Package ‘scone’

April 16, 2024

Version  1.26.0

Title   Single Cell Overview of Normalized Expression data

Description  SCONE is an R package for comparing and ranking the performance of
different normalization schemes for single-cell RNA-seq and other
high-throughput analyses.

License  Artistic-2.0

Depends  R (>= 3.4), methods, SummarizedExperiment

Imports  graphics, stats, utils, aroma.light, BiocParallel, class,
          cluster, compositions, diptest, edgeR, fpc, gplots, grDevices,
          hexbin, limma, matrixStats, mixtools, RColorBrewer, boot,
          rhdf5, RUVSeq, rARPACK, MatrixGenerics, SingleCellExperiment

Suggests BiocStyle, DT, ggplot2, knitr, miniUI, NMF, plotly, reshape2,
         rmarkdown, scran, scRNAseq, shiny, testthat, visNetwork,
         doParallel, batchtools, splatter, scater, kableExtra, mclust,
         TENxPBMCData

VignetteBuilder  knitr

biocViews  ImmunoOncology, Normalization, Preprocessing,
           QualityControl, GeneExpression, RNASeq, Software,
           Transcriptomics, Sequencing, SingleCell, Coverage

BugReports  https://github.com/YosefLab/scone/issues

RoxygenNote  7.1.2

Encoding  UTF-8

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

git_branch  RELEASE_3_18

git_last_commit  a7873ee

git_last_commit_date  2023-10-24

Repository  Bioconductor 3.18

Date/Publication  2024-04-15
Author  Michael Cole [aut, cph],
        Davide Risso [aut, cre, cph],
        Matteo Borella [ctb],
        Chiara Romualdi [ctb]
Maintainer  Davide Risso <risso.davide@gmail.com>

R topics documented:

  .likfn ................................................................. 3
  .parse_row ........................................................... 3
  .pzfn ................................................................. 4
  biplot_color ........................................................ 4
  biplot_interactive ................................................ 5
  CLR_FN ............................................................... 6
  control_genes ....................................................... 7
  DESEQ_FN ............................................................. 8
  estimate_ziber ...................................................... 9
  factor_sample_filter ............................................... 10
  fast_estimate_ziber ................................................ 12
  FQ_FN ................................................................. 13
  get_bio ............................................................... 14
  get_design ........................................................... 15
  get_negconruv ....................................................... 16
  get_normalized ..................................................... 17
  get_params ........................................................... 18
  get_qc ................................................................. 19
  get_scores ........................................................... 20
  impute_expectation ................................................ 21
  impute_null .......................................................... 22
  lm_adjust .............................................................. 22
  make_design ........................................................... 23
  metric_sample_filter ............................................... 24
  PsiNorm .............................................................. 26
  PSINORM_FN .......................................................... 27
  scone ................................................................. 28
  SconeExperiment-class ............................................. 31
  sconeReport .......................................................... 34
  scone_easybake ...................................................... 35
  score_matrix ........................................................ 38
  SCRAN_FN ............................................................. 41
  select_methods ...................................................... 41
  simple_FNR_params ................................................ 42
  SUM_FN ................................................................. 43
  TMM_FN ................................................................. 44
  UQ_FN ................................................................. 44

Index  46
.likfn

Likelihood Function of the Logistic Model

Description
Likelihood Function of the Logistic Model

Usage
.likfn(Z, X, Beta)

Arguments
Z   data matrix
X   sample-level values
Beta  gene-level values

.parse_row

Parse rows

Description
This function is used internally in scone to parse the variables used to generate the design matrices.

Usage
.parse_row(pars, bio, batch, ruv_factors, qc)

Arguments
pars  character. A vector of parameters corresponding to a row of workflow parameters.
bio   factor. The biological covariate.
batch factor. The batch covariate.
ruv_factors  list. A list containing the factors of unwanted variation (RUVg) for all upstream workflows.
qc  matrix. The principal components of the QC metric matrix.

Value
A list with the variables to be passed to make_design.
.pzfn  

*Posterior probability of detection*

**Description**

Posterior probability of detection

**Usage**

```r
.pzfn(Y, W, Alpha, X, Beta)
```

**Arguments**

- **Y**: detection matrix.
- **W**: sample-level drop-out coefficients.
- **Alpha**: gene-level drop-out features.
- **X**: sample-level expression features.
- **Beta**: gene-level sample coefficients.

---

biplot_color  

*Function for biplotting with no point labels and with points color-coded according to a quantitative variable. For example: the rank of normalization performance.*

**Description**

This function implements biplot for `prcomp` objects.

**Usage**

```r
biplot_color(
  x, y,
  rank = TRUE,
  ties_method = c("max", "min", "first", "last", "random"),
  choices = 1:2,
  expand = 1,
  ...
)
```
Arguments

- **x**: A `prcomp` object.
- **y**: Numeric. Quantitative values used to color the points. If rank is FALSE, all values must be positive integers and less than or equal to the length of y.
- **rank**: Logical. If TRUE (default) y will be transformed by the rank() function.
- **ties_method**: Character. Ties.method used by the rank() function.
- **choices**: Numeric. 2 principal components to plot. Default to first two PCs.
- **expand**: Numeric. Value used to adjust the spread of the arrows relative to the points.
- **...**: Arguments passed to plot.

Value

Invisibly returns scaled point coordinates used in plot.

Examples

```r
mat <- matrix(rnorm(1000), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")

pc <- prcomp(mat)

biplot_color(pc, rank(pc$x[,1]))
```

---

**biplot_interactive**  
Interactive biplot

Description

This is a wrapper around `biplot_color`, creating a shiny gadget to allow the user to select specific points in the graph.

Usage

```r
biplot_interactive(x, ...)
```

Arguments

- **x**: A `SconeExperiment` object.
- **...**: Passed to `biplot_color`.

Details

Since this is based on the shiny gadget feature, it will not work in static documents, such as vignettes or markdown / knitr documents. See `biplot_color` for more details on the internals.
CLR_FN

Centered log-ratio (CLR) normalization wrapper function

Description

Centered log-ratio (CLR) normalization wrapper function

Usage

CLR_FN(ei)

Arguments

ei 

Numerical matrix. (rows = genes, cols = samples).

Details

SCONE scaling wrapper for \texttt{clr}).

Value

CLR normalized matrix.

Examples

ei <- matrix(0:20, nrow = 7)
eo <- CLR_FN(ei)
control_genes

Data: Positive and Negative Control Genes

Description

Sets of "positive" and "negative" control genes, useful arguments for scone.

Details

These gene sets can be used as negative or positive controls, either for RUV factor normalization or for evaluation and ranking of the normalization workflows.

Gene set datasets are in the form of data.frame, with the first column containing the gene symbols and an (optional) second column containing additional information (such as cortical layer or cell cycle phase).

Note that the gene symbols follow the mouse conventions (i.e. capitalized) or the human conventions (i.e. all upper-case), based on the original publication. One can use the toupper, tolower, and toTitleCase functions to alter symbol conventions.

Mouse gene symbols in cortical_markers are transcribed from Figure 3 of Molyneaux et al. (2007): "laminar-specific expression of 66 genes within the neocortex."

Human gene symbols in housekeeping are derived from the list of "housekeeping" genes from the cDNA microarray analysis of Eisenberg and Levanon (2003): "[HK genes] belong to the class of genes that are EXPRESSED in all tissues." "... from 47 different human tissues and cell lines."

Human gene symbols in housekeeping_revised from Eisenberg and Levanon (2013): "This list provided ... is based on analysis of next-generation sequencing (RNA-seq) data. At least one variant of these genes is expressed in all tissues uniformly... The RefSeq transcript according to which we deemed the gene 'housekeeping' is given." Housekeeping exons satisfy "(i) expression observed in all tissues; (ii) low variance over tissues: standard-deviation [log2(RPKM)]<1; and (iii) no exceptional expression in any single tissue; that is, no log-expression value differed from the averaged log2(RPKM) by two (fourfold) or more." "We define a housekeeping gene as a gene for which at least one RefSeq transcript has more than half of its exons meeting the previous criteria (thus being housekeeping exons)."

