Package 'BRETIGEA'

May 5, 2021

Title Brain Cell Type Specific Gene Expression Analysis

Version 1.0.3

Author Andrew McKenzie [aut, cre],
Minghui Wang [aut],
Bin Zhang [aut]

Maintainer Andrew McKenzie <amckenz@gmail.com>

Description Analysis of relative cell type proportions in bulk gene expression data. Provides a well-validated set of brain cell type-specific marker genes derived from multiple types of experiments, as described in McKenzie (2018) <doi:10.1038/s41598-018-27293-5>. For brain tissue data sets, there are marker genes available for astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells, derived from each of human, mice, and combination human/mouse data sets. However, if you have access to your own marker genes, the functions can be applied to bulk gene expression data from any tissue. Also implements multiple options for relative cell type proportion estimation using these marker genes, adapting and expanding on approaches from the 'Cell-CODE' R package described in Chikina (2015) <doi:10.1093/bioinformatics/btv015>. The number of cell type marker genes used in a given analysis can be increased or decreased based on your preferences and the data set. Finally, provides functions to use the estimates to adjust for variability in the relative proportion of cell types across samples prior to downstream analyses.

Depends R (>= 3.0.0)

Suggests testthat, stats, utils, knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 6.1.0

License MIT + file LICENSE

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2021-05-05 04:40:02 UTC

R topics documented:

`aba_marker_expression` *Normalized FPKM expression data from a subset of the Allen Brain Atlas Aging, Dementia, and TBI study.*

Description

Filtered data set to only contain brain marker genes as well as ~100 other genes. The total data set can be downloaded by following the links in the original paper.

Usage

```
aba_marker_expression
```

Format

An object of class `data.frame` with 395 rows and 377 columns.

References

© 2016 Allen Institute for Cell Science. Aging, Dementia and TBI. Available from: http://aging.brain-map.org/

aba_pheno_data *Phenotype data from brain samples in the Allen Brain Atlas Aging, Dementia, and TBI study.*

Description

A subset of phenotype data to be used for correlations with the expression data in the tests and vignette, to validate that the estimated cell type proportions correspond to useful quantities.

Usage

```
aba_pheno_data
```

Format

An object of class `data.frame` with 377 rows and 4 columns.

References

© 2016 Allen Institute for Cell Science. Aging, Dementia and TBI. Available from: http://aging.brain-map.org/

adjustBrainCells *Estimate and adjust for brain cell type proportions in bulk expression data.*

Description

This function uses a linear model to adjust each row of gene expression for cell types. Other covariates can be included as well.

Usage

```
adjustBrainCells(
  inputMat,
  nMarker = 50,
  species = "combined",
  celltypes = c("ast", "end", "mic", "neu", "oli", "opc"),
  addMeans = FALSE,
  formula = NULL,
  verbose = FALSE
)
```

Arguments

inputMat Input gene expression data, with rows as features (e.g. genes) and samples as columns.

nMarker The number of marker genes (that are present in your expression data set) to use in estimating the surrogate cell type proportion variable for each cell type.

species By default, this function uses markers from combined human and mouse measurements, which are the most robust and reliable, as the gene expression patterns are very conserved between these two species. Other options are "human" and "mouse" for data specific to those species. Note that OPCs only have 500 gene symbols in this case, and are taken from only the Darmanis et al or Tasic et al data sets, respectively.

celltypes Character vector of which cell types to estimate and adjust for.

addMeans Whether the mean should be added to the residuals in the resulting adjusted gene expression table.

formula If you want to add additional covariates to be adjusted for, then you can supply them by adding an (optional) formula function here. The format should be "expression_data ~ $cov1 + factor($cov2) +", where $cov1 is a numeric covariate present in the current environment, and $cov2 is a factor variable also defined in the current environment.

verbose Whether to report the formula used for adjustment of each row.

Value

A list containing both a matrix of estimate surrogate proportion variables (SPVs), as well as a matrix of adjusted gene expression values.

