

# Package ‘PathwaySplice’

October 16, 2018

**Type** Package

**Title** An R Package for Unbiased Splicing Pathway Analysis

**Version** 1.4.0

**Author** Aimin Yan, Xi Chen, Lily Wang

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

**Depends** R (>= 3.5.0)

**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, EnrichmentBrowser, annotate, KEGGREST

**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

**git\_url** <https://git.bioconductor.org/packages/PathwaySplice>

**git\_branch** RELEASE\_3\_7

**git\_last\_commit** e8f9ac8

**git\_last\_commit\_date** 2018-04-30

**Date/Publication** 2018-10-15

## R topics documented:

compareResults . . . . .	2
enrichmentMap . . . . .	3
featureBasedData . . . . .	5
gmtGene2Cat . . . . .	6
lrTestBias . . . . .	7
makeGeneTable . . . . .	8
outKegg2Gmt . . . . .	8
runPathwaySplice . . . . .	9
<b>Index</b>	<b>12</b>

---

compareResults	<i>compareResults</i>
----------------	-----------------------

---

### 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 analysis 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 or 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')

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

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

## Not run:
# illustrate specification of output directory on windows systems
compareResults(20, res.adj, res.unadj, gene.based.table, type.boxplot='Only3', output.dir=tempdir())

output.dir <- '~/OutputTestPathwaySplice' #linux system
compareResults(20, res.adj, res.unadj, gene.based.table, output.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

<code>pathway.res</code>	Pathway analysis results, an object returned by <code>runPathwaySplice</code>
<code>n</code>	The top $n$ most significant gene sets are shown on enrichment map
<code>fixed</code>	If set to <code>FALSE</code> , will invoke <code>tkplot</code> (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 <code>XQuartz</code> installed to run this function. <code>tcltk</code> R package is also required, but in most distributions of R <code>tcltk</code> is already included
<code>node.label.font</code>	Font size of node label
<code>similarity.threshold</code>	Gene sets with Jaccard Coefficient $>$ <code>similarity.threshold</code> will be connected on the enrichment map
<code>scaling.factor</code>	Scaling factor that users can use to adjust the edge thickness of the network, which is based on value of $\sqrt{\text{JC coefficient} * 5} * \text{scaling.factor}$
<code>output.file.dir</code>	Output files directory, see Details section below.
<code>label.node.by.index</code>	Options for labeling nodes on network. <code>FALSE</code> indicates to label nodes by gene set names <code>TRUE</code> indicates to label nodes by the index of gene sets
<code>add.numSIGInCat</code>	Option for users to decide whether to add number of significant genes of each gene set to the nodes in enrichment map or not
<code>...</code>	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 number of significant genes in the gene set.

The Jaccard similarity coefficient ranges from 0 to 1.  $\text{JC}=0$  indicates there are no overlapping genes between two gene sets,  $\text{JC}=1$  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 function 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=tempdir())

enmap <- enrichmentMap(res,n=10,similarity.threshold=0.3,
                      label.node.by.index = FALSE, output.file.dir=tempdir())

## 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-Mercado<sup>1</sup>, 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 Ogawa<sup>1</sup>, J P Maciejewski, M Cazzola, K I Savage<sup>1</sup> and J Boulton<sup>1</sup> (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) *Deceiving 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](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](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 presence of selection bias, we fit the following logistic regression model:

$$\Pr(\text{a gene is significant}) \sim \text{number of features within the gene}$$

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	<i>makeGeneTable</i>
---------------	----------------------

---

### 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

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

### Value

Returns a genewised table with several variables (columns)

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

### Examples

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

---

outKegg2Gmt	<i>outKegg2gmt</i>
-------------	--------------------

---

### Description

This function obtains a .gmt file for KEGG pathways.

### Usage

```
outKegg2Gmt(organism.id, out.gmt.file)
```

### Arguments

<code>organism.id</code>	an identifier for the organism being studied, for example, "hsa" for "Homo sapiens"
<code>out.gmt.file</code>	name of the output .gmt file



**Details**

The function calls the `get.kegg.genesets` function in `EnrichmentBrowser` R package and modifies the resulting output into a `.gmt` file.

**Value**

Returns a `.gmt` file for KEGG pathways

**Examples**

```
## Not run:

data.dir <- tempdir()
outKegg2Gmt ("hsa",file.path(data.dir,"kegg.gmt.txt"))

kegg.pathways <- gmtGene2Cat(file.path(data.dir, "kegg.gmt.txt"),genomeID = "hg19")

result.kegg <- runPathwaySplice(genewise.table = gene.based.table.fdr,
                               genome = "hg19",
                               id = "ensGene",
                               gene2cat = kegg.pathways,
                               go.size.limit = c(5, 100),
                               method = "Wallenius",
                               use.genes.without.cat = TRUE)

## End(Not run)
```

---

<code>runPathwaySplice</code>	<i>runPathwaySplice</i>
-------------------------------	-------------------------

---

**Description**

This function identifies pathways that are enriched with significant 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**

<code>genewise.table</code>	data frame returned from function <code>makeGeneTable</code>
<code>genome</code>	Genome to be used, options are 'hg19' or 'mm10'
<code>id</code>	GeneID, options are 'entrezgene' or 'ensembl_gene_id'
<code>gene2cat</code>	Get sets to be tested, these are defined by users, can be obtained from <code>gmtGene2Cat</code> function
<code>test.cats</code>	Default gene ontology gene sets to be tested if <code>gene2cat</code> 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 the analysis. Default is set to FALSE, where genes not mapped to any tested categories are ignored in analysis. Set this option to TRUE if it's desired that all genes in <code>genewise.table</code> to be counted towards the total number of genes outside the category.
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 significant genes are then plotted against the gene bins.

### Value

`runPathwaySplice` returns a **tibble** with the following information:

<code>gene_set</code>	Name of the gene set. Note in this document we used the terms <code>gene_set</code> , <code>category</code> , and <code>pathway</code> interchangeably
<code>over_represented_pvalue</code>	P-value for the associated <code>gene_set</code> being over-represented among significant genes
<code>under_represented_pvalue</code>	P-value for the associated <code>gene_set</code> being under-represented among significant genes
<code>numSIGInCat</code>	The number of significant genes in the <code>gene_set</code>
<code>numInCat</code>	The total number of genes in the <code>gene_set</code>
<code>description</code>	Description of the gene <code>gene_set</code>
<code>ontology</code>	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.
<code>SIGgene_ensembl</code>	Ensembl gene ID of significant genes in the <code>gene_set</code>
<code>SIGgene_symbol</code>	Gene symbols of significant genes in the <code>gene_set</code>
<code>Ave_value_all_gene</code>	The average value for <code>numFeature</code> for all the genes in the <code>gene_set</code> , note that <code>numFeature</code> is the bias factor adjusted by <code>PathwaySplice</code>

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) *Deecting 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=tempfile())

# 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=tempfile())

## End(Not run)
```

# Index

## \*Topic **datasets**

featureBasedData, [5](#)

compareResults, [2](#)

enrichmentMap, [3](#)

featureBasedData, [5](#)

gmtGene2Cat, [6](#)

lrTestBias, [7](#)

makeGeneTable, [8](#)

outKegg2Gmt, [8](#)

runPathwaySplice, [9](#)