Package 'LRcell'

October 14, 2021

Type Package

Title Differential cell type change analysis using Logistic/linear Regression

Version 1.0.0

Date 2021-03-10

BugReports https://github.com/marvinquiet/LRcell/issues

GitURL https://github.com/marvinquiet/LRcell

Description The goal of LRcell is to identify specific sub-cell types that drives the changes observed in a bulk RNA-seq differential gene expression experiment. To achieve this, LRcell utilizes sets of cell marker genes acquired from single-cell RNA-sequencing (scRNA-seq) as indicators for various cell types in the tissue of interest. Next, for each cell type, using its marker genes as indicators, we apply Logistic Regression on the complete set of genes with differential expression p-values to calculate a cell-type significance p-value. Finally, these p-values are compared to predict which one(s) are likely to be responsible for the differential gene expression pattern observed in the bulk RNA-seq experiments. LRcell is inspired by the LRpath[@sartor2009lrpath] algorithm developed by Sartor et al., originally designed for pathway/gene set enrichment analysis. LRcell contains three major components: LRcell analysis, plot generation and marker gene selection. All modules in this package are written in R. This package also provides marker genes in the Prefrontal Cortex (pFC) human brain region, human PBMC and nine mouse brain regions (Frontal Cortex, Cerebellum, Globus Pallidus, Hippocampus, Entopeduncular, Posterior Cortex, Striatum, Substantia Nigra and Thalamus).

License MIT + file LICENSE

Encoding UTF-8

biocViews SingleCell, GeneSetEnrichment, Sequencing, Regression, GeneExpression, DifferentialExpression

Depends R (>= 4.1), ExperimentHub, AnnotationHub

Imports BiocParallel, dplyr, ggplot2, ggrepel, magrittr, stats, utils

RoxygenNote 7.1.1

Suggests LRcellTypeMarkers, BiocStyle, knitr, rmarkdown, roxygen2, testthat

VignetteBuilder knitr

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

git_branch RELEASE_3_13

git_last_commit 85370d4

git_last_commit_date 2021-05-19

Date/Publication 2021-10-14

Author Wenjing Ma [cre, aut] (<https://orcid.org/0000-0001-8757-651X>)

Maintainer Wenjing Ma <wenjing.ma@emory.edu>

R topics documented:

`enrich_posfrac_score` *Calculate enrichment scores for each cell type in a specific gene.*

Description

This function takes a specific gene expression, cell type annotation and a hyperparameter to calculate enrichment scores.

Usage

```
enrich_posfrac_score(gene, expr, annot, power = 1)
```

Arguments

`gene`	Gene name from the expression matrix.
`expr`	Complete expression matrix with rows as genes and columns as cells.
`annot`	Cell type annotation named vector with names as cell ids and values as cell types.
`power`	The penalty on fraction of cells expressing the genes

Value

Enrichment score list with cell type as names and enrichment score as values.

`example_gene_pvals` *Example gene_pvals named vector from mouse experiment.*

Description

A named vector containing gene symbols as name and p-values as values. This is from a mouse Alzheimer's disease model (GEO: GSE90693), specifically 6 months after treatment in Frontal Cortex brain region. In this dataset, we expect to see the Microglia as the most enriched cell type.

Usage

```
data(example_gene_pvals)
```

Format

A named vector with 23,420 items

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90693 'GSE90693_RawCountsData_TPR50_6months_A

`example_LRcell_res` *An example output of LRcell using data example_gene_pvals and mouse_FC_marker_genes.*

Description

An example output of LRcell using data example_gene_pvals and mouse_FC_marker_genes.

