Package 'SplicingFactory'

April 12, 2022

Type Package

Title Splicing Diversity Analysis for Transcriptome Data

biocViews Transcriptomics, RNASeq, DifferentialSplicing, AlternativeSplicing, TranscriptomeVariant

Version 1.2.0

Description The SplicingFactory R package uses transcript-level expression values to analyze splicing diversity based on various statistical measures, like Shannon entropy or the Gini index. These measures can quantify transcript isoform diversity within samples or between conditions. Additionally, the package analyzes the isoform diversity data, looking for significant changes between conditions.

RoxygenNote 7.1.1

Imports SummarizedExperiment, methods, stats

Suggests testthat, knitr, rmarkdown, ggplot2, tidyr

URL https://github.com/SU-CompBio/SplicingFactory

BugReports https://github.com/SU-CompBio/SplicingFactory/issues

Depends R (>= 4.1)

License GPL-3 + file LICENSE

Encoding UTF-8

LazyData true

VignetteBuilder knitr

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

git_branch RELEASE_3_14

git_last_commit 906d50d

git_last_commit_date 2021-10-26

Date/Publication 2022-04-12

Author Peter A. Szikora [aut],
Tamas Por [aut],
Endre Sebestyen [aut, cre] (<https://orcid.org/0000-0001-5470-2161>)

Maintainer Endre Sebestyen <endre.sebestyen@gmail.com>

R topics documented:

`calculate_difference` *Calculate splicing diversity changes between two conditions.*

Description

Calculate splicing diversity changes between two conditions.

Usage

```
calculate_difference(
  x,
  samples,
  control,
  method = "mean",
  test = "wilcoxon",
  randomizations = 100,
  pcorr = "BH",
  assayno = 1,
  verbose = FALSE,
  ...
)
```

Arguments

`x`	A `SummarizedExperiment` with splicing diversity values for each gene in each sample or a `data.frame` with gene names in the first column and splicing diversity values for each sample in additional columns.
`samples`	A vector of length one, specifying the column name of the `colData` annotation column from the `SummarizedExperiment` object, that should be used as the category column or a character vector with an equal length to the number of columns in the input dataset, specifying the category of each sample in the case of a `data.frame` input.

control	Name of the control sample category, defined in the samples vector, e.g. control = 'Normal' or control = 'WT'.
method	Method to use for calculating the average splicing diversity value in a condition. Can be 'mean' or 'median'.
test	Method to use for p-value calculation: use 'wilcoxon' for Wilcoxon rank sum test or 'shuffle' for a label shuffling test.
randomizations	Number of random shuffles, used for the label shuffling test (default = 100).
pcorr	P-value correction method applied to the Wilcoxon rank sum test or label shuffling test results, as defined in the p.adjust function.
assayno	An integer value. In case of multiple assays in a SummarizedExperiment input, the argument specifies the assay number to use for difference calculations.
verbose	If TRUE, the function will print additional diagnostic messages.
...	Further arguments to be passed on for other methods.

Details

The function calculates diversity changes between two sample conditions. It uses the output of the diversity calculation function, which is a SummarizedExperiment object of splicing diversity values. Additionally, it can use a data.frame as input, where the first column contains gene names, and all additional columns contain splicing diversity values for each sample. A vector of sample conditions also serves as input, used for aggregating the samples by condition.

It calculates the mean or median of the splicing diversity data per sample condition, the difference of these values and the log2 fold change of the two conditions. Furthermore, the user can select a statistical method to calculate the significance of the changes. The p-values and adjusted p-values are calculated using a Wilcoxon sum rank test or label shuffling test.

The function will exclude genes of low sample size from the significance calculation, depending on which statistical test is applied.

Value

A data.frame with the mean or median values of splicing diversity across sample categories and all samples, log2(fold change) of the two different conditions, raw and corrected p-values.

