Package ‘SC3’

October 12, 2016

Type Package

Title Single-Cell Consensus Clustering

Version 1.1.4

Date 2016-05-06

Author Vladimir Kiselev

Maintainer Vladimir Kiselev <vladimir.yu.kiselev@gmail.com>

Description Interactive tool for clustering and analysis of single cell RNA-Seq data.

License GPL-3

Imports graphics, stats, utils, methods, RSelenium, e1071, parallel, foreach, doParallel, doRNG, shiny, ggplot2, pheatmap (>= 1.0.8), RColorBrewer, colorspace, ROCR, robustbase, rrcov, cluster, WriteXLS, Rtsne

Depends R(>= 3.3)

LazyData TRUE

RoxygenNote 5.0.1

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

biocViews Classification, Clustering, DimensionReduction, SupportVectorMachine, RNASeq, Visualization, Transcriptomics, DataRepresentation, GUI, DifferentialExpression, GeneSetEnrichment, Transcription

NeedsCompilation no

URL https://github.com/hemberg-lab/SC3

BugReports https://github.com/hemberg-lab/SC3/issues

R topics documented:

`calculate_distance` *Calculate a distance matrix*

Description

Distance between the cells, i.e. columns, in the input expression matrix are calculated using the Euclidean, Pearson and Spearman metrics to construct distance matrices.

Usage

```
calculate_distance(data, method)
```

Arguments

`data`	expression matrix
`method`	one of the distance metrics: "spearman", "pearson", "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski"

Value

distance matrix

`cell_filter` *Cell filter*

Description

The cell filter should be used if the quality of data is low, i.e. if one suspects that some of the cells may be technical outliers with poor coverage. The cell filter removes cells containing fewer than cell.filter.genes.

Usage

```
cell_filter(data, cell.filter.genes)
```

Arguments

`data` expression matrix

`cell.filter.genes` minimum number of genes that must be expressed in each cell, default is 2,000.

Value

Filtered expression matrix

`consensus_matrix` *Calculate consensus matrix*

Description

Consensus matrix is calculated using the Cluster-based Similarity Partitioning Algorithm (CSPA). For each clustering result a binary similarity matrix is constructed from the corresponding cell labels: if two cells belong to the same cluster, their similarity is 1, otherwise the similarity is 0. A consensus matrix is calculated by averaging all similarity matrices.

Usage

```
consensus_matrix(clusts)
```

Arguments

`clusts` a list clustering labels (separated by a space)

Value

consensus matrix

`gene_filter` *Gene filter*

Description

The gene filter removes genes that are either expressed or absent (expression value is less than 2) in at least X The motivation for the gene filter is that ubiquitous and rare genes most often are not informative for the clustering.

Usage

```
gene_filter(data, fraction)
```

Arguments

`data`	expression matrix
`fraction`	fraction of cells (1 - X/100), default is 0.06.

Value

filtered expression matrix some genes were removed.

`get_data` *Import expression matrix*

Description

Import an input expression matrix.

Usage

```
get_data(name)
```

Arguments

`name`	name of an R object or a text file

Value

expression matrix

`get_de_genes` *Find differentially expressed genes*

Description

If the cell labels are available this functions allows a user to calculate differentially expressed genes manually.

Usage

```
get_de_genes(dataset, labels, p.val = 0.01)
```

Arguments

`dataset`	expression matrix
`labels`	cell labels corresponding to the columns of the expression matrix
`p.val`	p-value threshold, by default it is 0.01

Value

a numeric vector containing the differentially expressed genes and correspoding p-values

Examples

```
d <- get_de_genes(treutlein, colnames(treutlein))
head(d)
```

`get_marker_genes` *Find marker genes*

Description

If the cell labels are available this functions allows a user to calculate marker genes manually.

Usage

```
get_marker_genes(dataset, labels, auroc.threshold = 0.85, p.val = 0.01)
```

Arguments

`dataset`	expression matrix
`labels`	cell labels corresponding to the columns of the expression matrix
`auroc.threshold`	area under the ROC curve threshold, by default it is 0.85. Values close to 0.5 will include very weak marker genes, values close to 1 will only include very strong marker genes.
`p.val`	p-value threshold, by default it is 0.01

Value

data.frame containing the marker genes

Examples

```
d <- get_marker_genes(treutlein, colnames(treutlein))
head(d)
```

`get_outl_cells` *Find cell outliers*

Description

If the cell labels are available this functions allows a user to calculate cell outlier scores manually.

