Package 'countsimQC'

October 14, 2021

Type Package

Title Compare Characteristic Features of Count Data Sets

Version 1.10.0

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

RoxygenNote 7.1.1

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_13

git_last_commit a1da5e8

git_last_commit_date 2021-05-19

Date/Publication 2021-10-14

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

Maintainer Charlotte Soneson <charlottesoneson@gmail.com>

R topics documented:

countsimExample 2
countsimExample_dfmat 3
countsimQC-pkg 3
countsimQCReport 4
generateIndividualPlots 6

Index **8**

`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,
  ...
)
```

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.
...	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)
```

`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)
```

Index

∗ **datasets**
countsimExample, 2
countsimExample_dfmat, 3

countsimExample, 2
countsimExample_dfmat, 3
countsimQC-pkg, 3
countsimQCReport, 4

generateIndividualPlots, 6