Package 'scShapes'

December 19, 2024

Type Package

Title A Statistical Framework for Modeling and Identifying Differential Distributions in Single-cell RNA-sequencing Data

Version 1.13.0

Description We present a novel statistical framework for identifying differential distributions in single-cell RNA-sequencing (scRNA-seq) data between treatment conditions by modeling gene expression read counts using generalized linear models (GLMs). We model each gene independently under each treatment condition using error distributions Poisson (P), Negative Binomial (NB), Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) with log link function and model based normalization for differences in sequencing depth. Since all four distributions considered in our framework belong to the same family of distributions, we first perform a Kolmogorov-Smirnov (KS) test to select genes belonging to the family of ZINB distributions. Genes passing the KS test will be then modeled using GLMs. Model selection is done by calculating the Bayesian Information Criterion (BIC) and likelihood ratio test (LRT) statistic.

NeedsCompilation yes

Imports Matrix, stats, methods, pscl, VGAM, dgof, BiocParallel, MASS, emdbook, magrittr, utils

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.1

URL https://github.com/Malindrie/scShapes

BugReports https://github.com/Malindrie/scShapes/issues

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

biocViews RNASeq, SingleCell, MultipleComparison, GeneExpression

Depends R (>= 4.1)

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

git_branch devel

git_last_commit f5a9606

git_last_commit_date 2024-10-29

Repository Bioconductor 3.21

Date/Publication 2024-12-18

Author Malindrie Dharmaratne [cre, aut] (ORCID:
<https://orcid.org/0000-0002-1694-6496>)

Maintainer Malindrie Dharmaratne <malindrie@gmail.com>

Contents

`change_shape` *change_shape*

Description

This function returns a list of genes changing shape between conditions and the list of genes changing distribution from a unimodal to distribution to a zero-inflated distribution

Usage

```
change_shape(shapes_genes)
```

Arguments

`shapes_genes`	A dataframe consisting of distributions followed by each gene passing the KS test. Where rows are genes and each column the treatment condition.

Value

A list of two lists with genes changing distribution shape between conditions. The list "All" contains all the genes changing distribution. The list "Uni2ZI" contains the genes changing distribution form a unimodal distribution in one condition to a zero-inflated distribution in another condition.

`filter_counts` *filter_counts*

Description

This function is used to preprocess matrix of read counts to only keep genes with a certain number of nonzero entries.

Usage

```
filter_counts(counts, perc.zero = 0.1)
```

Arguments

`counts`	A non-negative integer matrix of scRNA-seq raw read counts. The rows of the matrix are genes and columns are samples/cells.
`perc.zero`	A numeric value between 0 and 1 that represents the proportion of zeros per gene in the processed dataset.

Value

An object of class `Matrix` with genes removed if they have more than `perc.zero` zeros.

Examples

```
# load toy  example data

data(scData)

# apply the filter_counts function to filter out genes if they have
# more than 10% zero

scData_filt <- filter_counts(scData$counts, perc.zero = 0.1)
```

fit_models *fit_models*

Description

This function is used to fit genes with GLM

Usage

```
fit_models(counts, cexpr, lib.size, formula = NULL, model = NULL, BPPARAM)
```

Arguments

counts	A non-negative integer matrix of scRNA-seq filtered read counts containing genes belonging to the family of ZINB distributions selected from ks_test.
cexpr	A dataframe that contains the covariate values. The rows of the dataframe are the corresponding samples/cells from the counts matrix from filter_counts. The cells of the dataframe are the covariates to be included in the GLM.
lib.size	A numeric vector that contains the total number of counts per cell from the counts matrix from filter_counts.
formula	A regression formula to fit the covariates in the ZINB GLM.
model	A specific model to fit (1:P, 2:NB, 3:ZIP, 4:ZINB, NULL:All)
BPPARAM	configuration parameter related to the method of parallel execution. For further information on how to set-up parallel execution refer to BiocParallel vignette.