Usage

```
get_outl_cells(dataset, labels, chisq.quantile = 0.9999)
```

Arguments

`dataset`	expression matrix
`labels`	cell labels corresponding to the columns of the expression matrix
`chisq.quantile`	a threshold of the chi-squared distribution used for cell outliers detection, default is 0.9999

Value

a numeric vector containing the cell labels and correspoding outlier scores ordered by the labels

Examples

```
d <- get_outl_cells(treutlein, colnames(treutlein))
head(d)
```

iwanthue *Generate a colour palette by k-means clustering of LAB colour space.*

Description

Generate a palette of distinct colours through k-means clustering of LAB colour space.

Usage

```
iwanthue(n, hmin = 0, hmax = 360, cmin = 0, cmax = 180, lmin = 0,
  lmax = 100, plot = FALSE, random = FALSE)
```

Arguments

`n`	Numeric. The number of colours to generate.
`hmin`	Numeric, in the range [0, 360]. The lower limit of the hue range to be clustered.
`hmax`	Numeric, in the range [0, 360]. The upper limit of the hue range to be clustered.
`cmin`	Numeric, in the range [0, 180]. The lower limit of the chroma range to be clustered.
`cmax`	Numeric, in the range [0, 180]. The upper limit of the chroma range to be clustered.
`lmin`	Numeric, in the range [0, 100]. The lower limit of the luminance range to be clustered.
`lmax`	Numeric, in the range [0, 100]. The upper limit of the luminance range to be clustered.
`plot`	Logical. Should the colour swatches be plotted (using `swatch`)?
`random`	Logical. If `TRUE`, clustering will be determined by the existing RNG state. If `FALSE`, the seed will be set to `1` for clustering, and on exit, the function will restore the pre-existing RNG state.

Details

Note that `iwanthue` currently doesn't support `hmin` greater than `hmax` (which should be allowed, since hue is circular).

Value

A vector of `n` colours (as hexadecimal strings), representing centers of clusters determined through k-means clustering of the LAB colour space delimited by `hmin`, `hmax`, `cmin`, `cmax`, `lmin` and `lmax`.

References

- R implementation of iwanthue by John Baumgartner
- iwanthue - colors for data scientists
- iwanthue on GitHub

See Also

swatch

norm_laplacian *Graph Laplacian calculation*

Description

Calculate graph Laplacian of a distance matrix

Usage

```
norm_laplacian(x, tau)
```

Arguments

x	adjacency/distance matrix
tau	regularization term

Value

graph Laplacian of the adjacency/distance matrix

sc3 *SC3 main function*

Description

Run SC3 clustering pipeline and starts an interactive session in a web browser.

Usage

```
sc3(filename, ks = 3:7, cell.filter = FALSE, cell.filter.genes = 2000,
  gene.filter = TRUE, gene.filter.fraction = 0.06, log.scale = TRUE,
  d.region.min = 0.04, d.region.max = 0.07, interactivity = TRUE,
  show.original.labels = FALSE, svm = FALSE, svm.num.cells = NA,
  n.cores = NA, seed = 1)
```

Arguments

`filename` either an R matrix / data.frame object OR a path to your input file containing an input expression matrix. The expression matrix must contain both colnames (cell IDs) and rownames (gene IDs).

`ks` a range of the number of clusters that needs to be tested. k.min is the minimum number of clusters (default is 3). k.max is the maximum number of clusters (default is 7).

`cell.filter` defines whether to filter cells that express less than cell.filter.genes genes (lowly expressed cells). By default it is FALSE. The cell filter should be used if the quality of data is low, i.e. if one suspects that some of the cells may be technical outliers with poor coverage. Filtering of lowly expressed cells usually improves clustering.

`cell.filter.genes`
if cell.filter is used then this parameter defines the minimum number of genes that have to be expressed in each cell (expression value > 1e-2). If there are fewer, the cell will be removed from the analysis. The default is 2000.

`gene.filter` defines whether to perform gene filtering or not. Boolean, default is TRUE.

`gene.filter.fraction`
fraction of cells (1 - X/100), default is 0.06. The gene filter removes genes that are either expressed or absent (expression value is less than 2) in at least X The motivation for the gene filter is that ubiquitous and rare genes most often are not informative for the clustering.

`log.scale` defines whether to perform log2 scaling or not. Boolean, default is TRUE.

`d.region.min` the lower boundary of the optimum region of d, default is 0.04.

`d.region.max` the upper boundary of the optimum region of d, default is 0.07.