Human gene symbols in cellcycle_genes from Macosko et al. (2015) and represent a set of genes marking G1/S, S, G2/M, M, and M/G1 phases.

References


DESEQ_FN

Examples

data(housekeeping)
data(housekeeping_revised)
data(cellcycle_genes)
data(cortical_markers)

| DESEQ_FN | Relative log-expression (RLE; DESeq) scaling normalization wrapper function |

Description

Relative log-expression (RLE; DESeq) scaling normalization wrapper function

Usage

DESEQ_FN(ei)

Arguments

ei Numerical matrix. (rows = genes, cols = samples).

Details

SCONE scaling wrapper for calcNormFactors).

Value

RLE normalized matrix.

Examples

ei <- matrix(0:20,nrow = 7)
eo <- DESEQ_FN(ei)
estimate_ziber

Parameter estimation of zero-inflated bernoulli model

Description

This function implements an expectation-maximization algorithm for a zero-inflated bernoulli model of transcript detection, modeling gene expression state (off of on) as a bernoulli draw on a gene-specific expression rate (Z in 0,1). Detection conditioned on expression is a logistic function of gene-level features. The bernoulli model is modeled numerically by a logistic model with an intercept.

Usage

```
estimate_ziber(
  x,
  fp_tresh = 0,
  gfeatM = NULL,
  bulk_model = FALSE,
  pos_controls = NULL,
  em_tol = 0.01,
  maxiter = 100,
  verbose = FALSE
)
```

Arguments

- **x** matrix. An expression data matrix (genes in rows, cells in columns)
- **fp_tresh** numeric. Threshold for calling a positive detection (D = 1). Default 0.
- **gfeatM** matrix. Numeric gene level determinants of drop-out (genes in rows, features in columns)
- **bulk_model** logical. Use median log-expression of gene in detected fraction as sole gene-level feature. Default FALSE. Ignored if gfeatM is specified.
- **pos_controls** logical. TRUE for all genes that are known to be expressed in all cells.
- **em_tol** numeric. Convergence threshold on log-likelihood.
- **maxiter** numeric. The maximum number of iterations. Default 100.
- **verbose** logical. Whether or not to print the value of the likelihood at each iteration.

Value

a list with the following elements:

- W coefficients of sample-specific logistic drop-out model
- Alpha intercept and gene-level parameter matrix
- X intercept
- Beta coefficient of gene-specific logistic expression model
- fnr_character the probability, per gene, of P(D=0|E=1)
- p_nodrop 1 - the probability P(drop|Y), useful as weights in weighted PCA
- expected_state the expected value E[Z] (1 = "on")
- loglik the log-likelihood
- convergence 0 if the algorithm converged and 1 if maxiter was reached

Examples

```r
mat <- matrix(rpois(1000, lambda = 3), ncol=10)
mat = mat * matrix(1-rbinom(1000, size = 1, prob = .01), ncol=10)
ziber_out = suppressWarnings(estimate_ziber(mat,
  bulk_model = TRUE,
  pos_controls = 1:10))
```

factor_sample_filter  

Factor-based Sample Filtering: Function to filter single-cell RNA-Seq libraries.

Description

This function returns a sample-filtering report for each cell in the input expression matrix, describing whether it passed filtering by factor-based filtering, using PCA of quality metrics.

Usage

```r
factor_sample_filter(
  expr,
  qual,
  gene_filter = NULL,
  max_exp_pcs = 5,
  qual_select_q_thresh = 0.01,
  force_metrics = NULL,
  good_metrics = NULL,
  min_qual_variance = 0.7,
  zcut = 1,
  mixture = TRUE,
  dip_thresh = 0.01,
  plot = FALSE,
  hist_breaks = 20
)
```
factor_sample_filter

Arguments

expr  matrix The data matrix (genes in rows, cells in columns).

qual  matrix Quality metric data matrix (cells in rows, metrics in columns).

gene_filter  Logical vector indexing genes that will be used for PCA. If NULL, all genes are used.

max_exp_pcs  numeric number of expression PCs used in quality metric selection. Default 5.

qual_select_q_thresh  numeric. q-value threshold for quality/expression correlation significance tests. Default 0.01

force_metrics  logical. If not NULL, indexes quality metric to be forcefully included in quality PCA.

good_metrics  logical. If not NULL, indexes quality metric that indicate better quality when of higher value.

min_qual_variance  numeric. Minimum proportion of selected quality variance addressed in filtering. Default 0.70

zcut  A numeric value determining threshold Z-score for sd, mad, and mixture sub-criteria. Default 1.

mixture  A logical value determining whether mixture modeling sub-criterion will be applied per primary criterion (quality score). If true, a dip test will be applied to each quality score. If a metric is multimodal, it is fit to a two-component normal mixture model. Samples deviating zcut sd’s from optimal mean (in the inferior direction), have failed this sub-criterion.

dip_thresh  A numeric value determining dip test p-value threshold. Default 0.05.

plot  logical. Should a plot be produced?

hist_breaks  hist() breaks argument. Ignored if ‘plot=FALSE’.

Details

None

Value

A logical, representing samples passing factor-based filter.

Examples

mat <- matrix(rpois(1000, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
qc = as.matrix(cbind(colSums(mat),colSums(mat > 0)))
rownames(qc) = colnames(mat)
colnames(qc) = c("NCOUNTS","NGENES")
mfilt = factor_sample_filter(expr = mat,
qc, plot = TRUE, qual_select_q_thresh = 1)
Description

This function implements Newton’s method for solving zero of Expectation-Maximization equation at the limit of parameter convergence: a zero-inflated bernoulli model of transcript detection, modeling gene expression state (off of on) as a bernoulli draw on a gene-specific expression rate (Z in 0,1). Detection conditioned on expression is a logistic function of gene-level features. The bernoulli model is modeled numerically by a logistic model with an intercept.

Usage

```r
fast_estimate_ziber(
  x,
  fp_tresh = 0,
  gfeatM = NULL,
  bulk_model = FALSE,
  pos_controls = NULL,
  rate_tol = 0.01,
  maxiter = 100,
  verbose = FALSE
)
```

Arguments

- `x`: matrix. An expression data matrix (genes in rows, cells in columns)
- `fp_tresh`: numeric. Threshold for calling a positive detection (D = 1). Default 0.
- `gfeatM`: matrix. Numeric gene level determinants of drop-out (genes in rows, features in columns)
- `bulk_model`: logical. Use median log-expression of gene in detected fraction as sole gene-level feature. Default FALSE. Ignored if gfeatM is specified.
- `pos_controls`: logical. TRUE for all genes that are known to be expressed in all cells.
- `rate_tol`: numeric. Convergence threshold on expression rates (0-1).
- `maxiter`: numeric. The maximum number of steps per gene. Default 100.
- `verbose`: logical. Whether or not to print the value of the likelihood at each iteration.

Value

A list with the following elements:

- W coefficients of sample-specific logistic drop-out model
- Alpha intercept and gene-level parameter matrix
- X intercept
• Beta coefficient of gene-specific logistic expression model
• fnr_character the probability, per gene, of P(D=0|E=1)
• p_nodrop 1 - the probability P(drop|Y), useful as weights in weighted PCA
• expected_state the expected value E[Z] (1 = "on")
• loglik the log-likelihood
• convergence for all genes, 0 if the algorithm converged and 1 if maxiter was reached

Examples

```r
mat <- matrix(rpois(1000, lambda = 3), ncol=10)
mat = mat * matrix(1-rbinom(1000, size = 1, prob = .01), ncol=10)
ziber_out = suppressWarnings(fast_estimate_ziber(mat,
bulk_model = TRUE,
pos_controls = 1:10))
```

---

**FQ_FN**

*Full-quantile normalization wrapper function*

**Description**

Full-quantile normalization wrapper function

**Usage**

FQ_FN(ei)

FQT_FN(ei)

**Arguments**

ei Numerical matrix. (rows = genes, cols = samples).

