Package 'PathAnalyser'

April 25, 2022

Title ER & HER2 Pathway Activity Analysis For Breast Cancer Transcriptomic Datasets

Version 1.0.0

Description PathAnalyser provides an assessment of molecular pathway activity in transcriptomic datasets. The classification algorithm employed by PathAnalyser implements Gene GSVA to classify samples in a transcriptomic dataset by pathway activity using a gene signature. Currently, the package provides built-in data for assessment of ER & HER2 pathway activity in breast cancer transcriptomic datasets, however the PathAnalyser functionality could also be applied to gene signatures and/or transcriptomic datasets in non-oncological contexts.

License MIT + file LICENSE

URL https://github.com/ozlemkaradeniz/PathAnalyser

BugReports https://github.com/ozlemkaradeniz/PathAnalyser/issues

Depends GSVA,
R (>= 2.10)

biocView

Imports edgeR,
ggfortify,
ggplot2,
limma,
plotly,
reader,
reshape2

Suggests BiocStyle,
knitr,
rmarkdown,
testthat,
magick,
qpdf

VignetteBuilder knitr

Encoding UTF-8

LazyData true

LazyDataCompression xz

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2.9000

R topics documented:

`calculate_accuracy` *Accuracy calculation of classification method*

Description

This method provides several classification evaluation metrics to assess the accuracy of predicted pathway classification. The accuracy calculation is performed using predicted pathway activity labels from the employed classification method for each sample and the corresponding true activity labels for the given pathway. A confusion matrix is created to display the classification accuracy decomposed into the distinct pathway activity classes in tabular form for the user. Additional classification evaluation statistics (such as sensitivity, specificity, recall, percentage of classified samples etc) is the optional feature that the user can specify.

Usage

```
calculate_accuracy(true_labels, predicted_labels, pathway, show_stats = FALSE)
```

Arguments

`true_labels`	a data frame, matrix or file name which contains a column named "sample" that consists of sample names / IDs and another column named after a specific pathway which contains the corresponding true pathway activity labels.
`predicted_labels`	a predicted labels data frame or matrix generated by the classification method yielding predicted pathway activity labels in a column called "class" for the samples in the "sample" column.
`pathway`	name of pathway used for classification (Note: this pathway name must be present in the true labels data frame / matrix / file name for classification evaluation and generation of the confusion matrix.)
`show_stats`	an optional flag to display additional statistical information using the confusion matrix and other classification evaluation metrics including: sensitivity, specificity, precision, false positive rate, false negative rate etc.

Value

confusion_matrix

Author(s)

Ozlem Karadeniz <ozlem.karadeniz.283@cranfield.ac.uk>

Examples

```
## Not run: calculate_accuracy(true_labels_df, predicted_labels_df, "ER",
show_stats= TRUE)
## End(Not run)
```

---

check_signature_vs_dataset

Validity check in gene signatures and gene expression datasets

---

Description

This function performs a validity check against the gene signature and gene expression data set and filters genes that are absent in expression data set and /or are not expressed in at least 10% of the total number of samples. A bar plot of mean-normalised counts for each gene is also displayed.

Usage

```
check_signature_vs_dataset(norm_data, sig_df, barplot = TRUE)
```

Arguments

norm_data	Normalized gene expression data matrix
sig_df	A signature data frame containing two columns: the first column contains a list of gene (represented by gene symbols) that are differentially expressed in a given pathway and their relative expression values are given in the second column, with 1 representing up-regulated genes and -1 representing down-regulated genes.
barplot	optional Boolean parameter to display bar plot of mean-normalised gene counts following pre-processing performed by the function (default=TRUE).

Value

A filtered gene expression data matrix with gene symbols as row names and sample names / IDs as column names

Author(s)

Yi-Hsuan Lee <yi-hsuan.lee@cranfield.ac.uk>

Examples

```
## Not run: check_signature_vs_dataset(norm_data, sig_df)
```

`classes_pca` *PCA plot visualising pathway-based classification of samples in a dataset*

Description

This function generates a PCA plot showing the clustering of samples in the normalised expression data set, coloured by predicted pathway activity labels: active (green), inactive (orange) and uncertain (blue). An ideal pathway-based classification would generate a PCA plot showing tight clustering of samples in each activity class and less overlap between classes.

