Package 'countsimQC'

May 3, 2024

Type Package

Title Compare Characteristic Features of Count Data Sets

Version 1.22.0

Description countsimQC provides functionality to create a comprehensive report comparing a broad range of characteristics across a collection of count matrices. One important use case is the comparison of one or more synthetic count matrices to a real count matrix, possibly the one underlying the simulations. However, any collection of count matrices can be compared.

License GPL (>=2)

Encoding UTF-8

Depends R (>= 3.5)

Imports rmarkdown (>= 2.5), edgeR, DESeq2 (>= 1.16.0), dplyr, tidyr, ggplot2, grDevices, tools, SummarizedExperiment, genefilter, DT, GenomeInfoDbData, caTools, randtests, stats, utils, methods, ragg

RoxygenNote 7.2.3

Suggests knitr, testthat

VignetteBuilder knitr

biocViews Microbiome, RNASeq, SingleCell, ExperimentalDesign, QualityControl, ReportWriting, Visualization, ImmunoOncology

URL https://github.com/csoneson/countsimQC

BugReports https://github.com/csoneson/countsimQC/issues

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

git_branch RELEASE_3_19

git_last_commit 87198d8

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-05-03

Author Charlotte Soneson [aut, cre] (<https://orcid.org/0000-0003-3833-2169>)

Maintainer Charlotte Soneson <charlottesoneson@gmail.com>

Contents

`calculateDispersionsddsList` *Calculate dispersions*

Description

Calculate the dispersions for each data set in a list of DESeqDataSets, using both edgeR and DESeq2.

Usage

```
calculateDispersionsddsList(ddsList, maxNForDisp)
```

Arguments

`ddsList`	A list of DESeqDataSets
`maxNForDisp`	If any data set contains more than `maxNForDisp` samples, `maxNForDisp` of them will be randomly sampled before the dispersions are calculated, in order to speed up calculations

Value

A list of the same length as the input list. Each element in the list is itself a list, containing a DGEList and a DESeqDataSet with calculated dispersions.

Author(s)

Charlotte Soneson

calculateFeatureCorrs *Calculate Spearman correlation between feature pairs*

Description

Calculate Spearman correlation between feature pairs

Usage

```
calculateFeatureCorrs(ddsList, maxNForCorr)
```

Arguments

ddsList	List of lists, with one element per data set. Each element is a list containing a DGEList and a DESeqDataSet, with calculated dispersions (e.g., output from calculateDispersionsddsList).
maxNForCorr	Maximal number of features to use for calculation of correlations. If the number of features in a data set exceeds maxNForCorr, maxNForCorr features will be randomly selected for calculation of correlations.

Value

A data frame with pairwise feature correlations for each data set

Author(s)

Charlotte Soneson

calculateSampleCorrs *Calculate Spearman correlation between sample pairs*

Description

Calculate Spearman correlation between sample pairs

Usage

```
calculateSampleCorrs(ddsList, maxNForCorr)
```

Arguments

ddsList	List of lists, with one element per data set. Each element is a list containing a DGEList and a DESeqDataSet, with calculated dispersions (e.g., output from calculateDispersionsddsList).
maxNForCorr	Maximal number of samples to use for correlation calculation. If the number of samples in a data set exceeds maxNForCorr, maxNForCorr samples will be randomly selected for calculation of correlations.

Value

A data frame with pairwise sample correlations for each data set

Author(s)

Charlotte Soneson

`calculateStats` *Calculate statistics for pairwise comparison of data sets*

Description

Calculate a range of statistics and p-values for comparison of two data sets.

Usage

```
calculateStats(
  df,
  ds1,
  ds2,
  column,
  subsampleSize,
  permute = FALSE,
  kmin,
  kfrac,
  xmin,
  xmax
)
```

Arguments

`df`	The input data frame. Must contain at least a column named 'dataset' and an additional column with values to use as the basis for the comparison.
`ds1`, `ds2`	The names of the two data sets to be compared.
`column`	The name of the column(s) of `df` to be used as the basis for the comparison.
`subsampleSize`	The number of observations for which certain time-consuming statistics will be calculated. The observations will be selected randomly among the rows of `df`.
`permute`	Whether to permute the dataset column of `df` before calculating the statistics.
`kmin`, `kfrac`	For statistics that require the extraction of k nearest neighbors of a given point, the number of neighbors will be max(kmin, kfrac * nrow(df)).
`xmin`, `xmax`	Smallest and largest value of `column`, used to normalize the x-axis when calculating the area between the eCDFs.

Value

A vector with statistics and p-values

Author(s)

Charlotte Soneson

`countsimExample` *Example list with three count data sets*

Description

A named list with three elements, each corresponding to a (real or simulated) count data set.

Usage

```
countsimExample
```

Format

A named list with three elements, each corresponding to a (real or simulated) count data set.

