.v6.0.symbols.gmt.txt')
hlp <- gmtGene2Cat(hallmark.local.pathways, genomeID='hg19')

## Not run:
#using url for pathways database linked to a website
hallmark.url.pathways <- paste0('https://raw.githubusercontent.com/SCCC-BBC',
                  '/PathwaySplice/development/inst/extdata',
                    '/h.all.v6.0.symbols.gmt.txt')
hup <- gmtGene2Cat(hallmark.url.pathways, genomeID='hg19')
## End(Not run)
```

lrTestBias *lrTestBias*

Description

This function tests presence of selection bias using logistic regression, and produces a boxplot that compares distributions of bias factors (e.g. number of exons) for significant genes and non-significant genes.

Usage

```
lrTestBias(genewise.table, boxplot.width = 0.1)
```

Arguments

genewise.table	A dataframe with genewise p-value for each gene, returned from makeGeneTable()
boxplot.width	width of boxplot

Details

To determine presentce of selection bias, we fit the following logistic regression model:

Pr(a gene is significant) ~ number of features within the gene

Here features refer to exon bins or splicing junction bins, depending on how genewise pvalues were obtained in the genewise.table

Value

Nothing to be returned

Examples

```
gene.based.table <- makeGeneTable(featureBasedData)
lrTestBias(gene.based.table)
```

---

makeGeneTable *makeGeneTable*

Description

This function obtains genewise p-values, by representing each gene with the smallest p-value among its features, and then determines genes status as significant or not.

Usage

```
makeGeneTable(feature.table, sig.threshold = 0.05, stat = "pvalue")
```

Arguments

feature.table	An featureBasedData object.
sig.threshold	Significance threshold used to determine whether the gene is significant or not
stat	The statistic used to select significant genes. Options are 'pvalue' or 'fdr'

Value

Returns a genewised table with several variables (columns)

geneID	Gene identifiers in ensembl gene IDs
geneWisePvalue	each gene is represented by the smallest p-value among its features
numFeature	number of gene features within the gene
fdr	false discovery rate for genewisePvalue
sig.gene	a gene is significant (1) or not (0)

Examples

```
data(featureBasedData)
gene.based.table <- makeGeneTable(featureBasedData)
```

---

runPathwaySplice *runPathwaySplice*

Description

This function identifies pathways that are enriched with signficant genes, while accounting for different number of gene features (e.g. exons) associated with each gene

Usage

```
runPathwaySplice(genewise.table, genome, id, gene2cat = NULL,
  test.cats = c("GO:CC", "GO:BP", "GO:MF"), go.size.limit = c(10, 200),
  method = "Wallenius", repcnt = 2000, use.genes.without.cat = FALSE,
  binsize = "auto", output.file = tempfile())
```

Arguments

genewise.table	data frame returned from function makeGeneTable
genome	Genome to be used, options are 'hg19' or 'mm10'
id	GeneID, options are 'entrezgene' or 'ensembl_gene_id'
gene2cat	Get sets to be tested, these are defined by users, can be obtained from gmtGene2Cat function
test.cats	Default gene ontology gene sets to be tested if gene2cat is not defined
go.size.limit	Size limit of the gene sets to be tested
method	the method used to calculate pathway enrichment p value. Options are 'Wallenius', 'Sampling', and 'Hypergeometric'
repcnt	Number of random samples to be calculated when 'Sampling' is used, this argument ignored unless method='Sampling'
use.genes.without.cat	Whether genes not mapped to any gene_set tested are included in analysis. Default is set to FALSE, where genes not mapped to any tested categories are ignored in analysis.
binsize	The number of genes in each gene bin in the bias plot
output.file	File name for the analysis result in .csv format.

Details

This function implements the methodology described in Young et al. (2011) to adjust for different number of gene features (column `numFeature` in `gene.based.table`). For example, gene features can be non-overlapping exon counting bins associated with each gene (Fig 1 in Anders et al. 2012). In the bias plot, the genes are grouped by `numFeature` in `genewise.table` into gene bins, the proportions of signficant genes are then plotted against the gene bins.

Value

runPathwaySplice returns a tibble with the following information:

`gene_set` Name of the gene set. Note in this document we used the terms gene_set, category, and pathway interchangeably

`over_represented_pvalue` P-vaue for the associated gene_set being over-represented among significant genes

`under_represented_pvalue` P-vaue for the associated gene_set being under-represented among significant genes

`numSIGInCat` The number of significant genes in the gene_set

`numInCat` The total number of genes in the gene_set

`description` Description of the gene gene_set

`ontology` The domain of the gene ontology terms if GO categories were tested. Go categories can be classified into three domains: cellular component, biological process, molecular function.

`SIGgene_ensembl` Ensembl gene ID of significant genes in the gene_set

`SIGgene_symbol` Gene symbols of signficant genes in the gene_set

`Ave_value_all_gene` The average value for `numFeature` for all the genes in the gene_set, note that `numFeature` is the bias factor adjusted by PathwaySplice

These information are also saved in the file `output.file`

References

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2011) *Gene ontology analysis for RNA-seq: accounting for selection bias*. Genome Biology 11:R14

Anders S, Reyes A, Huber W (2012) *Dececting differential usage of exons from RNA-seq data.* Genome Research 22(10): 2008-2017

Examples

```
gene.based.table <- makeGeneTable(featureBasedData)

res <- runPathwaySplice(gene.based.table,genome='hg19',id='ensGene',
                        test.cats=c('GO:BP'),
                        go.size.limit=c(5,30),
                        method='Wallenius',binsize=20)
## Not run:
# demonstrate how output file can be specified
```

```
res <- runPathwaySplice(gene.based.table,genome='hg19',id='ensGene',
                        test.cats=c('GO:BP'),
                        go.size.limit=c(5,30),
                        method='Wallenius',binsize=800,
                        output.file='C:/temp/test.csv')

# demonstrate using customized gene sets
dir.name <- system.file('extdata', package='PathwaySplice')
hallmark.local.pathways <- file.path(dir.name,'h.all.v6.0.symbols.gmt.txt')
hlp <- gmtGene2Cat(hallmark.local.pathways, genomeID='hg19')

res <- runPathwaySplice(gene.based.table,genome='hg19',id='ensGene',
                        gene2cat=hlp,
                        go.size.limit=c(5,200),
                        method='Wallenius',binsize=20,
                        output.file='C:/temp/test.csv')

## End(Not run)
```

Index