Usage

```
data(example_LRcell_res)
```

Format

A data frame with 81 rows as mouse FC sub-clusters and 8 variables:

ID The IDs of each marker genes, can be a cell type or cluster

genes_num How many marker genes are contributing to the analysis

coef The coefficients of Logistic Regression or Linear Regression

odds_ratio The odds ratio quantifies association in Logistic Regression

p.value The p-value calculated from the analysis

FDR The FDR after BH correction

lead_genes Genes that are contributing to the analysis

cell_type Cell typ ename

`get_markergenes` *Get top marker genes for each subcluster*

Description

Get top marker genes for each subcluster

Usage

```
get_markergenes(enriched.g, method = c("LR", "LiR"), topn = 100)
```

Arguments

`enriched.g`	A return from LRcell_gene_enriched_scores or from provided data
`method`	If LR, the return will be a list of genes; If LiR, the return will be a list of named vector with names as genes and values as enriched scores.
`topn`	Top N genes as marker genes.

Value

A list of top marker genes.

Examples

```
library(ExperimentHub)
eh <- ExperimentHub::ExperimentHub()
eh <- query(eh, "LRcellTypeMarkers")
# eh$title
enriched_genes <- eh[['EH4548']]
marker.g <- get_markergenes(enriched_genes, method="LR", topn=100)
```

LRcell *Cell-type enrichment analysis for preranked gene set.*

Description

This function wraps around LRcellCore in case of empty inputs of the marker gene file and brain region.

Usage

```
LRcell(
  gene.p,
  marker.g = NULL,
  species = c("mouse", "human"),
  region = NULL,
  method = c("LR", "LiR"),
  min.size = 5,
  sig.cutoff = 0.05
)
```

Arguments

gene.p	Named vector of gene-level pvalues from DEG analysis, i.e. DESeq2, LIMMA
marker.g	List of Cell-type specific marker genes derived from single-cell RNA-seq. The name of the list is cell-type or cluster name, the values are marker genes vectors or numeric named vectors. LRcell provides marker genes list in different human/mouse brains, but users could use their own marker gene list as input. default: NULL
species	Either ‘mouse‘ or ‘human‘, default: mouse.
region	Specific brain regions provided by LRcell. For mouse, LRcell provides 9 brain regions: c("FC", "HC", "PC", "GP", "STR", "TH", "SN", "ENT", "CB"). For human, LRcell provides c("pFC", "PBMC")
method	Either ‘logistic regression‘ or ‘linear regression‘. Logistic regression equally treats cell-type specific marker genes, however, if certain values could determine the importance of marker genes, linear regression can be performed, default: LR.
min.size	Minimal size of a marker gene set, will impact the balance of labels
sig.cutoff	Cutoff for input genes pvalues, default: 0.05.

Value

A list with LRcell results. Each item represents a marker gene input. Each item in this list is a statistics table. In the table, the row represents the name of marker genes, and the columns are:

- ID The IDs of each marker genes, can be a cell type or cluster;

- genes_num How many marker genes are contributing to the analysis;
- coef The coefficients of Logistic Regression or Linear Regression;
- odds_ratio The odds ratio quantifies association in Logistic Regression;
- p-value The p-value calculated from the analysis;
- FDR The FDR after BH correction.
- lead_genes Genes that are contributing to the analysis;

Examples

```
data(example_gene_pvals)
res <- LRcell(example_gene_pvals, species="mouse", region="FC", method="LR")
```

`LRcellCore` *Find most enriched cell types in bulk DE genes by Logistic Regression*

Description

This is a function which takes marker genes from single-cell RNA-seq as reference to calculate the enrichment of certain cell types in bulk DEG analysis. We assume that bulk DEG is derived from certain cell-type specific pattern.

Usage

```
LRcellCore(gene.p, marker.g, method, min.size = 5, sig.cutoff = 0.05)
```

Arguments

`gene.p`	Named vector of gene-level pvalues from DEG analysis, i.e. DESeq2, LIMMA
`marker.g`	List of Cell-type specific marker genes derived from single-cell RNA-seq. The name of the list is cell-type or cluster name, the values are marker genes vectors or numeric named vectors. LRcell provides marker genes list in different human/mouse brains, but users could use their own marker gene list as input. default: NULL
`method`	Either 'logistic regression' or 'linear regression'. Logistic regression equally treats cell-type specific marker genes, however, if certain values could determine the importance of marker genes, linear regression can be performed, default: LR.
`min.size`	Minimal size of a marker gene set, will impact the balance of labels
`sig.cutoff`	Cutoff for input genes' pvalues, default: 0.05.