Details

The `Original` data set represents a subset of 10,000 genes and 11 cells from the GSE74596 single-cell RNA-seq data set, obtained from the conquer repository (http://imlspenticton.uzh.ch:3838/conquer/). The `Sim1` and `Sim2` data sets similarly represent subsets of scRNA-seq data sets simulated with two different simulation methods, using the real GSE74596 data set as the basis for parameter estimation. Each data set is represented as a `DESeqDataSet` object.

Value

A named list with three elements, each corresponding to a (real or simulated) count data set.

`countsimExample_dfmat` *Example list with three count data sets in different formats*

Description

A named list with three elements, each corresponding to a (real or simulated) count data set. One of them is provided as a DESeqDataset, one as a count data frame and one as a count matrix.

Usage

```
countsimExample_dfmat
```

Format

A named list with three elements, each corresponding to a (real or simulated) count data set.

Details

The `Original` data set represents a subset of 10,000 genes and 11 cells from the GSE74596 single-cell RNA-seq data set, obtained from the conquer repository (http://imlspenticton.uzh.ch:3838/conquer/). The `Sim1` and `Sim2` data sets similarly represent subsets of scRNA-seq data sets simulated with two different simulation methods, using the real GSE74596 data set as the basis for parameter estimation.

Value

A named list with three elements, each corresponding to a (real or simulated) count data set.

`countsimQC-pkg` *countsimQC*

Description

countsimQC

`countsimQCReport` *Generate countsimQC report*

Description

Generate a report comparing a range of characteristics across a collection of one or more count data sets.

Usage

```
countsimQCReport(
  ddsList,
  outputFile,
  outputDir = "./",
  outputFormat = NULL,
  showCode = FALSE,
  rmdTemplate = NULL,
  forceOverwrite = FALSE,
  savePlots = FALSE,
  description = NULL,
  maxNForCorr = 500,
  maxNForDisp = Inf,
  calculateStatistics = TRUE,
  subsampleSize = 500,
  kfrac = 0.01,
  kmin = 5,
  permutationPvalues = FALSE,
```

```
  nPermutations = NULL,
  knitrProgress = FALSE,
  quiet = FALSE,
  ignorePandoc = FALSE,
  useRAGG = FALSE,
  dpi = 96,
  ...
)
```

Arguments

ddsList	Named list of DESeqDataSets or count matrices to compare. See the DESeq2 Bioconductor package (http://bioconductor.org/packages/release/bioc/html/DESeq2.html) for more information about the DESeqDataSet class. Each DESeqDataSet object in the list should contain a count matrix, a data frame with sample information and a design formula. The sample information and design formula will be used to calculate dispersions appropriately. If count matrices are provided, it is assumed that all columns represent replicate samples, and the design formula ~1 will be used.
outputFile	The file name of the final report. The extension must match the selected outputFormat (i.e., either .html or .pdf).
outputDir	The directory where the final report should be saved.
outputFormat	The output format of the report. If set to NULL or "html_document", an html report will be generated. If set to "pdf_document", a pdf report will be generated.
showCode	Whether or not to include the code in the final report.
rmdTemplate	The Rmarkdown (.Rmd) file that will be used as the template for generating the report. If set to NULL (default), the template provided with the countsimQC package will be used. See Details for more information.
forceOverwrite	Whether to force overwrite existing output files when saving the generated report and figures.
savePlots	Whether to save the ggplot objects for all the output figures, to allow additional fine-tuning and generation of individual plots. Note that the resulting file can be quite large, especially when many and/or large data sets are compared.
description	A string (of arbitrary length) describing the content of the generated report. This will be included in the beginning of the report. If set to NULL, a default description listing the number and names of the included data sets will be used.
maxNForCorr	The maximal number of samples (features) for which pairwise correlation coefficients will be calculated. If the number of samples (features) exceeds this number, they will be randomly subsampled.
maxNForDisp	The maximal number of samples that will be used to estimate dispersions. By default, all samples are used. This can be lowered to speed up calculations (and obtain approximate results) for large data sets.
calculateStatistics	Whether to calculate quantitative pairwise statistics for comparing data sets in addition to generating the plots.

subsampleSize	The number of randomly selected observations (samples, features or pairs of samples or features) for which certain (time-consuming) statistics will be calculated. Only used if calculateStatistics = TRUE.
kmin, kfrac	For statistics that require the extraction of the k nearest neighbors of a given point, the number of neighbors will be max(kmin, kfrac * nrow(df))
permutationPvalues	Whether to calculate permutation p-values for selected pairwise data set comparison statistics.
nPermutations	The number of permutations to perform when calculating permutation p-values for data set comparison statistics. Only used if permutationPvalues = TRUE.
knitrProgress	Whether to show the progress bar when the report is generated.
quiet	Whether to suppress warnings and progress messages when the report is generated.
ignorePandoc	Determines what to do if pandoc or pandoc-citeproc is missing (if Sys.which("pandoc") or Sys.which("pandoc-citeproc") is ""). If ignorePandoc is TRUE, only a warning is given. The figures will be generated, but not the final report. If ignorePandoc is FALSE (default), the execution stops immediately.
useRAGG	Logical scalar, indicating whether to use ragg_png as the graphics device in the report rather than the default png.
dpi	Numeric scalar, setting the dpi of the generated plots. Only used if useRAGG is TRUE.
...	Other arguments that will be passed to rmarkdown::render.