**Details**

SCONE "scaling" wrapper for `normalizeQuantileRank.matrix`.

Unlike FQ_FN, FQT_FN handles ties carefully (see `normalizeQuantiles` for details).

**Value**

Full-quantile normalized matrix.
get_bio

Examples

```r
ei <- matrix(0:20, nrow = 7)
eo <- FQ_FN(ei)

ei <- matrix(0:20, nrow = 7)
eo <- FQT_FN(ei)
```

---

**Description**

Get Factor of Biological Conditions and Batch

**Usage**

```r
get_bio(x)
get_batch(x)
```

```r
## S4 method for signature 'SconeExperiment'
get_bio(x)
```

```r
## S4 method for signature 'SconeExperiment'
get_batch(x)
```

**Arguments**

- `x` an object of class `SconeExperiment`.

**Value**

NULL or a factor containing bio or batch covariate.

**Examples**

```r
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat, bio = factor(rep(c(1,2), each = 5)),
                        batch = factor(rep(c(1,2), times = 5)))
bio = get_bio(obj)
batch = get_batch(obj)
```
get_design

Retrieve Design Matrix

Description

Given a SconeExperiment object created by a call to scone, it will return the design matrix of the selected method.

Usage

get_design(x, method)

## S4 method for signature 'SconeExperiment,character'
get_design(x, method)

## S4 method for signature 'SconeExperiment,numeric'
get_design(x, method)

Arguments

x a SconeExperiment object containing the results of scone.
method character or numeric. Either a string identifying the normalization scheme to be retrieved, or a numeric index with the rank of the normalization method to retrieve (according to scone ranking of normalizations).

Details

The numeric method will always return the design matrix corresponding to row method of the scone_params slot. This means that if scone was run with eval=TRUE, get_design(x, 1) will return the top ranked method. If scone was run with eval=FALSE, get_design(x, 1) will return the first normalization in the order saved by scone.

Value

The design matrix.

Functions

- get_design,SconeExperiment,character-method: If method is a character, it will return the design matrix corresponding to the normalization scheme specified by the character string. The string must be one of the row.names of the slot scone_params.
- get_design,SconeExperiment,numeric-method: If method is a numeric, it will return the design matrix according to the scone ranking.
Examples

```r
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat, bio = factor(rep(c(1,2),each = 5)),
                        batch = factor(rep(c(1,2),times = 5)))
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN),
             evaluate=TRUE, k_ruv=0, k_qc=0,
             adjust_batch = "yes", adjust_bio = "yes",
             eval_kclust=2, bpparam = BiocParallel::SerialParam())
design_top = get_design(res,1)
```

get_negconruv

Get Negative and Positive Controls

Description

Get Negative and Positive Controls

Usage

- `get_negconruv(x)`
- `get_negconeval(x)`
- `get_poscon(x)`

Arguments

- `x` an object of class `SconeExperiment`.

Value

NULL or a logical vector.

For `get_negconruv` the returned vector indicates which genes are negative controls to be used for RUV.
For `get_negconeval` the returned vector indicates which genes are negative controls to be used for evaluation.

For `get_poscon` the returned vector indicates which genes are positive controls to be used for evaluation.

Examples

```r
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat, negcon_ruv = 1:50 %in% 1:10,
                        negcon_eval = 1:50 %in% 11:20,
                        poscon = 1:50 %in% 21:30)
negcon_ruv = get_negconruv(obj)
negcon_eval = get_negconeval(obj)
poscon = get_poscon(obj)
```

---

**get_normalized**

*Retrieve Normalized Matrix*

---

**Description**

Given a `SconeExperiment` object created by a call to `scone`, it will return a matrix of normalized counts (in log scale if `log=TRUE`).

**Usage**

```r
get_normalized(x, method, ...)  
## S4 method for signature 'SconeExperiment,character'
get_normalized(x, method, log = FALSE)  
## S4 method for signature 'SconeExperiment,numeric'
get_normalized(x, method, log = FALSE)
```

**Arguments**

- `x` a `SconeExperiment` object containing the results of `scone`.
- `method` character or numeric. Either a string identifying the normalization scheme to be retrieved, or a numeric index with the rank of the normalization method to retrieve (according to `scone` ranking of normalizations).
- `...` additional arguments for specific methods.
- `log` logical. Should the data be returned in log-scale
Details

If `scone` was run with `return_norm="in_memory"`, this function simply retrieves the normalized data from the assays slot of object.

If `scone` was run with `return_norm="hdf5"`, this function will read the normalized matrix from the specified hdf5 file.

If `scone` was run with `return_norm="no"`, this function will compute the normalized matrix on the fly.

The numeric method will always return the normalization corresponding to row method of the `scone_params` slot. This means that if `scone` was run with `eval=TRUE`, `get_normalized(x, 1)` will return the top ranked method. If `scone` was run with `eval=FALSE`, `get_normalized(x, 1)` will return the first normalization in the order saved by `scone`.

Value

A matrix of normalized counts in log-scale.

Functions

- `get_normalized,SconeExperiment,character-method`: If method is a character, it will return the normalized matrix corresponding to the normalization scheme specified by the character string. The string must be one of the `row.names` of the slot `scone_params`.

- `get_normalized,SconeExperiment,numeric-method`: If method is a numeric, it will return the normalized matrix according to the `scone` ranking.

Examples

```r
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN),
              evaluate=TRUE, k_ruv=0, k_qc=0,
              eval_kclust=2, bpparam = BiocParallel::SerialParam())
top_norm = get_normalized(res,1)
```

---

**get_params**

Extract scone parameters

Description

Extract scone parameters
get_qc

Usage

get_params(x)

## S4 method for signature 'SconeExperiment'
get_params(x)

Arguments

x an object of class SconeExperiment.

Value

A data.frame containing workflow parameters for each scone workflow.

Examples

set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN),
                run = FALSE, k_ruv=0, k_qc=0, eval_kclust=2)
params = get_params(res)

get_qc

Get Quality Control Matrix

Description

Get Quality Control Matrix

Usage

get_qc(x)

## S4 method for signature 'SconeExperiment'
get_qc(x)

Arguments

x an object of class SconeExperiment.

Value

NULL or the quality control (QC) metric matrix.
Examples

```
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat,
    qc = cbind(colSums(mat),colSums(mat > 0)))
qc = get_qc(obj)
```

get_scores

Extract scone scores

Description

Extract scone scores

Usage

```
get_scores(x)
get_score_ranks(x)
```

## S4 method for signature 'SconeExperiment'
get_scores(x)

## S4 method for signature 'SconeExperiment'
get_score_ranks(x)

Arguments

x

an object of class SconeExperiment.

Value

get_scores returns a matrix with all (non-missing) scone scores, ordered by average score rank.

get_score_ranks returns a vector of average score ranks.

Examples

```
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN),
    evaluate=TRUE, k_ruv=0, k_qc=0,
    eval_kclust=2, bpparam = BiocParallel::SerialParam())
scores = get_scores(res)
score_ranks = get_score_ranks(res)
```
Description
This function is used to impute the data, weighted by probability of data coming from the zero-inflation part of the distribution.

Usage
impute_expectation(expression, impute_args)

Arguments
expression the data matrix (genes in rows, cells in columns)
impute_args arguments for imputation (see details)

Details
The imputation is carried out with the following formula: 
\[ y_{ij}^* = y_{ij} \times \Pr(\text{No Drop} \mid y_{ij}) + \mu_i \times \Pr(\text{Drop} \mid y_{ij}). \]

impute_args must contain 2 elements: 
1) p_nodrop = posterior probability of data not having resulted from drop-out (genes in rows, cells in columns)
2) mu = expected expression of dropped data (genes in rows, cells in columns)

Value
the imputed expression matrix.