Usage

```
classes_pca(norm_data, predicted_labels_df, pathway = "Pathway Activity")
```

Arguments

`norm_data`	A (logCPM) normalized gene expression data matrix, with row names consisting of the HUGO gene symbols and column names corresponding to the name / ID of each sample in the dataset
`predicted_labels_df`	A data frame containing the pathway activity labels predicted by classification algorithm for each sample in the dataset. The first column is called "sample" which contains the sample names in the data set and the second column is called "class" containing the corresponding predicted pathway activity of the sample.
`pathway`	The pathway name used in the title of the plot, default is a placeholder "Pathway Activity".

Author(s)

Yi-Hsuan Lee <yi-hsuan.lee@cranfield.ac.uk>

Examples

```
## Not run: classes_pca(norm_data, predicted_labels_df, 'ER')
```

`classify_gsva_abs` *Classification Using Absolute GSVA Score Thresholds*

Description

Classifies samples according to pathway activity by first ranking samples by their GSVA score and assessing evidence of expression consistency of each sample with the up-regulated gene-set and down-regulated gene-set of the gene signature using absolute GSVA score thresholds. GSVA scores generated by the GSVA algorithm provide a measure of expression abundance of the up-regulated and down-regulated gene-sets, which PathAnalyser uses to assess expression consistency with both parts (up-regulated and down-regulated gene-sets) of the gene signature using the user-supplied absolute GSVA score thresholds as thresholds for expression consistency with each part of the signature.

Usage

```
classify_gsva_abs(
  expr_mat,
  sig_df,
  up_thresh.low,
  up_thresh.high,
  dn_thresh.low,
  dn_thresh.high
)
```

Arguments

expr_mat	Normalised expression data set matrix comprising the expression levels of genes (rows) for each sample (columns) in a data set. Row names are gene symbols and column names are sample IDs / names. Gene expression matrices can contain normalised (logCPM transformed) RNASeq or microarray transcriptomic data.
sig_df	Gene expression signature for a specific pathway given as data frame with the first column named "gene" containing a list of genes that are the most differentially expressed when the given pathway is active and the second column named "expression" containing their corresponding expression: -1 for down-regulated genes and 1 for up-regulated genes.
up_thresh.low	Number denoting the absolute GSVA score threshold for categorizing a sample as having inconsistent expression with the up-regulated gene-set from the gene signature.
up_thresh.high	Number denoting the absolute GSVA score threshold for categorizing a sample as having consistent expression with the up-regulated gene set from the gene signature.
dn_thresh.low	Number denoting the absolute GSVA score threshold for categorizing a sample as having consistent expression with the down-regulated gene-set from the gene signature.
dn_thresh.high	Number denoting the absolute GSVA score threshold for categorizing a sample as having inconsistent expression with the down-regulated gene-set from the gene signature.

Details

Four thresholds are specified by the user:

1. "up_thresh.low" - a GSVA score threshold for considering a sample as having inconsistent expression with up-regulated gene-set.
2. "up_thresh.high" - a GSVA score threshold for considering a sample as having expression consistent with the up-regulated gene-set up-regulated gene set
3. "dn_thresh.low" - a GSVA score threshold for considering a sample as having consistent expression with down-regulated gene-set
4. "dn_thresh.high" - a GSVA score threshold for considering a sample as having inconsistent expression with the down-regulated gene-set. Samples that have consistent expression with both the up-regulated and down-regulated gene-sets of the signature are classified as "Active", those with inconsistent expression with both parts of the signature are classified as "Inactive" and the rest of the samples are classified as "Uncertain".

Value

A data frame containing a list of samples as the first column and their classified pathway activity (Active, Inactive or Uncertain) in the second column.

Author(s)

Anisha Thind <a.thind@cranfield.ac.uk>

Examples

```
# Default thresholds for up-regulated and down-regulated gene-sets
## Not run: classes_df <- classify_gsva_abs(ER_dataset, ER_sig, up_thresh.low=-0.25,
     up_thresh.high=0.25, dn_thresh.low=-0.25, dn_thresh.high=0.35)
## End(Not run)
```

---

classify_gsva_percent *Sample classification according to pathway activity using a percentile threshold for assessing expression consistency with both the up-regulated and down-regulated gene-set of a gene signature.*

Description

Classifies samples according to pathway activity by first ranking samples by their expression abundance of the up-regulated gene set and then the down-regulated gene-set using GSVA scores generated by the GSVA algorithm as measures of expression abundance. Samples are then assessed for expression consistency with both the up-regulated and down-regulated gene-sets using percentile thresholds during the pathway activity sample classification.

