Package ‘SigFuge’

December 22, 2016

Type Package
Title SigFuge
Version 1.12.0
Date 2014-09-22
Author Patrick Kimes, Christopher Cabanski
Maintainer Patrick Kimes <patrick.kimes@gmail.com>
Description Algorithm for testing significance of clustering in RNA-seq data.
License GPL-3
Imports ggplot2, matlab, reshape, sigclust
Depends R (>= 3.1.1), GenomicRanges
Suggests org.Hs.eg.db, prebsdata, Rsamtools (>= 1.17.0),
       TxDb.Hsapiens.UCSC.hg19.knownGene, BiocStyle
biocViews Clustering, Visualization, RNASeq
NeedsCompilation no

R topics documented:

SigFuge-package ................................................. 2
geneAnnot .................................................. 2
geneDepth ................................................... 3
SFfigure ...................................................... 3
SFlabels ..................................................... 5
SFnormalize .................................................. 6
SFPval ....................................................... 7

Index 8
SigFuge-package

SigFuge

Description
Tests significance of clustering in RNA-seq data.

Details

*SFpval* computes a p-value for significance of clustering for RNA-seq data, and *SFfigure* produces accompanying figures.

Author(s)
Patrick Kimes <pkimes@live.unc.edu>

geneAnnot

CDKN2A gene locus annotation

Description
A dataset containing the annotations for the CDKN2A locus.

Usage
data(geneAnnot)

Format
A GRanges object

Source
**geneDepth**  

*Coverage matrix across CDKN2A gene locus*

**Description**  
A dataset containing read depths for 179 lung squamous cell carcinoma samples across the CDKN2A locus.

**Usage**  
```r
data(geneDepth)
```

**Format**  
A $2078 \times 179$ data.frame of read depth (coverage). Each column corresponds to a sample and each row to a base position along the CDKN2A locus. These RNA-Seq read counts are a subset from 179 lung squamous cell tumor samples sequenced as part of the Cancer Genome Atlas.

**Source**  

---

**SFfigure**  
*Plot expression as curves*

**Description**  
Function for producing various figures corresponding to the SigFuge functional data approach to studying RNA-seq data as expression curves along base positions. The primary input for the function is a read count matrix and GRanges. The default behavior is to identify clusters based on applying SFlabels to a normalized version of the data produced by SFnormalize. If specified, the function will compute a p-value for the significance of the labels by calling the SFpval function.

**Usage**  
```r
SFfigure(data, locusname, annot = c(), flip.fig = 1,
label.exon = 1, print.n = 1, data.labels = 0,
label.colors = c(), flag = 1, lplots = 2,
log10 = 1, summary.type = "median",
savestr = c(), titlestr = c(), pval = 1)
```

**Arguments**  
- `data`: a $d \times n$ matrix or data.frame of read counts at $d$ base positions for $n$ samples.
- `locusname`: a character string specifying gene or locus name to be used in figure title.
- `annot`: a GRanges object or data.frame including annotation information for locus, including:
  - `start`: start of contiguous genomic regions
SFfigure

- end of contiguous genomic regions
- seqname chromosome name for genomic region
- strand strandedness of sequence

flip.fig an indicator whether to flip the plotting direction if strand == "-" when annotation information is provided.

label.exon an indicator whether to print the exon boundaries to the figure.

print.n an indicator whether to print cluster sizes.

data.labels a \( n \times 1 \) vector of class labels to use instead of calculating SigFuge labels

label.colors a \( K \times 3 \) matrix of RBG colors specifying cluster colors for \( K \) clusters. ggplot2 default colors are used if not specified. If using SigFuge default labels, \( K = 3 \) even if no low expression samples are flagged.

flag a \( n \times 1 \) logical vector of samples flagged as low expression. If flag == 1, default low expression cutoffs are used. If flag == 0, no samples are flagged as low expression (equivalent to setting flag = rep(0, n)).

lplots a specification of which figures to output
- 1: curves in single panel, random colors
- 2: curves in single panel, colored by cluster
- 3: curves in \( K \) panels, separated and colored by cluster
- 4: curves in \( n \) panels, colored by cluster (single sample per panel)
- 5: cluster medians in single panel, colored by cluster

log10 an indicator whether the y-axis (read depth) should be log10 transformed. Default is to plot on log-scale.

summary.type a character string specifying which summary statistic should be used when plotting clusters in lplots == 2, 3, and 5. Options: "median" (default) or "mean".

savestr a string specifying the file name for resulting figures. Extensions can also be specified in savestr. If no extension is specified figures will be saved as pdfs. If length(lplots) > 1, figures will be saved as paste0(savestr,"_x") for x in lplots with the appropriate extension. If no savestr is specified, function will return a list containing the created ggplot objects.

titlestr a string specifying figure title. If unspecified, default is titlestr=paste(locusname," locus, SigFuge analysis").

pval an indicator whether the SFpval should be computed. If pval == 1, the p-value is added to the title, i.e. (titlestr=paste0(titlestr, ", p-value = ", p)).