Examples
mat <- matrix(rpois(1000, lambda = 3), ncol=10)
mat = mat * matrix(1-rbinom(1000, size = 1, prob = .01), ncol=10)
mu = matrix(rep(3/ppois(0,lambda = 3,lower.tail = FALSE),1000),ncol = 10)
p_false = 1 / ( 1 + ppois(0, lambda = 3, lower.tail = TRUE ) / (0.01 * ppois(0, lambda = 3, lower.tail = FALSE) ) )
p_nodrop = matrix(rep(1-p_false,1000),ncol = 10)
p_nodrop[mat > 0] = 1
impute_args = list()
impute_args = list(mu = mu, p_nodrop = p_nodrop)
imat = impute_expectation(mat,impute_args = impute_args)
**impute_null**  
*Null or no-op imputation*

**Description**
Null or no-op imputation

**Usage**
`impute_null(expression, impute_args)`

**Arguments**
- `expression` the data matrix (genes in rows, cells in columns)
- `impute_args` arguments for imputation (not used)

**Value**
the imputed expression matrix.

**Examples**
```r
mat <- matrix(rpois(1000, lambda = 5), ncol=10)
imat = impute_null(mat)
```

---

**lm_adjust**  
*Linear Adjustment Normalization*

**Description**
Given a matrix with log expression values and a design matrix, this function fits a linear model and removes the effects of the batch factor as well as of the linear variables encoded in W.

**Usage**
`lm_adjust(log_expr, design_mat, batch = NULL, weights = NULL)`

**Arguments**
- `log_expr` matrix. The log gene expression (genes in row, samples in columns).
- `design_mat` matrix. The design matrix (usually the result of make_design).
- `batch` factor. A factor with the batch information, identifying batch effect to be removed.
- `weights` matrix. A matrix of weights.
Details
The function assumes that the columns of the design matrix corresponding to the variable for which expression needs to be adjusted, start with either the word "batch" or the letter "W" (case sensitive). Any other covariate (including the intercept) is kept.

Value
The corrected log gene expression.

Examples
```r
set.seed(141)
bio = as.factor(rep(c(1,2), each = 2))
batch = as.factor(rep(c(1,2),2))
design_mat = make_design(bio, batch, W = NULL)

log_expr = matrix(rnorm(20), ncol = 4)
adjusted_log_expr = lm_adjust(log_expr = log_expr,
                              design_mat = design_mat,
                              batch = batch)
```

make_design

Make a Design Matrix

Description
This function builds a design matrix for the Adjustment Normalization Step, in which covariates are two (possibly nested) categorical factors and one or more continuous variables.

Usage
```r
make_design(bio, batch, W, nested = FALSE)
```

Arguments
- **bio**: factor. The biological covariate.
- **batch**: factor. The batch covariate.
- **W**: numeric. Either a vector or matrix containing one or more continuous covariates (e.g. RUVg factors).
- **nested**: logical. Whether or not to consider a nested design (see details).

Details
If nested=TRUE a nested design is used, i.e. the batch variable is assumed to be nested within the bio variable. Here, nested means that each batch is composed of samples from only *one* level of bio, while each level of bio may contain multiple batches.
Value

The design matrix.

Examples

```r
bio = as.factor(rep(c(1,2),each = 2))
batch = as.factor(rep(c(1,2),2))
design_mat = make_design(bio, batch, W = NULL)
```

---

**Metric-based Sample Filtering: Function to filter single-cell RNA-Seq libraries.**

Description

This function returns a sample-filtering report for each cell in the input expression matrix, describing which filtering criteria are satisfied.

Usage

```r
metric_sample_filter(expr, nreads = colSums(expr), ralign = NULL, gene_filter = NULL, pos_controls = NULL, scale. = FALSE, glen = NULL, AUC_range = c(0, 15), zcut = 1, mixture = TRUE, dip_thresh = 0.05, hard_nreads = 25000, hard_ralign = 15, hard_breadth = 0.2, hard_auc = 10, suff_nreads = NULL, suff_ralign = NULL, suff_breadth = NULL, suff_auc = NULL, plot = FALSE, hist_breaks = 10,
...)
```
**Arguments**

expr
matrix The data matrix (genes in rows, cells in columns).

nreads
A numeric vector representing number of reads in each library. Default to 'colSums' of 'expr'.

ralign
A numeric vector representing the proportion of reads aligned to the reference genome in each library. If NULL, filtered_ralign will be returned NA.

gene_filter
A logical vector indexing genes that will be used to compute library transcriptome breadth. If NULL, filtered_breadth will be returned NA.

pos_controls
A logical, numeric, or character vector indicating positive control genes that will be used to compute false-negative rate characteristics. If NULL, filtered_fnr will be returned NA.

scale
logical. Will expression be scaled by total expression for FNR computation? Default = FALSE

glen
Gene lengths for gene-length normalization (normalized data used in FNR computation).

AUC_range
An array of two values, representing range over which FNR AUC will be computed (log(expr_units)). Default c(0,15)

zcut
A numeric value determining threshold Z-score for sd, mad, and mixture sub-criteria. Default 1. If NULL, only hard threshold sub-criteria will be applied.

mixture
A logical value determining whether mixture modeling sub-criterion will be applied per primary criterion (metric). If true, a dip test will be applied to each metric. If a metric is multimodal, it is fit to a two-component normal mixture model. Samples deviating zcut sd’s from optimal mean (in the inferior direction), have failed this sub-criterion.

dip_thresh
A numeric value determining dip test p-value threshold. Default 0.05.

hard_nreads
numeric. Hard (lower bound on) nreads threshold. Default 25000.

hard_ralign

hard_breadth
numeric. Hard (lower bound on) breadth threshold. Default 0.2.

hard_auc

suff_nreads
numeric. If not null, serves as an overriding upper bound on nreads threshold.

suff_ralign
numeric. If not null, serves as an overriding upper bound on ralign threshold.

suff_breadth
numeric. If not null, serves as an overriding upper bound on breadth threshold.

suff_auc
numeric. If not null, serves as an overriding lower bound on fnr auc threshold.

plot
logical. Should a plot be produced?

hist_breaks
hist() breaks argument. Ignored if ‘plot=FALSE’.

... Arguments to be passed to methods.

**Details**

For each primary criterion (metric), a sample is evaluated based on 4 sub-criteria: 1) Hard (encoded) threshold 2) Adaptive thresholding via sd’s from the mean 3) Adaptive thresholding via mad’s from the median 4) Adaptive thresholding via sd’s from the mean (after mixture modeling) A sample must pass all sub-criteria to pass the primary criterion.
Value
A list with the following elements:

- filtered_nreads Logical. Sample has too few reads.
- filtered_ralign Logical. Sample has too few reads aligned.
- filtered_breadth Logical. Sample has too few genes detected (low breadth).
- filtered_fnr Logical. Sample has a high FNR AUC.

Examples

mat <- matrix(rpois(1000, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
qc = as.matrix(cbind(colSums(mat),colSums(mat > 0)))
rownames(qc) = colnames(mat)
colnames(qc) = c("NCOUNTS","NGENES")
mfilt = metric_sample_filter(expr = mat, nreads = qc[,"NCOUNTS"],
plot = TRUE, hard_nreads = 0)

PsiNorm

PsiNorm: scaling normalization based on the Pareto distribution

Description
Normalization of a raw counts matrix using the estimate of the shape parameter of the Pareto distribution.

Usage

PsiNorm(x, ...)

## S4 method for signature 'SummarizedExperiment'
PsiNorm(x, whichAssay = 1, assayName = "PsiNorm")

## S4 method for signature 'SingleCellExperiment'
PsiNorm(x, whichAssay = "counts")

## S4 method for signature 'ANY'
PsiNorm(x)

Arguments

x A SingleCellExperiment/SummarizedExperiment object or a matrix=like object with genes in rows and samples in columns.
...
generic argument
whichAssay if x is a SingleCellExperiment/SummarizedExperiment the assay with the counts to normalize (default to 1).
assayName if x is a SummarizedExperiment the name of the assay in which to save the normalized data (default to "PsiNorm").

Value

If the input is a SingleCellExperiment object the function returns the same object adding as size-Factors those computed by PsiNorm. If the object is a SummarizedExperiment object, the function returns the same object adding an assay with the normalized count matrix. If the input is a matrix-like object PsiNorm returns a matrix with the same dimensions containing the normalized counts.

