Package ‘OutSplice’

October 10, 2024

Type Package

Title Comparison of Splicing Events between Tumor and Normal Samples

Version 1.4.0

Description An easy to use tool that can compare splicing events in tumor and normal tissue samples using either a user generated matrix, or data from The Cancer Genome Atlas (TCGA). This package generates a matrix of splicing outliers that are significantly over or underexpressed in tumors samples compared to normal denoted by chromosome location. The package also will calculate the splicing burden in each tumor and characterize the types of splicing events that occur.

Encoding UTF-8

License GPL-2

Depends R(>= 4.3)

Imports AnnotationDbi (>= 1.60.0), GenomicRanges (>= 1.49.0), GenomicFeatures (>= 1.50.2), IRanges (>= 2.32.0), org.Hs.eg.db (>= 3.16.0), TxDb.Hsapiens.UCSC.hg19.knownGene (>= 3.2.2), TxDb.Hsapiens.UCSC.hg38.knownGene (>= 3.16.0), S4Vectors (>= 0.36.0)

Suggests BiocStyle, knitr, rmarkdown, testthat

VignetteBuilder knitr

biocViews AlternativeSplicing, DifferentialExpression, DifferentialSplicing, GeneExpression, RNASeq, Software, VariantAnnotation

BugReports https://github.com/GuoLabUCSD/OutSplice/issues

URL https://github.com/GuoLabUCSD/OutSplice

RoxygenNote 7.2.3

git_url https://git.bioconductor.org/packages/OutSplice

git_branch RELEASE_3_19

git_last_commit 311bef9

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-10-09

Author Joseph Bendik [aut] (<https://orcid.org/0000-0003-0877-5639>),
Sandhya Kalavacherla [aut] (<https://orcid.org/0000-0003-0485-9042>),
Michael Considine [aut] (<https://orcid.org/0000-0002-8666-4857>),
Bahman Afsari [aut] (<https://orcid.org/0000-0001-8717-7199>),
Michael F. Ochs [aut],
Joseph Califano [aut] (<https://orcid.org/0000-0002-4715-6761>),
Daria A. Gaykalova [aut] (<https://orcid.org/0000-0001-5037-0147>),
Elana Fertig [aut] (<https://orcid.org/0000-0003-3204-342X>),
Theresa Guo [cre, aut] (<https://orcid.org/0000-0002-1689-3275>)

Maintainer Theresa Guo <twguo@health.ucsd.edu>

Contents

OutSplice 2
outspliceAnalysis 3
outspliceTCGA 6
plotJunctionData 8

Index **11**

`OutSplice` *OutSplice package overview*

Description

An easy to use tool that can compare splicing events in tumor and normal tissue samples using either a user generated matrix, or data from The Cancer Genome Atlas (TCGA). This package generates a matrix of splicing outliers that are significantly over or underexpressed in tumors samples compared to normal denoted by chromosome location. The package also will calculate the splicing burden in each tumor, characterize the types of splicing events that occur, and allows the user to create waterfall plots of event expression.

Details

Below are the available functions provided by OutSplice:

- `outspliceAnalysis`
- `outspliceTCGA`
- `plotJunctionData`

Value

Please see the man pages for each function.

outspliceAnalysis *Alternative Splicing Analyses*

Description

Analyzes differential splicing events between tumor and normal samples.

Usage

```
outspliceAnalysis(
  junction,
  gene_expr,
  rawcounts,
  sample_labels,
  saveOutput = FALSE,
  output_file_prefix = NULL,
  dir = NULL,
  filterSex = TRUE,
  annotation = "org.Hs.eg.db",
  TxDb = "TxDb.Hsapiens.UCSC.hg38.knownGene",
  offsets_value = 1e-05,
  correction_setting = "fdr",
  p_value = 0.05,
  use_junc_col = 1,
  use_gene_col = 1,
  use_rc_col = 1,
  use_labels_col = 1
)
```

Arguments

`junction` A character string giving the path to a tab separated text file with raw junction counts. One column should include all of the junctions to be looked at by OutSplice (Ex: chr1: 1-100). Each proceeding column is a sample with the raw count information for each corresponding junction. The header row contains the name of the junction column, and the names of the samples.

