# Package 'ccRemover'

October 12, 2022

**Type** Package

**Title** Removes the Cell-Cycle Effect from Single-Cell RNA-Sequencing Data

**Version** 1.0.4

**Description** Implements a method for identifying and removing the cell-cycle effect from scRNA-Seq data. The description of the method is in Barron M. and Li J. (2016) <doi:10.1038/srep33892>. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. Submitted. Different from previous methods, ccRemover implements a mechanism that formally tests whether a component is cell-cycle related or not, and thus while it often thoroughly removes the cell-cycle effect, it preserves other features/signals of interest in the data.

**Depends** R (>= 2.10.0)

**License** GPL-3

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.0.1

**Suggests** knitr, rmarkdown

**VignetteBuilder** knitr

**Imports** stats, utils

**NeedsCompilation** no

**Author** Jun Li [aut, cre], Martin Barron [aut]

**Maintainer** Jun Li <jun.li@nd.edu>

**Repository** CRAN

**Date/Publication** 2017-08-19 12:23:32 UTC

## R topics documented:

---

`bootstrap_diff` *Calculates the difference in the loading score for cell-cycle and control genes*

---

## Description

This function is only used internally inside ccRemover. The function calcualtes the average load difference on the cell-cycle and control genes. Bootstrap resampling is then used to provide a score for each component. Please see the original manuscript for the mathematical details.

## Usage

```
bootstrap_diff(xy, xn, nboot = 200, bar = TRUE)
```

## Arguments

| | |
|---|---|
| `xy` | The data for the genes which are annotated to the cell-cycle, i.e. those genes for which "if_cc" is `TRUE`. |
| `xn` | The data for the genes which are not annotated to the cell-cycle, control genes, genes for which "if_cc" is `FALSE` |
| `nboot` | The number of bootstrap repititions to be carried out on each iteration to determine the significance of each component. |
| `bar` | Whether to display a progress bar or not. The progress bar will not work in R-markdown enviornments so this option may be turned off. The default value is `TRUE`. |

## Value

A data frame containing the loadings for each component on the cell-cycle and control genes as well as the difference between the loadings and the bootstrapped statistic for each loading.

ccRemover *Removes the effect of the cell-cycle*

## Description

ccRemover returns a data matrix with the effects of the cell-cycle removed.

## Usage

```
ccRemover(dat, cutoff = 3, max_it = 4, nboot = 200, ntop = 10,
  bar = TRUE)
```

## Arguments

| | |
|---|---|
| dat | A list containing a data frame , x, that contains gene expression measurements with each column representing a sample and each row representing a gene and a logical vector, if_cc, that indicates which of the genes/rows are related to the cell-cycle or factor of interest.<br>It is recommended that the elements of x are log-transformed and centered for each gene. For example if x contains TPM measurements then we suggest the following two-steps: Step 1: dat$x <- log(dat$x + 1) Step 2: dat$x - rowMeans(dat$x) ccRemover requires that the samples have been properly normalized for sequencing depth and we recommend doing so prior to applying the above steps.<br>The if_cc vector must be the same length as the number of rows in x and have elements equal to TRUE for genes which are related to the cell-cycle and and elements equal to FALSE for genes which are unrelated to the cell-cycle. |
| cutoff | The significance cutoff for identifying sources of variation related to the cell-cycle. The default value is 3, which roughly corresponds to a p-value of 0.01. |
| max_it | The maximum number of iterations for the algorithm. The default value is 4. |
| nboot | The number of bootstrap repititions to be carried out on each iteration to determine the significance of each component. |
| ntop | The number of components considered tested at each iteration as cell-cycle effects. The default value if 10 |
| bar | Whether to display a progress bar or not. The progress bar will not work in R-markdown enviornments so this option may be turned off. The default value is TRUE. |

## Details

Implements the algorithm described in Barron, M. & Li, J. "Identifying and removing the cell-cycle effect from scRNA-Sequencing data" (2016), Scientific Reports. This function takes a normalized, log-transformed and centered matrix of scRNA-seq expression data and a list of genes which are known to be related to the cell-cycle effect. It then captures the main sources of variation in the data and determines which of these are related to the cell-cycle before removing those that are. Please see the original manuscript for further details.

## Value

A data matrix with the effects of the cell-cycle removed.

## Examples

```
set.seed(10)
# Load in example data
data(t.cell_data)
head(t.cell_data[,1:5])
# Center data and select small sample for example
t_cell_data_cen <- t(scale(t(t.cell_data[,1:20]), center=TRUE, scale=FALSE))
# Extract gene names
gene_names <- rownames(t_cell_data_cen)
# Determine which genes are annotated to the cell-cycle
cell_cycle_gene_indices <- gene_indexer(gene_names,
species = "mouse", name_type = "symbol")
# Create "if_cc" vector
if_cc <- rep(FALSE,nrow(t_cell_data_cen))
if_cc[cell_cycle_gene_indices] <- TRUE
# Move data into list
dat <- list(x=t_cell_data_cen, if_cc=if_cc)
# Run ccRemover
## Not run:
 xhat <- ccRemover(dat, cutoff = 3, max_it = 4, nboot = 200, ntop = 10)

## End(Not run)
# Run ccRemover with reduced bootstrap repetitions for example only
xhat <- ccRemover(dat, cutoff = 3, max_it = 4, nboot = 20, ntop = 10)
head(xhat[,1:5])
# Run ccRemover with more compoents considered
## Not run:
xhat <- ccRemover(dat, cutoff = 3, max_it = 4, nboot = 200, ntop = 15)

## End(Not run)
```

---

# dat *A simulated scRNA-Seq data.*

---

## Description

This data contains expression levels (log-transformed and centered) for 50 cells and 2000 genes. The 50 cells are randomly assigned to two cell types and three cell-cycle stages. 400 genes are assigned as cell-cycle genes, and the other 1600 genes are control genes. For descriptions of how we generated this data, please refer to the paper.