Examples

```
# data.frame with splicing diversity values
x <- data.frame(Genes = letters[seq_len(10)], matrix(runif(80), ncol = 8))

# sample categories
samples <- c(rep('Healthy', 4), rep('Pathogenic', 4))

# To calculate the difference of splicing diversity changes between the
# 'Healthy' and 'Pathogenic' condition together with the significance values,
# using mean and Wilcoxon rank sum test, use:
calculate_difference(x, samples, control = 'Healthy', method = 'mean', test = 'wilcoxon')
```

`calculate_diversity` *Main function for calculating splicing diversity*

Description

Main function for calculating splicing diversity

Usage

```
calculate_diversity(
  x,
  genes = NULL,
  method = "laplace",
  norm = TRUE,
  tpm = FALSE,
  assayno = 1,
  verbose = FALSE
)
```

Arguments

`x`	A numeric `matrix`, `data.frame`, `tximport` list, `DGEList`, `SummarizedExperiment` or `ExpressionSet`.
`genes`	Character vector with equal length to the number of rows of the input dataset with transcript-level expression values. The values in `x` are grouped into genes based on this vector.
`method`	Method to use for splicing diversity calculation, including naive entropy (`naive`), Laplace entropy (`laplace`), Gini index (`gini`), Simpson index (`simpson`) and inverse Simpson index (`invsimpson`). The default method is Laplace entropy.
`norm`	If `TRUE`, the entropy values are normalized to the number of transcripts for each gene. The normalized entropy values are always between 0 and 1. If `FALSE`, genes cannot be compared to each other, due to possibly different maximum entropy values.
`tpm`	In the case of a tximport list, TPM values or raw read counts can serve as an input. If `TRUE`, TPM values will be used, if `FALSE`, read counts will be used.
`assayno`	An integer value. In case of multiple assays in a `SummarizedExperiment` input, the argument specifies the assay number to use for diversity calculations.
`verbose`	If `TRUE`, the function will print additional diagnostic messages, besides the warnings and errors.

Details

The function is intended to process transcript-level expression data from RNA-seq or similar datasets.

Given a N x M matrix or similar data structure, where the N rows are transcripts and the M columns are samples, and a vector of gene ids, used for aggregating the transcript level data, the function

calculates transcript diversity values for each gene in each sample. These diversity values can be used to investigate the dominance of a specific transcript for a gene, the diversity of transcripts in a gene, and analyze changes in diversity.

There are a number of diversity values implemented in the package. These include the following:

- Naive entropy: Shannon entropy using the transcript frequencies as probabilities. 0 entropy means a single dominant transcript, higher values mean a more diverse set of transcripts for a gene.
- Laplace entropy: Shannon entropy where the transcript frequencies are replaced by a Bayesian estimate, using Laplace's prior.
- Gini index: a measure of statistical dispersion originally used in economy. This measurement ranges from 0 (complete equality) to 1 (complete inequality). A value of 1 (complete inequality) means a single dominant transcript.
- Simpson index: a measure of diversity, characterizing the number of different species (transcripts of a gene) in a dataset. Originally, this measurement calculates the probability that randomly selected individuals belong to different species. Simpson index ranges between 0 and 1; the higher the value, the higher the diversity.
- Inverse Simpson index: Similar concept as the Simpson index, although a higher inverse-Simpson index means greater diversity. It ranges between 1 and the total number of transcripts for a gene.

The function can calculate the gene level diversity index using any kind of expression measure, including raw read counts, FPKM, RPKM or TPM values, although results may vary.

Value

Gene-level splicing diversity values in a `SummarizedExperiment` object.

Examples

```
# matrix with RNA-seq read counts
x <- matrix(rpois(60, 10), ncol = 6)
colnames(x) <- paste0("Sample", 1:6)

# gene names used for grouping the transcript level data
gene <- c(rep("Gene1", 3), rep("Gene2", 2), rep("Gene3", 3), rep("Gene4", 2))

# calculating normalized Laplace entropy
result <- calculate_diversity(x, gene, method = "laplace", norm = TRUE)
```

`calculate_entropy` *Calculate entropy for a vector of transcript-level expression values of one gene.*

Description

Calculate entropy for a vector of transcript-level expression values of one gene.