Value

A dataframe of LRcell statistics as described in `LRcell`.

Examples

```
data(mouse_FC_marker_genes)
data(example_gene_pvals)
res <- LRcellCore(example_gene_pvals, mouse_FC_marker_genes, method="LR")
```

`LRcell_gene_enriched_scores`

Find most enriched cell types in bulk DE genes by Logistic Regression

---

Description

This is a function which takes marker genes from single-cell RNA-seq as reference to calculate the enrichment of certain cell types in bulk DEG analysis. This algorithm borrows from Marques et al, 2016 (https://science.sciencemag.org/content/352/6291/1326.long).

Usage

```
LRcell_gene_enriched_scores(
  expr,
  annot,
  power = 1,
  parallel = TRUE,
  n.cores = 4
)
```

Arguments

`expr`	Expression matrix with rows as genes and columns as cells, can be an object of Matrix or dgCMatrix or a dataframe.
`annot`	Cell type annotation named vector with names as cell ids and values as cell types.
`power`	The penalty on fraction of cells expressing the genes.
`parallel`	Whether to run it in parallel.
`n.cores`	How many cores to use in parallel mode.

Value

A numeric matrix with rows as genes and columns as cell types, values are gene enrichment scores.

mouse_celltypes *Mapping between subclusters and cell types in Mouse Brain*

Description

A named vector containing the subclusters as name and cell types as values in Mouse Brain. The cell types are pre-annotated by the dataset, which includes: Endothelial, FibroblastLike, Mural, Oligodendrocytes, Polydendrocytes, Astrocytes and Microglia.

Usage

```
data(mouse_celltypes)
```

Format

A named vector with 565 subclusters:

Named vector with name as subclusters and values as cell types.

Source

http://dropviz.org/ under tab 'data'

mouse_FC_marker_genes *Example marker genes from mouse FC brain region.*

Description

A list of marker genes with names indicating cell types. We selected top 100 enriched genes from each subcluster as marker genes list.

Usage

```
data(mouse_FC_marker_genes)
```

Format

A named vector with 81 subclusters in mouse Frontal Cortex:

Named vector with name as subclusters and values as marker genes.

Source

Caculated from gene enrichment scores

plot_manhattan_enrich *Manhattan plot for the enrichment of cell types*

Description

This function draws out the LRcell result dataframe. In this function, we take LRcell result dataframe and added cell types according to

Usage

```
plot_manhattan_enrich(lrcell_res, sig.cutoff = 0.05, label.topn = 5)
```

Arguments

lrcell_res	LRcell result dataframe.
sig.cutoff	The p-value cutoff showing significance result of LRcell.
label.topn	A numeric number showing how many significant cell types will be labeled.

Value

A ggplot2 object

Examples

```
data(example_LRcell_res)
plot_manhattan_enrich(example_LRcell_res)
```

plot_marker_dist *Plot marker genes distribution on DE gene rank*

Description

This function draws out the marker gene distribution for a certain cell type (or cluster) on the DE gene rank list.

Usage

```
plot_marker_dist(markers, gene.p, colour = "red")
```

Arguments

markers	Vector of marker genes from a cell type or cluster of interest.
gene.p	Named vector of gene-level pvalues from DEG analysis, i.e. DESeq2, LIMMA
colour	Users can define the bar color they want on the ggplot2 object.

Value

A ggplot2 object

Examples

```
data(example_gene_pvals)
data(mouse_FC_marker_genes)
Oligos_markers <- mouse_FC_marker_genes[["FC_9-5.Oligodendrocytes_5"]]
plot_marker_dist(Oligos_markers, example_gene_pvals)
```

Index