Examples

```
brain_cells_adjusted = adjustBrainCells(aba_marker_expression,
  nMarker = 50, species = "combined")
expression_data_adj = brain_cells_adjusted$expression
cor_mic_unadj = cor.test(as.numeric(aba_marker_expression["AIF1", ]),
  as.numeric(aba_pheno_data$ihc_iba1_ffpe), method = "spearman")
cor_mic_adj = cor.test(expression_data_adj["AIF1", ],
  as.numeric(aba_pheno_data$ihc_iba1_ffpe), method = "spearman")
```

---

adjustCells *Adjust for estimated cell type proportions in bulk gene expression data.*

Description

This function uses a linear model to adjust each row of gene expression for cell types, taking the residuals from the linear model for downstream analysis. Other covariates can be included as well, if they are defined in the current environment, through the use of the formula argument.

Usage

```
adjustCells(
  inputMat,
  cellSPV,
  celltypes = NULL,
  addMeans = FALSE,
  formula = NULL,
  verbose = FALSE
)
```

Arguments

`inputMat`	Input gene expression data, with rows as features (e.g. genes) and samples as columns.
`cellSPV`	Estimated matrix of surrogate proportion variables for each cell type, with samples as rows and columns as cell types.
`celltypes`	Character vector of which cell types to use. Must correspond to column names in the cellSPV data frame. If NULL, all of the columns of SPV will be used for adjustment.
`addMeans`	Whether the mean should be added to the residuals in the resulting adjusted gene expression table.
`formula`	If you want to add additional covariates to be adjusted for, then you can supply them by adding an (optional) formula function here. The format should be "expression_data ~ $cov1 + factor($cov2) +", where $cov1 is a numeric covariate present in the current environment, and $cov2 is a factor variable also defined in the current environment.
`verbose`	Whether to print out the formula used for adjustment of each row.

Value

A matrix of adjusted gene expression values.

Examples

```
svp_res = brainCells(inputMat = aba_marker_expression, nMarker = 10,
  species = "human", celltypes = c("ast", "neu", "oli"))
str(svp_res)
adjust_res = adjustCells(inputMat = aba_marker_expression,
  cellSPV = svp_res, addMeans = FALSE)
str(adjust_res)
```

brainCells *Estimate brain cell type proportions in bulk expression data with marker genes.*

Description

This function uses marker genes estimated in a meta-analysis of brain cell type-associated RNA expression data sets, and uses them as input for the findCells cell type proportion estimation procedure pipeline.

Usage

```
brainCells(
  inputMat,
  nMarker = 50,
  species = "combined",
  data_set = "mckenzie",
  celltypes = c("ast", "end", "mic", "neu", "oli", "opc"),
  method = "SVD",
  scale = TRUE
)
```

Arguments

inputMat	Gene expression data frame or matrix, with rownames corresponding to gene names, some of which are marker genes, and columns corresponding to samples.
nMarker	The number of marker genes (that are present in your expression data set) to use in estimating the surrogate cell type proportion variable for each cell type.
species	By default, this function uses markers from combined human and mouse measurements, which are the most robust and reliable, as the gene expression patterns are very conserved between these two species. Other options are "human" and "mouse" for data specific to those species. Note that OPCs only have 500 gene symbols in this case, and are taken from only the Darmanis et al or Tasic et al data sets, respectively.
data_set	Which data set the data should be derived from. Options are "mckenzie" (default), "kelley". Note that the "kelley" data set will ignore the "species" argument.
celltypes	Character vector of which cell types to estimate.
method	To estimate the cell type proportions, can either use "PCA" or "SVD".
scale	Whether or not to scale the gene expression data from each marker gene prior to using it as an input for dimension reduction.

Value

A sample-by-cell type matrix of estimate cell type proportion variables.

