Package 'condcomp'

April 15, 2019

Title Condition Comparison in scRNA-seq Data

Version 1.1.1

Description For a given clustered data, which can also be split into two conditions, this package provides a way to perform a condition comparison on said clustered data. The comparison is performed on each cluster. Several statistics are used and, when analysed in conjunction, they might give some insight regarding the heterogeneity of some of the clusters.

Depends R (>= 3.5.0)

License GPL (>=2) | file LICENSE

Encoding UTF-8

LazyData true

Imports cluster, ggplot2, ggrepel, outliers

RoxygenNote 6.0.1

Suggests testthat, knitr, rmarkdown, BiocStyle, Matrix, Seurat, monocle, HSMMSingleCell

biocViews ImmunoOncology, Clustering, SingleCell, Visualization

VignetteBuilder knitr

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

git_branch master

git_last_commit 8d03758

git_last_commit_date 2019-01-04

Date/Publication 2019-04-14

Author Diogo P. P. Branco [aut, cre]

Maintainer Diogo P. P. Branco <diogo.pp.branco@gmail.com>

R topics documented:

condcomp *Comparison of data conditions in a clustering*

Description

Performs a condition comparison on a given `clustering`. The comparison is performed on each cluster separately between each condcition (`cond`). Several statistics are used and, when analysed in conjunction, they might give some insight regarding the heterogeneity of some of the clusters.

Usage

```
condcomp(clustering, cond, dmatrix, n = 1000, remove.na = TRUE)
```

Arguments

`clustering`	A clustering of the data.
`cond`	A factor indicating the condition which each data point is subject to.
`dmatrix`	A distance matrix describing the data to be analysed.
`n`	The number of random silhouettes to be performed. Keep in mind that the computation of several random silhouettes is the bottleneck of this process.
`remove.na`	Logical. Remove lines with NA (i.e. clusters which the silhouette could not be computed).

Details

For a given cluster, several metrics are computed, see the 'Return' section for details about each metric. Some metrics make use of Random Silhouettes, which is defined as follows: given a labeled data set, assign a random label (from the set of labels) to each data point without changing the original ratio of groups. Then compute the `silhouette` index for this data considering these randomly assigned labels, the average silhouette width is the Random Silhouette for the data (with randomly assigned labels). Being an stochastic process, the Monte Carlo approach is applied which gives a vector of several Random Silhouettes.

Value

A data frame with various statistics regarding data heterogeneity inside each cluster.

Each row of the data frame contains several metrics regarding the conditions found in an specific cluster. For now only a maximum of two conditions are supported. These metrics are described below:

x_perc Numeric. The percentage of data points belonging to condition 'x'.

x_ratio Numeric. The ratio of data points belonging to condition 'x'. For example, considering another condition 'y', the 'x_ratio' would be computed as x_perc/y_perc.

true_sil Numeric. True silhouette. The silhouette for the data in this cluster considering the conditions, as defined by the parameter `cond`, as groups.

zscore Numeric. The Z-score computed based on the `silhouette`. See the 'Details' section.

pval Numeric. The p-value for 'true_sil'. Computed from the number of Random Silhouettes (see 'Details') that are greater than the 'true_sil' for this cluster.

iqr Factor. Interquartile Range (IQR) based outlier detection. Considering the vector including the random silhouettes (see 'Details') and the 'true_sil', the method checks whether 'true_sil' is an outlier in said vector. This will be set to 'Diff' in case 'true_sil' is an outlier or 'Same' otherwise.

Examples

```
clustering <- iris$Species
dmatrix <- as.matrix(dist(iris[-length(iris)]))
# Suppose the conditions are 'young' and 'old' fish
cond <- sample(c("young", "old"), nrow(iris), replace=TRUE)
comp <- condcomp(clustering, cond, dmatrix=dmatrix, n=10)
```

---

condcompPlot *Plots the data frame of comparison of data conditions*

Description

This function takes the output from condcomp and plots some of its attributes in a scatter plot.

Usage

```
condcompPlot(ccomp, col = ccomp$iqr, main = NULL, legend.title = "IQR")
```

Arguments

ccomp	A data frame output from condcomp.
col	Color parameter to be used. The default is to color according to the IQR column of ccomp.
main	Character vector (or expression) giving plot title.
legend.title	Character vector giving the legend title.

Details

The first condition ratio that appears in the data frame will be plotted in the y-axis (-log10 scale), whereas the Z-score will be plotted along the x-axis. Each group will be colored by their respective IQR value as shown in the legend.

Value

A ggplot2 object.

Examples

```
clustering <- iris$Species
dmatrix <- as.matrix(dist(iris[-length(iris)]))
# Suppose the conditions are 'young' and 'old' fish
cond <- sample(c("young", "old"), nrow(iris), replace=TRUE)
comp <- condcomp(clustering, cond, dmatrix=dmatrix, n=10)
condcompPlot(comp)
```

Index