Details

When the function is called, the template file (specified by rmdTemplate) will be copied into the output folder, and rmarkdown::render will be called to generate the final report. If there is already a .Rmd file with the same name in the output folder, the function will raise an error and stop, to avoid overwriting the existing file. The reason for this behaviour is that the copied template in the output folder will be deleted once the report is generated.

Value

No value is returned, but a report is generated in the outputDir directory.

Author(s)

Charlotte Soneson

Examples

```
## Load example data
data(countsimExample)
## Not run:
## Generate report
countsimQCReport(countsimExample, outputDir = "./",
                 outputFile = "example.html")
```

```
## End(Not run)
```

`defaultStats` *Return a vector of NA scores*

Description

Return a vector of NA scores

Usage

```
defaultStats(n, withP = FALSE)
```

Arguments

`n`	Number of columns to use for the comparison
`withP`	Whether or not to include p-value columns

Value

A vector with NA values for all applicable statistics

Author(s)

Charlotte Soneson

`defineTableDesc` *Define table descriptions*

Description

Generate the text that describes the content of the tables generated by `makeDF`.

Usage

```
defineTableDesc(
  calculateStatistics,
  subsampleSize,
  kfrac,
  kmin,
  obstype,
  aspect,
  minvalue,
  maxvalue,
```

```
  permutationPvalues,
  nPermutations,
  nDatasets
)
```

Arguments

`calculateStatistics`	Whether or not statistics and p-values are calculated
`subsampleSize`	The number of observations for which certain (time-consuming) statistics will be calculated
`kmin`, `kfrac`	For statistics that require the extraction of k nearest neighbors of a given point, the number of neighbors will be max(kmin, kfrac * nrow(df))
`obstype`	The type of observation (e.g., sample, feature, sample pair)
`aspect`	The name of the aspect of interest
`minvalue`, `maxvalue`	The minimal and maximal value of the aspect of interest, used for scaling of the x axis when calculating the area between the eCDFs
`permutationPvalues`	Whether or not to calculate p-values of statistics via permutation
`nPermutations`	The number of permutations (only used if permutationPvalues = TRUE)
`nDatasets`	The number of data sets that are being compared

Value

A list with two text strings in markdown format: one for tables based on a single data column, and one for tables based on two data columns

Author(s)

Charlotte Soneson

`generateIndividualPlots` *Generate individual plots from countsimQCReport output*

Description

Generate separate plots for all evaluation criteria using the collection of ggplot objects that can be saved when generating a countsimQC report (by setting `savePlots = TRUE`).

Usage

```
generateIndividualPlots(
  ggplotsRds,
  device = "png",
  outputDir = "./",
  nDatasets = 2
)
```

Arguments

ggplotsRds	The path to a .rds file generated by countsimQCReport by setting savePlots = TRUE, or the list of plots stored in this file.
device	One of "eps", "ps", "tex" (pictex), "pdf", "jpeg", "tiff", "png", "bmp", "svg" or "wmf" (windows only) (will be provided to the ggsave function from the ggplot2 package).
outputDir	The output directory where the plots should be generated.
nDatasets	The number of data sets that are compared in the figures. This is needed to set the size of the plots correctly.

Value

Nothing is returned, but plots are generated in the designated output directory.

Author(s)

Charlotte Soneson

Examples

```
## Load example data
data(countsimExample)
## Not run:
## Generate report
countsimQCReport(countsimExample, outputDir = "./",
                 outputFile = "example.html", savePlots = TRUE)
## Generate individual plots
generateIndividualPlots("example_ggplots.rds", nDatasets = 3)

## End(Not run)
```

makeDF *Construct data frame with pairwise statistics*

Description

Construct a data frame containing statistics and p-values for pairwise comparison of data sets.

Usage

```
makeDF(
  df,
  column,
  permutationPvalues,
  nPermutations,
  subsampleSize,
  kmin,
  kfrac
)
```

Arguments

df	The input data frame. Must contain at least a column named 'dataset' and an additional column with values
column	The name of the column(s) of df to be used as the basis for the comparison
permutationPvalues	Whether or not to calculate p-values of statistics via permutation
nPermutations	The number of permutations (only used if permutationPvalues = TRUE)
subsampleSize	The number of observations for which certain (time-consuming) statistics will be calculated. The observations will be selected randomly among the rows of df
kmin, kfrac	For statistics that require the extraction of k nearest neighbors of a given point, the number of neighbors will be max(kmin, kfrac * nrow(df))

Value

A data table with statistics and p-values for pairwise comparisons of data sets, based on the provided column

Author(s)

Charlotte Soneson

Index