Examples

```
ct_res = brainCells(aba_marker_expression, nMarker = 50, species = "combined")
cor.test(ct_res[, "mic"], as.numeric(aba_pheno_data$ihc_iba1_ffpe), method = "spearman")
```

BRETIGEA *BRETIGEA: BRain cEll Type specIfic Gene Expression Analysis*

Description

This package provides two major functions: 1) A function to estimate cell type surrogate proportion variables based on marker genes. 2) A function to adjust bulk gene expression data for estimated cell type proportions, as covariates. It also provides data containing estimated marker genes for six major brain cell types, i.e. astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells (OPCs), as well as wrapper functions to use this data on the two major package functions.

findCells *Find cell type proportions from bulk gene expression data using marker genes.*

Description

Input a gene expression matrix and your own data frame of marker genes, and this function will estimate cell type proportions in your data set using one of the SVD or PCA dimension reduction approaches.

Usage

```
findCells(inputMat, markers, nMarker = 50, method = "SVD", scale = TRUE)
```

Arguments

inputMat	Numeric gene expression data frame or matrix, with rownames corresponding to gene names, some of which are marker genes, and columns corresponding to samples.
markers	Data frame with marker genes in one column (named "marker") and the cell type that that gene symbol corresponds to in another column (named "cell").
nMarker	The number of marker genes (that are present in your expression data set) to use in estimating the surrogate cell type proportion variable for each cell type.
method	To estimate the cell type proportions, can either use PCA or SVD.
scale	Whether or not to scale the gene expression data from each marker gene prior to using it as an input for dimension reduction.

Value

A sample-by-cell type matrix of estimate cell type proportion variables.

References

Chikina M, Zaslavsky E, Sealfon SC. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinformatics. 2015;31(10):1584-91.

Examples

```
cell_type_proportions = findCells(aba_marker_expression,
 markers = markers_df_brain, nMarker = 10)
str(cell_type_proportions)
```

---

kelley_df_brain *Marker genes estimated from*

Description

Top 1000 marker genes from each of four major brain cell types (ie astrocytes, microglia, neurons, and oligodendrocytes) estimated via correlation-based analysis of bulk brain tissue. Contains marker data averaged across all brain regions as well as markers derived from individual brain regions.

Usage

```
kelley_df_brain
```

Format

An object of class `data.frame` with 84000 rows and 2 columns.

References

Kelley KW, Nakao-inoue H, Molofsky AV, Oldham MC. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nat Neurosci. 2018;21(9):1171-1184.

`markers_df_brain` *Marker genes estimated from a meta-analysis of brain cell gene expression data from both humans and mice.*

Description

Top 1000 marker genes from each of the six major brain cell types (ie astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and OPCs) estimated from a meta-analysis of brain cell gene expression data from both humans and mice.

Usage

```
markers_df_brain
```

Format

An object of class `data.frame` with 6000 rows and 2 columns.

References

Mckenzie AT, Wang M, Hauberg ME, et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Sci Rep. 2018;8(1):8868.

`markers_df_human_brain` *Marker genes estimated from a meta-analysis of brain cell gene expression data from humans only.*

Description

Top 1000 marker genes from each of the six major brain cell types (ie astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and OPCs) estimated from a meta-analysis of brain cell gene expression data from humans only.

Usage

```
markers_df_human_brain
```

Format

An object of class `data.frame` with 5500 rows and 2 columns.

References

Mckenzie AT, Wang M, Hauberg ME, et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Sci Rep. 2018;8(1):8868.

markers_df_mouse_brain

Marker genes estimated from a meta-analysis of brain cell gene expression data from mice only.

Description

Top 1000 marker genes from each of the six major brain cell types (ie astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and OPCs) estimated from a meta-analysis of brain cell gene expression data from mice only.

Usage

```
markers_df_mouse_brain
```

Format

An object of class `data.frame` with 5430 rows and 2 columns.

References

Mckenzie AT, Wang M, Hauberg ME, et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Sci Rep. 2018;8(1):8868.

Index