# Package 'rnaseqcomp'

October 17, 2020

**Version** 1.18.0

**Title** Benchmarks for RNA-seq Quantification Pipelines

**Description** Several quantitative and visualized benchmarks for RNA-seq quantification pipelines. Two-condition quantifications for genes, transcripts, junctions or exons by each pipeline with necessary meta information should be organized into numeric matrices in order to proceed the evaluation.

**Author** Mingxiang Teng and Rafael A. Irizarry

**Maintainer** Mingxiang Teng <tengmx@gmail.com>

**Depends** R (>= 3.2.0)

**Imports** RColorBrewer, methods

**VignetteBuilder** knitr

**Suggests** BiocStyle, knitr, rmarkdown

**URL** https://github.com/tengmx/rnaseqcomp

**License** GPL-3

**biocViews** RNASeq, Visualization, QualityControl

**RoxygenNote** 6.1.1

**git_url** https://git.bioconductor.org/packages/rnaseqcomp

**git_branch** RELEASE_3_11

**git_last_commit** 62cc381

**git_last_commit_date** 2020-04-27

**Date/Publication** 2020-10-16

## R topics documented:

---

`check_rnaseqcomp` *Sanity Check of S4 rnaseqcomp Class*

---

## Description

This function always checks if the elements are valid to create a S4 rnaseqcomp object. Specifically, check if `quantData` is a list of matrices, if `condInfo` has the correct length and levels, etc.

## Usage

```
check_rnaseqcomp(object)
```

## Arguments

| | |
|---|---|
| `object` | A object of S4 rnaseqcomp class |

## Value

TRUE, or character if error happens.

---

`plot2TX` *Estimate And Plot Transcript Proportion Difference*

---

## Description

For any compared two replicates in each cell line, the proportion of one transcript for genes that only include two annotated transcripts can be different even flipped. This function estimates and plots the proportion difference stratefied by detrended logsignal. Means of absolute difference will be reported for three levels of detrened logsignals. Average is used when multiple two-replicate comparisons included.

## Usage

```
plot2TX(dat, genes, step = 0.5, thresholds = c(1, 6), plotcell = 1,
  ...)
```

## Arguments

| | |
|---|---|
| `dat` | A `rnaseqcomp` S4 class object. |
| `genes` | A vector of gene names corresponding to quantified transcripts. Note that `length(genes)` should equal to `nrow(dat@quantData[[1]])`. |
| `step` | A number specifying the resolution on detrended logsignal for calculation and plotting the proportion difference. (default: 0.5) |
| `thresholds` | A vector of two numbers define cutoffs for three levels of detreded log signals, where one number summary will be generated. (default: c(1, 6)) |
| `plotcell` | 1 or 2 indicating which cell line will be plotted. If values other than 1 and 2, both cell lines will be plotted. This value won't affect estimation for both cell lines. (default: 1) |
| `...` | Parameters for base function `plot`. |

## Value

| | |
|---|---|
| plot | 2TX plots of quantification pipelines for selected cell line by plotcell. |
| list | A list of two matrices indicating the mean and standard error of absolute proportion differences. Valuesa are based on average of two cell lines. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
plot2TX(dat,genes=simdata$meta$gene)
```

---

# plotFC *Estimate And Plot Fold Change Accuracy*

---

## Description

For each pipeline, differential expression is estimated by fold change on mean signals across replicates of cell lines. For features that are truely differential expressed, their fold changes levels are summarized based on different levels of detrended logsignals.

## Usage

```
plotFC(dat, positive, fcsign, constant = 0.5, loessspan = 1/3,
  thresholds = c(1, 6), ...)
```

## Arguments

| | |
|---|---|
| dat | A rnaseqcomp S4 class object. |
| positive | A logical vector with length equivalent to row number of matrices in dat@quantData. TRUE means true differential and FALSE means true non-differential, while missing value NA means unknown. |
| fcsign | A numeric vector with length equivalent to row number of matrices in dat@quantData. Only values 1, -1, 0, NA are allowed. 1 means upregulated in second cell line, -1 means downregulated in second cell line, and 0 means no change. If elements in fcsign is NA or correspond to NA in positive, these elements will be ignored in estimation. |
| constant | A numeric constant that is added to quantifications before fold changes calculation. (default: 0.5) |
| loessspan | A numeric number indicating span used for loess smooth. Details see loess.smooth function. (Default: 1/3) |
| thresholds | A numeric vector defining cutoffs on fold changes as the points to make threshold averaging on ROC curves. (default: seq(12, 0, len = 300)) |
| ... | Parameters for base function plot. |

## Value

| | |
|---|---|
| plot | Fold change plots for all the quantification pipelines. |
| list | A list of two numeric vectors indicating median and standard error of fold changes in three different levels of detrended logsignals. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
## only select the true differential that have exact fold changes
simdata$meta$fcsign[simdata$meta$fcstatus == "off.on"] <- NA
plotFC(dat,simdata$meta$positive,simdata$meta$fcsign)
```

---

# plotNE *Estimate And Plot Express And Non-express Features*

---

## Description

For each cell line, any compared two replicates might have a portion of transcripts that express in one replicate but not the other, depending on what cutoff is used to define non-express. This function estimate and plot the proportion of disagreement using multiple cutoffs. Average is used when multiple two-replicate comparisons included.

