

# Package ‘SigFuge’

May 21, 2024

**Type** Package

**Title** SigFuge

**Version** 1.43.0

**Date** 2021-11-21

**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, ImmunoOncology

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

**git\_branch** devel

**git\_last\_commit** 2ed6c60

**git\_last\_commit\_date** 2024-04-30

**Repository** Bioconductor 3.20

**Date/Publication** 2024-05-20

## Contents

SigFuge-package . . . . .	2
geneAnnot . . . . .	2
geneDepth . . . . .	3
SFfigure . . . . .	3
SFlabels . . . . .	5
SFnormalize . . . . .	6
SFpval . . . . .	7
<b>Index</b>	<b>8</b>

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**

The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489: 519-525.

---

`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**

```
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**

The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489: 519-525.

---

`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**

```
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**

<code>data</code>	a $d \times n$ matrix or <code>data.frame</code> of read counts at $d$ base positions for $n$ samples.
<code>locusname</code>	a character string specifying gene or locus name to be used in figure title.
<code>annot</code>	a <code>GRanges</code> object or <code>data.frame</code> including annotation information for locus, including: <ul style="list-style-type: none"> <li>• <code>start</code> start of contiguous genomic regions</li> <li>• <code>end</code> end of contiguous genomic regions</li> <li>• <code>seqname</code> chromosome name for genomic region</li> <li>• <code>strand</code> strandedness of sequence</li> </ul>
<code>flip.fig</code>	an indicator whether to flip the plotting direction of the locus if <code>strand == "-"</code> when annotation information is provided.
<code>label.exon</code>	an indicator whether to print the exon boundaries to the figure.
<code>print.n</code>	an indicator whether to print cluster sizes.
<code>data.labels</code>	a $n \times 1$ vector of class labels to use instead of calculating SigFuge labels
<code>label.colors</code>	a $K \times 3$ matrix of RGB colors specifying cluster colors for $K$ clusters. <code>ggplot2</code> default colors are used if not specified. If using SigFuge default labels, $K = 3$ even if no low expression samples are flagged.
<code>flag</code>	a $n \times 1$ logical vector of samples flagged as low expression. If <code>flag == 1</code> , default low expression cutoffs are used. If <code>flag == 0</code> , no samples are flagged as low expression (equivalent to setting <code>flag = rep(0, n)</code> ).
<code>lplots</code>	a specification of which figures to output <ul style="list-style-type: none"> <li>• 1: curves in single panel, random colors</li> <li>• 2: curves in single panel, colored by cluster</li> <li>• 3: curves in <math>K</math> panels, separated and colored by cluster</li> <li>• 4: curves in <math>n</math> panels, colored by cluster (single sample per panel)</li> <li>• 5: cluster medians in single panel, colored by cluster</li> </ul>
<code>log10</code>	an indicator whether the y-axis (read depth) should be log10 transformed. Default is to plot on log-scale.
<code>summary.type</code>	a character string specifying which summary statistic should be used when plotting clusters in <code>lplots == 2, 3, and 5</code> . Options: "median" (default) or "mean".
<code>savestr</code>	a string specifying the file name for resulting figures. Extensions can also be specified in <code>savestr</code> . If no extension is specified figures will be saved as pdfs. If <code>length(lplots) &gt; 1</code> , figures will be saved as <code>paste0(savestr, "_x")</code> for $x$ in <code>lplots</code> with the appropriate extension. If no <code>savestr</code> is specified, function will return a list containing the created <code>ggplot</code> objects.
<code>titlestr</code>	a string specifying figure title. If unspecified, default is <code>titlestr=paste(locusname, "locus, SigFuge analysis")</code> .
<code>pval</code>	an indicator whether the <code>SFpval</code> should be computed. If <code>pval == 1</code> , the p-value is added to the title, i.e. ( <code>titlestr=paste0(titlestr, ", p-value = ", p)</code> ).

**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)

# only use first 50 samples
mdata <- geneDepth[,1:50]

# make plot
locusname <- "CDKN2A"
SFfigure(mdata, locusname, geneAnnot, flag=1,
  lplots=3, savestr=paste0(locusname, ".pdf"), titlestr="CDKN2A locus, LUSC samples",
  pval=1)

mySFs <- SFfigure(mdata, locusname, geneAnnot, flag=1,
  lplots=1, savestr=c(), titlestr="CDKN2A locus, LUSC samples not saved",
  pval=0)
mySFs$plot1
```

---

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**

```
SFlabels(normData)
```

**Arguments**

- |          |   |
|----------|---|
| normData | a list containing <ul style="list-style-type: none"> <li>• <code>data.norm</code> a <math>d \times (n - m)</math> matrix of normalized read counts at <math>d</math> positions for <math>(n - m)</math> samples where <math>n</math> is the total number of samples and <math>m</math> is the number of low expression samples.</li> <li>• <code>flag</code> a <math>n \times 1</math> logical vector of flagged samples with <math>\sum \text{flag} = m</math>.</li> </ul> |
|----------|---|

**Value**

SFlabels returns a  $n \times 1$  vector of class labels.

**Author(s)**

Patrick Kimes <pkimes@live.unc.edu>

**Examples**

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

---

SFnormalize

*SigFuge normalize read counts*

---

**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**

```
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 <code>flag == 1</code> , default low expression cutoffs are used. If <code>flag == 0</code> , no samples are flagged as low expression (equivalent to setting <code>flag = zeros(n, 1)</code> ).

**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**

```
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**

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

**Arguments**

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

**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**

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

# Index

\* **datasets**

geneAnnot, [2](#)

geneDepth, [3](#)

\* **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](#)