Usage

```
calculate_entropy(x, norm = TRUE, pseudocount = 0)
```

Arguments

x	Vector of expression values.
norm	If TRUE, the entropy values are normalized to the number of transcripts for each gene. The normalized entropy values are always between 0 and 1. If FALSE, genes cannot be compared to each other, due to possibly different maximum entropy values.
pseudocount	Pseudocount added to each transcript expression value. Default is 0, while Laplace entropy uses a pseudocount of 1.

Details

The function calculates an entropy value as part of different diversity calculations. Given a vector of transcript-level expression values of a gene, this function characterizes the diversity of splicing isoforms for a gene. If there only a single transcript, the diversity value will be NaN, as it cannot be calculated. If the expression of the given gene is 0, the diversity value will be NA.

Value

A single gene-level entropy value.

Examples

```
# read counts for the transcripts of a single gene with 5 transcripts
x <- rnbinom(5, size = 10, prob = 0.4)
# calculate non-normalized naive entropy value
entropy <- calculate_entropy(x, norm = FALSE)
# calculate Laplace-entropy, also normalized for transcript number
# (the default)
norm_laplace_entropy <- calculate_entropy(x, pseudocount = 1)
```

---

calculate_fc *Calculate splicing diversity changes between two conditions.*

Description

Calculate splicing diversity changes between two conditions.

Usage

```
calculate_fc(x, samples, control, method = "mean")
```

Arguments

x	A matrix with the splicing diversity values.
samples	Character vector with an equal length to the number of columns in the input dataset, specifying the category of each sample.
control	Name of the control sample category, defined in the samples vector, e.g. control = 'Normal' or control = 'WT'.
method	Method to use for calculating the average splicing diversity value in a condition. Can be 'mean' or 'median'.

Details

The function uses a matrix of splicing diversity values in order to calculate mean or median differences and log2 fold changes between two conditions.

Value

A data.frame with mean or median value of splicing diversity across sample categories, the difference between these values and the log2 fold change values.

---

calculate_gini *Calculate Gini coefficient for a vector of transcript-level expression values of one gene.*

Description

Calculate Gini coefficient for a vector of transcript-level expression values of one gene.

Usage

```
calculate_gini(x)
```

Arguments

x	Vector of expression values.

Details

The function calculates a Gini coefficient as part of different diversity calculations. Given a vector of transcript-level expression values of a gene, this function characterize the diversity of splicing isoforms for a gene. If there only one single transcript, the resulted index will be NaN, as diversity cannot be calculated. If the expression of the given gene is 0, the diversity index will be NA.

Value

A single gene-level Gini coefficient.

Examples

```
# read counts for the transcripts of a single gene with 5 transcripts
x <- rnbinom(5, size = 10, prob = 0.4)
# calculate Gini index
gini <- calculate_gini(x)
```

`calculate_inverse_simpson` *#' Calculate inverse Simpson index for a vector of transcript-level expression values of one gene.*

Description

#' Calculate inverse Simpson index for a vector of transcript-level expression values of one gene.

Usage

```
calculate_inverse_simpson(x)
```

Arguments

x	Vector of expression values.

Details

The function calculates an inverse Simpson index as part of different diversity calculations. Given a vector of transcript-level expression values of a gene, this function characterize the diversity of splicing isoforms for a gene. If there only one single transcript, the resulted index will be NaN, as diversity cannot be calculated. If the expression of the given gene is 0, the diversity index will be NA.

Value

A single gene-level inverse Simpson index.

Examples

```
# read counts for the transcripts of a single gene with 5 transcripts
x <- rnbinom(5, size = 10, prob = 0.4)
# calculate inverse Simpson index
invsimpson <- calculate_inverse_simpson(x)
```

`calculate_method` *Calculate diversity values for a matrix of transcripts.*

Description

Calculate diversity values for a matrix of transcripts.