## Usage

```
plotNE(dat, steps = seq(-0.5, 12, 0.5), Ks = 0:3,
  pchK = seq_along(Ks) - 1, plotcell = 1, ...)
```

## Arguments

| | |
|---|---|
| dat | A rnaseqcomp S4 class object. |
| steps | A numeric vector specifying log-scale cutoffs to be used for calculation and plotting. (default: seq(-0.5, 12, 0.5)) |
| Ks | A numeric vector specifying which cutoffs to be highlighted and to which the reported proportions to be corresponding. |
| pchK | Plot styles of highlight points corresponding to Ks. (default: seq_along(Ks) - 1) |
| plotcell | 1 or 2 indicating which cell line will be plotted. If values other than 1 and 2, both cell lines will be plotted. This value won't affect estimation for both cell lines. (default: 1) |
| ... | Parameters for base function plot. |

## Value

| | |
|---|---|
| plot | NE plots of quantification pipelines for selected cell line by plotcell. |
| NE | A list of two matrices. The first matrix gives the proportion of disagreement and the second matrix gives the proportion of both replicates under (non-express) correspoinding cutoff Ks. Values are based on average of two cell lines. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
plotNE(dat)
```

---

# plotROC *Estimate And Plot Differential Expression*

---

## Description

For each pipeline, differential expression is first estimated by fold change on 1 vs. 1 comparison between cell lines. ROC curves then are made by comparing fold changes with predefined true differentials. Then, ROC curves from multiple 1 vs. 1 comparisons are averaged using threshold averaging strategy. Standardized partial area under the curve (pAUC) is reported for each pipeline.

## Usage

```
plotROC(dat, positive, fcsign, cut = 1, constant = 0.5,
  thresholds = seq(12, 0, len = 300), arrow = FALSE, ...)
```

## Arguments

| | |
|---|---|
| dat | A rnaseqcomp S4 class object. |
| positive | A logical vector with length equivalent to row number of matrices in dat@quantData. TRUE means true differential and FALSE means true non-differential, while missing value NA means unknown. |
| fcsign | A numeric vector with length equivalent to row number of matrices in dat@quantData. Only values 1, -1, 0 are allowed. 1 means upregulated in second cell line, -1 means downregulated in second cell line, and 0 means no change. If elements in fcsign correspond to NA in positive, these elements will be ignored in estimation. |
| cut | A numeric cutoff used to decide if fold change should be estimated. For a 1 vs 1 comparison, if features have signals less than cut in both samples, their fold changes will be set to 0. (default: 1) |
| constant | A numeric constant that is added to quantifications before fold changes calculation. (default: 0.5) |

| | |
|---|---|
| thresholds | A numeric vector defining cutoffs on fold changes as the points to make threshold averaging on ROC curves. (default: seq(12, 0, len = 300)) |
| arrow | A logical indicating if error bars should be added to the averaged ROC curves. (default: FALSE) |
| ... | Parameters for base function plot. |

## Value

| | |
|---|---|
| plot | ROC plots for all the quantification pipelines. |
| pAUC | A numeric vector indicating pipeline accuracy. This is standardized partial AUC based on ranges chosen on false positive rate. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
plotROC(dat,simdata$meta$positive,simdata$meta$fcsign)
```

---

# plotSD *Estimate And Plot Median Standard Deviation*

---

## Description

For each cell line in each pipeline, the standard deviation of detrend logsignals are calculated for individual features. Then, loess smooth on standard deviation are plotted stratified by detrended log signals for select cell line. The median of standard deviation at three different levels of detrend logsignals are reported.

