Package ‘HTSFilter’

March 22, 2017

Type Package

Title Filter replicated high-throughput transcriptome sequencing data

Version 1.14.1

Date 2016-11-23

Author Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic

Maintainer Andrea Rau <andrea.rau@jouy.inra.fr>

Depends R (>= 3.3.0)

Imports edgeR (>= 3.9.14), DESeq2 (>= 1.10.1), DESeq (>= 1.22.1), BiocParallel (>= 1.4.3), Biobase, utils, stats, grDevices, graphics, methods

Suggests EDASeq (>= 2.1.4), BiocStyle, testthat

Description This package implements a filtering procedure for replicated transcriptome sequencing data based on a global Jaccard similarity index in order to identify genes with low, constant levels of expression across one or more experimental conditions.

License Artistic-2.0

LazyLoad yes

biocViews Sequencing, RNASeq, Preprocessing, DifferentialExpression, GeneExpression, Normalization

RoxygenNote 5.0.1

NeedsCompilation no

R topics documented:

HTSFilter-package .......................... 2
HTSBasicFilter .................................. 3
HTSFilter ........................................... 6
HTSFilter-deprecated ................................ 9
normalizeData ..................................... 11
sultan .............................................. 12

Index ........................................ 14
HTSFilter-package  
Filter replicated high-throughput transcriptome sequencing data

Description
This package implements a filtering procedure for replicated transcriptome sequencing data based on a global Jaccard similarity index in order to identify genes with low, constant levels of expression across one or more experimental conditions.

Details

Package: HTSFilter
Type: Package
Version: 1.13.2
Date: 2016-10-13
License: Artistic-2.0
LazyLoad: yes

Author(s)
Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic
Maintainer: Andrea Rau <andrea.rau@jouy.inra.fr>

References

Examples

```r
library(Biobase)
data("sultan")
conds <- pData(sultan)$cell.line

########################################################################
## Matrix or data.frame
########################################################################
filter <- HTSFilter(exprs(sultan), conds, s.len=25, plot=FALSE)

########################################################################
## DGEExact
########################################################################
```

## Matrix or data.frame

```r
filter <- HTSFilter(exprs(sultan), conds, s.len=25, plot=FALSE)
```
library(edgeR)
dge <- DGEList(counts=exprs(sultan), group=conds)
dge <- calcNormFactors(dge)
dge <- estimateCommonDisp(dge)
dge <- estimateTagwiseDisp(dge)
et <- exactTest(dge)
et <- HTSFilter(et, DGEList=dge, s.len=25, plot=FALSE)$filteredData
## topTags(et)

########################################################################
## DESeq2
########################################################################

library(DESeq2)
.dds <- DESeqDataSetFromMatrix(countData = exprs(sultan),
    colData = data.frame(cell.line = conds),
    design = ~ cell.line)
## Not run:
## .dds <- DESeq(dds)
## filter <- HTSFilter(dds, s.len=25, plot=FALSE)$filteredData
## class(filter)
## res <- results(filter, independentFiltering=FALSE)

## HTSBasicFilter
Implement basic filters for transcriptome sequencing data.

Description
Implement a variety of basic filters for transcriptome sequencing data.

Usage
HTSBasicFilter(x, ...)

## S4 method for signature 'matrix'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq", "none"))

## S4 method for signature 'data.frame'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq", "none"))

## S4 method for signature 'DGEList'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq",
    "pseudo.counts", "none"))

## S4 method for signature 'DGEExact'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq",
    "none"))
## S4 method for signature 'DGEGLM'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq", "none"))

## S4 method for signature 'DGELRT'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("TMM", "DESeq", "none"))

## S4 method for signature 'DESeqDataSet'
HTSBasicFilter(x, method, cutoff.type = "value",
    cutoff = 10, length = NA, normalization = c("DESeq", "TMM", "none"),
    pAdjustMethod = "BH")

### Arguments

**x**  
A numeric matrix or data.frame representing the counts of dimension \((g \times n)\), for \(g\) genes in \(n\) samples, a DGEList object, a DGEExact object, a DGEGLM object, a DGELRT object, or a DESeqDataSet object.