Usage

```
classify_gsva_percent(expr_mat, sig_df, percent_thresh = 25)
```

Arguments

expr_mat	Normalised expression data set matrix comprising the expression levels of genes (rows) for each sample (columns) in a data set. Row names are gene symbols and column names are sample IDs / names. Gene expression matrices can contain normalised (logCPM transformed) RNASeq or microarray transcriptomic data.
sig_df	Gene expression signature for a specific pathway given as data frame with the first column named "gene" containing a list of genes that are the most differentially expressed when the given pathway is active and the second column named "expression" containing their corresponding expression: -1 for down-regulated genes and 1 for up-regulated genes.
percent_thresh	Percentile threshold (0-100) of samples for checking consistency of gene expression of a sample with first the up-regulated and then down-regulated gene-set of the gene signature (default= 25% (quartile)). For example, using the 25% percentile threshold samples ranked in the top 25% and bottom 25% of the up-regulated and down-regulated gene-sets respectively, would be considered as "Active". Likewise, samples ranked in the bottom 25% and top 25% of the up-regulated and down-regulated gene-set of the gene signature would be classified as "Inactive".

Value

A data frame with the first column named "sample" containing sample names and the second column named "class" containing their corresponding predicted pathway activity classes (Active, Inactive or Uncertain).

Author(s)

Anisha Thind <a.thind@cranfield.ac.uk>

Examples

```
# default using quartile threshold (25th percentile)
## Not run: classes_df <- classify_gsva_percent(ER_data_mat, ER_sig)
# custom percentile threshold e.g. 30th percentile
## Not run: classes_df <- classify_gsva_percent(ER_data_mat, ER_sig,
       percent_thresh=30)
## End(Not run)
```

ER_GEO_microarr *ER microarray data set obtained from GEO*

Description

A matrix containing microarray data for 60 human breast tumour samples (30 estrogen receptor (ER) positive and 30 ER negative samples were selected at random) and were extracted from GSE31448 dataset (Sabatier et al. 2011) using the GEO query library. Multi-gene probes were excluded and only cancer samples from human breast cancer tumours were selected for constructing this matrix.

Usage

```
ER_GEO_microarr
```

Format

A matrix containing 21,656 HUGO gene symbols (row names) and 60 breast cancer tumour samples IDs (columns) which are given as sample accession numbers.

Source

GSE31448 series link on GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31448).

References

Sabatier, R., Finetti, P., Adelaide, J., Guille, A., Borg, J.P., Chaffanet, M., Lane, L., Birnbaum, D. and Bertucci, F., 2011. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. PloS one, 6(11), p.e27656. doi: https://doi.org/10.1371/journal.pone.0027656

ER_sig_df *ER gene signature*

Description

A list of up-regulated and down-regulated genes constituting the gene signature of ER pathway activation. The ER signature was obtained from the sensitivity to endocrine therapy (SET) genomic index defined using a list of genes co-expressed with the estrogen receptor (ESR1 or ER) from microarray data (437 profiles) originating from newly diagnosed breast cancer independent of outcome and treatment (Symmans et al., 2010).

Usage

```
ER_sig_df
```

Format

A data frame with 160 rows (gene names) and 2 variables:

gene The HUGO gene symbols of a highly differentially expressed gene when ER pathway is active.

expression The change in expression of the gene when the pathway is active i.e. -1 for down-regulated genes and 1 for up-regulated genes.

Source

SET index defined by Symmans et al. 2010 (see https://pubmed.ncbi.nlm.nih.gov/20697068/).

References

Symmans, W.F., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F., Regitnig, P., Daxenbichler, G., Desmedt, C., Domont, J., Marth, C. and Delaloge, S., 2010. Genomic index of sensitivity to endocrine therapy for breast cancer. Journal of clinical oncology, 28(27), p.4111. doi: https://doi.org/10.1200/jco.2010.28.4273

ER_TCGA_RNAseq *ER RNA-seq gene expression data set from TCGA*

Description

A gene expression matrix containing RNA-seq raw read counts for 60 human primary breast tumour samples (30 estrogen receptor (ER) positive and 30 ER negative samples were selected at random). This data set is a subset of a much larger data set containing 1,101 primary breast tumour samples collected from The Cancer Genome Atlas (TCGA).

Usage

```
ER_TCGA_RNAseq
```

Format

A matrix containing 20,124 HUGO gene symbols (row names) and 60 breast cancer tumour samples IDs (columns) given in the form of TCGA barcodes for each sample. For further information on TCGA bar code semantics, please see the NIH GDC documentation https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/.