`gene_expr` A character string giving the path to a tab separated file with normalized gene expression data. One column should include all of the entrez ids for each gene, and each proceeding column should be a sample with the normalized expression values for each gene. The file header row contains the name of the entrez id column, and the names of the samples.

`rawcounts` A character string giving the path to a tab separated text file with the reads per million counts for each sample. This file can either include a row with the total counts per sample, or multiple rows with raw counts per gene per sample that will be summed automatically by OutSplice. One of the columns should include the user defined row names and the subsequent columns are the sample's

rawcount information. The header row contains the name of the row names column and the names of the samples.

`sample_labels` A character string giving the path to a tab separated text file with a matrix of tumor and normal labels (T/F) for each sample. One of the columns should include the names of the samples, and the other column should include "T" for tumors and "F" for normals. The header row contains user defined column names.

`saveOutput` A boolean representing whether or not to save the results to an R data file and tab separated files. Default is FALSE. Optional.

`output_file_prefix` A character string giving the name of the prefix the user would like to use for the output data file. Default is NULL. Optional.

`dir` A character string giving the path to the directory the user would like to save output to. Default is NULL. Optional.

`filterSex` A boolean representing whether or not to include junctions found on the sex chromosomes. Default is TRUE. Optional.

`annotation` A connection or a character string giving the name of the Bioconductor library the user would like to use containing the genome wide annotation. Default is "org.Hs.eg.db". Optional.

`TxDb` A character string giving the name of the Bioconductor library the user would like to use that will expose the annotation database as a TxDb object. Default is "TxDb.Hsapiens.UCSC.hg38.knownGene". Optional.

`offsets_value` The minimum expression value needed to call an event an outlier after normalizing event expression with gene expression. Default is 0.00001. Optional.

`correction_setting` Option to designate how to correct significance. The available options are: "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", and "none". Default is "fdr". Optional.

`p_value` Set the alpha value for the significance threshold. Default is 0.05. Optional.

`use_junc_col` An integer indicating which column in the junction matrix contains the junction regions in your matrices. Default is 1. Optional

`use_gene_col` An integer indicating which column in the gene_expr matrix contains the entrez ids of your genes. Default is 1. Optional

`use_rc_col` An integer indicating which column in the rawcounts matrix contains the row names. Default is 1. Optional

`use_labels_col` An integer indicating which column in the sample_labels matrix contains the sample names. Default is 1. Optional

Value

A list containing the below data.

- FisherAnalyses: Data Frame of junction events containing the number of under/over-expressed outliers in the tumor group (Num_UE_Outliers/Num_OE_Outliers), the Fisher p-value for under/over-expressed events (FisherP1/FisherP2), and a ranking of the under/over expressed events (UE_Rank/OE_Rank)

- ASE.type: significant junction events labeled by type (skipping, insertion, or deletion)
- geneAnnotations: object containing gene names corresponding to each junction region
- junc.Outliers: list containing the logical matrices TumorOverExpression and TumorUnderExpression. "True" indicates an over-expressed event in TumorOverExpression, or an under-expressed event in TumorUnderExpression.
- junc.RPM: junction counts in reads per million following a division of the junction counts input by the total rawcounts for each sample
- junc.RPM.norm: junction counts normalized by each event's total gene expression value
- gene_expr: gene expression values for each junction event
- splice_burden: matrix containing the number of Fisher-P significant over-expressed, under-expressed, and total number of outliers per sample
- NORM.gene_expr.norm: Median of junction data normalized by gene expression for normal samples only (Used for Junction Plotting Only)
- pheno: Phenotypes of Samples (Tumor or Normal)
- pvalues: Junction Fisher P-values

References

Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and Harvard. doi:10.7908/C11G0KM9

Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015 Jan 29;517(7536):576-82. doi: 10.1038/nature14129. PMID: 25631445; PMCID: PMC4311405.