Usage

```
calculate_method(x, genes, method, norm = TRUE, verbose = FALSE)
```

Arguments

`x`	An input `matrix`, or `data.frame` containing transcript-level expression values.
`genes`	Character vector with equal length to the number of rows of the input dataset with transcript-level expression values. The values in `x` are grouped into genes based on this vector.
`method`	Method to use for splicing diversity calculation, including naive entropy (`naive`), Laplace entropy (`laplace`), Gini index (`gini`), Simpson index (`simpson`) and inverse Simpson index (`invsimpson`). The default method is Laplace entropy.
`norm`	If `TRUE`, the entropy values are normalized to the number of transcripts for each gene. The normalized entropy values are always between 0 and 1. If `FALSE`, genes cannot be compared to each other, due to possibly different maximum entropy values.
`verbose`	If `TRUE`, the function will print additional diagnostic messages, besides the warnings and errors.

Details

The function calculates diversity values on a matrix of transcript-level expression values, aggregated by the genes defined in the `genes` parameter.

Value

Gene-level splicing diversity values in a `data.frame`, where each row belongs to a gene and each column belongs to a sample from the data, in addition to the first column, containing gene names, given in the ‘genes‘ parameter.

`calculate_simpson` *#' Calculate Simpson index for a vector of transcript-level expression values of one gene.*

Description

#' Calculate Simpson index for a vector of transcript-level expression values of one gene.

Usage

```
calculate_simpson(x)
```

Arguments

x	Vector of expression values.

Details

The function calculates a Simpson index as part of different diversity calculations. Given a vector of transcript-level expression values of a gene, this function characterize the diversity of splicing isoforms for a gene. If there only one single transcript, the resulted index will be NaN, as diversity cannot be calculated. If the expression of the given gene is 0, the diversity index will be NA.

Value

A single gene-level Simpson index.

Examples

```
# read counts for the transcripts of a single gene with 5 transcripts
x <- rnbinom(5, size = 10, prob = 0.4)
# calculate Simpson index
simpson <- calculate_simpson(x)
```

`label_shuffling` *Calculate p-values using label shuffling.*

Description

Calculate p-values using label shuffling.

Usage

```
label_shuffling(
  x,
  samples,
  control,
  method,
  randomizations = 100,
  pcorr = "BH"
)
```

Arguments

x	A matrix with the splicing diversity values.
samples	Character vector with an equal length to the number of columns in the input dataset, specifying the category of each sample.
control	Name of the control sample category, defined in the samples vector, e.g. control = 'Normal' or control = 'WT'.
method	Method to use for calculating the average splicing diversity value in a condition. Can be 'mean' or 'median'.
randomizations	The number of random shuffles.
pcorr	P-value correction method applied to the results, as defined in the p.adjust function.

Value

Raw and corrected p-values.

---

tcga_brca_luma_dataset

TCGA Luminal A breast cancer dataset

Description

Data from The Cancer Genome Atlas, downloaded on 08th September, 2020. It contains transcript level read counts of 20 patients with Luminal A type breast cancer (primary tumor and solid normal samples).

Usage

```
data(tcga_brca_luma_dataset)
```

Format

A data frame with 996 rows and 41 columns. The first column contains gene names, all additional columns contain RNA-sequencing read counts for samples.

Source

TCGA Legacy

References

The Cancer Genome Atlas Network (2012) Nature 490, 61–70 (doi:10.1038/nature11412)

`wilcoxon` *Calculate p-values using Wilcoxon rank sum test.*

Description

Calculate p-values using Wilcoxon rank sum test.

Usage

```
wilcoxon(x, samples, pcorr = "BH", paired = FALSE, exact = FALSE)
```

Arguments

`x`	A `matrix` with the splicing diversity values.
`samples`	Character vector with an equal length to the number of columns in the input dataset, specifying the category of each sample.
`pcorr`	P-value correction method applied to the results, as defined in the `p.adjust` function.
`paired`	If `TRUE`, the Wilcox-test will be paired, and therefore it will be a signed rank test instead of the rank sum test.
`exact`	If `TRUE`, an exact p-value will be computed.

Value

Raw and corrected p-values in a matrix.

Index