Author(s)

Matteo Borella and Davide Risso

Examples

m <- matrix(c(1,0,2,0,2,9,3,0), ncol=2)
sce <- SingleCellExperiment::SingleCellExperiment(assays=list(counts=m))

sce <- PsiNorm(sce) # SingleCellExperiment object
norm.matrix <- PsiNorm(m) # normalized matrix object

PSINORM_FN PsiNorm normalization wrapper

Description

PsiNorm normalization wrapper

Usage

PSINORM_FN(ei)

Arguments

ei Numerical matrix. (rows = genes, cols = samples).

Details

SCONE scaling wrapper for PsiNorm.

Value

PsiNorm normalized matrix.

Examples

ei <- matrix(c(1,0,2,0,2,9,3,0), ncol=2)
eo <- PSINORM_FN(ei)
Normalize Expression Data and Evaluate Normalization Performance

Description

This function applies and evaluates a variety of normalization schemes with respect to a specified SconeExperiment containing scRNA-Seq data. Each normalization consists of three main steps:

- **Impute**: Replace observations of zeroes with expected expression values.
- **Scale**: Match sample-specific expression scales or quantiles.
- **Adjust**: Adjust for sample-level batch factors / unwanted variation.

Following completion of each step, the normalized expression matrix is scored based on SCONE’s data-driven evaluation criteria.

Usage

```r
scone(x, ...)

## S4 method for signature 'SconeExperiment'
scone(
x, 
imputation = list(none = impute_null),
impute_args = NULL,
zero = c("none", "preadjust", "postadjust", "strong"),
scaling,
k_ruv = 5,
k_qc = 5,
adjust_bio = c("no", "yes", "force"),
adjust_batch = c("no", "yes", "force"),
run = TRUE,
evaluate = TRUE,
eval_pcs = 3,
eval_proj = NULL,
eval_proj_args = NULL,
eval_kclust = 2:10,
verbose = FALSE,
stratified_pam = FALSE,
stratified_cor = FALSE,
stratified_rle = FALSE,
return_norm = c("no", "in_memory", "hdf5"),
hdf5file,
bpparam = BiocParallel::bpparam()
)
```
Arguments

\textbf{x} \quad \text{a} \ SconeExperiment \ \text{object.}

... \quad \text{see specific S4 methods for additional arguments.}

imputation \quad \text{list or function. (A list of) function(s) to be used for imputation. By default only scone::impute_null is included.}

impute_args \quad \text{arguments passed to all imputation functions.}

zero \quad \text{character. Zero-handling option, see Details.}

scaling \quad \text{list or function. (A list of) function(s) to be used for scaling normalization step.}

k_ruv \quad \text{numeric. The maximum number of factors of unwanted variation. Adjustment step models will include a range of 1 to k_ruv factors of unwanted variation. If 0, RUV adjustment will not be performed.}

k_qc \quad \text{numeric. The maximum number of quality metric PCs. Adjustment step models will include a range of 1 to k_qc quality metric PCs. If 0, QC factor adjustment will not be performed.}

adjust_bio \quad \text{character. If 'no', bio will not be included in Adjustment step models; if 'yes', both models with and without 'bio' will be run; if 'force', only models with 'bio' will be run.}

adjust_batch \quad \text{character. If 'no', batch will not be included in Adjustment step models; if 'yes', both models with and without 'batch' will be run; if 'force', only models with 'batch' will be run.}

run \quad \text{logical. If FALSE the normalization and evaluation are not run, but normalization parameters are returned in the output object for inspection by the user.}

evaluate \quad \text{logical. If FALSE the normalization methods will not be evaluated.}

eval_pcs \quad \text{numeric. The number of principal components to use for evaluation. Ignored if evaluate=FALSE.}

eval_proj \quad \text{function. Projection function for evaluation (see score_matrix for details). If NULL, PCA is used for projection.}

eval_proj_args \quad \text{list. List of arguments passed to projection function as eval_proj_args.}

eval_kclust \quad \text{numeric. The number of clusters (> 1) to be used for pam tightness evaluation. If an array of integers, largest average silhouette width (tightness) will be reported. If NULL, tightness will be returned NA.}

verbose \quad \text{logical. If TRUE some messages are printed.}

stratified_pam \quad \text{logical. If TRUE then maximum ASW for PAM_SIL is separately computed for each biological-cross-batch stratum (accepting NAs), and a weighted average is returned as PAM_SIL.}

stratified_cor \quad \text{logical. If TRUE then cor metrics are separately computed for each biological-cross-batch stratum (accepts NAs), and weighted averages are returned for EXP_QC_COR, EXP_UV_COR, & EXP_WV_COR. Default FALSE.}

stratified_rle \quad \text{logical. If TRUE then rle metrics are separately computed for each biological-cross-batch stratum (accepts NAs), and weighted averages are returned for RLE_MED & RLE_IQR. Default FALSE.}
return_norm character. If "no" the normalized values will not be returned with the output object. This will create a much smaller object and may be useful for large datasets and/or when many combinations are compared. If "in_memory" the normalized values will be returned as part of the output. If "hdf5" they will be written on file using the rhdf5 package.

hdf5file character. If return_norm="hdf5", the name of the file onto which to save the normalized matrices.

bpparam object of class bpparamClass that specifies the back-end to be used for computations. See bpparam for details.

Details

If run=FALSE only the scone_params slot of the output object is populated with a data.frame, each row corresponding to a set of normalization parameters.

If x has a non-empty scone_params slot, only the subset of normalizations specified in scone_params are performed and evaluated.

The zero arguments supports 3 zero-handling options:

- none: Default. No special zero-handling.
- preadjust: Restore prior zero observations to zero following Impute and Scale steps.
- postadjust: Set prior zero observations and all negative expression values to zero following the Adjust Step.
- strong: Apply both preadjust and postadjust options.

Evaluation metrics are defined in score_matrix. Each metric is assigned a +/- signature for conversion to scores: Positive- signature metrics increase with improving performance, including BIO_SIL, PAM_SIL, and EXP_WV_COR. Negative-signature metrics decrease with improving performance, including BATCH_SIL, EXP_QC_COR, EXP_UV_COR, RLE_MED, and RLE_IQR. Scores are computed so that higer-performing methods are assigned higher scores.

Note that if one wants to include the unnormalized data in the final comparison of normalized matrices, the identity function must be included in the scaling list argument. Analogously, if one wants to include non-imputed data in the comparison, the SconeExperiment object will not contain the normalized counts.

If return_norm="hdf5", the normalized matrices will be written to the hdf5file file. This must be a string specifying (a path to) a new file. If the file already exists, it will return error. In this case, the SconeExperiment object will not contain the normalized counts.

If return_norm="no" the normalized matrices are computed to compute the scores and then discarded.

In all cases, the normalized matrices can be retrieved via the get_normalized function.

Value

A SconeExperiment object with the log-scaled normalized data matrix as elements of the assays slot, if return_norm is "in_memory", and with the performance metrics and scores.
SconeExperiment-class

See Also
get_normalized, get_design

Examples

```r
mat <- matrix(rpois(1000, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
no_results <- scone(obj, scaling=list(none=identity, uq=UQ_FN, deseq=DESEQ_FN),
                   run=FALSE, k_ruv=0, k_qc=0, eval_kclust=2)

results <- scone(obj, scaling=list(none=identity, uq=UQ_FN, deseq=DESEQ_FN),
                  run=TRUE, k_ruv=0, k_qc=0, eval_kclust=2,
                  bpparam = BiocParallel::SerialParam())

results_in_memory <- scone(obj, scaling=list(none=identity, uq=UQ_FN, deseq=DESEQ_FN),
                          k_ruv=0, k_qc=0, eval_kclust=2,
                          return_norm = "in_memory",
                          bpparam = BiocParallel::SerialParam())
```

SconeExperiment-class

Class SconeExperiment

Description

Objects of this class store, at minimum, a gene expression matrix and a set of covariates (sample metadata) useful for running `scone`. These include, the quality control (QC) metrics, batch information, and biological classes of interest (if available).