Guo T, Sakai A, Afsari B, Considine M, Danilova L, Favorov AV, Yegnasubramanian S, Kelley DZ, Flam E, Ha PK, Khan Z, Wheelan SJ, Gutkind JS, Fertig EJ, Gaykalova DA, Califano J. A Novel Functional Splice Variant of AKT3 Defined by Analysis of Alternative Splice Expression in HPV-Positive Oropharyngeal Cancers. Cancer Res. 2017 Oct 1;77(19):5248-5258. doi: 10.1158/0008-5472.CAN-16-3106. Epub 2017 Jul 21. PMID: 28733453; PMCID: PMC6042297.

Liu C, Guo T, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Califano JA. A novel splice variant of LOXL2 promotes progression of human papillomavirus-negative head and neck squamous cell carcinoma. Cancer. 2020 Feb 15;126(4):737-748. doi: 10.1002/cncr.32610. Epub 2019 Nov 13. PMID: 31721164.

Liu C, Guo T, Xu G, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Khan Z, Fisch KM, Califano J. Characterization of Alternative Splicing Events in HPV-Negative Head and Neck Squamous Cell Carcinoma Identifies an Oncogenic DOCK5 Variant. Clin Cancer Res. 2018 Oct 15;24(20):5123-5132. doi: 10.1158/1078-0432.CCR-18-0752. Epub 2018 Jun 26. PMID: 29945995; PMCID: PMC6440699.

M. F. Ochs, J. E. Farrar, M. Considine, Y. Wei, S. Meshinchi, and R. J. Arceci. Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. IEEE/ACM Trans Comput Biol Bioinform, 11: 520-32, 2014. PMCID: PMC4156935

Examples

```
junction <- system.file("extdata", "HNSC_junctions.txt.gz", package = "OutSplice")
gene_expr <- system.file("extdata", "HNSC_genes_normalized.txt.gz", package = "OutSplice")
```

```
rawcounts <- system.file("extdata", "Total_Rawcounts.txt", package = "OutSplice")
sample_labels <- system.file("extdata", "HNSC_pheno_table.txt", package = "OutSplice")
output_file_prefix <- "OutSplice_Example"
TxDb_hg19 <- "TxDb.Hsapiens.UCSC.hg19.knownGene"
dir <- paste0(tempdir(), "/")
results <- outspliceAnalysis(junction, gene_expr, rawcounts, sample_labels, saveOutput = TRUE, output_file_prefix
message("Output is located at: ", dir)
```

`outspliceTCGA` *Alternative Splicing Analysis for TCGA Data*

Description

Analyze differential splicing events between tumor and normal samples for TCGA formatted datasets. Examples of TCGA file formats can be viewed here (https://gdac.broadinstitute.org/)

Usage

```
outspliceTCGA(
  junction,
  gene_expr,
  rawcounts,
  saveOutput = FALSE,
  output_file_prefix = NULL,
  dir = NULL,
  filterSex = TRUE,
  annotation = "org.Hs.eg.db",
  TxDb = "TxDb.Hsapiens.UCSC.hg19.knownGene",
  offsets_value = 1e-05,
  correction_setting = "fdr",
  p_value = 0.05
)
```

Arguments

`junction` A character string giving the path to a tab separated text file with raw junction counts.

`gene_expr` A character string giving the path to a tab separated file with normalized gene expression data.

`rawcounts` A character string giving the path to a tab separated text file with the reads per million counts for each sample.

`saveOutput` A boolean representing whether or not to save the results to an R data file and tab separated files. Default is FALSE. Optional.

`output_file_prefix` A character string giving the name of the prefix the user would like to use for the output data file. Default is NULL. Optional.

`dir` A character string giving the path to the directory the user would like to save output to. Default is NULL. Optional.

`filterSex` A boolean representing whether or not to include junctions found on the sex chromosomes. Default is TRUE. Optional.

`annotation` A connection or a character string giving the name of the Bioconductor library the user would like to use containing the genome wide annotation. Default is "org.Hs.eg.db". Optional.

`TxDb` A character string giving the name of the Bioconductor library the user would like to use that will expose the annotation database as a TxDb object. Default is "TxDb.Hsapiens.UCSC.hg19.knownGene". Optional.

`offsets_value` The minimum expression value needed to call an event an outlier after normalizing event expression with gene expression. Default is 0.00001 Optional.