## Usage

```
data(dat)
```

### Format

A list that contains the following attributes (only x and if.cc are used by ccRemover.main.)

x the data matrix. rows are genes, and columns are cells. These should be treated as log-transformed and centered (each row has mean 0) expression levels.

if.cc a vector of values FALSE's or TRUE's, denoting whether the genes are cell-cycle related or control.

n the number of cells. n=ncol(x).

p the number of genes. p=nrow(x).

pc the number of cell-cycle genes. pc=sum(if.cc).

ct cell types. a vector of values 1 and 2.

cc cell-cycle stages. a vector of values 1, 2, or 3.

### Value

A simulated dataset used to demonstrate the application of ccRemover

---

## gene_indexer *Identifies genes annotated to the cell-cycle*

---

### Description

Determines which of the genes contained in the dataset are annotated ti the cell-cycle. This is a preprocessing function for ccRemover. Genes can be either mouse or human and either official gene symbols, Ensembl, Entrez or Unigene IDs.

### Usage

```
gene_indexer(gene_names, species = NULL, name_type = NULL)
```

### Arguments

| | |
|---|---|
| gene_names | A vector containing the gene names for the dataset. |
| species | The species which the gene names are from. Either "human" or "mouse". |
| name_type | The type of gene name considered either, Ensembl gene IDS ("ensembl"), offical gene symbols (symbol), Entrez gene IDS ("entrez"), or Unigene IDS (unigene). |

### Value

A vector containg the indices of genes which are annotated to the cell-cycle

## Examples

```
set.seed(10)
# Load in example data
data(t.cell_data)
head(t.cell_data[,1:5])
# Center example data
t_cell_data_cen <- t(scale(t(t.cell_data), center=TRUE, scale=FALSE))
# Extract gene names
gene_names <- rownames(t_cell_data_cen)
# Determine which genes are annotated to the cell-cycle
cell_cycle_gene_indices <- gene_indexer(gene_names = gene_names,
species = "mouse", name_type = "symbol")
# Create "if_cc" vector
if_cc <- rep(FALSE,nrow(t_cell_data_cen))
if_cc[cell_cycle_gene_indices] <- TRUE

# Can allow the function to automatically detect the name type
cell_cycle_gene_indices <- gene_indexer(gene_names = gene_names,
species = NULL, name_type = NULL)
```

---

# get_diff *Calculates the average load difference between the cell-cycle genes and control genes for each component.*

---

## Description

This is an interal function for use by "bootstrap_diff" only.

## Usage

```
get_diff(xy, xn)
```

## Arguments

| | |
|---|---|
| xy | The data for the genes which are annotated to the cell-cycle, i.e. those genes for which "if_cc" is TRUE. |
| xn | The data for the genes which are not annotated to the cell-cycle, control genes, genes for which "if_cc" is FALSE |

## Value

A data frame containing the loadings for each component on the cell-cycle and control genes.

## hello *Hello, World!*

### Description

Prints 'Hello, world!'.

### Usage

```
hello()
```

### Examples

```
hello()
```

## HScc_genes *Homo Sapien genes which are annotated to the cell-cycle*

### Description

This data set contains Homo Sapien genes which are annotated to the cell-cycle. These genes were retrieved from biomart and are intended for use with the "gene_indexer" function. The data set contains the gene names in four different formats, Ensemble Gene IDs (1838 values), HGNC symbols (1740 values), Entrez Gene IDs (1744 values) and Unigene IDs (1339).

### Usage

```
data("HScc_genes")
```

### Format

A data set that contains with the following attributes

`human_cell_cycle_genes` A data frame with four columns corresponding to each of the different ID formats.

### Value

A data set containing genes annotated to the cell-cycle in different ID formats

## MMcc_genes *Mus Musculus genes which are annotated to the cell-cycle*

### Description

This data set contains Mus Musculus genes which are annotated to the cell-cycle. These genes were retrieved from biomart and are intended for use with the "gene_indexer" function. The data set contains the gene names in three different formats, Ensemble Gene IDs (1433 values), MGI symbols (1422 values), Entrez Gene IDs (1435 values) and Unigene IDs (1102 values).

### Usage

```
data("MMcc_genes")
```

### Format

A data set that contains with the following attributes

`mouse_cell_cycle_genes` A data frame with four columns corresponding to each of the different ID formats.

### Value

A dataset containing genes annotated to the cell-cycle in different ID formats

## t.cell_data *Single-Cell RNA-Seq data from differentiating T-helper cells*

### Description

This data contains expression levels (log-transformed normalized count values) for 81 cells and 14,147 genes. The data was normalized using ERCC spike-ins. This data was generated by Mahata, B. et al (2014). The processed data was retrieved from the supplementary material of Buettner et al. (2015), for descriptions of how the data was processed, please refer to their paper.

### Usage

```
data(t.cell_data)
```

### Format

A data set that contains with the following attributes

`t.cell_data` the data matrix. rows are cells, and columns are genes. These should be treated as log-transformed and normalized

### Value

A scRNA-Seq dataset with gene expression levels for 187 T-helper cells

# Index