Source

https://portal.gdc.cancer.gov).

---

gsva_scores_dist *GSVA score density distribution plot*

Description

Plots GSVA scores distribution across samples after performing GSVA algorithm on samples in a dataset (gene expression matrix) for the up-regulated and down-regulated gene-sets of a given gene expression signature.

Usage

```
gsva_scores_dist(expr_mat, sig_df)
```

Arguments

expr_mat	Normalised expression data set matrix comprising the expression levels of genes (rows) for each sample (columns) in a data set. Row names are gene symbols and column names are sample IDs / names. Gene expression matrices can contain normalised (logCPM transformed) RNASeq or microarray transcriptomic data.
sig_df	Gene expression signature for a specific pathway given as data frame with the first column named "gene" containing a list of genes that are the most differentially expressed when the given pathway is active, and the second column named "expression" containing their corresponding expression in the gene signature: -1 for down-regulated genes and 1 for up-regulated genes.

Value

A density plot displaying distribution of GSVA scores obtained for the samples using up-regulated and down-regulated gene-sets from the gene signature

Author(s)

Anisha Thind <a.thind@cranfield.ac.uk>

Examples

```
## Not run: gsva_scores_dist(ER_dataset, ER_sig)
```

HER2_GEO_microarr *HER2 microarray data set obtained from GEO*

Description

A matrix containing microarray data for 60 human breast tumour samples (30 ERBB2 (HER2) positive and 30 HER2 negative samples were selected at random) and were extracted from GSE31448 dataset (Sabatier et al. 2011) using the GEO query library.Multi-gene probes were excluded and only cancer samples from human breast cancer tumours were selected for constructing this matrix.

Usage

```
HER2_GEO_microarr
```

Format

A matrix containing 21,656 HUGO gene symbols (row names) and 60 breast cancer tumour samples IDs (columns) which are given as sample accession numbers.

Source

GSE31448 series link on GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31448).

References

Sabatier, R., Finetti, P., Adelaide, J., Guille, A., Borg, J.P., Chaffanet, M., Lane, L., Birnbaum, D. and Bertucci, F., 2011. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. PloS one, 6(11), p.e27656. doi: https://doi.org/10.1371/journal.pone.0027656

HER2_sig_df *HER2 gene signature*

Description

A list of up-regulated and down-regulated genes of the ERBB2 (HER2) sub type (which is characterized by high expression of ERBB2) of 344 primary breast tumors from lymph node-negative patients (Smid et al. 2008).

Usage

```
HER2_sig_df
```

Format

A data frame with 156 rows (gene names) and 2 variables:

gene The HUGO gene symbols of a highly differentially expressed gene when HER2 pathway is active.

expression The change in expression of the gene when the pathway is active i.e. -1 for down-regulated genes and 1 for up-regulated genes.

Source

MSigDB (For up-regulated component of the: https://www.gsea-msigdb.org/gsea/msigdb/cards/SMID_BREAST_CANCER_ERBB2_UP.html)

MSigDB (For down-regulated component of the signature: https://www.gsea-msigdb.org/gsea/msigdb/cards/SMID_BREAST_CANCER_ERBB2_DN.html)

References

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A.M., Yu, J., Klijn, J.G., Foekens, J.A. and Martens, J.W., 2008. Subtypes of breast cancer show preferential site of relapse. Cancer research, 68(9), pp.3108-3114. doi: https://doi.org/10.1158/0008-5472.CAN-07-5644

HER2_TCGA_RNAseq *HER2 RNA-seq gene expression data set from TCGA*

Description

A gene expression matrix containing RNA-seq raw read counts for 60 primary human breast tumour samples (30 human epidermal growth receptor (HER2) positive and 30 HER2 negative samples were selected at random). This data set is a subset of a much larger data set containing 1,101 primary breast tumour samples collected from The Cancer Genome Atlas (TCGA).

Usage

```
HER2_TCGA_RNAseq
```

Format

A matrix containing 20,124 HUGO gene symbols (row names) and 60 breast cancer tumour samples IDs (columns) given in the form of TCGA barcodes for each sample. For further information on TCGA bar code semantics, please see the NIH GDC documentation https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/.

Source

TCGA https://portal.gdc.cancer.gov).

---

log_cpm_transform *Log CPM transformation of RNA-seq raw count data by using log CPM*

Description

Performs a logCPM transformation of RNA-seq raw count data using the counts per million (CPM) method from the edgeR library. In addition to the log CPM transformation, the function can also plot a boxplot as sanity check for logCPM transformation of the gene expression matrix. The first series of boxplots display the distribution of the raw counts for each sample in thousands, while the second series of boxplots show the distribution of the logCPM normalised gene expression matrix.