`correction_setting` Option to designate how to correct significance. The available options are: "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", and "none". Default is "fdr". Optional.

`p_value` Set the alpha value for the significance threshold. Default is 0.05. Optional.

Value

A list containing the below data.

- FisherAnalyses: Data Frame of junction events containing the number of under/over-expressed outliers in the tumor group (Num_UE_Outliers/Num_OE_Outliers), the Fisher p-value for under/over-expressed events (FisherP1/FisherP2), and a ranking of the under/over expressed events (UE_Rank/OE_Rank). This function will also output tab sepaerated text files and an R data file with the following:
- ASE.type: significant junction events labeled by type (skipping, insertion, or deletion)
- geneAnnotations: object containing gene names corresponding to each junction region
- junc.Outliers: list containing the logical matrices TumorOverExpression and TumorUnderExpression. "True" indicates an over-expressed event in TumorOverExpression, or an under-expressed event in TumorUnderExpression.
- junc.RPM: junction counts in reads per million following a division of the junction counts input by the total rawcounts for each sample
- junc.RPM.norm: junction counts normalized by each event's total gene expression value
- gene_expr: gene expression values for each junction event
- splice_burden: matrix containing the number of Fisher-P significant over-expressed, under-expressed, and total number of outliers per sample
- NORM.gene_expr.norm: Median of junction data normalized by gene expression for normal samples only (Used for Junction Plotting)
- pheno: Phenotypes of Samples (Tumor or Normal)
- pvalues: Junction Fisher P-values

References

Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and Harvard. doi:10.7908/C11G0KM9

Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015 Jan 29;517(7536):576-82. doi: 10.1038/nature14129. PMID: 25631445; PMCID: PMC4311405.

Guo T, Sakai A, Afsari B, Considine M, Danilova L, Favorov AV, Yegnasubramanian S, Kelley DZ, Flam E, Ha PK, Khan Z, Wheelan SJ, Gutkind JS, Fertig EJ, Gaykalova DA, Califano J. A Novel Functional Splice Variant of AKT3 Defined by Analysis of Alternative Splice Expression in HPV-Positive Oropharyngeal Cancers. Cancer Res. 2017 Oct 1;77(19):5248-5258. doi: 10.1158/0008-5472.CAN-16-3106. Epub 2017 Jul 21. PMID: 28733453; PMCID: PMC6042297.

Liu C, Guo T, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Califano JA. A novel splice variant of LOXL2 promotes progression of human papillomavirus-negative head and neck squamous cell carcinoma. Cancer. 2020 Feb 15;126(4):737-748. doi: 10.1002/cncr.32610. Epub 2019 Nov 13. PMID: 31721164.

Liu C, Guo T, Xu G, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Khan Z, Fisch KM, Califano J. Characterization of Alternative Splicing Events in HPV-Negative Head and Neck Squamous Cell Carcinoma Identifies an Oncogenic DOCK5 Variant. Clin Cancer Res. 2018 Oct 15;24(20):5123-5132. doi: 10.1158/1078-0432.CCR-18-0752. Epub 2018 Jun 26. PMID: 29945995; PMCID: PMC6440699.

M. F. Ochs, J. E. Farrar, M. Considine, Y. Wei, S. Meshinchi, and R. J. Arceci. Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. IEEE/ACM Trans Comput Biol Bioinform, 11: 520-32, 2014. PMCID: PMC4156935

Examples

```
junction <- system.file("extdata", "TCGA_HNSC_junctions.txt.gz", package = "OutSplice")
gene_expr <- system.file("extdata", "TCGA_HNSC_genes_normalized.txt.gz", package = "OutSplice")
rawcounts <- system.file("extdata", "Total_Rawcounts.txt", package = "OutSplice")
output_file_prefix <- "TCGA_OutSplice_Example"
dir <- paste0(tempdir(), "/")
results_TCGA <- outspliceTCGA(junction, gene_expr, rawcounts, saveOutput = TRUE, output_file_prefix, dir, filterSe
message("Output is located at: ", dir)
```

`plotJunctionData` *Create Plots of Junction Expression*

Description

Create Bar and Waterfall plots of raw junction expression, overall gene expression, and junction expression normalized by gene expression for splicing events found by OutSplice.