## Usage

```
plotSD(dat, constant = 0.5, loessspan = 1/3, thresholds = c(1, 6),
  plotcell = 1, ...)
```

## Arguments

| | |
|---|---|
| dat | A rnaseqcomp S4 class object. |
| constant | A numeric pseudo-constant to be added on all the signals before transferred to log scale. (default: 0.5) |
| loessspan | A numeric number indicating span used for loess smooth. Details see loess.smooth function. (Default: 1/3) |
| thresholds | A vector of two numbers define cutoffs for three levels of detreded log signals. (default: c(1, 6)) |
| plotcell | 1 or 2 indicating which cell line will be plotted. If values other than 1 and 2, both cell lines will be plotted. This value won't affect estimation for both cell lines. (default: 1) |
| ... | Parameters for base function plot. |

## Value

| | |
|---|---|
| plot | SD plots of quantification pipelines for selected cell line by plotcell. |
| list | A list of two matrices of median and standard error of standard deviations. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
plotSD(dat)
```

---

# rnaseqcomp-class *rnaseqcomp*

---

## Description

This is a S4 class to organize data ready for benchmark summarization. There are 5 S3 objects inside this class. quantData documents a list of data matrices ready for evaluation by functions plotSD, plotNE, plot2TX or plotROC. condInfo is a factor corresponding to columns of quantData matrices, indicating to which cell lines each sample belongs. repInfo is a factor corresponding to columns of quantData matrices indicating replicate information. repInfo is a legacy from previous versions, and doesn't have too much meanings in current version. refMed is the median log2 signal of calibration references. scaler is a number that point to the median log2 signal of reference pipeline. refMed and scaler were used to calibrate and generate quantData.

---

# signalCalibrate *Quantification Filtering And Calibration*

---

## Description

This is the function to do any pre-filtering or pre-processing analysis for downstream benchmark estimation and visualization. Pre-filtering includes row selection (e.g. protein coding genes) of quantification table; pre-processing includes calculation on a set of rows as calibration reference (e.g. house keeping genes) across different quantification pipelines, calibration of quantifications across all the pipelines based on given cutoffs from selected pipelines.

## Usage

```
signalCalibrate(quantData, condInfo, repInfo, evaluationFeature,
  calibrationFeature, unitReference, unitCutoff = 0,
  calibrationFeature2 = NULL, fixMedian = 4.776)
```

## Arguments

| | |
|---|---|
| quantData | A list of quantification matrices each with rows by features (transcripts, genes, junctions or exons) and columns by samples. Names of the list should be provided. The sizes of each element should be the same. Missing data will be set to 0. |
| condInfo | A factor documenting condition information of samples, corresponding to the columns of each element in quantData. |
| repInfo | A factor documenting replicate information of samples, corresponding to the columns of each element in quantData. |
| evaluationFeature | A logical vector corresponding to the rows of each element in quantData, providing which features should be considered for downstream evaluation, e.g. protein coding genes. |
| calibrationFeature | A logical vector corresponding to the rows of each element in quantData, providing which features should be considered as calibration reference, e.g. house keeping genes. |
| unitReference | A numeric number specifying which pipeline will be selected as reference pipeline, i.e. the index of one element in quantData. |
| unitCutoff | A numeric number for signal cutoff on reference pipeline specified by unitReference (default: 0). Equivalent effects of cutoffs will be applied to other pipelines accordingly. |
| calibrationFeature2 | A logical vector corresponding to the rows of each element in quantData, providing which features should be considered as references for calibration across different datasets. Default NULL means no calibration needed. |
| fixMedian | A numeric number specifying the median of detrend logsignals for features specified by calibrationFeature2. When comparing across datasets, those features will be calibrated to have the same median as fixMedian, while other features calibrated accordingly. The default is 4.776, which was calculated based on one ENCODE dataset used in our web tool. |

## Details

In the functions plotSD and plot2TX, detrended signals with value 0 will be at the same level as value 1 for giving pipeline by unitReference.

## Value

A rnaseqcomp S4 class object

| | |
|---|---|
| quantData | A filtered and calibrated list of quantifications for downstream analysis. |
| condInfo | A factor documenting sample condition information. |
| repInfo | A factor documenting sample replicate information. |
| refMed | A list of numeric vectors giving the log scale medians of calibration features in different pipelines. |
| scaler | A number that was used for scaling quantifications onto reference pipeline. |

## Examples

```
data(simdata)
condInfo <- factor(simdata$samp$condition)
repInfo <- factor(simdata$samp$replicate)
evaluationFeature <- rep(TRUE, nrow(simdata$meta))
calibrationFeature <- simdata$meta$house & simdata$meta$chr == 'chr1'
unitReference <- 1
dat <- signalCalibrate(simdata$quant, condInfo, repInfo, evaluationFeature,
calibrationFeature, unitReference, calibrationFeature2 = calibrationFeature)
```

---

# simdata *Example of Quantifications on Simulation Data*

---

## Description

This dataset include quantifications on 15776 transcripts on two cell lines each with 8 replicates. The true differential expressed transcripts were simulated. Quantifications from two pipelines (RSEM and FluxCapacitor) are included in this dataset at `simdata$quant`. Meta information of transcripts is included at `simdata$meta`, inlcuding if they belongs to house keeping genes and their true fold change status. Sample information is included at `simdata$samp`.

## Format

A list of objects including list of two 15776*16 quantification matrices, one 15776*6 data frame with meta information and one 16*3 data frame with sample information.

# Index