The typical way of creating SconeExperiment objects is via a call to the `SconeExperiment` function or to the `scone` function. If the object is a result to a `scone` call, it will contain the results, e.g., the performance metrics, scores, and normalization workflow comparisons. (See Slots for a full list).

This object extends the `SummarizedExperiment` class.

The constructor SconeExperiment creates an object of the class SconeExperiment.

Usage

```r
SconeExperiment(object, ...)
```

## S4 method for signature 'SummarizedExperiment'
SconeExperiment(
  object, 
  which_qc = integer(),
)
which_bio = integer(),
which_batch = integer(),
which_negconruv = integer(),
which_negconeval = integer(),
which_poscon = integer(),
is_log = FALSE
)

## S4 method for signature 'matrix'
SconeExperiment(
  object,
  qc,
  bio,
  batch,
  negcon_ruv = NULL,
  negcon_eval = negcon_ruv,
  poscon = NULL,
  is_log = FALSE
)

Arguments

object Either a matrix or a SummarizedExperiment containing the raw gene expression.

... see specific S4 methods for additional arguments.

which_qc index that specifies which columns of ‘colData’ correspond to QC measures.

which_bio index that specifies which column of ‘colData’ corresponds to ‘bio’.

which_batch index that specifies which column of ‘colData’ corresponds to ‘batch’.

which_negconruv index that specifies which column of ‘rowData’ has information on negative controls for RUV.

which_negconeval index that specifies which column of ‘rowData’ has information on negative controls for evaluation.

which_poscon index that specifies which column of ‘rowData’ has information on positive controls.

is_log are the expression data in log scale?

qc numeric matrix with the QC measures.

bio factor with the biological class of interest.

batch factor with the batch information.

negcon_ruv a logical vector indicating which genes to use as negative controls for RUV.

negcon_eval a logical vector indicating which genes to use as negative controls for evaluation.

poscon a logical vector indicating which genes to use as positive controls.
Details

The QC matrix, biological class, and batch information are stored as elements of the 'colData' of the object.

The positive and negative control genes are stored as elements of the 'rowData' of the object.

Value

A SconeExperiment object.

Slots

which qc integer. Index of columns of 'colData' that contain the QC metrics.

which_bio integer. Index of the column of 'colData' that contains the biological classes information (it must be a factor).

which_batch integer. Index of the column of 'colData' that contains the batch information (it must be a factor).

which_negconruv integer. Index of the column of 'rowData' that contains a logical vector indicating which genes to use as negative controls to infer the factors of unwanted variation in RUV.

which_negconeval integer. Index of the column of 'rowData' that contains a logical vector indicating which genes to use as negative controls to evaluate the performance of the normalizations.

which_poscon integer. Index of the column of 'rowData' that contains a logical vector indicating which genes to use as positive controls to evaluate the performance of the normalizations.

hdf5_pointer character. A string specifying to which file to write / read the normalized data.

imputation_fn list of functions used by scone for the imputation step.

scaling_fn list of functions used by scone for the scaling step.

scone_metrics matrix. Matrix containing the "raw" performance metrics. See scone for a description of each metric.

scone_scores matrix. Matrix containing the performance scores (transformed metrics). See scone for a discussion on the difference between scores and metrics.

scone_params data.frame. A data frame containing the normalization schemes applied to the data and compared.

scone_run character. Whether scone was run and in which mode ("no", "in_memory", "hdf5").

is_log logical. Are the expression data in log scale?

nested logical. Is batch nested within bio? (Automatically set by scone).

rezero logical. TRUE if scone was run with zero="preadjust" or zero="strong".

fixzero logical. TRUE if scone was run with zero="postadjust" or zero="strong".

impute_args list. Arguments passed to all imputation functions.

See Also

generate normalization, get_params, get_batch, get_bio, get_design, get_negconeval, get_negconruv, get_poscon, get_qc, get_scores, and get_score_ranks to access internal fields, select_methods for subsetting by method, and scone for running scone workflows.
Examples

```r
set.seed(42)
nrows <- 200
ncols <- 6
counts <- matrix(rpois(nrows * ncols, lambda=10), nrows)
rowdata <- data.frame(poscon=c(rep(TRUE, 10), rep(FALSE, nrows-10)))
coldata <- data.frame(bio=gl(2, 3))
se <- SummarizedExperiment(assays=SimpleList(counts=counts),
                             rowData=rowdata, colData=coldata)

scone1 <- SconeExperiment(assay(se), bio=coldata$bio, poscon=rowdata$poscon)

scone2 <- SconeExperiment(se, which_bio=1L, which_poscon=1L)
```

---

**sconeReport**  
*SCONE Report Browser: Browse Evaluation of Normalization Performance*

**Description**

This function opens a shiny application session for visualizing performance of a variety of normalization schemes.

**Usage**

```r
sconeReport(
  x,
  methods,
  qc,
  bio = NULL,
  batch = NULL,
  poscon = character(),
  negcon = character(),
  eval_proj = NULL,
  eval_proj_args = NULL
)
```

**Arguments**

- **x**: a `SconeExperiment` object
- **methods**: character specifying the normalizations to report.
- **qc**: matrix. QC metrics to be used for QC evaluation report. Required.
- **bio**: factor. A biological condition (variation to be preserved). Default NULL.
- **batch**: factor. A known batch variable (variation to be removed). Default NULL.
poscon character. Genes to be used as positive controls for evaluation. These genes should be expected to change according to the biological phenomenon of interest. Default empty character.

negcon character. Genes to be used as negative controls for evaluation. These genes should be expected not to change according to the biological phenomenon of interest. Default empty character.

eval_proj function. Projection function for evaluation (see score_matrix for details). If NULL, PCA is used for projection.

eval_proj_args list. List of args passed to projection function as eval_proj_args.

Value
An object that represents the SCONE report app.

Examples
set.seed(101)
mat <- matrix(rpois(1000, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN, deseq=DESEQ_FN),
             evaluate=TRUE, k_ruv=0, k_qc=0, eval_kclust=2,
             bpparam = BiocParallel::SerialParam())
qc = as.matrix(cbind(colSums(mat),colSums(mat > 0)))
rownames(qc) = colnames(mat)
colnames(qc) = c("NCOUNTS","NGENES")
## Not run:
sconeReport(res,rownames(get_params(res)), qc = qc)
## End(Not run)
out_dir = getwd(),
seed = 112233,
filt_cells = TRUE,
filt_genes = TRUE,
always_keep_genes = NULL,
fnr_maxiter = 1000,
norm_impute = c("yes", "no", "force"),
norm_scaling = c("none", "sum", "deseq", "tmm", "uq", "fq", "detect"),
norm_rezero = FALSE,
norm_k_max = NULL,
norm_qc_expl = 0.5,
norm_adjust_bio = c("yes", "no", "force"),
norm_adjust_batch = c("yes", "no", "force"),
eval_dim = NULL,
eval_expr_expl = 0.1,
eval_poscon = NULL,
eval_negcon = negcon,
eval_max_kclust = 10,
eval_stratified_pam = TRUE,
report_num = 13,
out_rda = FALSE,
... )

Arguments

expr matrix. The expression data matrix (genes in rows, cells in columns).
qc data frame. The quality control (QC) matrix (cells in rows, metrics in columns) to be used for filtering, normalization, and evaluation.
biocfactor. The biological condition to be modeled in the Adjustment Step as variation to be preserved. If adjust_bio="no", it will not be used for normalization, but only for evaluation.
batch factor. The known batch variable to be included in the adjustment model as variation to be removed. If adjust_batch="no", it will not be used for normalization, but only for evaluation.
negcon character. The genes to be used as negative controls for filtering, normalization, and evaluation. These genes should be expressed uniformly across the biological phenomenon of interest. Default NULL.
verbose character. Verbosity level: higher level is more verbose. Default "0".
out_dir character. Output directory. Default getwd().
seed numeric. Random seed. Default 112233.
filt_cells logical. Should cells be filtered? Set to FALSE if low quality cells have already been excluded. If cells are not filtered, then initial gene filtering (the one that is done prior to cell filtering) is disabled as it becomes redundant with the gene filtering that is done after cell filtering. Default TRUE.
filt_genes logical. Should genes be filtered post-sample filtering? Default TRUE.
always_keep_genes logical. A character vector of gene names that should never be excluded (e.g., marker genes). Default NULL.