**...**  
Additional optional arguments

**method**  
Basic filtering method to be used: "mean", "sum", "rpkm", "variance", "cpm", "max", "cpm.mean", "cpm.sum", "cpm.variance", "cpm.max", "rpkm.mean", "rpkm.sum", "rpkm.variance", or "rpkm.max"

**cutoff.type**  
Type of cutoff to be used: a numeric value indicating the number of samples to be used for filtering (when method = "cpm" or "rpkm"), or one of "value", "number", or "quantile"

**cutoff**  
Cutoff to be used for chosen filter

**length**  
Optional vector of length \(n\) containing the lengths of each gene in \(x\); optional except in the case of method = "rpkm"

**normalization**  
Normalization method to be used to correct for differences in library sizes, with choices "TMM" (Trimmed Mean of M-values), "DESeq" (normalization method proposed in the DESeq package), "pseudo.counts" (pseudo-counts obtained via quantile-quantile normalization in the edgeR package, only available for objects of class DGEList and DGEExact), and "none" (to be used only if user is certain no normalization is required, or if data have already been pre-normalized by an alternative method)

**pAdjustMethod**  
The method used to adjust p-values, see ?p.adjust

### Details

This function implements a basic filter for high-throughput sequencing data for a variety of filter types: mean, sum, RPKM, variance, CPM, maximum, mean CPM values, the sum of CPM values, the variance of CPM values, maximum CPM value, mean RPKM values, the sum of RPKM values, the variance of RPKM values, or the maximum RPKM value. The filtering criteria used may be for a given cutoff value, a number of genes, or a given quantile value.

### Value

- filteredData An object of the same class as \(x\) containing the data that passed the filter
• on A binary vector of length \( g \), where 1 indicates a gene with normalized expression greater than the optimal filtering threshold \( s_{\text{optimal}} \) in at least one sample (irrespective of condition labels), and 0 indicates a gene with normalized expression less than or equal to the optimal filtering threshold in all samples
• normFactor A vector of length \( n \) giving the estimated library sizes estimated by the normalization method specified in normalization
• removedData A matrix containing the filtered data
• filterCrit A vector or matrix containing the criteria used to perform filtering

Author(s)
Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic

References

Examples
```r
library(Biobase)
data("sultan")
conds <- pData(sultan)$cell.line

########################################################################
## Matrix or data.frame
########################################################################
## Filter genes with total (sum) normalized gene counts < 10
filter <- HTSBasicFilter(exprs(sultan), method="sum", cutoff.type="value",
cutoff = 10)

########################################################################
## DGEExact
########################################################################
library(edgeR)
## Filter genes with CPM values less than 100 in more than 2 samples
dge <- DGEList(counts=exprs(sultan), group=conds)
dge <- calcNormFactors(dge)
filter <- HTSBasicFilter(dge, method="cpm", cutoff.type=2, cutoff=100)

########################################################################
## DESeq2
########################################################################
library(DESeq2)
.dds <- DESeqDataSetFromMatrix(countData = exprs(sultan),
   colData = data.frame(cell.line = conds),
   design = ~ cell.line)
```
## Not run: Filter genes with mean normalized gene counts < 40% quantile
##
## dds <- DESeq(dds)
## filter <- HTSBasicFilter(dds, method="mean", cutoff.type="quantile",
## cutoff = 0.4)
## res <- results(filter, independentFiltering=FALSE)

### HTSFilter

**Calculate data-based filtering threshold for replicated transcriptome sequencing data.**

**Description**

Calculate a data-based filtering threshold for replicated transcriptome sequencing data through the pairwise Jaccard similarity index between pairs of replicates within each experimental condition.

**Usage**

```r
HTSFilter(x, ...)  
```

## S4 method for signature 'matrix'

```r
HTSFilter(x, conds, s.min = 1, s.max = 200,  
           s.len = 100, loess.span = 0.3, normalization = c("TMM", "DESeq", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```

## S4 method for signature 'data.frame'

```r
HTSFilter(x, conds, s.min = 1, s.max = 200,  
           s.len = 100, loess.span = 0.3, normalization = c("TMM", "DESeq", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```

## S4 method for signature 'DGEList'

```r
HTSFilter(x, s.min = 1, s.max = 200, s.len = 100,  
           loess.span = 0.3, normalization = c("TMM", "DESeq", "pseudo.counts", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```

## S4 method for signature 'DGEExact'

```r
HTSFilter(x, DGEList, s.min = 1, s.max = 200,  
           s.len = 100, loess.span = 0.3, normalization = c("TMM", "DESeq", "pseudo.counts", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```

## S4 method for signature 'DGEGLM'

```r
HTSFilter(x, DGEGLM, s.min = 1, s.max = 200,  
           s.len = 100, loess.span = 0.3, normalization = c("TMM", "DESeq", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```

## S4 method for signature 'DGEGLRT'