Value

SFfigure returns a figure that is saved to the current working directory if a savestr is specified. Else, a list containing the plots is returned.

Author(s)

Patrick Kimes <pkimes@live.unc.edu>

Examples

```
# load data
data(geneAnnot)
data(geneDepth)
```
## SFlabels

_SFLabels_ Calculate SigFuge labels

**Description**

Function for producing vector of SigFuge labels using 2-means clustering on non-low expression normalized data and combining with low expression flags. Typically, _SFlabels_ is used by passing output from _SFnormalize_.

**Usage**

```r
SFlabels(normData)
```

**Arguments**

- `normData` a list containing
  - `data.norm` a \( d \times (n-m) \) matrix of normalized read counts at \( d \) positions for \( (n-m) \) samples where \( n \) is the total number of samples and \( m \) is the number of low expression samples.
  - `flag` a \( n \times 1 \) logical vector of flagged samples with \( \sum flag = m \).

**Value**

_SFLabels_ returns a \( n \times 1 \) vector of class labels.

**Author(s)**

Patrick Kimes <pkimes@live.unc.edu>

**Examples**

```r
data(geneDepth)
normalizedData <- SFnormalize(geneDepth)
labels <- SFlabels(normalizedData)
```
**SFnormalize**

**Description**

Function for normalizing read count data as specified in the SigFuge method. The normalization procedure is applied prior to SigFuge clustering to remove the effect of sample-locus specific expression from the analysis. This allows the method to identify clusters based on expression patterns across the genomic locus. It is recommended to flag and remove low expression samples from the normalization and analysis since their shapes may be overwhelmed by noise. A threshold based method for identifying low expression samples is included in the function, but users may also specify their own flags for low expression samples.

**Usage**

```r
SFnormalize(data, flag = 1)
```

**Arguments**

- `data` a $d \times n$ matrix of read counts at $d$ positions for $n$ samples.
- `flag` a $n \times 1$ logical vector of samples flagged as low expression. If `flag == 1`, default low expression cutoffs are used. If `flag == 0`, no samples are flagged as low expression (equivalent to setting `flag = zeros(n, 1)`).

**Value**

`SFnormalize` returns a list containing:

- `data.norm` a $d \times (n - m)$ matrix of normalized read counts where $m$ is the number of low expression samples.
- `flag` a $n \times 1$ logical vector of flagged samples.

**Author(s)**

Patrick Kimes <pkimes@live.unc.edu>

**Examples**

```r
data(geneDepth)
depthnorm <- SFnormalize(geneDepth, flag = 1)
```
**SFpval**

**Calculate SigFuge p-value**

**Description**

Function for computing significance of clustering p-value. p-value is obtained from **sigclust**, a simulation based procedure for testing significance of clustering in high dimension low sample size (HDLSS) data.

The SigClust hypothesis test is given:

- H0: data generated from single Gaussian
- H1: data not generated from single Gaussian

**Usage**

```r
SFpval(data, normalize = 1, flag = 1)
```

**Arguments**

- `data` a $d \times n$ matrix of read counts at $d$ positions for $n$ samples.
- `normalize` a $n \times 1$ logical vector of flagged samples.
- `flag` a $n \times 1$ logical vector of samples flagged as low expression. If `flag == 1`, default low expression cutoffs are applied to `data`. If `flag == 0`, no samples are flagged as low expression (equivalent to setting `flag = zeros(n,1)`).

**Value**

`SFpval` returns an object of class `sigclust-class`. Available slots are described in detail in the `sigclust` package. Primarily, we make use of @pvalnorm.

**Author(s)**

Patrick Kimes <pkimes@live.unc.edu>

**Examples**

```r
data(geneDepth)
SFout <- SFpval(geneDepth, normalize = 1, flag = 1)
SFout@pvalnorm
```
Index

*Topic datasets
  geneAnnot, 2
geneDepth, 3

*Topic package
  SigFuge-package, 2

  geneAnnot, 2
geneDepth, 3

  SFfigure, 2, 3
  SFlabels, 5, 5
  SFnormalize, 5, 6
  SFpval, 2, 7
  sigclust, 7
  SigFuge (SigFuge-package), 2
  SigFuge-package, 2