fnr_maxiter numeric. Maximum number of iterations in EM estimation of expression posteriors. If 0, then FNR estimation is skipped entirely, and as a consequence no imputation will be performed, disregarding the value of the "norm_impute" argument. Default 1000.

norm_impute character. Should imputation be included in the comparison? If 'force', only imputed normalizations will be run. Default "yes."


norm_rezero logical. Restore prior zeroes and negative values to zero following normalization. Default FALSE.

norm_k_max numeric. Max number (norm_k_max) of factors of unwanted variation modeled in the Adjustment Step. Default NULL.

norm_qc_expl numeric. In automatic selection of norm_k_max, what fraction of variation must be explained by the first norm_k_max PCs of qc? Default 0.5. Ignored if norm_k_max is not NULL.

norm_adjust_bio character. If 'no' it will not be included in the model; if 'yes', both models with and without 'bio' will be run; if 'force', only models with 'bio' will be run. Default "yes."

norm_adjust_batch character. If 'no' it will not be modeled in the Adjustment Step; if 'yes', both models with and without 'batch' will be run; if 'force', only models with 'batch' will be run. Default "yes."

eval_dim numeric. The number of principal components to use for evaluation. Default NULL.

eval_expr_expl numeric. In automatic selection of eval_dim, what fraction of variation must be explained by the first eval_dim PCs of expr? Default 0.1. Ignored if eval_dim is not NULL.

eval_poscon character. The genes to be used as positive controls for evaluation. These genes should be expected to change according to the biological phenomenon of interest.

eval_negcon character. Alternative negative control gene list for evaluation only.

eval_max_kclust numeric. The max number of clusters (> 1) to be used for pam tightness evaluation. If NULL, tightness will be returned NA.

eval_stratified_pam logical. If TRUE then maximum ASW for PAM_SIL is separately computed for each biological-cross-batch condition (accepting NAs), and a weighted average is returned as PAM_SIL. Default TRUE.

Logical. If TRUE, `sconeResults.Rda` file with the object that the `scone` function returns is saved in the `out_dir` (may be very large for large datasets, but useful for post-processing) Default FALSE.

Extra params passed to the `metric_sample_filter` and `scone` when they’re called by `easybake`.

Details

"ADD DESCRIPTION"

Value

Directory structure "ADD DESCRIPTION"

Examples

```r
set.seed(101)
mat <- matrix(rpois(1000, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN, deseq=DESEQ_FN),
             evaluate=TRUE, k_ruv=0, k_qc=0, eval_kclust=2,
             bpparam = BiocParallel::SerialParam())
qc = as.matrix(cbind(colSums(mat),colSums(mat > 0)))
rownames(qc) = colnames(mat)
colnames(qc) = c("NREADS","RALIGN")
## Not run:
scone_easybake(mat, qc = as.data.frame(qc), verbose = "2",
               norm_adjust_bio= "no",
               norm_adjust_batch= "no", norm_k_max = 0,
               fnr_maxiter = 0, filt_cells=FALSE, filt_genes=FALSE,
               eval_stratified_pam = FALSE,
               out_dir="/scone_out")
## End(Not run)
```
score_matrix

Usage

score_matrix(
  expr,
  eval_pcs = 3,
  eval_proj = NULL,
  eval_proj_args = NULL,
  eval_kclust = NULL,
  bio = NULL,
  batch = NULL,
  qc_factors = NULL,
  uv_factors = NULL,
  wv_factors = NULL,
  is_log = FALSE,
  stratified_pam = FALSE,
  stratified_cor = FALSE,
  stratified_rle = FALSE
)

Arguments

expr matrix. The expression data matrix (genes in rows, cells in columns).
eval_pcs numeric. The number of principal components to use for evaluation (Default 3).
  Ignored if !is.null(eval_proj).
eval_proj function. Projection function for evaluation (see Details). If NULL, PCA is used for projection.
eval_proj_args list. List of arguments passed to projection function as eval_proj_args (see Details).
eval_kclust numeric. The number of clusters (> 1) to be used for pam tightness (PAM_SIL) evaluation. If an array of integers, largest average silhouette width (tightness) will be reported in PAM_SIL. If NULL, PAM_SIL will be returned NA.
bio factor. A known biological condition (variation to be preserved), NA is allowed. If NULL, condition ASW, BIO_SIL, will be returned NA.
batch factor. A known batch variable (variation to be removed), NA is allowed. If NULL, batch ASW, BATCH_SIL, will be returned NA.
qc_factors Factors of unwanted variation derived from quality metrics. If NULL, qc correlations, EXP_QC_COR, will be returned NA.
uv_factors Factors of unwanted variation derived from negative control genes (evaluation set). If NULL, uv correlations, EXP_UV_COR, will be returned NA.
wv_factors Factors of wanted variation derived from positive control genes (evaluation set). If NULL, wv correlations, EXP_WV_COR, will be returned NA.
is_log logical. If TRUE the expr matrix is already logged and log transformation will not be carried out prior to projection. Default FALSE.
stratified_pam logical. If TRUE then maximum ASW is separately computed for each biological-cross-batch stratum (accepts NAs), and a weighted average silhouette width is returned as PAM_SIL. Default FALSE.
stratified_cor logical. If TRUE then cor metrics are separately computed for each biological-cross-batch stratum (accepts NAs), and weighted averages are returned for EXP_QC_COR, EXP_UV_COR, & EXP_WV_COR. Default FALSE.

stratified_rle logical. If TRUE then rle metrics are separately computed for each biological-cross-batch stratum (accepts NAs), and weighted averages are returned for RLE_MED & RLE_IQR. Default FALSE.

Details

Users may specify their own eval_proj function that will be used to compute Clustering and Correlation metrics. This eval_proj() function must have 2 input arguments:

- e matrix. log-transformed (+ pseudocount) expression data (genes in rows, cells in columns).
- eval_proj_args list. additional function arguments, e.g. prior data weights.

and it must output a matrix representation of the original data (cells in rows, factors in columns). The value of eval_proj_args is passed to the user-defined function from the eval_proj_args argument of the main score_matrix() function call.

Value

A list with the following metrics:

- BIO_SIL Average silhouette width by biological condition.
- BATCH_SIL Average silhouette width by batch condition.
- PAM_SIL Maximum average silhouette width from PAM clustering (see stratified_pam argument).
- EXP_QC_COR Coefficient of determination between expression pcs and quality factors (see stratified_cor argument).
- EXP_UV_COR Coefficient of determination between expression pcs and negative control gene factors (see stratified_cor argument).
- EXP_WV_COR Coefficient of determination between expression pcs and positive control gene factors (see stratified_cor argument).
- RLE_MED The mean squared median Relative Log Expression (RLE) (see stratified_rle argument).
- RLE_IQR The variance of the inter-quartile range (IQR) of the RLE (see stratified_rle argument).

Examples

```r
set.seed(141)
bio = as.factor(rep(c(1,2),each = 2))
bio = as.factor(rep(c(1,2),each = 2))
log_expr = matrix(rnorm(20),ncol = 4)

scone_metrics = score_matrix(log_expr, 
  bio = bio, batch = batch,
  eval_kclust = 2, is_log = TRUE)
```
**SCRN_FN**

*Simple deconvolution normalization wrapper*

**Description**

Simple deconvolution normalization wrapper

**Usage**

```
SCRN_FN(ei)
```

**Arguments**

- **ei**
  
  Numerical matrix. (rows = genes, cols = samples).

**Details**

SCONE scaling wrapper for `computeSumFactors`.

**Value**

scran normalized matrix.