```r
HTSFilter(x, DGEGLRT, s.min = 1, s.max = 200,  
           s.len = 100, loess.span = 0.3, normalization = c("TMM", "DESeq", "none"),  
           plot = TRUE, plot.name = NA, parallel = FALSE,  
           BPPARAM = bpparam())
```
### HTSFilter

"none"), plot = TRUE, plot.name = NA, parallel = FALSE, BPPARAM = bpparam())

```r
## S4 method for signature 'DESeqDataSet'
HTSFilter(x, s.min = 1, s.max = 200, s.len = 100,
  loess.span = 0.3, normalization = c("DESeq", "TMM", "none"),
  plot = TRUE, plot.name = NA, pAdjustMethod = "BH", parallel = FALSE,
  BPPARAM = bpparam()))
```

#### Arguments

- **x**: A numeric matrix or data.frame representing the counts of dimension \((g \times n)\), for \(g\) genes in \(n\) samples, a DGEList object, a DGEEExact object, a DGELM object, a DGELRT object, or a DESeqDataSet object.
- **...**: Additional optional arguments
- **conds**: Vector of length \(n\) identifying the experimental condition of each of the \(n\) samples; required when sQuote(x) is a numeric matrix.
- **s.min**: Minimum value of filtering threshold to be considered, with default value equal to 1.
- **s.max**: Maximum value of filtering threshold to be considered, with default value equal to 200.
- **s.len**: Length of sequence of filtering thresholds to be considered (from \(s.min\) to \(s.max\)) for the calculation of the global similarity index.
- **loess.span**: Span of the loess curve to be fitted to the filtering thresholds and corresponding global similarity indices, with default value equal to 0.3.
- **normalization**: Normalization method to be used to correct for differences in library sizes, with choices "TMM" (Trimmed Mean of M-values), "DESeq" (normalization method proposed in the DESeq package), "pseudo.counts" (pseudo-counts obtained via quantile-quantile normalization in the edgeR package, only available for objects of class DGEList and DGEEexact), and "none" (to be used only if user is certain no normalization is required, or if data have already been pre-normalized by an alternative method).
- **plot**: If "TRUE", produce a plot of the calculated global similarity indices against the filtering threshold with superimposed loess curve.
- **plot.name**: If plot = "TRUE", the name of the PDF file to be saved to the current working directory. If plot.name = NA, the plot is drawn in the current window.
- **parallel**: If FALSE, no parallelization. If TRUE, parallel execution using BiocParallel (see next argument BPPARAM). A note on running in parallel using BiocParallel: it may be advantageous to remove large, unneeded objects from the current R environment before calling the function, as it is possible that R’s internal garbage collection will copy these files while running on worker nodes.
- **BPPARAM**: Optional parameter object passed internally to bplapply when parallel=TRUE. If not specified, the parameters last registered with register will be used.
- **DGEList**: Object of class DGELIST, to be used when filtering objects of class DGEEExact.
- **DGELM**: Object of class DGELM, to be used when filtering objects of class DGELRT.
- **pAdjustMethod**: The method used to adjust p-values, see ?p.adjust.
HTSFilter

Details

The Jaccard similarity index, which measures the overlap of two sets, is calculated as follows. Given two binary vectors, each of length \( n \), we define the following values:

- \( a \) = the number of attributes with a value of 1 in both vectors
- \( b \) = the number of attributes with a value of 1 in the first vector and 0 in the second
- \( c \) = the number of attributes with a value of 0 in the first vector and 1 in the second
- \( d \) = the number of attributes with a value of 0 in both vectors

We note that all attributes fall into one of these four quantities, so \( a + b + c + d = n \). Given these quantities, we may calculate the Jaccard similarity index between the two vectors as follows:

\[
J = \frac{a}{a + b + c}.
\]

Value

- filteredData An object of the same class as \( x \) containing the data that passed the filter
- on A binary vector of length \( g \), where 1 indicates a gene with normalized expression greater than the optimal filtering threshold \( s_{\text{optimal}} \) in at least one sample (irrespective of condition labels), and 0 indicates a gene with normalized expression less than or equal to the optimal filtering threshold in all samples
- \( s \) The optimal filtering threshold as identified by the global similarity index
- indexValues A matrix of dimension \( (s_{\text{len}} \times 2) \) giving the tested filtering thresholds and the corresponding global similarity indices. Note that the threshold values are equally spaced on the log scale, and thus unequally spaced on the count scale (i.e., we test more threshold values at very low levels of expression, and fewer at very high levels of expression).
- normFactor A vector of length \( n \) giving the estimated library sizes estimated by the normalization method specified in normalization
- removedData A matrix containing the filtered data

Author(s)

Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic

References


Examples

library(Biobase)
data("sultan")
conds <- pData(sultan)$cell.line

########################################################################
## Matrix or data.frame

### Matrix or data.frame
library(edgeR)
dge <- DGEList(counts=exprs(sultan), group=conds)
dge <- calcNormFactors(dge)
dge <- estimateCommonDisp(dge)
et <- exactTest(dge)
et <- HTSFilter(et, DGEList=dge, s.len=25, plot=FALSE)$filteredData

## topTags(et)

library(DESeq2)
dds <- DESeqDataSetFromMatrix(countData = exprs(sultan),
   colData = data.frame(cell.line = conds),
   design = ~ cell.line)

## Not run:
## dds <- DESeq(dds)
## filter <- HTSFilter(dds, s.len=25, plot=FALSE)$filteredData
## class(filter)
## res <- results(filter, independentFiltering=FALSE)

HTSFilter-deprecated  Deprecated functions in package ‘HTSFilter’