Value

A list of models fitted by 'glm'

Examples

```
data(scData)

# apply the fit_models function to subset genes belonging to the
# family of ZINB distributions, selceted from ks_test function.

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates,
lib.size=scData$lib_size, BPPARAM=bpparam())
```

gof_model *gof_model*

Description

This function is used to perform the likelihood ratio test on the models chosen based on the BIC values from best_model to check for model adequacy.

Usage

```
gof_model(lbic, cexpr, lib.size, formula = NULL, BPPARAM)
```

Arguments

lbic	A list of genes together with filtered read counts based on the selected distribution from best_model. Output from best_model.
cexpr	A dataframe that contains the covariate values. The rows of the dataframe are the corresponding samples/cells from the counts matrix from filter_counts. The cells of the dataframe are the covariates to be included in the GLM.
lib.size	A numeric vector that contains the total number of counts per cell from the counts matrix from filter_counts.
formula	A regression formula to fit the covariates in the ZINB GLM.
BPPARAM	configuration parameter related to the method of parallel execution. For further information on how to set-up parallel execution refer to BiocParallel vignette.

Value

A list of genes with the p-values from performing the GOF tests.

Examples

```
data(scData)

# apply the gof_model function to perform the likelihood ratio
# test on the models selected by using the lbic_model function

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())
scData_bicvals <- model_bic(scData_models)
scData_least.bic <- lbic_model(scData_bicvals, scData$counts)

scData_gof <- gof_model(scData_least.bic, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())
```

ks_sig *ks_sig*

Description

This function is used to select genes significant from the ks_test.

Usage

```
ks_sig(ks.pval.unadj)
```

Arguments

ks.pval.unadj	The output from ks_test which is a list of p-values from the KS test with gene names.

Value

List object containing the significant gene indices from the KS test, their adjusted p-values

Examples

```
data(scData)

# apply the ks_test function to subset genes belonging to the
# family of ZINB distributions.

library(BiocParallel)
scData_KS <- ks_test(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size, BPPARAM=bpparam())

# apply the ks_sig function to select genes significant from
# the Kolmogorov Smirnov test.

scData_KS_sig <- ks_sig(scData_KS)
```

ks_test *ks_test*

Description

This function is used to perform Kolmogorv-Smirnov test on the filtered sparse counts matrix from filter_counts to select genes belonging to the family of ZINB distributions

Usage

```
ks_test(counts, cexpr, lib.size, formula = NULL, BPPARAM)
```

Arguments

counts	A non-negative integer matrix of scRNA-seq filtered read counts from filter_counts. The rows of the matrix are genes and columns are samples/cells.
cexpr	A dataframe that contains the covariate values. The rows of the dataframe are the corresponding samples/cells from the counts matrix from filter_counts. The cells of the dataframe are the covariates to be included in the GLM.
lib.size	A numeric vector that contains the total number of counts per cell from the counts matrix from filter_counts.
formula	A regression formula to fit the covariates in the ZINB GLM.
BPPARAM	configuration parameter related to the method of parallel execution. For further information on how to set-up parallel execution refer to BiocParallel vignette.

Value

List object containing the p-values from the KS test.

Examples

```
#' # load toy  example data

data(scData)

# apply the ks_test function to subset genes belonging to the
# family of ZINB distributions.

library(BiocParallel)
scData_KS <- ks_test(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size, BPPARAM=bpparam())
```

---

lbic_model *lbic_model*

Description

This function is used to select the best fit model for each gene based on the least BIC value

Usage

```
lbic_model(bic.value, counts)
```

Arguments

bic.value	A dataframe of BIC values from fitting GLM using error distributions P, NB, ZIP, ZINB; the output from model_bic.
counts	A non-negative integer matrix of scRNA-seq filtered read counts containing genes belonging to the family of ZINB distributions selected from ks_test. The rows of the matrix are genes and columns are samples/cells.