`interactivity` defines whether a browser interactive window should be open after all computation is done. By default it is TRUE. This option can be used to separate clustering calculations from visualisation, e.g. long and time-consuming clustering of really big datasets can be run on a farm cluster and visualisations can be done using a personal laptop afterwards. If interactivity is FALSE then all clustering results will be saved as "sc3.interactive.arg" list. To run interactive visulisation with the precomputed clustering results please use sc3_interactive(sc3.interactive.arg).

`show.original.labels`
if cell labels in the dataset are not unique, but represent clusters expected from the experiment, they can be visualised by setting this parameter to TRUE. The default is FALSE.

`svm` if TRUE then an SVM prediction will be used. The default is FALSE.

`svm.num.cells` number of training cells to be used for SVM prediction. The default is NA. If the svm parameter is TRUE and svn.num.cells is not provided, then the defaults of SC3 will be used: if number of cells is more than 5000, then svn.num.cells = 1000, otherwise svn.num.cells = 20 percent of the total number of cells

`n.cores` defines the number of cores to be used on the user's machine. Default is NA.

`seed` sets seed for the random number generator, default is 1. Can be used to check the stability of clustering results: if the results are the same after changing the seed several time, then the clustering solution is stable.

Value

Opens a browser window with an interactive shine app and visualize all precomputed clusterings.

Examples

```
sc3(treutlein, 3:7, interactivity = FALSE, n.cores = 2)
```

`sc3_interactive` *SC3 interactive function*

Description

Runs interactive session of SC3 based on precomputed objects

Usage

```
sc3_interactive(input.param)
```

Arguments

`input.param`	parameters precomputed by sc3() with interactivity = FALSE (sc3.interactive.arg).

Value

Opens a browser window with an interactive shiny app and visualize all precomputed clusterings.

`StabilityIndex` *Calculate the stability index of the obtained clusters when changing k*

Description

Stability index shows how stable each cluster is accross the selected range of k. The stability index varies between 0 and 1, where 1 means that the same cluster appears in every solution for different k.

Usage

```
StabilityIndex(stab.res, k)
```

Arguments

`stab.res`	internal matrix of precomputed clustering results
`k`	current value of the number of clusters k

Details

Formula (imagine a given cluster with is split into N clusters when k is changed, and in each of the new clusters there are given_cells of the given cluster and also some extra_cells from other clusters): SI = sum_over_ks(sum_over_clusters_N(given_cells/(given_cells + extra_cells)))/N(corrects for stability of each cluster)/N(corrects for the number of clusters)/length(ks)

Value

a numeric vector containing a stability index of each cluster

`support_vector_machines` *Run support vector machines (SVM) prediction*

Description

Train an SVM classifier on training cells and then classify study cells using the classifier.

Usage

```
support_vector_machines(train, study, kern)
```

Arguments

`train`	expression matrix with training cells
`study`	expression matrix with study cells
`kern`	kernel to be used with SVM

Value

classification of study cells

`swatch` *Plot colour swatches for a vector of colours*

Description

Plot named colour swatches for a vector of colours.

Usage

```
swatch(x)
```

Arguments

`x` a vector of colours, specified as: colour names (i.e. colour names returned by `colors()`); numeric indices into `palette()`, or hexadecimal strings in the form `"#RRGGBB"`, where `RR`, `GG`, and `BB` are pairs of hexadecimal digits representing red, green, and blue components, in the range `00` to `FF`.

Value

`NULL`. The colour swatch is plotted to the active plotting device.

See Also

`iwanthue`

`transformation` *Distance matrix transformation*

Description

All distance matrices are transformed using either principal component analysis (PCA), multidimensional scaling (MDS) or by calculating the eigenvectors of the graph Laplacian (Spectral). The columns of the resulting matrices are then sorted in descending order by their corresponding eigenvalues.

Usage

```
transformation(dists, method)
```

Arguments

`dists` distance matrix

`method` transformation method: either "pca", "mds", "spectral" or "spectral_reg", where "spectral_reg" calculates graph Laplacian with regularization (tau = 1000)

Value

transformed distance matrix

`treutlein`	*Single cell RNA-Seq data extracted from a publication by Treutlein et al.*

Description

Single cell RNA-Seq data extracted from a publication by Treutlein et al.

Usage

```
treutlein
```

Format

An object of class `matrix` with 23271 rows and 80 columns.

Value

blah blah

Source

`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52583`

Columns represent cells, rows represent genes expression values. Colnames respresent indexes of cell clusters (known information based on the experimental protocol). There are 80 cells and 5 clusters in this dataset.

Index