Description

These functions are provided for compatibility with older versions of ‘HTSFilter’ only, and will be
defunct at the next release.

Usage

## S4 method for signature ‘CountDataSet’
HTSFilter(x, conds = NA, s.min = 1, s.max = 200,
   s.len = 100, loess.span = 0.3, normalization = c("DESeq", "TMM",
   "none"), plot = TRUE, plot.name = NA, parallel = FALSE,
   BPPARAM = bpparam())

## S4 method for signature ‘CountDataSet’
HTSBasicFilter(x, method, cutoff.type = "value",
cutoff = 10, length = NA, normalization = c("DESeq", "TMM", "none"))
Arguments

x  A numeric matrix or data.frame representing the counts of dimension \((g \times n)\), for \(g\) genes in \(n\) samples, a DGEList object, a DGEExact object, a DGEGLM object, a DGELRT object, or a DESeqDataSet object.

conds  Vector of length \(n\) identifying the experimental condition of each of the \(n\) samples; required when \texttt{sQuote(x)} is a numeric matrix.

s.min  Minimum value of filtering threshold to be considered, with default value equal to 1

s.max  Maximum value of filtering threshold to be considered, with default value equal to 200

s.len  Length of sequence of filtering thresholds to be considered (from s.min to s.max) for the calculation of the global similarity index

loess.span  Span of the loess curve to be fitted to the filtering thresholds and corresponding global similarity indices, with default value equal to 0.3

normalization  Normalization method to be used to correct for differences in library sizes, with choices ”TMM” (Trimmed Mean of M-values), ”DESeq” (normalization method proposed in the DESeq package), ”pseudo.counts” (pseudo-counts obtained via quantile-quantile normalization in the edgeR package, only available for objects of class DGEList and DGEExact), and ”none” (to be used only if user is certain no normalization is required, or if data have already been pre-normalized by an alternative method)

plot  If “TRUE”, produce a plot of the calculated global similarity indices against the filtering threshold with superimposed loess curve

plot.name  If plot = “TRUE”, the name of the PDF file to be saved to the current working directory. If plot.name = NA, the plot is drawn in the current window

parallel  If FALSE, no parallelization. If TRUE, parallel execution using BiocParallel (see next argument BPPARAM). A note on running in parallel using BiocParallel: it may be advantageous to remove large, unneeded objects from the current R environment before calling the function, as it is possible that R’s internal garbage collection will copy these files while running on worker nodes.

BPPARAM  Optional parameter object passed internally to bplapply when parallel=TRUE. If not specified, the parameters last registered with register will be used.

method  Basic filtering method to be used: “mean”, “sum”, “rpkm”, “variance”, “cpm”, “max”, “cpm.mean”, “cpm.sum”, “cpm.variance”, “cpm.max”, “rpkm.mean”, “rpkm.sum”, “rpkm.variance”, or “rpkm.max”

cutoff.type  Type of cutoff to be used: a numeric value indicating the number of samples to be used for filtering (when method = “cpm” or “rpkm”), or one of “value”, “number”, or “quantile”

cutoff  Cutoff to be used for chosen filter

length  Optional vector of length \(n\) containing the lengths of each gene in \(x\); optional except in the case of method = “rpkm”

Details

The following functions are deprecated and will be made defunct. Objects of type ‘CountDataSet’ from the original ‘DESeq’ package will no longer be supported; users should make use of ‘DESeqDataSet’ objects from the ‘DESeq2’ package:

- HTSFilter.CountDataSet: HTSFilter
- HTSBasicFilter.CountDataSet: HTSBasicFilter
**normalizeData**

**Value**

- filteredData An object of the same class as `x` containing the data that passed the filter
- on A binary vector of length `g`, where 1 indicates a gene with normalized expression greater than the optimal filtering threshold `s.optimal` in at least one sample (irrespective of condition labels), and 0 indicates a gene with normalized expression less than or equal to the optimal filtering threshold in all samples
- `s` The optimal filtering threshold as identified by the global similarity index
- indexValues A matrix of dimension `(s.len x 2)` giving the tested filtering thresholds and the corresponding global similarity indices. Note that the threshold values are equally spaced on the log scale, and thus unequally spaced on the count scale (i.e., we test more threshold values at very low levels of expression, and fewer at very high levels of expression).
- normFactor A vector of length `n` giving the estimated library sizes estimated by the normalization method specified in `normalization`
- removedData A matrix containing the filtered data

**Author(s)**

Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic

**References**


---

**Description**

Normalize count-based measures of transcriptome sequencing data using the Trimmed Means of M-values (TMM) or DESeq approach.

**Usage**

`normalizeData(data, normalization)`

**Arguments**

- `data` numeric matrix representing the counts of dimension `(g x n)`, for `g` genes in `n` samples.
- `normalization` Normalization method to be used to correct for differences in library sizes, with choices “TMM” (Trimmed Mean of M-values), “DESeq” (normalization method proposed in the DESeq package), and “none”
Value

- data.norm A numeric matrix representing the normalized counts of dimension \((g \times n)\), for \(g\) genes in \(n\) samples.
- norm.factor A vector of length \(n\) giving the estimated library sizes estimated by the normalization method specified in normalization

Author(s)

Andrea Rau, Melina Gallopin, Gilles Celeux, and Florence Jaffrezic

References


Examples