Value

A list of genes chosen to be following one of the 4 distributions P, NB, ZIP, ZINB based on the least BIC value and the corresponding subset of counts from `filter_counts`

Examples

```
data(scData)

# apply the lbic_model function to select the model with the least
# BIC value on the matrix of BIC values obtained after running
# model_bic function.

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())

scData_bicvals <- model_bic(scData_models)

scData_least.bic <- lbic_model(scData_bicvals, scData$counts)
```

`logo` *A Statistical Framework for Modeling and Identifying Differential Distributions in Single-cell RNA-sequencing Data*

Description

A Statistical Framework for Modeling and Identifying Differential Distributions in Single-cell RNA-sequencing Data

Usage

```
logo()
```

Value

... Reports the logo of the package:scShapes.

`model_bic` *model_bic*

Description

This function is used to calculate the Bayesian information criterion of the models fitted in `fit_models`.

Usage

```
model_bic(fit_list)
```

Arguments

fit_list	A list of models fitted from fit_models

Value

A dataframe containing the BIC values for each distribution type (P, NB, ZIP, ZINB).

Examples

```
data(scData)

# apply the model_bic function to calculate the BIC values on the models
# obtained after running fit_models function.

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())

scData_bicvals <- model_bic(scData_models)
```

---

model_param *model_param*

Description

This function returns model parameters based on the best fit distribution as selected by distr_fit and models fitted by fit_models

Usage

```
model_param(fit.model, gof.res, model = NULL)
```

Arguments

fit.model	A list of models fitted by 'glm' from fit_models.
gof.res	A list of selected model distributions for genes select_model.
model	A specific model to fit (1:P, 2:NB, 3:ZIP, 4:ZINB, NULL:All)

Value

A list of model parameters estimated. Estimated model parameters include mean (for all 4 models), theta (over-dispersion parameter for NB & ZINB models), pi (zero-inflation parameter for ZIP & ZINB models).

Examples

```
data(scData)

# apply the model_param function to extract the parameters of the best fit
# model obtained by running the select_model function

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())
scData_bicvals <- model_bic(scData_models)
scData_least.bic <- lbic_model(scData_bicvals, scData$counts)
scData_gof <- gof_model(scData_least.bic, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())
scData_fit <- select_model(scData_gof)

scData_params <- model_param(scData_models, scData_fit, model=NULL)
```

---

scData *Sample data for analysis*

Description

Toy example data list of scRNA-seq counts, information on covariates, and library sizes for randomly generated 20 genes to illustrate how to use the functions of the package scShapes

Usage

```
data(scData)
```

Format

A list of three lists labeled 'counts', 'covariates', 'lib_size'. 'counts' an RNA-seq counts matrix of 20 genes and 500 cells; 'covariates' a dataframe of covariates corresponding to the RNA-seq counts where rows are cells of the counts matrix; 'lib_size' a numeric vector of library sizes corresponding to the columns of the RNA-seq counts matrix.

Value

An RData object

Examples

```
# load toy  example data

data(scData)
```

select_model *select_model*

Description

This function is used to select the distribution of best fit for scRNA-seq count data

Usage

```
select_model(lrt.value)
```

Arguments

lrt.value	A list of genes with the p-values from performing the GOF tests from gof_model

Value

A list of selected model distributions for genes scShapes selects.

Examples

```
data(scData)

# apply the select_model function to the best fit model from the results of
# the gof_model function

library(BiocParallel)
scData_models <- fit_models(counts=scData$counts, cexpr=scData$covariates,
                            lib.size=scData$lib_size, BPPARAM=bpparam())
scData_bicvals <- model_bic(scData_models)
scData_least.bic <- lbic_model(scData_bicvals, scData$counts)
scData_gof <- gof_model(scData_least.bic, cexpr=scData$covariates, lib.size=scData$lib_size,
BPPARAM=bpparam())

scData_fit <- select_model(scData_gof)
```

Index