Usage

```
log_cpm_transform(dataset, boxplot = TRUE)
```

Arguments

dataset	An unnormalised gene expression matrix containing raw RNA-seq (integer) counts with gene symbols as row names and sample IDs as column names.
boxplot	Optional argument that is a boolean (TRUE or FALSE) indicating whether boxplots displaying before and after log CPM transformation should be displayed. (Default=TRUE)

Value

A logCPM transformed gene expression data matrix with gene symbols as row names and samples names / IDs as column names.

Author(s)

Rishabh Kaushik and Taniya Pal <rishabh.kaushik.126@cranfield.ac.uk, taniya.pal.094@cranfiled.ac.uk>

Examples

```
## Not run: log_cpm_transform(data.matrix)
```

---

pam50 *Gene signature for ER and HER2 extracted from PAM50*

Description

A list of two data frames containing the gene names and their corresponding expression values constituting ER and HER gene signatures obtained from PAM50 gene signature (Parker et al. 2009). The ER signature was obtained by combining the list of genes for the Luminal A and B intrinsic subtypes (centroids) predicted by PAM50 classifier in "pam50" list featured in the genefu R BiocManager package. Genes that were the most informative as demonstrated by a highly deviant centroid index for either Luminal A or B (estrogen receptor positive breast cancer subtypes) were selected as part of the ER signature for PAM50. Similarly, genes that were highly deviant in their centroid index in the HER2 intrinsic subtype (centroid) from PAM50 were selected as the HER2 signature for PAM50 using pam50 from genefu.

Usage

```
pam50
```

Format

A list holding two data frames: one for ER signature, the other for HER2 signature extracted from the PAM50 gene signature:

ER data frame containing HUGO gene symbols and their relative expression value (-1 for down-regulated genes and 1 for up-regulated genes) in the ER signature

HER2 data frame containing gene names their relative expression value (-1 for down-regulated genes and 1 for up-regulated genes) in the HER2 signature

Source

PAM50 publication: https://pubmed.ncbi.nlm.nih.gov/19204204/).

genefu: https://bioconductor.org/packages/release/bioc/html/genefu.html

References

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. and Quackenbush, J.F., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology, 27(8), p.1160. doi: https://doi.org/10.1200/JCO.2008.18.1370

---

read_expression_data *Reading gene expression data from file*

Description

Reads gene expression matrix data file which is either tab/comma/white-space value separated text file. After reading in the input file, any gene rows that contain NAs are removed and the data set is screened for duplicates. Duplicate genes are reduced to single gene row entries, with each value corresponding to the mean expression value for each sample for the duplicated gene entry. If duplicate samples are detected then the only the first sample of the duplicates are retained. Finally, the data frame is converted to a numerical matrix, where row names represent gene symbols and columns represent sample names or IDs.

Usage

```
read_expression_data(file)
```

Arguments

file	Path for gene expression matrix file, which is either tab / comma / white-space value separated. The first column must contain gene symbols and the subsequent column names should be the sample name or ID.

Value

A numerical matrix containing gene symbols/IDs as row names and sample IDs as column names.

Author(s)

Taniya Pal <taniya.pal.094@cranfield.ac.uk>

Examples

```
## Not run: read_expr_data("/Users/taniyapal/Documents/Group Project/TCGA_unannotated.txt")
```

`read_signature` *Reads up-regulated and down-regulated gene signatures from gene set files*

Description

Reads up and down regulated signature files provided either in gene set file format (.grp) or gene matrix transposed format (.gmt) creating a data frame which the first column named "gene" containing the gene symbols and the second column called "expression" containing the corresponding expression value for a gene in the gene signature, where 1 signifies up-regulation and -1 represents down-regulation of the gene in the gene signature.

Usage

```
read_signature(up_sig_file, down_sig_file)
```

Arguments

`up_sig_file`	Up-regulated gene-set format file
`down_sig_file`	Down-regulated gene-set format file

Value

A data frame containing both up regulated and down regulated signature files and signifying their up or down expression with +1 and -1 respectively.

Author(s)

Taniya Pal <taniya.pal.094@cranfield.ac.uk>

Examples

```
## Not run: read_sign_data("ESR1_UP.v1._UP.csv","ESR1_DN.v1_DN.csv" )
```