**Examples**

```r
ei <- matrix(0:76,nrow = 7)
eo <- SCRN_FN(ei)
```

---

**select_methods**

*Get a subset of normalizations from a SconeExperiment object*

**Description**

This method let a user extract a subset of normalizations. This is useful when the original dataset is large and/or many normalization schemes have been applied.

In such cases, the user may want to run scone in mode `return_norm = "no"`, explore the results, and then select the top performing methods for additional exploration.

**Usage**

```
select_methods(x, methods)
```

```r
## S4 method for signature 'SconeExperiment,character'
select_methods(x, methods)
```

```r
## S4 method for signature 'SconeExperiment,numeric'
select_methods(x, methods)
```
Arguments

x a SconeExperiment object.

methods either character or numeric specifying the normalizations to select.

Details

The numeric method will always return the normalization corresponding to the methods rows of the scone_params slot. This means that if scone was run with eval=TRUE, select_methods(x, 1:3) will return the top three ranked method. If scone was run with eval=FALSE, it will return the first three normalization in the order saved by scone.

Value

A SconeExperiment object with selected method data.

Functions

• select_methods,SconeExperiment,character-method: If methods is a character, it will return the subset of methods named in methods (only perfect match). The string must be a subset of the row.names of the slot scone_params.

• select_methods,SconeExperiment,numeric-method: If methods is a numeric, it will return the subset of methods according to the scone ranking.

Examples

```r
set.seed(42)
mat <- matrix(rpois(500, lambda = 5), ncol=10)
colnames(mat) <- paste("X", 1:ncol(mat), sep="")
obj <- SconeExperiment(mat)
res <- scone(obj, scaling=list(none=identity, uq=UQ_FN),
    evaluate=TRUE, k_ruv=0, k_qc=0,
    eval_kclust=2, bpparam = BiocParallel::SerialParam())
select_res = select_methods(res,1:2)
```

---

**simple_FNR_params**

Fit Simple False-Negative Model

Description

Fits a logistic regression model of false negative observations as a function of expression level, using a set of positive control (ubiquitously expressed) genes

Usage

```r
simple_FNR_params(expr, pos_controls, fn_tresh = 0.01)
```
**SUM_FN**

**Description**

Sum scaling normalization function

**Usage**

```r
SUM_FN(ei)
```

**Arguments**

- `ei` Numerical matrix. (rows = genes, cols = samples).

**Details**

SCONE scaling by library size or summed expression.

**Value**

Sum-scaled normalized matrix.

**Arguments**

- `expr` matrix A matrix of transcript-proportional units (genes in rows, cells in columns).
- `pos_controls` A logical, numeric, or character vector indicating control genes that will be used to compute false-negative rate characteristics. User must provide at least 2 control genes.
- `fn_tresh` Inclusive threshold for negative detection. Default 0.01. fn_tresh must be non-negative.

**Details**

logit(Probability of False Negative) ~ a + b*(median log-expr)

**Value**

A matrix of logistic regression coefficients corresponding to glm fits in each sample (a and b in columns 1 and 2 respectively). If the a & b fit does not converge, b is set to zero and only a is estimated.

**Examples**

```r
mat <- matrix(rpois(10000, lambda = 3), ncol=10)
mat = mat * matrix(1-rbinom(1000, size = 1, prob = .01), ncol=10)
fnr_out = simple_FNR_params(mat,pos_controls = 1:10)
```
**Examples**

```r
ei <- matrix(0:20, nrow = 7)
eo <- SUM_FN(ei)
```

<table>
<thead>
<tr>
<th>TMM_FN</th>
<th>Weighted trimmed mean of M-values (TMM) scaling normalization wrapper function</th>
</tr>
</thead>
</table>

**Description**

Weighted trimmed mean of M-values (TMM) scaling normalization wrapper function

**Usage**

```r
TMM_FN(ei)
```

**Arguments**

- `ei` Numerical matrix. (rows = genes, cols = samples).

**Details**

SCONE scaling wrapper for `calcNormFactors`.

**Value**

TMM normalized matrix.

**Examples**

```r
ei <- matrix(0:20, nrow = 7)
eo <- TMM_FN(ei)
```

<table>
<thead>
<tr>
<th>UQ_FN</th>
<th>Upper-quartile (UQ) scaling normalization wrapper function</th>
</tr>
</thead>
</table>

**Description**

Upper-quartile (UQ) scaling normalization wrapper function

**Usage**

```r
UQ_FN(ei)
```
Arguments

ei Numerical matrix. (rows = genes, cols = samples).

Details

SCONE scaling wrapper for \texttt{calcNormFactors}).

Value

UQ normalized matrix.

Examples

\begin{verbatim}
  ei <- matrix(0:20,nrow = 7)
  eo <- UQ_FN(ei)
\end{verbatim}
Index

* internal
  .likfn, 3
  .parse_row, 3
  .pzfn, 4
  .likfn, 3
  .parse_row, 3
  .pzfn, 4
biplot_color, 4, 5
biplot_interactive, 5
bpparam, 30
calcNormFactors, 8, 44, 45
cellcycle_genes (control_genes), 7
clr, 6
CLR_FN, 6
computeSumFactors, 41
control_genes, 7
cortical_markers (control_genes), 7
DESEQ_FN, 8
estimate_ziber, 9
factor_sample_filter, 10
fast_estimate_ziber, 12
FQ_FN, 13
FQT_FN (FQ_FN), 13
get_batch, 33
get_batch (get_bio), 14
get_batch, SconeExperiment-method (get_bio), 14
get_bio, 14, 33
get_bio, SconeExperiment-method (get_bio), 14
get_design, 15, 31, 33
get_design, SconeExperiment, character-method (get_design), 15
get_design, SconeExperiment, numeric-method (get_design), 15
get_negconeval, 33
get_negconeval (get_negconruv), 16
get_negconeval, SconeExperiment-method (get_negconruv), 16
get_negconruv, 16, 33
get_negconruv, SconeExperiment-method (get_negconruv), 16
get_normalized, 17, 30, 31, 33
get_normalized, SconeExperiment, character-method (get_normalized), 17
get_normalized, SconeExperiment, numeric-method (get_normalized), 17
get_params, 18, 33
get_params, SconeExperiment-method (get_params), 18
get_poscon, 33
get_poscon (get_negconruv), 16
get_poscon, SconeExperiment-method (get_negconruv), 16
get_qc, 19, 33
get_qc, SconeExperiment-method (get_qc), 19
get_score_ranks, 33
get_score_ranks (get_scores), 20
get_score_ranks, SconeExperiment-method (get_scores), 20
get_scores, 20, 33
get_scores, SconeExperiment-method (get_scores), 20
housekeeping (control_genes), 7
housekeeping_revised (control_genes), 7
impute_expectation, 21
impute_null, 22
lm_adjust, 22
make_design, 23
metric_sample_filter, 24
normalizeQuantileRank.matrix, 13
normalizeQuantiles, 13

prcomp, 4, 5
PsiNorm, 26, 27
PsiNorm, ANY-method (PsiNorm), 26
PsiNorm, SingleCellExperiment-method (PsiNorm), 26
PsiNorm, SummarizedExperiment-method (PsiNorm), 26
PSINORM_FN, 27

scone, 7, 15, 17, 18, 28, 31, 33, 42
scone, SconeExperiment-method (scone), 28
scone_easybake, 35
SconeExperiment, 5, 6, 14–17, 19, 20, 29–31, 33
SconeExperiment (SconeExperiment-class), 31
SconeExperiment, matrix-method (SconeExperiment-class), 31
SconeExperiment, SummarizedExperiment-method (SconeExperiment-class), 31
SconeExperiment-class, 31
sconeReport, 34
score_matrix, 29, 30, 35, 38
SCRAN_FN, 41
select_methods, 33, 41
select_methods, SconeExperiment, character-method (select_methods), 41
select_methods, SconeExperiment, numeric-method (select_methods), 41
simple_FNR_params, 42
SUM_FN, 43
SummarizedExperiment, 31, 32

TMM_FN, 44
tolower, 7
toTitleCase, 7
toupper, 7

UQ_FN, 44