```r
library(Biobase)
data("sultan")
normData <- normalizeData(exprs(sultan), norm="DESeq")
```

sultan

RNA-seq data from humans in Sultan et al. (2008)

Description

This dataset represents RNA-seq data from humans in two conditions (Ramos B cell line and HEK293T), with two biological replicates per condition. The ExpressionSet was downloaded from the ReCount online resource.

Usage

data(sultan)

Format

An ExpressionSet named sultan.eset containing the phenotype data and expression data for the Sultan et al. (2008) experiment. Phenotype data may be accessed using the pData function, and expression data may be accessed using the exprs function.

Value

Object of class ‘ExpressionSet’. Matrix of counts can be accessed after loading the ‘Biobase’ package and calling exprs(sultan)).
Source

ReCount online resource (http://bowtie-bio.sourceforge.net/recount).

References

data_blah.com


Index

*Topic datasets
  sultan, 12

*Topic methods
  HTSBasicFilter, 3
  HTSFilter, 6
  HTSFilter-deprecated, 9
  normalizeData, 11

*Topic package
  HTSFilter-package, 2

HTSBasicFilter, 3, 10
HTSBasicFilter,CountDataSet-method
  (HTSFilter-deprecated), 9
HTSBasicFilter,data.frame-method
  (HTSBasicFilter), 3
HTSBasicFilter,DESeqDataSet-method
  (HTSBasicFilter), 3
HTSBasicFilter,DGEExact-method
  (HTSBasicFilter), 3
HTSBasicFilter,DGEGlm-method
  (HTSBasicFilter), 3
HTSBasicFilter,DGEList-method
  (HTSBasicFilter), 3
HTSBasicFilter,DGEList-method
  (HTSBasicFilter), 3
HTSBasicFilter,DGELRT-method
  (HTSBasicFilter), 3
HTSBasicFilter,matrix-method
  (HTSBasicFilter), 3

HTSFilter, 6, 10
HTSFilter,CountDataSet-method
  (HTSFilter-deprecated), 9
HTSFilter,data.frame-method
  (HTSFilter), 6
HTSFilter,DESeqDataSet-method
  (HTSFilter), 6
HTSFilter,DGEExact-method (HTSFilter), 6
HTSFilter,DGEGlm-method (HTSFilter), 6
HTSFilter,DGEList-method (HTSFilter), 6
HTSFilter,DGEList-method (HTSFilter), 6
HTSFilter,DGELRT-method (HTSFilter), 6
HTSFilter,matrix-method (HTSFilter), 6
HTSFilter-deprecated, 9
HTSFilter-methods (HTSFilter), 6
HTSFilter-package, 2

normalizeData, 11