Usage

```
plotJunctionData(
  data_file,
  NUMBER = 1,
  junctions = NULL,
  tail = NULL,
  p_value = 0.05,
  GENE = FALSE,
  SYMBOL = NULL,
  makepdf = FALSE,
  pdffile = NULL,
  tumcol = "red",
  normcol = "blue"
)
```

Arguments

data_file	An R data file containing OutSplice Output.
NUMBER	An integer indicating the number of junctions to plot. This can be top number of junctions (over or under expressed), or can be specific junctions in a list. Default is 1.
junctions	A vector of user specified junctions that should be plotted. Default is NULL.
tail	A character string indicating either "RIGHT" to plot the top over expressed junctions, or "LEFT" to plot the top under expressed junctions. Default is NULL.
p_value	Set the alpha value for the significance threshold.
GENE	A boolean indicating whether to plot junctions by a specific gene. TRUE means you will pick all the junctions mapping to a certain gene. FALSE means you do not pick based on the gene. Default is NULL.
SYMBOL	The HGNC gene symbol of the gene to be graphed. Default is NULL.
makepdf	A boolean specifying whether or not to save plots to a PDF. Default is FALSE.
pdffile	A character string giving the file path to the desired pdf. Default is NULL.
tumcol	A character string defining the color of the tumor samples in the plots. Default is red.
normcol	A character string defining the color of the normal samples in the plots. Default is blue.

Value

NULL. Displays or saves a pdf containing waterfall plots of junction expression.

References

Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and Harvard. doi:10.7908/C11G0KM9

Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015 Jan 29;517(7536):576-82. doi: 10.1038/nature14129. PMID: 25631445; PMCID: PMC4311405.

Guo T, Sakai A, Afsari B, Considine M, Danilova L, Favorov AV, Yegnasubramanian S, Kelley DZ, Flam E, Ha PK, Khan Z, Wheelan SJ, Gutkind JS, Fertig EJ, Gaykalova DA, Califano J. A Novel Functional Splice Variant of AKT3 Defined by Analysis of Alternative Splice Expression in HPV-Positive Oropharyngeal Cancers. Cancer Res. 2017 Oct 1;77(19):5248-5258. doi: 10.1158/0008-5472.CAN-16-3106. Epub 2017 Jul 21. PMID: 28733453; PMCID: PMC6042297.

Liu C, Guo T, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Califano JA. A novel splice variant of LOXL2 promotes progression of human papillomavirus-negative head and neck squamous cell carcinoma. Cancer. 2020 Feb 15;126(4):737-748. doi: 10.1002/cncr.32610. Epub 2019 Nov 13. PMID: 31721164.

Liu C, Guo T, Xu G, Sakai A, Ren S, Fukusumi T, Ando M, Sadat S, Saito Y, Khan Z, Fisch KM, Califano J. Characterization of Alternative Splicing Events in HPV-Negative Head and Neck Squamous Cell Carcinoma Identifies an Oncogenic DOCK5 Variant. Clin Cancer Res. 2018 Oct 15;24(20):5123-5132. doi: 10.1158/1078-0432.CCR-18-0752. Epub 2018 Jun 26. PMID: 29945995; PMCID: PMC6440699.

M. F. Ochs, J. E. Farrar, M. Considine, Y. Wei, S. Meshinchi, and R. J. Arceci. Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. IEEE/ACM Trans Comput Biol Bioinform, 11: 520-32, 2014. PMCID: PMC4156935

Examples

```
data_file <- system.file("extdata", "OutSplice_Example_2023-01-06.RDa", package = "OutSplice")
ecm1_junc <- "chr1:150483674-150483933"
pdf <- "ecm1_expression.pdf"
pdf_output <- paste0(tempdir(), "/", pdf)
plotJunctionData(data_file, NUMBER = 1, junctions = ecm1_junc, tail = NULL, p_value = 0.05, GENE = FALSE, SYMBOL = N
message("Output is located at: ", pdf_output)
```

Index

OutSplice, 2
outspliceAnalysis, *2*, 3
outspliceTCGA, *2*, 6

plotJunctionData, *2*, 8